

# Alibaba Cloud

Machine Learning Platform for  
AI

Visualized modeling in Machine  
Learning Studio

Document Version: 20210519

# Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.
2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.
3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.
4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).
5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.
6. Please directly contact Alibaba Cloud for any errors of this document.

# Document conventions

| Style                                                                                            | Description                                                                                                                       | Example                                                                                                                                                                                                 |
|--------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  <b>Danger</b>  | A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results. |  <b>Danger:</b><br>Resetting will result in the loss of user configuration data.                                       |
|  <b>Warning</b> | A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results. |  <b>Warning:</b><br>Restarting will cause business interruption. About 10 minutes are required to restart an instance. |
|  <b>Notice</b>  | A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.      |  <b>Notice:</b><br>If the weight is set to 0, the server no longer receives new requests.                              |
|  <b>Note</b>  | A note indicates supplemental instructions, best practices, tips, and other content.                                              |  <b>Note:</b><br>You can use Ctrl + A to select all files.                                                            |
| >                                                                                                | Closing angle brackets are used to indicate a multi-level menu cascade.                                                           | Click <b>Settings</b> > <b>Network</b> > <b>Set network type</b> .                                                                                                                                      |
| <b>Bold</b>                                                                                      | Bold formatting is used for buttons, menus, page names, and other UI elements.                                                    | Click <b>OK</b> .                                                                                                                                                                                       |
| <code>Courier font</code>                                                                        | Courier font is used for commands                                                                                                 | Run the <code>cd /d C:/window</code> command to enter the Windows system folder.                                                                                                                        |
| <i>Italic</i>                                                                                    | Italic formatting is used for parameters and variables.                                                                           | <code>bae log list --instanceid</code><br><i>Instance_ID</i>                                                                                                                                            |
| [ ] or [a b]                                                                                     | This format is used for an optional value, where only one item can be selected.                                                   | <code>ipconfig [-all -t]</code>                                                                                                                                                                         |
| { } or {a b}                                                                                     | This format is used for a required value, where only one item can be selected.                                                    | <code>switch {active stand}</code>                                                                                                                                                                      |

# Table of Contents

|                                                                |     |
|----------------------------------------------------------------|-----|
| 1. Overview                                                    | 08  |
| 2. Authorization                                               | 09  |
| 3. General Machine learning components                         | 10  |
| 3.1. Data source and target                                    | 10  |
| 3.2. Data preprocessing                                        | 13  |
| 3.2.1. Sampling and Filter                                     | 13  |
| 3.2.2. Data Merge                                              | 21  |
| 3.2.3. Append ID Column                                        | 27  |
| 3.2.4. Split                                                   | 30  |
| 3.2.5. Missing Data Imputation                                 | 33  |
| 3.2.6. Normalization                                           | 39  |
| 3.2.7. Standardization                                         | 45  |
| 3.2.8. Data Type Conversion                                    | 52  |
| 3.2.9. KV to Table                                             | 57  |
| 3.2.10. Table to KV                                            | 61  |
| 3.3. Feature engineering                                       | 66  |
| 3.3.1. Feature transformation                                  | 66  |
| 3.3.2. Feature importance evaluation                           | 88  |
| 3.3.3. Feature Importance Filtering                            | 95  |
| 3.3.4. Feature generation                                      | 96  |
| 3.3.5. Anomaly Detection                                       | 110 |
| 3.3.6. Use the Binning component to implement the discretiz... | 112 |
| 3.4. Statistical analysis                                      | 115 |
| 3.4.1. Histogram (Multiple Columns)                            | 116 |
| 3.4.2. Pearson Coefficient                                     | 119 |
| 3.4.3. Percentile                                              | 121 |

|                                                  |     |
|--------------------------------------------------|-----|
| 3.4.4. Whole Table Statistics                    | 126 |
| 3.4.5. Discrete Feature Analysis                 | 130 |
| 3.4.6. One Sample T Test                         | 136 |
| 3.4.7. Chi-square Goodness of Fit Test           | 138 |
| 3.4.8. Data Pivoting                             | 141 |
| 3.4.9. Covariance                                | 143 |
| 3.4.10. Empirical Probability Density Chart      | 144 |
| 3.4.11. Box Plot                                 | 149 |
| 3.4.12. Scatter Plot                             | 156 |
| 3.4.13. Correlation Coefficient Matrix           | 162 |
| 3.4.14. Normality Test                           | 165 |
| 3.4.15. Lorenz Curve                             | 169 |
| 3.5. Machine Learning                            | 173 |
| 3.5.1. Binary classification                     | 173 |
| 3.5.2. Multiclass classification                 | 205 |
| 3.5.3. Clustering                                | 233 |
| 3.5.4. Regression                                | 247 |
| 3.5.5. Evaluation                                | 275 |
| 3.5.6. Recommendation                            | 291 |
| 3.5.7. Prediction                                | 299 |
| 3.6. Time series                                 | 303 |
| 3.6.1. x13_arima                                 | 303 |
| 3.6.2. x13_auto_arima                            | 311 |
| 3.7. Text analysis                               | 321 |
| 3.7.1. Word Splitting                            | 321 |
| 3.7.2. Convert Row, Column, and Value to KV Pair | 327 |
| 3.7.3. String Similarity                         | 331 |
| 3.7.4. BERT Embedding                            | 334 |

|                                          |     |
|------------------------------------------|-----|
| 3.7.5. String Similarity - Top N         | 337 |
| 3.7.6. Deprecated Word Filter            | 340 |
| 3.7.7. N-gram Counting                   | 341 |
| 3.7.8. Text Summarization                | 343 |
| 3.7.9. Keyword Extraction                | 346 |
| 3.7.10. Sentence Splitting               | 351 |
| 3.7.11. Semantic Vector Distance         | 353 |
| 3.7.12. Word Frequency Statistics        | 356 |
| 3.7.13. TF-IDF                           | 357 |
| 3.7.14. PLDA                             | 359 |
| 3.7.15. Word Splitting (Generate Models) | 362 |
| 3.7.16. Word2Vec                         | 366 |
| 3.7.17. Doc2Vec                          | 369 |
| 3.7.18. Conditional Random Field         | 373 |
| 3.7.19. Document Similarity              | 379 |
| 3.7.20. PMI                              | 384 |
| 3.8. Network analysis                    | 389 |
| 3.8.1. k-Core                            | 389 |
| 3.8.2. Single-source Shortest Path       | 392 |
| 3.8.3. Page Rank                         | 395 |
| 3.8.4. Label Propagation Clustering      | 398 |
| 3.8.5. Label Propagation Classification  | 403 |
| 3.8.6. Modularity                        | 407 |
| 3.8.7. Maximum Connected Subgraph        | 409 |
| 3.8.8. Vertex Clustering Coefficient     | 412 |
| 3.8.9. Edge Clustering Coefficient       | 416 |
| 3.8.10. Counting Triangle                | 419 |
| 3.8.11. Tree Depth                       | 423 |

|                                                              |     |
|--------------------------------------------------------------|-----|
| 3.9. Tools                                                   | 426 |
| 3.10. Financials                                             | 429 |
| 4.Components in deep learning frameworks                     | 449 |
| 4.1. Overview and enabling of deep learning frameworks       | 449 |
| 4.2. Caffe                                                   | 449 |
| 4.3. MXNet                                                   | 456 |
| 4.4. PyTorch                                                 | 460 |
| 4.5. Deep learning code downloads                            | 463 |
| 5.AutoML                                                     | 464 |
| 5.1. Use AutoML for automatic parameter tuning               | 464 |
| 5.2. Use AutoML for automatic feature engineering            | 474 |
| 6.Periodic scheduling                                        | 479 |
| 6.1. Import PAI models to OSS for a scheduling task          | 479 |
| 6.2. Perform offline scheduling                              | 480 |
| 7.Use the online notebook feature of Machine Learning Studio | 483 |
| 8.Configure and use global variables                         | 485 |
| 9.FastNN                                                     | 490 |

# 1. Overview

Machine Learning Studio provides a visual experiment and development environment for machine learning. It allows you to develop AI services without the need to write code. In addition, Machine Learning Studio provides multiple proven machine learning algorithms for you to meet your business requirements in different scenarios, such as product recommendation, financial risk management, and advertising prediction.

You can log on to Machine Learning Studio by using an Alibaba Cloud account or a Resource Access Management (RAM) user. If you use a RAM user, you must first use your Alibaba Cloud account to authorize the RAM user. For more information, see [Authorization](#).

Machine Learning Studio allows you to use a template to create an experiment or manually create an experiment. Template-based experiment creation is simple. After you create an experiment by using a template and run the experiment, you can deploy models. If you manually create an experiment, the system provides more than one hundred algorithm components and supports access to various data sources, such as MaxCompute table data and Object Storage Service (OSS) data.

When you train models, the system allows you to use Automated Machine Learning (AutoML) to automatically tune parameters and export models in the Predictive Model Markup Language (PMML) format. This helps you obtain the optimal models.

Machine Learning Studio provides the following algorithm components:

- Traditional machine learning components

These components include data preprocessing, feature engineering, statistical analysis, time series, text analysis, and network analysis.

- Deep learning framework components

These components include TensorFlow, Caffe, MXNet, and PyTorch.



## 2. Authorization

Machine Learning Studio shares projects with DataWorks. Before you use a RAM user to create MaxCompute or Machine Learning Studio projects, you must authorize the RAM user.

### Authorization in the RAM console

1. Log on to the [RAM console](#).
2. In the left-side navigation pane, choose **Identities > Users**.
3. Authorize a RAM user.
  - i. On the **Users** page, find the RAM user whom you want to authorize, and click **Add Permissions** in the **Actions** column.
  - ii. (Optional) In the **Add Permissions** panel, click **Custom Policy**.
  - iii. In the **Select Policy** field, enter **AliyunDataWorksFullAccess**.
  - iv. Click **AliyunDataWorksFullAccess** in the **Authorization Policy Name** column to make it appear in the **Selected** list.
  - v. Click **OK**.
4. Create an AccessKey pair for the RAM user.
  - i. On the **Users** page, find the RAM user and click its name in the **User Logon Name** column.
  - ii. In the **User AccessKeys** section, click **Create AccessKey**.
5. Log on to the DataWorks console by using your Alibaba Cloud account and grant permissions on the MaxCompute project to the RAM user. For more information, see [Add workspace members](#).

### OSS authorization

Machine Learning Studio supports OSS. You must use your Alibaba Cloud account to authorize service linked roles.

1. Log on to the [Machine Learning Platform for AI \(API\) console](#).
2. In the left-side navigation pane, choose **Model Training > Studio-Modeling Visualization**.
3. On the **PAI Visualization Modeling** page, find your project and click **Machine Learning** in the **Operation** column.
4. In the left-side navigation pane, choose **Settings > General**.
5. In the **OSS Authorization** section of the **General** page, select **Authorize Machine Learning Platform for AI to access my OSS resources**. For other parameters, use the default values.

### GPU resource authorization

The underlying layer of PAI-TensorFlow uses GPU resources. You must authorize the GPU resources.

1. Log on to the [PAI console](#).
2. In the left-side navigation pane, choose **Model Training > Studio-Modeling Visualization**.
3. On the **PAI Visualization Modeling** page, turn on **Open GPU** for your project.

## 3.General Machine learning components

### 3.1. Data source and target

This topic describes the data source and target components provided by Machine Learning Studio. The components include Write MaxCompute Table, Read MaxCompute Table, OSS Data Synchronization, and MySQL Data Synchronization.

#### Read MaxCompute Table

The Read MaxCompute Table component reads data from MaxCompute tables. By default, it reads the table data of the current project. To read data from tables across projects, you must prefix the table names with the related project names. You can configure the component only in the Machine Learning Platform for AI (PAI) console.

| Tab          | Parameter  | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------|------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Select Table | Table Name | <p>The name of the MaxCompute table from which you want to read data. If you want to read data from a MaxCompute table in another project, specify this parameter in the <b>Project name.Table name</b> format, such as <b>tianchi_project.weibo_data</b>.</p> <div> <b>Notice</b></div> <p>If the fields in the MaxCompute table are modified, the modifications cannot be synchronized to the algorithm platform of Machine Learning Studio. You must manually configure the MaxCompute data source again.</p> |
|              | Partition  | If the input table is a partitioned table, the system automatically selects the <b>Partition</b> check box.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|              | Parameters | The specific partition. You can specify only one partition. Specify this parameter in the <b>dt=@@{yyyyMMdd}</b> or <b>dt=@@{yyyyMMdd-1d}</b> format. <b>@@{yyyyMMdd}</b> indicates the current date and <b>@@{yyyyMMdd-1d}</b> indicates the day before the current date. If this parameter is not specified, the full table is specified.                                                                                                                                                                                                                                                         |
|              |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |

| Tab                | Parameter            | Description                                                                            |
|--------------------|----------------------|----------------------------------------------------------------------------------------|
| Fields Information | Source Table Columns | After you specify <b>Table Name</b> , the system automatically reads the table schema. |

## Write MaxCompute Table

The Write MaxCompute Table component writes data to a MaxCompute table. This component cannot be used to partition a table. You can configure the component only in the PAI console.

| Parameter      | Description                                                                                                                                                                                                      |
|----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| New Table Name | The name of the MaxCompute table to which you want to write data. If you want to write data to a partitioned table, you must create the table before you use the Write MaxCompute Table component to write data. |
| Partition      | Specifies whether the table to which you want to write data is a partitioned table.                                                                                                                              |
| Lifecycle      | The value of this parameter must be an integer. If this parameter is not specified, the specified table does not have a lifecycle.                                                                               |

## MySQL Data Synchronization

The MySQL Data Synchronization component synchronizes MySQL data to a MaxCompute project. You can configure the component only in the PAI console.

| Parameter | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Instance  | <p>The name of your ApsaraDB RDS for MySQL instance. To obtain the name of your ApsaraDB RDS for MySQL instance, perform the following steps:</p> <ol style="list-style-type: none"><li>1. Log on to the by using an Alibaba Cloud account.</li><li>2. In the left-side navigation pane, click <b>Instances</b>.</li><li>3. On the <b>Basic Information</b> tab of the <b>Instances</b> page, view the name of your ApsaraDB RDS for MySQL instance in the <b>Instance ID/Name</b> column.</li></ol> |

| Parameter                      | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Database                       | The name of your ApsaraDB RDS for MySQL database. To obtain the name of your ApsaraDB RDS for MySQL database, perform the following steps: <ol style="list-style-type: none"><li>1. Configure a whitelist. For more information, see .</li><li>2. Connect to the database. For more information, see .</li><li>3. View the <b>name</b>, <b>table</b>, and <b>table schema</b> of the database.</li></ol>                                                                                                                                                                                                           |
| Table                          | The table that contains the data you want to synchronize.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Username                       | The username that is used to log on to your ApsaraDB RDS for MySQL database. To obtain the username of your ApsaraDB RDS for MySQL database, perform the following steps: <ol style="list-style-type: none"><li>1. On the <b>Basic Information</b> tab of the <b>Instances</b> page, find your ApsaraDB RDS for MySQL instance and click its name in the <b>Instance ID/Name</b> column.</li><li>2. In the left-side navigation pane of the instance details page that appears, click <b>Accounts</b>.</li><li>3. On the <b>Accounts</b> tab, view the username of your ApsaraDB RDS for MySQL database.</li></ol> |
| Password                       | The password that is used to log on to your ApsaraDB RDS for MySQL database.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Source Columns                 | By default, all columns in the database are synchronized.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Stale Data Threshold           | The maximum number of dirty data records that are allowed. Default value: 0.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Data Synchronization Bandwidth | Default value: 1. Unit: MB/s.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |

## OSS Data Synchronization

The OSS Data Synchronization component synchronizes text from an OSS bucket to a MaxCompute table. You can configure the component only in the PAI console.

| Parameter    | Description                      |
|--------------|----------------------------------|
| OSS Endpoint | The endpoint of your OSS bucket. |

| Parameter                      | Description                                                                                                                                                                                                                                                                                                                                                            |
|--------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| OSS AccessID                   | The AccessKey ID of your Alibaba Cloud account that has OSS activated.                                                                                                                                                                                                                                                                                                 |
| OSS AccessKey                  | The AccessKey secret of your Alibaba Cloud account that has OSS activated.                                                                                                                                                                                                                                                                                             |
| Bucket                         | The name of your OSS bucket.                                                                                                                                                                                                                                                                                                                                           |
| Object                         | The OSS object that you want to synchronize.                                                                                                                                                                                                                                                                                                                           |
| Column Mapping                 | The mappings between fields in the OSS object and those in the MaxCompute table. Specify this parameter in the <b>Index:Name</b> format. <b>Index</b> indicates the field in the OSS object, and <b>Name</b> indicates the field in the MaxCompute table. Separate multiple fields with commas (,), such as <b>0:label,1:s_width,2:s_length,3:v_width,4:v_length</b> . |
| Text Delimiter                 | The delimiter that is used to separate the text in the OSS object. Commas (,) are used by default.                                                                                                                                                                                                                                                                     |
| Data Compression Format        | Valid values: <b>No Compression</b> , <b>GZIP</b> , <b>ZIP</b> , and <b>BZIP2</b> .                                                                                                                                                                                                                                                                                    |
| Text Encoding Format           | The encoding mode of the OSS text. Only <b>UTF-8</b> is supported.                                                                                                                                                                                                                                                                                                     |
| Data Synchronization Bandwidth | Default value: <b>1</b> . Unit: MB/s.                                                                                                                                                                                                                                                                                                                                  |
| Stale Data Threshold           | The maximum number of dirty data records that are allowed. Default value: <b>0</b> .                                                                                                                                                                                                                                                                                   |

#### Note

You can log on to the [OSS console](#) to query OSS-related information.

## 3.2. Data preprocessing

### 3.2.1. Sampling and Filter

This topic describes the Sampling and Filter components provided by Machine Learning Studio. The Sampling and Filter components include Random Sampling, Weighted Sampling, Filtering and Mapping, and Stratified Sampling.

## Random Sampling

The Random Sampling component samples data randomly. The samples are independent of each other.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter              | Description                                                                                                        |
|--------------------|------------------------|--------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Sample Size            | The value must be a positive integer.                                                                              |
|                    | Sample Ratio           | The value is a floating-point number. Valid values: (0,1).                                                         |
|                    | Return Samples         | By default, this check box is not selected. If you select this check box, sampling with replacement is enabled.    |
|                    | Random Seed            | The value is automatically generated by the system.                                                                |
| Tuning             | Number of Cores        | The value must be a positive integer. The value is automatically allocated by the system.                          |
|                    | Core Memory Allocation | The value must be a positive integer. Valid values: (1,65536). The value is automatically allocated by the system. |

- PAI command

```
PAI -name WeightedSample
  -project algo_public \
  -Dlifecycle="28" \
  -DoutputTableName="test2" \
  -DprobCol="previous" \
  -Dreplace="false" \
  -DsampleSize="500" \
  -DinputPartitions="pt=20150501" \
  -DinputTableName="bank_data_partition";
```

| Parameter      | Required | Description                  | Default value    |
|----------------|----------|------------------------------|------------------|
| inputTableName | Yes      | The name of the input table. | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                        | Default value           |
|----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| inputTablePartitions | No       | <p>The partitions in the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multiple-level partitions</li> </ul> <p><b>Note</b> If you specify multiple partitions, separate these partitions with commas (,).</p> | No default value        |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                      | No default value        |
| sampleSize           | No       | <p>The number of samples.</p> <p><b>Note</b></p> <ul style="list-style-type: none"> <li>If both sampleSize and sampleRatio are empty, an error is returned.</li> <li>If both sampleSize and sampleRatio are not empty, sampleSize takes precedence.</li> </ul>                                                                     | No default value        |
| sampleRatio          | No       | The sampling ratio, which is a floating-point number. Valid values: (0,1)                                                                                                                                                                                                                                                          | No default value        |
| replace              | No       | Specifies whether to enable sampling with replacement. The value is of the BOOLEAN type.                                                                                                                                                                                                                                           | false                   |
| randomSeed           | No       | The random seed, which is a positive integer.                                                                                                                                                                                                                                                                                      | Automatically allocated |
| lifecycle            | No       | The lifecycle of the output table. Valid values: [1,3650].                                                                                                                                                                                                                                                                         | No default value        |

| Parameter      | Required | Description                                                   | Default value           |
|----------------|----------|---------------------------------------------------------------|-------------------------|
| coreNum        | No       | The number of cores, which is a positive integer.             | Automatically allocated |
| memSizePerCore | No       | The memory size of each core, in MB. Valid values: (1,65536). | Automatically allocated |

## Weighted Sampling

The Weighted Sampling component generates sampling data based on the values of weighted columns. The values of the weighted columns must be of the DOUBLE or BIGINT type. Weighted columns are sampled based on their values. For example, if the values of two weighted columns are 1.2 and 1.0, the weighted column with value 1.2 is preferentially sampled.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter              | Description                                                                                                                                                                                                              |
|--------------------|------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Number of Samples      | The value must be a positive integer.                                                                                                                                                                                    |
|                    | Sample Ratio           | The value is a floating-point number. Valid values: (0,1).                                                                                                                                                               |
|                    | Return Samples         | By default, this check box is not selected. If you select this check box, sampling with replacement is enabled.                                                                                                          |
|                    | Weight Column          | You can select weighted columns from the drop-down list. The values of the weighted columns must be of the DOUBLE or BIGINT type. Each value represents the weight of an existing record. Normalization is not required. |
|                    | Random Number Seed     | The value is automatically generated by the system.                                                                                                                                                                      |
| Tuning             | Number of Cores        | The value must be a positive integer. The value is automatically allocated by the system.                                                                                                                                |
|                    | Core Memory Allocation | The value must be a positive integer. Valid values: (1,65536). The value is automatically allocated by the system.                                                                                                       |

- PAI command



```
PAI -name WeightedSample
-project algo_public \
-Dlifecycle="28" \
-DoutputTableName="test2" \
-DprobCol="previous" \
-Dreplace="false" \
-DsampleSize="500" \
-DinputPartitions="pt=20150501" \
-DinputTableName="bank_data_partition";
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                        | Default value    |
|----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                       | No default value |
| inputTablePartitions | No       | <p>The partitions in the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multiple-level partitions</li> </ul> <p><b>Note</b> If you specify multiple partitions, separate these partitions with commas (,).</p> | All partitions   |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                      | No default value |
| sampleSize           | No       | <p>The number of samples.</p> <p><b>Note</b></p> <ul style="list-style-type: none"> <li>If both sampleSize and sampleRatio are empty, an error is returned.</li> <li>If both sampleSize and sampleRatio are not empty, sampleSize takes precedence.</li> </ul>                                                                     | No default value |
| sampleRatio          | No       | The sampling ratio, which is a floating-point number. Valid values: (0,1).                                                                                                                                                                                                                                                         | No default value |
| probCol              | Yes      | The columns to be weighted. Each value represents the weight of an existing record. Normalization is not required. The values of the weighted columns must be of the DOUBLE or BIGINT type.                                                                                                                                        | No default value |

| Parameter      | Required | Description                                                                              | Default value                                                         |
|----------------|----------|------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|
| replace        | No       | Specifies whether to enable sampling with replacement. The value is of the BOOLEAN type. | false, which indicates that sampling with replacement is not enabled. |
| randomSeed     | No       | The random seed, which is a positive integer.                                            | Automatically allocated                                               |
| lifecycle      | No       | The lifecycle of the output table. Valid values: [1,3650].                               | No default value                                                      |
| coreNum        | No       | The number of cores, which is a positive integer.                                        | Automatically allocated                                               |
| memSizePerCore | No       | The memory size of each core, in MB. Valid values: (1,65536).                            | Automatically allocated                                               |

## Filtering and Mapping

The Filtering and Mapping component uses the expressions of filter conditions to filter data. This component allows you to modify the names of the columns that you want to filter.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Parameter                | Description                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Mapping Rules</b>     | The columns that you want to filter. By default, all columns are selected. You can also modify the names of the columns.                                                                                                                                                                                                                                                                                                                                |
| <b>Filter Conditions</b> | <p>Similar to the WHERE clause in SQL statements, the WHERE clause you specified, such as age&gt;40, is used to filter data.</p> <div> <p> <b>Note</b> Only the following operators are supported:</p> <ul style="list-style-type: none"> <li>=</li> <li>!=</li> <li>&gt;</li> <li>&lt;</li> <li>&gt;=</li> <li>&lt;=</li> <li>like</li> <li>rlike</li> </ul> </div> |

- PAI command

#### PAI -name Filter

```
-project algo_public \
-DoutTableName="test_9" \
-DinputPartitions="pt=20150501" \
-DinputTableName="bank_data_partition" \
-Dfilter="age>=40";
```

| Parameter       | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|-----------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| outputTableName | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| inputPartitions | No       | The partitions in the input table. The partitions are used for training. If an entire table is selected, the value is None.                                                                                                                                                                                                                                                                                                                                           |
| inputTableName  | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| filter          | No       | <p>Similar to the WHERE clause in SQL statements, the WHERE clause you specified, such as age&gt;40, is used to filter data.</p> <div> <p> <b>Note</b> Only the following operators are supported:</p> <ul style="list-style-type: none"> <li>◦ =</li> <li>◦ !=</li> <li>◦ &gt;</li> <li>◦ &lt;</li> <li>◦ &gt;=</li> <li>◦ &lt;=</li> <li>◦ like</li> <li>◦ rlike</li> </ul> </div> |

## Stratified Sampling

The Stratified Sampling component collects random samples of some data or a proportion of data from a dataset.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter             | Description                                                |
|--------------------|-----------------------|------------------------------------------------------------|
| Fields Setting     | Stratification Column | The columns that are used for stratification.              |
| Parameters Setting | Sample Size           | The value must be a positive integer.                      |
|                    | Sampling Fraction     | The value is a floating-point number. Valid values: (0,1). |
|                    |                       |                                                            |

| Tab           | Parameter                   | Description                                                                                                        |
|---------------|-----------------------------|--------------------------------------------------------------------------------------------------------------------|
|               | <b>Random Seed</b>          | The value is automatically generated by the system. The default value is 1234567.                                  |
| <b>Tuning</b> | <b>Cores</b>                | The value must be a positive integer. The value is automatically allocated by the system.                          |
|               | <b>Memory Size per Core</b> | The value must be a positive integer. Valid values: (1,65536). The value is automatically allocated by the system. |

- PAI command

```
PAI -name StratifiedSample
  -project algo_public \
  -DinputTableName="test_input" \
  -DoutputTableName="test_output" \
  -DstrataColName="label" \
  -DsampleSize="A:200,B:300,C:500" \
  -DrandomSeed=1007 \
  -Dlifecycle=30;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                            | Default value    |
|----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                           | No default value |
| inputTablePartitions | No       | <p>The partitions in the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multiple-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate these partitions with commas (,).</p> | All partitions   |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                          | No default value |
| strataColName        | Yes      | The name of the stratified column, which is used as the key for stratification.                                                                                                                                                                                                                                                                                                                                        | No default value |

| Parameter      | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Default value           |
|----------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| sampleSize     | No       | <p>The number of samples.</p> <ul style="list-style-type: none"> <li>◦ If the value is a positive integer, it indicates the number of samples at each stratum.</li> <li>◦ If the value is a string, the format is strata0:n0, strata1:n1, which indicates the number of samples that need to be configured for each stratum.</li> </ul> <div> <p> <b>Note</b></p> <ul style="list-style-type: none"> <li>◦ If both sampleSize and sampleRatio are empty, an error is returned.</li> <li>◦ If both sampleSize and sampleRatio are not empty, sampleSize takes precedence.</li> </ul> </div> | No default value        |
| sampleRatio    | No       | <p>The sampling ratio.</p> <ul style="list-style-type: none"> <li>◦ If the value is a number, it must be 0 or 1, and the value indicates the sampling ratio of each stratum.</li> <li>◦ If the value is a string, the format is strata0:r0, strata1:r1, which indicates the sampling ratio that needs to be configured for each stratum.</li> </ul>                                                                                                                                                                                                                                                                                                                         | No default value        |
| randomSeed     | No       | The random seed, which is a positive integer.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 123456                  |
| lifecycle      | No       | The lifecycle of the output table. Valid values: [1,3650].                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | No default value        |
| coreNum        | No       | The number of cores, which is a positive integer.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Automatically allocated |
| memSizePerCore | No       | The memory size of each core, in MB. Valid values: (1,65536).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Automatically allocated |

## 3.2.2. Data Merge

This topic describes the Data Merge components provided by Machine Learning Studio. The Data Merge components include JOIN, Merge Columns, and Merge Rows (UNION).

### Join

The JOIN component merges two tables by associating the information in the tables and determines the output columns. This component is similar to the JOIN statement of SQL.

You can configure the component only in the Machine Learning Platform for AI console.

| Parameter                                  | Description                                                                                             |
|--------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Join Type                                  | The join type. Valid values: Left Join, Inner Join, Right Join, and Full Join.                          |
| Join Condition                             | The join condition, which is in the format of equation. You can manually add or delete join conditions. |
| Select Output Columns from the Left Table  | The output columns you selected from the left table.                                                    |
| Select Output Columns from the Right Table | The output columns you selected from the right table.                                                   |

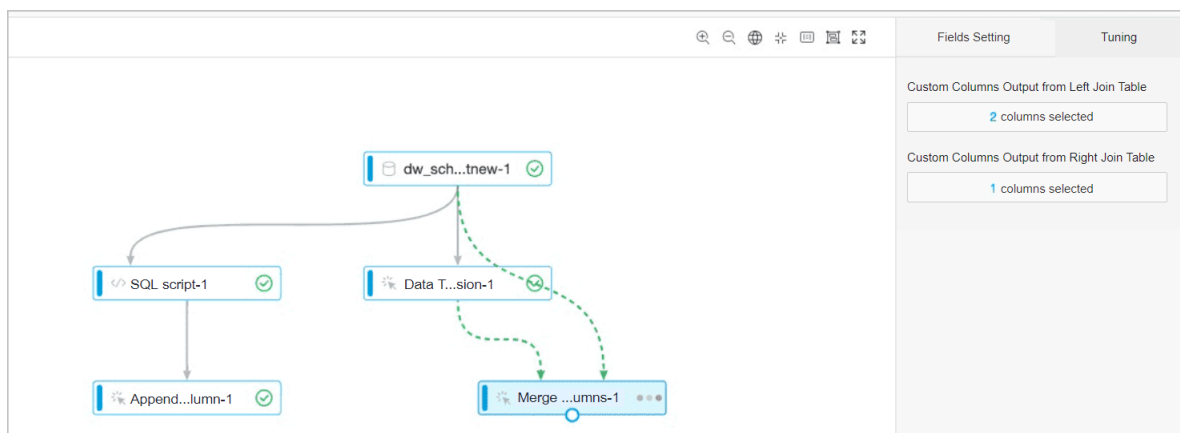
## Merge Columns

The Merge Columns component merges two tables by column. If this component is used, the two tables must have the same number of rows. If one of the two tables has partitions, the partitioned table must connect to the second input port.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

After you select the columns to be merged from the left table, the generated result is saved in the specified columns in the right table.



- PAI command

```
PAI -name appendColumns
  -project algo_public
  -DinputTableName=maple_test_appendcol_basic_input1,maple_test_appendcol_basic_input2
  -DoutputTableName=maple_test_appendcol_setOutCol_output
  -DoutputTableColNames=x0,x1,x2,x3,x4,x5,x6,x7,x8,x9;
```

| Parameter               | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Default value    |
|-------------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableNames         | Yes      | The names of the input tables. The names are separated by commas (,).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | No default value |
| outputTableName         | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | No default value |
| selectedColumnNamesList | No       | <p>The list of columns selected from the input table.</p> <ul style="list-style-type: none"> <li>◦ The names of columns in a table are separated by commas (,).</li> <li>◦ The names of columns in different tables are separated by semicolons (;).</li> </ul> <div> <p> <b>Note</b> If you select all columns in the two tables, the parameter value must be enclosed in quotation marks. Otherwise, the semicolon (;) is processed as the terminator. If all columns in a table are selected, all column names can be omitted, but the semicolons (;) must be retained.</p> </div> | No default value |
| inputPartitionsInfoList | No       | <p>The list of partitions selected from the input table.</p> <ul style="list-style-type: none"> <li>◦ Partitions in a table are separated by forward slashes (/).</li> <li>◦ Partitions in different tables are separated by semicolons (;).</li> </ul> <div> <p> <b>Note</b> If you select all partitions, the parameter value must be enclosed in quotation marks. Otherwise, the semicolon (;) is processed as the terminator. If a table is not partitioned, all partition names can be omitted, but the semicolons (;) must be retained.</p> </div>                            | No default value |
| autoRenameCol           | No       | Specifies whether to automatically rename the columns in the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | false            |

| Parameter              | Required | Description                                                                                                                                                                                                                                                                                                                                     | Default value           |
|------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| outputTableColumnNames | No       | <p>The new names of the columns in the output table. If this parameter is not specified, the column names in the original table are used.</p> <div>  <b>Notice</b> If the autoRenameCol parameter is set to true, this parameter is negligible.         </div> | No default value        |
| lifecycle              | No       | The lifecycle of the output table, which must be a positive integer.                                                                                                                                                                                                                                                                            | No default value        |
| coreNum                | No       | The number of cores. It must be a positive integer and is used together with the memSizePerCore parameter. Valid values: [1,9999].                                                                                                                                                                                                              | Automatically allocated |
| memSizePerCore         | No       | The memory size of each core, in MB. It must be a positive integer. Valid values: [1024,64 x 1024].                                                                                                                                                                                                                                             | Automatically allocated |

#### Example



- Source table

### maple\_test\_appendcol\_basic\_input1

| col0:bigint | col1:double | col2:string | col3:Datetime       | col4:Boolean |
|-------------|-------------|-------------|---------------------|--------------|
| 10          | 0.0         | aaaa        | 2015-10-01 00:00:00 | TRUE         |
| 11          | 1.0         | aaaa        | 2015-10-01 00:00:00 | FALSE        |
| 12          | 2.0         | aaaa        | 2015-10-01 00:00:00 | TRUE         |
| 13          | 3.0         | aaaa        | 2015-10-01 00:00:00 | TRUE         |
| 14          | 4.0         | aaaa        | 2015-10-01 00:00:00 | TRUE         |

### maple\_test\_appendcol\_basic\_input2

| col10:bigint | col11:double | col12:string | col13:Datetime      | col14:Boolean |
|--------------|--------------|--------------|---------------------|---------------|
| 110          | 10.0         | 2aaaa        | 2015-10-01 00:00:00 | TRUE          |
| 111          | 11.0         | 2aaaa        | 2015-10-01 00:00:00 | FALSE         |
| 112          | 12.0         | 2aaaa        | 2015-10-01 00:00:00 | TRUE          |
| 113          | 13.0         | 2aaaa        | 2015-10-01 00:00:00 | TRUE          |
| 114          | 14.0         | 2aaaa        | 2015-10-01 00:00:00 | FALSE         |

- PAI command

```
PAI -name appendColumns
  -project algo_public
  -DinputTableNames=maple_test_appendcol_basic_input1,maple_test_appendcol_basic_input2
  -DoutputTableName=maple_test_appendcol_setOutCol_output
  -DoutputTableColNames=x0,x1,x2,x3,x4,x5,x6,x7,x8,x9;
```

- Output table

### maple\_test\_appendcol\_setOutCol\_output

| x0 | x1 | x2   | x3                  | x4    | x5  | x6 | x7    | x8                  | x9    |
|----|----|------|---------------------|-------|-----|----|-------|---------------------|-------|
| 10 | 0  | aaaa | 2015-10-01 00:00:00 | true  | 110 | 10 | 2aaaa | 2015-10-01 00:00:00 | true  |
| 11 | 1  | aaaa | 2015-10-01 00:00:00 | false | 111 | 11 | 2aaaa | 2015-10-01 00:00:00 | false |
| 12 | 2  | aaaa | 2015-10-01 00:00:00 | true  | 112 | 12 | 2aaaa | 2015-10-01 00:00:00 | true  |
| 13 | 3  | aaaa | 2015-10-01 00:00:00 | true  | 113 | 13 | 2aaaa | 2015-10-01 00:00:00 | true  |
| 14 | 4  | aaaa | 2015-10-01 00:00:00 | true  | 114 | 14 | 2aaaa | 2015-10-01 00:00:00 | false |

## Merge Rows (UNION)

The Merge Rows (UNION) component merges the data of two tables by row. If this component is used, the numbers and data types of the output fields selected from the left and right tables must be the same. This component integrates the functions of UNION and UNION ALL.

You can configure the component only in the Machine Learning Platform for AI console.

| Parameter                                  | Description                                                                                                                |
|--------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|
| <b>Columns Output from Left Join Table</b> | When you merge tables, the numbers and data types of the columns selected from the left and right tables must be the same. |

| Parameter                                    | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|----------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>WHERE Conditions for Left Join Table</b>  | <p>Similar to the WHERE clause in SQL statements, the WHERE clause you specified, such as age&gt;40, is used to filter data.</p> <div> <p> <b>Note</b> Only the following operators are supported:</p> <ul style="list-style-type: none"> <li>• =</li> <li>• !=</li> <li>• &gt;</li> <li>• &lt;</li> <li>• &gt;=</li> <li>• &lt;=</li> <li>• like</li> <li>• rlike</li> </ul> </div>   |
| <b>Columns Output from Right Join Table</b>  | <p>When you merge tables, the numbers and types of the columns selected from the left and right tables must be the same.</p>                                                                                                                                                                                                                                                                                                                                            |
| <b>WHERE Conditions for Right Join Table</b> | <p>Similar to the WHERE clause in SQL statements, the WHERE clause you specified, such as age&gt;40, is used to filter data.</p> <div> <p> <b>Note</b> Only the following operators are supported:</p> <ul style="list-style-type: none"> <li>• =</li> <li>• !=</li> <li>• &gt;</li> <li>• &lt;</li> <li>• &gt;=</li> <li>• &lt;=</li> <li>• like</li> <li>• rlike</li> </ul> </div> |
| <b>Remove Duplicates</b>                     | <p>Removes duplicate rows in the generated table. By default, this check box is selected, which indicates that duplicate rows are removed.</p>                                                                                                                                                                                                                                                                                                                          |

### 3.2.3. Append ID Column

This topic describes the Append ID Column component provided by Machine Learning Studio. This component allows you to append an ID column to the first column of a data table.

#### Background information

The Append ID Column component can be used for the data tables that have a maximum of 1000000000 x 1023 rows.

#### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter               | Description                                                                           |
|--------------------|-------------------------|---------------------------------------------------------------------------------------|
| Parameters Setting | All checked by default. | By default, all columns are selected. Extra columns do not affect prediction results. |
|                    | ID Column               | The default value of this parameter is append_id.                                     |
|                    | Tuning                  | Number of Computing Cores                                                             |
|                    |                         | Memory Size per Core                                                                  |

- PAI command

```
PAI -name AppendId
-project algo_public
-DinputTableName=maple_test_appendid_basic_input
-DoutputTableName=maple_test_appendid_basic_output;
```

| Parameter        | Required | Description                                                                                                                                                                                                                                      | Default value    |
|------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName   | Yes      | The name of the input table.                                                                                                                                                                                                                     | No default value |
| selectedColNames | No       | The names of the columns selected from the input table for training. Column names are separated by commas (,). The names of the INT and DOUBLE types are supported. If the names are in sparse format, columns of the STRING type are supported. | All columns      |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                      | Default value           |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| inputTablePartitions | No       | <p>The partitions in the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"><li>◦ Partition_name=value</li><li>◦ name1=value1/name2=value2: multiple-level partitions</li></ul> <div><p> <b>Note</b> If you specify multiple partitions, separate these partitions with commas (,).</p></div> | All partitions          |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                    | No default value        |
| IDColName            | No       | The name of the appended ID column.                                                                                                                                                                                                                                                                                                                                                                                              | append_id               |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                                                                               | No default value        |
| coreNum              | No       | The number of cores.                                                                                                                                                                                                                                                                                                                                                                                                             | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core.                                                                                                                                                                                                                                                                                                                                                                                                    | Automatically allocated |

## Example

```
PAI -name AppendId
-project algo_public
-DinputTableName=maple_test_appendid_basic_input
-DoutputTableName=maple_test_appendid_basic_output;
```

- Generated data

| col0 | col1 | col2 | col3                         | col4 |
|------|------|------|------------------------------|------|
| 10   | 0.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | true |

| col0 | col1 | col2 | col3                         | col4  |
|------|------|------|------------------------------|-------|
| 11   | 1.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | false |
| 12   | 2.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | true  |
| 13   | 3.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | true  |
| 14   | 4.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | true  |

- Output table

| append_id | col0 | col1 | col2 | col3                         | col4  |
|-----------|------|------|------|------------------------------|-------|
| 0         | 10   | 0.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | true  |
| 1         | 11   | 1.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | false |
| 2         | 12   | 2.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | true  |
| 3         | 13   | 3.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | true  |
| 4         | 14   | 4.0  | aaaa | Thu Oct 01 00:00:00 CST 2015 | true  |

## 3.2.4. Split

This topic describes the Split component provided by Machine Learning Studio. This component randomly splits data to generate the datasets for training and testing.

### Split

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                                         | Operation                                                                                                                                                                                                                         |
|--------------------|---------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Splitting Method                                  | <ul style="list-style-type: none"> <li>Split by Ratio</li> <li>Split by Threshold</li> </ul>                                                                                                                                      |
|                    | Splitting Fraction                                | Valid values: (0,1)                                                                                                                                                                                                               |
|                    | Random Seed                                       | The random seed, which is automatically generated.                                                                                                                                                                                |
|                    | ID Column (Do Not Split Columns with the Same ID) | The ID columns you want to split.<br><div>  <b>Note</b> This parameter is available only when you select <b>Advanced Options</b>.         </div> |
|                    | Threshold Columns                                 | The columns that contain the threshold value. Columns of the STRING type are not supported.                                                                                                                                       |
|                    | Threshold                                         | The data specified by <b>Splitting Fraction</b> must be excluded.                                                                                                                                                                 |
| Tuning             | Cores                                             | The number of cores, which are automatically allocated for training based on the data volume that you entered.                                                                                                                    |
|                    | Memory Size per Core                              | The memory size of each core, in MB. The memory size is automatically allocated based on the data volume that you entered.                                                                                                        |

- PAI command

```
PAI -name split -project algo_public
-DinputTableName=wbpc
-Doutput1TableName=wbpc_split1
-Doutput2TableName=wbpc_split2
-Dfraction=0.25;
```

| Parameter      | Required | Description                  | Default value    |
|----------------|----------|------------------------------|------------------|
| inputTableName | Yes      | The name of the input table. | No default value |

| Parameter             | Required | Description                                                                                                                                                                                                                                                                                                                                   | Default value                              |
|-----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|
| inputTablePartitions  | No       | <p>The partitions selected from the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multiple-level partitions</li> </ul> <p><b>Note</b> If you specify multiple partitions, separate these partitions with commas (,).</p> | All partitions                             |
| output1TableName      | Yes      | Output table 1.                                                                                                                                                                                                                                                                                                                               | No default value                           |
| output1TablePartition | No       | The names of the partitions in output table 1.                                                                                                                                                                                                                                                                                                | Output table 1 is a non-partitioned table. |
| output2TableName      | Yes      | Output table 2.                                                                                                                                                                                                                                                                                                                               | No default value                           |
| output2TablePartition | No       | The names of the partitions in output table 2.                                                                                                                                                                                                                                                                                                | Output table 2 is a non-partitioned table. |
| fraction              | No       | The percentage of the split data that is allocated to output table 1. Valid values: (0,1).                                                                                                                                                                                                                                                    | No default value                           |
| randomSeed            | No       | The random seed, which must be a positive integer.                                                                                                                                                                                                                                                                                            | Automatically allocated                    |
| idColName             | No       | The ID column. Columns with the same ID cannot be split.                                                                                                                                                                                                                                                                                      | No default value                           |
| thresholdColName      | No       | The column that contains the threshold value. Columns of the STRING type are not supported.                                                                                                                                                                                                                                                   | No default value                           |
| threshold             | No       | The threshold.                                                                                                                                                                                                                                                                                                                                | No default value                           |



| Parameter      | Required | Description                                                   | Default value           |
|----------------|----------|---------------------------------------------------------------|-------------------------|
| lifecycle      | No       | The lifecycle of the output table. Valid values: [1,3650].    | No default value        |
| coreNum        | No       | The number of cores.                                          | Automatically allocated |
| memSizePerCore | No       | The memory size of each core, in MB. Valid values: (1,65536). | Automatically allocated |

## 3.2.5. Missing Data Imputation

The Missing Data Imputation component allows you to use a list of values to impute missing values in the input table.

### Background information

- This component can replace a numeric null value with a maximum, minimum, average, or custom value.
- This component can also replace a null string, empty string, null and empty string, or specified value with a custom value.
- The missing values you want to impute can be null values, empty strings, or custom values.

If you use empty strings to impute missing values, the destination column for which you want to impute missing values must be of the STRING type.

- You can replace a numeric null value with a custom value. Alternatively, you can directly replace the numeric null value with a maximum, minimum, or average value.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab            | Parameter         | Description                                                                                                                                                              |
|----------------|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Columns to Impute | By default, all columns are selected. Extra columns do not affect the prediction result.                                                                                 |
|                | Original Value    | <ul style="list-style-type: none"><li>◦ Null (Numerical and String)</li><li>◦ Empty String</li><li>◦ Null and Empty String (String)</li><li>◦ Custom (String)</li></ul>  |
|                | Replace With      | <ul style="list-style-type: none"><li>◦ Minimum (Numerical)</li><li>◦ Maximum (Numerical)</li><li>◦ Custom (Numerical)</li><li>◦ Custom (Numerical and String)</li></ul> |

| Tab | Parameter | Description                                                                                                                                                                                                |
|-----|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|     | configs   | <p>The ID column.</p> <div>  <b>Note</b> This parameter is available only when you select <b>Advanced Options</b>. </div> |
|     | Tuning    | Cores                                                                                                                                                                                                      |
|     |           | Memory Size per Core                                                                                                                                                                                       |

- PAI command

```
PAI -name FillMissingValues
  -project algo_public
  -Dconfigs="poutcome,null-empty,testing" \
  -DoutputTableName="test_3"
  -DinputPartitions="pt=20150501"
  -DinputTableName="bank_data_partition";
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                             | Default value    |
|----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                            | No default value |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multiple-level partitions</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate these partitions with commas (.). </div> | All partitions   |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                           | No default value |

| Parameter           | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Default value                              |
|---------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|
| configs             | Yes      | <p>The values that you want to use to impute missing values.</p> <p>For example, in <code>col1.null, 3.14; col2.empty, hello; col3.empty-null, world, null</code> indicates a null value, and <code>empty</code> indicates an empty string.</p> <ul style="list-style-type: none"> <li>◦ If you use empty strings, the destination column for which you want to impute missing values must be of the STRING type.</li> <li>◦ If you use a maximum, minimum, or average value, variables can be used and named as min, max, or mean.</li> <li>◦ If you use custom values, user-defined is used in a specified format, such as <code>col4.user-defined, string123</code>.</li> </ul> | No default value                           |
| outputParaTableName | No       | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Output table 1 is a non-partitioned table. |
| inputParaTableName  | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | No default value                           |
| lifecycle           | No       | The lifecycle of the output table. Valid values: [1,3650].                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | No default value                           |
| coreNum             | No       | The number of cores, which must be a positive integer.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Automatically allocated                    |
| memSizePerCore      | No       | The memory size of each core, in MB. Valid values: (1, 65536).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Automatically allocated                    |

## Example

1. Execute the following SQL statements to generate test data:

```
drop table if exists fill_missing_values_test_input;
create table fill_missing_values_test_input(
  col_string string,
  col_bigint bigint,
  col_double double,
  col_boolean boolean,
  col_datetime datetime);
insert overwrite table fill_missing_values_test_input
select
  *
from
(
  select
    '01' as col_string,
    10 as col_bigint,
    10.1 as col_double,
    True as col_boolean,
    cast('2016-07-01 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    cast(null as string) as col_string,
    11 as col_bigint,
    10.2 as col_double,
    False as col_boolean,
    cast('2016-07-02 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '02' as col_string,
    cast(null as bigint) as col_bigint,
    10.3 as col_double,
    True as col_boolean,
    cast('2016-07-03 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '03' as col_string,
    12 as col_bigint,
    cast(null as double) as col_double,
    False as col_boolean,
    cast('2016-07-04 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '04' as col_string,
    13 as col_bigint,
    10.4 as col_double,
    cast(null as boolean) as col_boolean,
    cast('2016-07-05 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '05' as col_string,
```

```

14 as col_bigint,
10.5 as col_double,
True as col_boolean,
cast(null as datetime) as col_datetime
from dual
) tmp;

```

Input data:

|      | col_string | col_bigint | col_double | col_boolean | col_datetime        |
|------|------------|------------|------------|-------------|---------------------|
| 04   | 13         | 10.4       | NULL       |             | 2016-07-05 10:00:00 |
| 02   | NULL       | 10.3       | true       |             | 2016-07-03 10:00:00 |
| 03   | 12         | NULL       | false      |             | 2016-07-04 10:00:00 |
| NULL | 11         | 10.2       | false      |             | 2016-07-02 10:00:00 |
| 01   | 10         | 10.1       | true       |             | 2016-07-01 10:00:00 |
| 05   | 14         | 10.5       | true       | NULL        |                     |

2. Run the following commands:

```

drop table if exists fill_missing_values_test_input_output;
drop table if exists fill_missing_values_test_input_model_output;
PAI -name FillMissingValues
-project algo_public
-Dconfigs="col_double,null,mean;col_string,null-empty,str_type_empty;col_bigint,null,max;col_boolean,null,true;col_datetime,null,2016-07-06 10:00:00"
-DoutputParaTableName="fill_missing_values_test_input_model_output"
-Dlifecycle="28"
-DoutputTableName="fill_missing_values_test_input_output"
-DinputTableName="fill_missing_values_test_input";
drop table if exists fill_missing_values_test_input_output_using_model;
drop table if exists fill_missing_values_test_input_output_using_model_model_output;
PAI -name FillMissingValues
-project algo_public
-DoutputParaTableName="fill_missing_values_test_input_output_using_model_model_output"
-DinputParaTableName="fill_missing_values_test_input_model_output"
-Dlifecycle="28"
-DoutputTableName="fill_missing_values_test_input_output_using_model"
-DinputTableName="fill_missing_values_test_input";

```

3. View the return results.

- fill\_missing\_values\_test\_input\_output

```

+-----+-----+-----+-----+-----+
| col_string | col_bigint | col_double | col_boolean | col_datetime |
+-----+-----+-----+-----+-----+
| 04 | 13 | 10.4 | true | 2016-07-05 10:00:00 |
| 02 | 14 | 10.3 | true | 2016-07-03 10:00:00 |
| 03 | 12 | 10.3 | false | 2016-07-04 10:00:00 |
| str_type_empty | 11 | 10.2 | false | 2016-07-02 10:00:00 |
| 01 | 10 | 10.1 | true | 2016-07-01 10:00:00 |
| 05 | 14 | 10.5 | true | 2016-07-06 10:00:00 |
+-----+-----+-----+-----+-----+

```

- fill\_missing\_values\_test\_input\_model\_output

```

+-----+-----+
| feature | json |
+-----+-----+
| col_string | {"name": "fillMissingValues", "type": "string", "paras":{"missing_value_type": "null-empty", "replaced_value": "str_type_empty"}} |
| col_bigint | {"name": "fillMissingValues", "type": "bigint", "paras":{"missing_value_type": "null", "replaced_value": 14}} |
| col_double | {"name": "fillMissingValues", "type": "double", "paras":{"missing_value_type": "null", "replaced_value": 10.3}} |
| col_boolean | {"name": "fillMissingValues", "type": "boolean", "paras":{"missing_value_type": "null", "replaced_value": 1}} |
| col_datetime | {"name": "fillMissingValues", "type": "datetime", "paras":{"missing_value_type": "null", "replaced_value": 1467770400000}} |
+-----+-----+

```

- fill\_missing\_values\_test\_input\_output\_using\_model

```

+-----+-----+-----+-----+-----+
| col_string | col_bigint | col_double | col_boolean | col_datetime |
+-----+-----+-----+-----+-----+
| 04 | 13 | 10.4 | true | 2016-07-05 10:00:00 |
| 02 | 14 | 10.3 | true | 2016-07-03 10:00:00 |
| 03 | 12 | 10.3 | false | 2016-07-04 10:00:00 |
| str_type_empty | 11 | 10.2 | false | 2016-07-02 10:00:00 |
| 01 | 10 | 10.1 | true | 2016-07-01 10:00:00 |
| 05 | 14 | 10.5 | true | 2016-07-06 10:00:00 |
+-----+-----+-----+-----+-----+

```

- fill\_missing\_values\_test\_input\_output\_using\_model\_model\_output

```

+-----+-----+
| feature | json |
+-----+-----+
| col_string | {"name": "fillMissingValues", "type": "string", "paras":{"missing_value_type": "null-empty", "replaced_value": "str_type_empty"}} |
| col_bigint | {"name": "fillMissingValues", "type": "bigint", "paras":{"missing_value_type": "null", "replaced_value": 14}} |
| col_double | {"name": "fillMissingValues", "type": "double", "paras":{"missing_value_type": "null", "replaced_value": 10.3}} |
| col_boolean | {"name": "fillMissingValues", "type": "boolean", "paras":{"missing_value_type": "null", "replaced_value": 1}} |
| col_datetime | {"name": "fillMissingValues", "type": "datetime", "paras":{"missing_value_type": "null", "replaced_value": 1467770400000}} |
+-----+-----+

```

## 3.2.6. Normalization

This topic describes the Normalization component provided by Machine Learning Studio.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter                | Description                                                                                                                                                                        |
|----------------|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | All Selected by Default  | All columns are selected by default. Additional columns do not affect the prediction result.                                                                                       |
|                | Reserve Original Columns | Specifies whether to reserve original columns. Column names are prefixed with normalized_ after normalization. Only the columns of the DOUBLE or BIGINT data type can be reserved. |
| Tuning         | Cores                    | The number of cores. The system automatically allocates the cores used for training based on the volume of input data.                                                             |
|                | Memory Size per Core     | The memory size of each core. The system automatically allocates the memory size based on the volume of input data. Unit: MB.                                                      |

- PAI command
  - Command for dense data

```

PAI -name Normalize
  -project algo_public
  -DkeepOriginal="true"
  -DoutputTableName="test_4"
  -DinputTablePartitions="pt=20150501"
  -DinputTableName="bank_data_partition"
  -DselectedColNames="emp_var_rate,euribor3m"

```

- Command for sparse data

```
PAI -name Normalize
-project projectxlib4
-DkeepOriginal="true"
-DoutputTableName="kv_norm_output"
-DinputTableName=kv_norm_test
-DselectedColNames="f0,f1,f2"
-DenableSparse=true
-DoutputParaTableName=kv_norm_model
-DkvIndices=1,2,8,6
-DitemDelimiter=",";
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                               | Default value    |
|----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                              | No default value |
| selectedColNames     | No       | The names of the columns selected from the input table for training. If you specify multiple columns, separate the column names with commas (,). Columns of the INT or DOUBLE data type can be used for training. However, if the input data is in the sparse format, only columns of the STRING data type can be used for training.                                                                                                      | All columns      |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>◦ Partition_name=value</li> <li>◦ name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | All partitions   |



| Parameter           | Required | Description                                                                                                                                                                                                                                                              | Default value                  |
|---------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------|
| outputTableName     | Yes      | The name of the output table.                                                                                                                                                                                                                                            | No default value               |
| outputParaTableName | No       | The name of the output parameter table.                                                                                                                                                                                                                                  | Non-partitioned output table 1 |
| inputParaTableName  | Yes      | The name of the input parameter table.                                                                                                                                                                                                                                   | No default value               |
| keepOriginal        | No       | Specifies whether to reserve original columns. Valid values: <ul style="list-style-type: none"><li>◦ true: Rename the normalized columns with the normalized_ prefix and reserve original columns.</li><li>◦ false: Reserve all columns without renaming them.</li></ul> | false                          |
| lifecycle           | No       | The lifecycle of the output table. Valid values: [1,3650].                                                                                                                                                                                                               | No default value               |
| coreNum             | No       | The number of cores used in computing. The value of this parameter must be a positive integer.                                                                                                                                                                           | Automatically allocated        |
| memSizePerCore      | No       | The memory size of each core. Unit: MB. Valid values: (1,65536).                                                                                                                                                                                                         | Automatically allocated        |
| enableSparse        | No       | Specifies whether to support the input data in the sparse format. Valid values: <ul style="list-style-type: none"><li>◦ true</li><li>◦ false</li></ul>                                                                                                                   | false                          |
| itemDelimiter       | No       | The delimiter used between key-value pairs.                                                                                                                                                                                                                              | ,                              |
| kvDelimiter         | No       | The delimiter used between keys and values.                                                                                                                                                                                                                              | :                              |

| Parameter | Required | Description                                                                                             | Default value    |
|-----------|----------|---------------------------------------------------------------------------------------------------------|------------------|
| kvIndices | No       | The feature indexes that require normalization in the table that contains data in the key-value format. | No default value |

## Example

- Data generation

```
drop table if exists normalize_test_input;
create table normalize_test_input(
  col_string string,
  col_bigint bigint,
  col_double double,
  col_boolean boolean,
  col_datetime datetime);
insert overwrite table normalize_test_input
select
  *
from
(
  select
    '01' as col_string,
    10 as col_bigint,
    10.1 as col_double,
    True as col_boolean,
    cast('2016-07-01 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    cast(null as string) as col_string,
    11 as col_bigint,
    10.2 as col_double,
    False as col_boolean,
    cast('2016-07-02 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '02' as col_string,
    cast(null as bigint) as col_bigint,
    10.3 as col_double,
    True as col_boolean,
    cast('2016-07-03 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '03' as col_string,
    12 as col_bigint,
    cast(null as double) as col_double,
    False as col_boolean,
```

```

        cast('2016-07-04 10:00:00' as datetime) as col_datetime
    from dual
union all
    select
        '04' as col_string,
        13 as col_bigint,
        10.4 as col_double,
        cast(null as boolean) as col_boolean,
        cast('2016-07-05 10:00:00' as datetime) as col_datetime
    from dual
union all
    select
        '05' as col_string,
        14 as col_bigint,
        10.5 as col_double,
        True as col_boolean,
        cast(null as datetime) as col_datetime
    from dual
) tmp;

```

- PAI commands

```

drop table if exists normalize_test_input_output;
drop table if exists normalize_test_input_model_output;
PAI -name Normalize
    -project algo_public
    -DoutputParaTableName="normalize_test_input_model_output"
    -Dlifecycle="28"
    -DoutputTableName="normalize_test_input_output"
    -DinputTableName="normalize_test_input"
    -DselectedColNames="col_double,col_bigint"
    -DkeepOriginal="true";
drop table if exists normalize_test_input_output_using_model;
drop table if exists normalize_test_input_output_using_model_model_output;
PAI -name Normalize
    -project algo_public
    -DoutputParaTableName="normalize_test_input_output_using_model_model_output"
    -DinputParaTableName="normalize_test_input_model_output"
    -Dlifecycle="28"
    -DoutputTableName="normalize_test_input_output_using_model"
    -DinputTableName="normalize_test_input";

```

- Input

normalize\_test\_input

| col_string | col_bigint | col_double | col_boolean | col_datetime        |
|------------|------------|------------|-------------|---------------------|
| 01         | 10         | 10.1       | true        | 2016-07-01 10:00:00 |
| NULL       | 11         | 10.2       | false       | 2016-07-02 10:00:00 |

| col_string | col_bigint | col_double | col_boolean | col_datetime        |
|------------|------------|------------|-------------|---------------------|
| 02         | NULL       | 10.3       | true        | 2016-07-03 10:00:00 |
| 03         | 12         | NULL       | false       | 2016-07-04 10:00:00 |
| 04         | 13         | 10.4       | NULL        | 2016-07-05 10:00:00 |
| 05         | 14         | 10.5       | true        | NULL                |

- Output

- normalize\_test\_input\_output

| col_string | col_bigint | col_double | col_boolean | col_datetime        | normalized_col_bigint | normalized_col_double |
|------------|------------|------------|-------------|---------------------|-----------------------|-----------------------|
| 01         | 10         | 10.1       | true        | 2016-07-01 10:00:00 | 0.0                   | 0.0                   |
| NULL       | 11         | 10.2       | false       | 2016-07-02 10:00:00 | 0.25                  | 0.24999999999999999   |
| 02         | NULL       | 10.3       | true        | 2016-07-03 10:00:00 | NULL                  | 0.500000000000000022  |
| 03         | 12         | NULL       | false       | 2016-07-04 10:00:00 | 0.5                   | NULL                  |
| 04         | 13         | 10.4       | NULL        | 2016-07-05 10:00:00 | 0.75                  | 0.750000000000000011  |
| 05         | 14         | 10.5       | true        | NULL                | 1.0                   | 1.0                   |

- normalize\_test\_input\_model\_output

| feature    | json                                                                         |
|------------|------------------------------------------------------------------------------|
| col_bigint | {"name": "normalize", "type": "bigint", "paras": {"min": 10, "max": 14}}     |
| col_double | {"name": "normalize", "type": "double", "paras": {"min": 10.1, "max": 10.5}} |

- normalize\_test\_input\_output\_using\_model

| col_string | col_bigint | col_double         | col_boolean | col_datetime        |
|------------|------------|--------------------|-------------|---------------------|
| 01         | 0.0        | 0.0                | true        | 2016-07-01 10:00:00 |
| NULL       | 0.25       | 0.2499999999999999 | false       | 2016-07-02 10:00:00 |
| 02         | NULL       | 0.5000000000000000 | true        | 2016-07-03 10:00:00 |
| 03         | 0.5        | NULL               | false       | 2016-07-04 10:00:00 |
| 04         | 0.75       | 0.7500000000000000 | NULL        | 2016-07-05 10:00:00 |
| 05         | 1.0        | 1.0                | true        | NULL                |

- normalize\_test\_input\_output\_using\_model\_model\_output

| feature    | json                                                                         |
|------------|------------------------------------------------------------------------------|
| col_bigint | {"name": "normalize", "type": "bigint", "paras": {"min": 10, "max": 14}}     |
| col_double | {"name": "normalize", "type": "double", "paras": {"min": 10.1, "max": 10.5}} |

## 3.2.7. Standardization

This topic describes the Standardization component provided by Machine Learning Studio.

### Background information

- You can standardize one or more columns in a table and save the generated data to a new table.
- The following formula is used for standardization:  $(X - \text{Mean}) / (\text{Standard deviation})$ 
  - Mean: the mean of samples.
  - Standard deviation: the standard deviation of samples. The standard deviation is used when samples are used to calculate the total deviation. To make the value obtained after standardization closer to the mean, you must moderately increase the calculated standard

deviation by using the formula  $\frac{1}{N-1}$ .

- The formula used to calculate the standard deviation of samples is

$$S = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2}$$

$\bar{X}$  represents the mean of samples  $X_1, X_2, \dots$ , and  $X_n$ .

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter                | Description                                                                                                                      |
|----------------|--------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | All Selected by Default  | By default, all columns are selected. Additional columns do not affect the prediction result.                                    |
|                | Reserve Original Columns | Column names are prefixed with stdized_ after standardization. The DOUBLE and BIGINT data types are supported.                   |
| Tuning         | Cores                    | The number of cores. The system automatically allocates the number of cores used for training based on the volume of input data. |
|                | Memory Size per Core     | The memory size of each core. The system automatically allocates the memory size based on the volume of input data. Unit: MB.    |

- PAI command
  - Command for dense data

```
PAI -name Standardize
  -project algo_public
  -DkeepOriginal="false"
  -DoutputTableName="test_5"
  -DinputTablePartitions="pt=20150501"
  -DinputTableName="bank_data_partition"
  -DselectedColNames="euribor3m,pdays"
```

◦ Command for sparse data

```
PAI -name Standardize
-project projectxlib4
-DkeepOriginal="true"
-DoutputTableName="kv_standard_output"
-DinputTableName=kv_standard_test
-DselectedColNames="f0,f1,f2"
-DenableSparse=true
-DoutputParaTableName=kv_standard_model
-DkvIndices=1,2,8,6
-DitemDelimiter=",";
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                        | Default value    |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                       | No default value |
| selectedColNames     | No       | The names of columns selected from the input table for training. Separate multiple names with commas (,). The selected columns can contain data of the INT or DOUBLE type. If data in the input table is in sparse format, the selected columns can contain data of the STRING type.                                                                                                                                                               | All columns      |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <div> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> </div> | All partitions   |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                      | No default value |

| Parameter           | Required | Description                                                                                                                                                                                                                                                                | Default value           |
|---------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| outputParaTableName | Yes      | The output table of parameters.                                                                                                                                                                                                                                            | No default value        |
| inputParaTableName  | No       | The input table of parameters.                                                                                                                                                                                                                                             | No default value        |
| keepOriginal        | No       | Specifies whether to reserve original columns. Valid values: <ul style="list-style-type: none"> <li>◦ true: Rename the standardized columns with the stdized_ prefix and reserve original columns.</li> <li>◦ false: Reserve all columns without renaming them.</li> </ul> | false                   |
| lifecycle           | No       | The lifecycle of the output table.                                                                                                                                                                                                                                         | No default value        |
| coreNum             | No       | The number of cores.                                                                                                                                                                                                                                                       | Automatically allocated |
| memSizePerCore      | No       | The memory size of each core.                                                                                                                                                                                                                                              | Automatically allocated |
| enableSparse        | No       | Specifies whether to support the input data in sparse format. Valid values: <ul style="list-style-type: none"> <li>◦ true</li> <li>◦ false</li> </ul>                                                                                                                      | false                   |
| itemDelimiter       | No       | The delimiter used between key-value pairs.                                                                                                                                                                                                                                | ,                       |
| kvDelimiter         | No       | The delimiter used between keys and values.                                                                                                                                                                                                                                | :                       |
| kvIndices           | No       | The feature indexes that require standardization in the table that contains data in the key-value format.                                                                                                                                                                  | No default value        |



## Example

Detailed example

```
drop table if exists standardize_test_input;
create table standardize_test_input(
  col_string string,
  col_bigint bigint,
  col_double double,
  col_boolean boolean,
  col_datetime datetime);
insert overwrite table standardize_test_input
select
  *
from
(
  select
    '01' as col_string,
    10 as col_bigint,
    10.1 as col_double,
    True as col_boolean,
    cast('2016-07-01 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    cast(null as string) as col_string,
    11 as col_bigint,
    10.2 as col_double,
    False as col_boolean,
    cast('2016-07-02 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '02' as col_string,
    cast(null as bigint) as col_bigint,
    10.3 as col_double,
    True as col_boolean,
    cast('2016-07-03 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '03' as col_string,
    12 as col_bigint,
    cast(null as double) as col_double,
    False as col_boolean,
    cast('2016-07-04 10:00:00' as datetime) as col_datetime
  from dual
  union all
  select
    '04' as col_string,
    13 as col_bigint,
    10.4 as col_double,
    cast(null as boolean) as col_boolean,
    cast('2016-07-05 10:00:00' as datetime) as col_datetime
  from dual
```

```

from dual
union all
select
  '05' as col_string,
  14 as col_bigint,
  10.5 as col_double,
  True as col_boolean,
  cast(null as datetime) as col_datetime
from dual
) tmp;

```

- PAI commands

```

drop table if exists standardize_test_input_output;
drop table if exists standardize_test_input_model_output;
PAI -name Standardize
  -project algo_public
  -DoutputParaTableName="standardize_test_input_model_output"
  -Dlifecycle="28"
  -DoutputTableName="standardize_test_input_output"
  -DinputTableName="standardize_test_input"
  -DselectedColNames="col_double,col_bigint"
  -DkeepOriginal="true";
drop table if exists standardize_test_input_output_using_model;
drop table if exists standardize_test_input_output_using_model_model_output;
PAI -name Standardize
  -project algo_public
  -DoutputParaTableName="standardize_test_input_output_using_model_model_output"
  -DinputParaTableName="standardize_test_input_model_output"
  -Dlifecycle="28"
  -DoutputTableName="standardize_test_input_output_using_model"
  -DinputTableName="standardize_test_input";

```

- Input

standardize\_test\_input

| col_string | col_bigint | col_double | col_boolean | col_datetime        |
|------------|------------|------------|-------------|---------------------|
| 01         | 10         | 10.1       | true        | 2016-07-01 10:00:00 |
| NULL       | 11         | 10.2       | false       | 2016-07-02 10:00:00 |
| 02         | NULL       | 10.3       | true        | 2016-07-03 10:00:00 |
| 03         | 12         | NULL       | false       | 2016-07-04 10:00:00 |
| 04         | 13         | 10.4       | NULL        | 2016-07-05 10:00:00 |
| 05         | 14         | 10.5       | true        | NULL                |

|            |            |            |             |              |
|------------|------------|------------|-------------|--------------|
| col_string | col_bigint | col_double | col_boolean | col_datetime |
|------------|------------|------------|-------------|--------------|

- Output

- standardize\_test\_input\_output

| col_string | col_bigint | col_double | col_boolean | col_datetime        | stdized_col_bigint  | stdized_col_double  |
|------------|------------|------------|-------------|---------------------|---------------------|---------------------|
| 01         | 10         | 10.1       | true        | 2016-07-01 10:00:00 | -1.2649110640673518 | -1.2649110640683832 |
| NULL       | 11         | 10.2       | false       | 2016-07-02 10:00:00 | -0.6324555320336759 | -0.6324555320341972 |
| 02         | NULL       | 10.3       | true        | 2016-07-03 10:00:00 | NULL                | 0.0                 |
| 03         | 12         | NULL       | false       | 2016-07-04 10:00:00 | 0.0                 | NULL                |
| 04         | 13         | 10.4       | NULL        | 2016-07-05 10:00:00 | 0.6324555320336759  | 0.6324555320341859  |
| 05         | 14         | 10.5       | true        | NULL                | 1.2649110640673518  | 1.2649110640683718  |

- standardize\_test\_input\_model\_output

| feature    | json                                                                                          |
|------------|-----------------------------------------------------------------------------------------------|
| col_bigint | {"name": "standardize", "type": "bigint", "paras": {"mean": 12, "std": 1.58113883008419}}     |
| col_double | {"name": "standardize", "type": "double", "paras": {"mean": 10.3, "std": 0.1581138830082909}} |

- `standardize_test_input_output_using_model`

| col_string | col_bigint                  | col_double                  | col_boolean | col_datetime           |
|------------|-----------------------------|-----------------------------|-------------|------------------------|
| 01         | -<br>1.2649110640673<br>515 | -<br>1.2649110640683<br>83  | true        | 2016-07-01<br>10:00:00 |
| NULL       | -<br>0.6324555320336<br>758 | -<br>0.6324555320341<br>971 | false       | 2016-07-02<br>10:00:00 |
| 02         | NULL                        | 0.0                         | true        | 2016-07-03<br>10:00:00 |
| 03         | 0.0                         | NULL                        | false       | 2016-07-04<br>10:00:00 |
| 04         | 0.6324555320336<br>758      | 0.6324555320341<br>858      | NULL        | 2016-07-05<br>10:00:00 |
| 05         | 1.2649110640673<br>515      | 1.2649110640683<br>716      | true        | NULL                   |

- `standardize_test_input_output_using_model_model_output`

| feature    | json                                                                                          |
|------------|-----------------------------------------------------------------------------------------------|
| col_bigint | {"name": "standardize", "type": "bigint", "paras": {"mean": 12, "std": 1.58113883008419}}     |
| col_double | {"name": "standardize", "type": "double", "paras": {"mean": 10.3, "std": 0.1581138830082909}} |

## 3.2.8. Data Type Conversion

This topic describes the Data Type Conversion component provided by Machine Learning Studio. You can use the Data Type Conversion component to convert features of all data types into features of the STRING, DOUBLE, or INT data type. This component also allows you to replace missing values if exceptions occur during data type conversion.

### Background information

- You can convert the data types of table fields.
- You can convert multiple data types of table fields at the same time.
- You can select whether to reserve original columns.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter                                   | Description                                                                                                              |
|----------------|---------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Convert to Double Type Columns              | The columns whose data types need to be converted into the DOUBLE data type.                                             |
|                | Default Imputed Value When Conversion Fails | The default value that is imputed when conversion to the DOUBLE data type fails.                                         |
|                | Convert to Int Type Columns                 | The columns whose data types need to be converted into the INT data type.                                                |
|                | Default Imputed Value When Conversion Fails | The default value that is imputed when conversion to the INT data type fails.                                            |
|                | Convert to String Type Columns              | The columns whose data types need to be converted into the STRING data type.                                             |
|                | Default Imputed Value When Conversion Fails | The default value that is imputed when conversion to the STRING data type fails.                                         |
|                | Reserve Original Columns                    | Specifies whether to reserve original columns. Add prefix typed_ to column names.                                        |
|                | Memory Size per Node                        | The memory size of each core. Unit: MB. Valid values: 1024 to 64 × 1024.                                                 |
|                | Cores                                       | The number of cores used in computing. This parameter is used with <b>Memory Size per Node</b> . Valid values: [1,9999]. |

- PAI command

```

pai -project algo_public
  -name type_transform_v1
  -DinputTable=type_test
  -Dcols_to_string="f0"
  -Ddefault_double_value=0.0
  -DoutputTable=type_test_output;

```

| Parameter  | Required | Description                  | Default value    |
|------------|----------|------------------------------|------------------|
| inputTable | Yes      | The name of the input table. | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                         | Default value    |
|----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | All partitions   |
| outputTable          | Yes      | The output table of data type conversion.                                                                                                                                                                                                                                                                                                                                                                                           | No default value |
| reserveOldFeat       | No       | Specifies whether to reserve original columns.                                                                                                                                                                                                                                                                                                                                                                                      | No default value |
| cols_to_double       | No       | The feature columns whose data types need to be converted into DOUBLE.                                                                                                                                                                                                                                                                                                                                                              | No default value |
| cols_to_string       | No       | The feature columns whose data types need to be converted into STRING.                                                                                                                                                                                                                                                                                                                                                              | No default value |
| cols_to_int          | No       | The feature columns whose data types need to be converted into INT.                                                                                                                                                                                                                                                                                                                                                                 | No default value |
| default_int_value    | No       | The value that is imputed when cols_to_int is not specified.                                                                                                                                                                                                                                                                                                                                                                        | 0                |
| default_double_value | No       | The value that is imputed when cols_to_double is not specified.                                                                                                                                                                                                                                                                                                                                                                     | 0.0              |

| Parameter            | Required | Description                                                                              | Default value           |
|----------------------|----------|------------------------------------------------------------------------------------------|-------------------------|
| default_string_value | No       | The value that is imputed when cols_to_string is not specified.                          | ""                      |
| coreNum              | No       | The number of cores. This parameter is used with memSizePerCore. Valid values: [1,9999]. | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core. Unit: MB. Valid values: [1024,64 × 1024].                  | Automatically allocated |
| lifecycle            | No       | The lifecycle of the output table.                                                       | 7                       |

## Example

- Test data

```
create table transform_test as
select * from
(
  select true as f0,2.0 as f1,1 as f2 from dual union all
  select false as f0,3.0 as f1,1 as f2 from dual union all
  select false as f0,4.0 as f1,1 as f2 from dual union all
  select true as f0,3.0 as f1,1 as f2 from dual union all
  select false as f0,3.0 as f1,1 as f2 from dual union all
  select false as f0,4.0 as f1,1 as f2 from dual union all
  select true as f0,3.0 as f1,1 as f2 from dual union all
  select false as f0,5.0 as f1,1 as f2 from dual union all
  select false as f0,3.0 as f1,1 as f2 from dual union all
  select true as f0,4.0 as f1,1 as f2 from dual union all
  select false as f0,3.0 as f1,1 as f2 from dual union all
  select true as f0,4.0 as f1,1 as f2 from dual
)tmp;
```

- Training data

| f0    | f1  | f2 |
|-------|-----|----|
| false | 3.0 | 1  |
| false | 3.0 | 1  |
| true  | 2.0 | 1  |
| true  | 4.0 | 1  |
| false | 4.0 | 1  |

| f0    | f1  | f2 |
|-------|-----|----|
| false | 3.0 | 1  |
| false | 3.0 | 1  |
| true  | 3.0 | 1  |
| false | 4.0 | 1  |
| true  | 4.0 | 1  |
| false | 5.0 | 1  |
| true  | 3.0 | 1  |

- PAI command for training

```

pai -project projectxlib4
  -name type_transform_v1
  -DinputTable=transform_test
  -Dcols_to_double=f0
  -Dcols_to_int=f1
  -Dcols_to_string=f2
  -DoutputTable=trans_test_output;

```

- Output

Output table

| f0  | f1 | f2 |
|-----|----|----|
| 0.0 | 3  | 1  |
| 0.0 | 3  | 1  |
| 1.0 | 2  | 1  |
| 1.0 | 4  | 1  |
| 0.0 | 4  | 1  |
| 0.0 | 3  | 1  |
| 1.0 | 3  | 1  |
| 0.0 | 4  | 1  |
| 0.0 | 3  | 1  |
| 0.0 | 5  | 1  |
| 1.0 | 3  | 1  |



| f0  | f1 | f2 |
|-----|----|----|
| 1.0 | 4  | 1  |

## 3.2.9. KV to Table

You can use the KV to Table component to convert a table in the key-value format into a table in a common format. The key is converted into a column name of the table, and the value is converted into the value of the column that corresponds to the row.

### Background information

Definition of the key-value table format: The key represents the column name index, and the value supports data of the BIGINT, DOUBLE, or STRING type. Users can import a custom key\_map table into this component. The key\_map table contains mappings of column names and keys. Regardless of whether a key\_map table is imported, this component exports a key\_map table to record the mappings of column names and keys after conversion, such as 1:10, 2:20, and 3:30.

Definition of the key\_map table format: The key\_map table format contains mappings of column names and indexes and data type information, which correspond to the col\_name, col\_index, and col\_datatype parameters. The values of the three parameters must be of the STRING data type. If you do not specify col\_datatype, the default value of this parameter is double.

| col_name | col_index | col_datatype |
|----------|-----------|--------------|
| col1     | 1         | bigint       |
| col2     | 2         | double       |

### KV to Table

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                       | Description                                                                                                                             |
|--------------------|---------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | KV Column                       | The names of the columns that contain data in the key-value format.                                                                     |
|                    | Appended Columns                | The names of appended columns.                                                                                                          |
|                    | KV Delimiter                    | The delimiter used between keys and values. Colons (:) are used by default.                                                             |
|                    | KV Pair Delimiter               | The delimiter used between key-value pairs. Commas (,) are used by default.                                                             |
| Parameters Setting | Reserve the First 1,200 Columns | Specifies whether to reserve only the first 1,200 columns if the number of columns in the key_map table after conversion exceeds 1,200. |

| Tab    | Parameter            | Description                                                                                                                      |
|--------|----------------------|----------------------------------------------------------------------------------------------------------------------------------|
| Tuning | Computing Cores      | The number of cores. The system automatically allocates the number of cores used for training based on the volume of input data. |
|        | Memory Size per Core | The memory size of each core. The system automatically allocates the memory size based on the volume of input data. Unit: MB.    |

- PAI command

```
PAI -name KVToTable
  -project algo_public
  -DinputTableName=test
  -DoutputTableName=test_out
  -DoutputKeyMapTableName=test_keymap_out
  -DkvColName=kv;
```

| Parameter             | Required | Description                                                         | Default value    |
|-----------------------|----------|---------------------------------------------------------------------|------------------|
| inputTableName        | Yes      | The name of the input table.                                        | No default value |
| kvColName             | Yes      | The names of the columns that contain data in the key-value format. | No default value |
| outputTableName       | Yes      | The name of the output table.                                       | No default value |
| outputKeyMapTableName | Yes      | The name of the output index table.                                 | No default value |
| inputKeyMapTableName  | No       | The name of the input index table.                                  | No default value |
| appendColName         | No       | The names of appended columns.                                      | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                 | Default value           |
|----------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"><li>Partition_name=value</li><li>name1=value1/name2=value2: multi-level partitions</li></ul> <div><p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p></div> | All partitions          |
| kvDelimiter          | No       | The delimiter used between keys and values.                                                                                                                                                                                                                                                                                                                                                                                                 | :                       |
| itemDelimiter        | No       | The delimiter used between key-value pairs.                                                                                                                                                                                                                                                                                                                                                                                                 | ,                       |
| top1200              | No       | <p>Specifies whether to reserve only the first 1,200 columns. Valid values:</p> <ul style="list-style-type: none"><li>true</li><li>false</li></ul>                                                                                                                                                                                                                                                                                          | true                    |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                                                                                          | No default value        |
| coreNum              | No       | The number of cores used in computing. Set the value to a positive integer.                                                                                                                                                                                                                                                                                                                                                                 | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core. Unit: MB. Valid values: (100,64 × 1024).                                                                                                                                                                                                                                                                                                                                                                      | Automatically allocated |

## Example

- Data generation

```
drop table if exists test;
create table test as
select
*
from
(
select '1:1,2:2,3:-3.3' as kv from dual
union all
select '1:10,2:20,3:-33.3' as kv from dual
) tmp;
```

- PAI command

```
PAI -name KVTtoTable
-project algo_public
-DinputTableName=test
-DoutputTableName=test_out
-DoutputKeyMapTableName=test_keymap_out
-DkvColName=kv;
```

- Output
  - Output table

```
+-----+-----+-----+
| kv_1  | kv_2  | kv_3  |
+-----+-----+-----+
| 1.0   | 2.0   | -3.3   |
| 10.0  | 20.0  | -33.3  |
+-----+-----+-----+
```

- Output mapping table

```
+-----+-----+-----+
| col_name | col_index | col_type |
+-----+-----+-----+
| kv_1     | 1         | double   |
| kv_2     | 2         | double   |
| kv_3     | 3         | double   |
+-----+-----+-----+
```

## Algorithm scale

The converted columns include the appended columns and the columns converted by using the KV to Table component. The columns converted by using the KV to Table component are exported before the appended columns. If the number of converted columns exceeds the maximum number of columns of the key\_map table, and the top1200 parameter is set to true, the maximum number of columns is exported. Otherwise, an error is reported. A maximum of 1,200 columns can be exported.

The number of data records cannot exceed 100 million.

## FAQ

- Q: Which columns are converted if the input includes a key\_map table?

A: The converted columns are the columns whose keys exist in both the key\_map and key-value tables.

- Q: What is the data type of the key column after conversion if the input includes a key\_map table?

A: The data type of the converted key column is the same as that of the key\_map table. If the data type of the key\_map table is not specified, the data type of the converted key column is DOUBLE.

- Q: What is the naming convention for the converted key column if a key\_map table is imported?

A: The naming convention is the names of columns in key:value format + "" + key.

The following characters are not supported:

%&()\*+-. /;<>=?

- Q: Why are column names contradictory to each other?

A: If you specify an appended column, and the name of the appended column is the same as that of the converted key column, an error is reported.

- Q: Which types of columns can be converted?

A: Only the columns of the numeric data types can be converted.

- Q: What would happen if the length of the column name exceeds 128 characters?

A: Only the first 128 characters of the column name are reserved.

- Q: What do I do if a row contains duplicate keys?

A: Calculate the sum of the values that correspond to these keys.

## 3.2.10. Table to KV

This topic describes the Table to KV component provided by Machine Learning Studio and how to use the component to convert a common table into a table in the key-value format.

### Limits

- The output table after conversion does not show the null values in the input table. You can specify the columns that you want to reserve in the output table. The specified columns are retained in their original formats.
- If the input includes a key\_map table, the converted columns are the columns whose keys exist in both the key\_map and key-value tables.
- If the input includes a key\_map table and its data type is different from that of the input table, the key\_map table in the output uses the data type that you specified.
- The columns that you want to convert into the key-value format in the input table must be of the BIGINT or DOUBLE data type.

### Table to KV

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab | Parameter | Description |
|-----|-----------|-------------|
|-----|-----------|-------------|

| Tab                | Parameter              | Description                                                                                                                        |
|--------------------|------------------------|------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Columns to Convert     | The names of the columns that need to be converted.                                                                                |
|                    | Reserved Columns       | The names of the columns that do not need to be converted.                                                                         |
|                    | KV Delimiter           | The delimiter used between keys and values. Colons (:) are used by default.                                                        |
|                    | KV Pair Delimiter      | The delimiter used between key-value pairs. Commas (,) are used by default.                                                        |
| Parameters Setting | Convert Columns to IDs | Specifies whether to convert the columns into IDs. Valid values: <ul style="list-style-type: none"> <li>Yes</li> <li>No</li> </ul> |
| Tuning             | Cores                  | The number of cores. The system automatically allocates the cores used for training based on the volume of input data.             |
|                    | Memory Size            | The memory size of each core. The system automatically allocates the memory size based on the volume of input data. Unit: MB.      |

- PAI command

```
PAI -name TableToKV
  -project algo_public
  -DinputTableName=maple_tabletokv_basic_input
  -DoutputTableName=maple_tabletokv_basic_output
  -DselectedColNames=col0,col1,col2
  -DappendColNames=rowid;
```

| Parameter      | Required | Description                  | Default value    |
|----------------|----------|------------------------------|------------------|
| inputTableName | Yes      | The name of the input table. | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                              | Default value    |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTablePartitions | No       | The partitions selected from the input table for training. Specify this parameter in the <b>Partition_name=value</b> format or the <b>name1=value1/name2=value2;</b> format (multi-level partitions). If you specify multiple partitions, separate them with commas (,). | All partitions   |
| selectedColNames     | No       | The names of the selected columns for conversion. The data types of the selected columns must be BIGINT or DOUBLE.                                                                                                                                                       | Full table       |
| appendColNames       | No       | The names of the columns that you want to reserve. The specified columns are retained in their original formats.                                                                                                                                                         | No default value |
| outputTableName      | Yes      | The name of the output table after conversion.                                                                                                                                                                                                                           | No default value |
| kvDelimiter          | No       | The delimiter used between keys and values.                                                                                                                                                                                                                              | :                |
| itemDelimiter        | No       | The delimiter used between key-value pairs.                                                                                                                                                                                                                              | ,                |
| convertColToIndexId  | No       | Specifies whether to convert columns into IDs. Valid values: <ul style="list-style-type: none"> <li>1: Yes</li> <li>0: No</li> </ul>                                                                                                                                     | 0                |

| Parameter             | Required                                       | Description                                                                                                                                                                                                     | Default value           |
|-----------------------|------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| inputKeyMapTableName  | No                                             | The name of the input index table.<br><br>This parameter is valid only when <code>convertColToIndexId</code> is set to 1. If this parameter is not specified, the system automatically calculates a set of IDs. | ""                      |
| outputKeyMapTableName | Determined by <code>convertColToIndexId</code> | The name of the output index table. This parameter is required only when <code>convertColToIndexId</code> is set to 1.                                                                                          | No default value        |
| lifecycle             | No                                             | The lifecycle of the output table. The value of this parameter must be a positive integer.                                                                                                                      | No default value        |
| coreNum               | No                                             | The number of cores. The value of this parameter must be a positive integer. Valid values: <code>[1,9999]</code> . This parameter is used with <code>memSizePerCore</code> .                                    | Automatically allocated |
| memSizePerCore        | No                                             | The memory size of each core. Unit: MB. The value of this parameter must be a positive integer. Valid values: <code>[1024,64 × 1024]</code> .                                                                   | Automatically allocated |

## Example 1

- Data generation

| rowid | kv                     |
|-------|------------------------|
| 0     | col0:1,col1:1.1,col2:2 |



| rowid | kv                     |
|-------|------------------------|
| 1     | col0:0,col1:1.2,col2:3 |
| 2     | col0:1,col1:2.3        |
| 3     | col0:1,col1:0.0,col2:4 |

- PAI command

```
PAI -name TableToKV
  -project algo_public
  -DinputTableName=maple_tabletokv_basic_input
  -DoutputTableName=maple_tabletokv_basic_output
  -DselectedColNames=col0,col1,col2
  -DappendColNames=rowid;
```

- Output

Output table *maple\_tabletokv\_basic\_output*

| rowid:bigint | kv:string |
|--------------|-----------|
| 0            | 1:1.1,2:2 |
| 1            | 1:1.2,2:3 |
| 2            | 1:2.3     |
| 3            | 1:0.0,2:4 |

## Example 2

- PAI command

```
PAI -name TableToKV
  -project projectxlib4 -DinputTableName=maple_tabletokv_basic_input
  -DoutputTableName=maple_tabletokv_basic_output
  -DselectedColNames=col0,col1,col2 -DappendColNames=rowid
  -DconvertColToIndexId=1
  -DinputKeyMapTableName=maple_test_tabletokv_basic_map_input
  -DoutputKeyMapTableName=maple_test_tabletokv_basic_map_output;
```

- Output

Output table *maple\_test\_tabletokv\_basic\_map\_output*

| col_name:string | col_index:string | col_datatype:string |
|-----------------|------------------|---------------------|
| col1            | 1                | bigint              |
| col2            | 2                | double              |

## 3.3. Feature engineering

### 3.3.1. Feature transformation

This topic describes the Principal Component Analysis (PCA), Feature Scaling, Feature Discretization, Singular-value Decomposition (SVD), and Feature Anomaly Smoothing components provided by Machine Learning Studio for feature transformation.

#### PCA

PCA is a multivariate statistical method used to explore the internal structures of multiple variables and how they correlate to each other based on a few principal components.

You can use PCA to export a few principal components that are unrelated to each other from original variables. These principal components retain as much information about the original variables as possible and are used as new comprehensive metrics.

PCA can reduce dimensions and noises for variables.

 **Note** PCA supports only data in the dense format.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter                  | Description                                                                                                                                        |
|--------------------|----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns            | The columns selected from the input table for analysis.                                                                                            |
|                    | Appended Columns           | The columns appended to the table after dimension reduction.                                                                                       |
| Parameters Setting | Data Size Ratio            | The information retaining ratio after dimension reduction.                                                                                         |
|                    | Feature Decomposition Mode | The method used to decompose features. Valid values: <ul style="list-style-type: none"><li>CORR</li><li>COVAR_SAMP</li><li>COVAR_POP</li></ul>     |
|                    | Data Conversion Method     | The method used to convert data types. Valid values: <ul style="list-style-type: none"><li>Simple</li><li>Sub-Mean</li><li>Normalization</li></ul> |

| Tab    | Parameter            | Description                                                                                                                                                             |
|--------|----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Tuning | Lifecycle            | The lifecycle of the output table. The value of this parameter must be a positive integer.                                                                              |
|        | Cores                | The number of cores. This parameter is used with <b>Memory Size per Node</b> . The value of this parameter must be a positive integer. Valid values: <i>[1, 9999]</i> . |
|        | Memory Size per Node | Unit: MB. The memory size of each core. The value of this parameter must be a positive integer. Valid values: <i>[1024, 64 × 1024]</i> .                                |

- PAI command

```
PAI -name PrinCompAnalysis
  -project algo_public
  -DinputTableName=bank_data
  -DeigOutputTableName=pai_temp_2032_17900_2
  -DprincompOutputTableName=pai_temp_2032_17900_1
  -DselectedColNames=pdays,previous,emp_var_rate,cons_price_idx,cons_conf_idx,euribor3m,nr_employed
  -DtransType=Simple
  -DcalcuType=CORR
  -DcontriRate=0.9;
```

| Parameter          | Required | Description                                                                                                                                                         | Default value    |
|--------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName     | Yes      | The input table used for principal component analysis.                                                                                                              | No default value |
| selectedColNames   | Yes      | The columns selected from the input table for analysis.<br><br>Separate multiple columns with commas (.). The columns of the INT or DOUBLE data type are supported. | No default value |
| eigOutputTableName | Yes      | The output table of feature vectors and feature values.                                                                                                             | No default value |

| Parameter               | Required | Description                                                                                                                                                                                              | Default value           |
|-------------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| princompOutputTableName | Yes      | The output table after dimension and noise reduction of principal components.                                                                                                                            | No default value        |
| transType               | No       | The method used to transform the original table to a principal component analysis table. Valid values: <ul style="list-style-type: none"> <li>Simple</li> <li>Sub-Mean</li> <li>Normalization</li> </ul> | <i>Simple</i>           |
| calcuType               | No       | The method used to decompose features of the original table. Valid values: <ul style="list-style-type: none"> <li>CORR</li> <li>COVAR_SAMP</li> <li>COVAR_POP</li> </ul>                                 | <i>CORR</i>             |
| contriRate              | No       | The information retaining ratio after dimension reduction. Valid values: <i>(0, 1)</i> .                                                                                                                 | <i>0.9</i>              |
| remainColumns           | No       | The fields retained from the original table after dimension reduction.                                                                                                                                   | No default value        |
| coreNum                 | No       | The number of cores. This parameter is used with memSizePerCore. The value of this parameter must be a positive integer. Valid values: <i>[1, 9999]</i> .                                                | Automatically allocated |
| memSizePerCore          | No       | The memory size of each core. Unit: MB. The value of this parameter must be a positive integer. Valid values: <i>[1024, 64 × 1024]</i> .                                                                 | Automatically allocated |

| Parameter | Required | Description                                                                                | Default value    |
|-----------|----------|--------------------------------------------------------------------------------------------|------------------|
| lifecycle | No       | The lifecycle of the output table. The value of this parameter must be a positive integer. | No default value |

### PCA output example

- Data table after dimension reduction

| prin0 ▲        | prin1 ▲            | prin2 ▲            | prin3 ▲            |
|----------------|--------------------|--------------------|--------------------|
| 95.78807909... | -42.95729950747... | -67.75761249391427 | -78.22763106620326 |
| 94.55356656... | -42.09090844438... | -68.61978267691727 | -78.13608278651054 |
| 89.91510009... | -47.88349039874... | -70.2131838386143  | -79.3464187998396  |
| 92.45302559... | -41.19278330986... | -69.35650482673981 | -78.92704651042823 |
| 89.76236928... | -46.44677428041... | -66.95927201734735 | -77.10906366265652 |
| 95.94066194... | -42.65895343179... | -69.30780740061341 | -78.59136424888288 |
| 92.33542049... | -41.16439652633... | -69.07389275424413 | -78.73867394929054 |
| 92.33143893... | -41.16317033059... | -69.07404676926707 | -78.74043328209397 |
| 90.14859611... | -46.10294878666... | -69.21610483502704 | -77.68690680501652 |

- Table of feature values and feature vectors

| prin_name ▲ | pdays ▲   | previous ▲   | emp_var_rate ▲   | cons_price_idx ▲   | cons_conf_idx ▲   | eurbor3m ▲    | nr_employed ▲   | eigenvalue ▲   | contributionrate ▲   | sumcontributionrate ▲ |
|-------------|-----------|--------------|------------------|--------------------|-------------------|---------------|-----------------|----------------|----------------------|-----------------------|
| prin0       | 0.2289... | -0.307801... | 0.49043713976... | 0.3676221023847... | 0.103704168633... | 0.49327195... | 0.4723413876... | 3.864397508... | 0.5520567869804135   | 0.5520567869804135    |
| prin1       | 0.6651... | -0.511030... | -0.1750362148... | -0.306844720053... | -0.38529019595... | -0.1519122... | 0.0083997057... | 1.333469023... | 0.190495574717606... | 0.7425523616980203    |
| prin2       | 0.1344... | -0.248552... | -0.1027813425... | -0.355514661117... | 0.882898436451... | 0.01908078... | -0.057503451... | 0.953306884... | 0.136186697819829... | 0.8787390595178498    |
| prin3       | -0.216... | 0.2034146... | 0.07125284421... | -0.732942136387... | -0.16599040564... | 0.21796215... | 0.5426739826... | 0.426458636... | 0.060922662369309... | 0.9396617218871589    |

## Feature Scaling

The Feature Scaling component has the following characteristics:

- Supports common scaling functions such as *log2*, *log10*, *ln*, *abs*, and *sqrt*.
- Supports data in the dense and sparse formats.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter       | Description                                                                                                                                      |
|----------------|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Scaled Features | The features that you want to scale.                                                                                                             |
|                | Label Column    | If this parameter is specified, the x-y histogram that displays the relationship between the features and objective variables can be visualized. |
|                |                 |                                                                                                                                                  |

| Tab                | Parameter        | Description                                                                                                                                                                             |
|--------------------|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Scaling Function | <b>Feature Scaling</b> supports the following scaling functions: <ul style="list-style-type: none"> <li>◦ log2</li> <li>◦ log10</li> <li>◦ ln</li> <li>◦ abs</li> <li>◦ sqrt</li> </ul> |

- PAI command

```
PAI -name fe_scale_runner -project algo_public
-Dlifecycle=28
-DscaleMethod=log2
-DscaleCols=nr_employed
-DinputTable=pai_dense_10_1
-DoutputTable=pai_temp_2262_20380_1;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                     | Default value                     |
|----------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTable           | Yes      | The name of the input table.                                                                                                                                                                                                                                                    | No default value                  |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in the <b>Partition_name=value</b> format or the <b>name1=value1/name2=value2;</b> format (multi-level partitions). If you specify multiple partitions, separate them with commas (,).</p> | All partitions in the input table |
| outputTable          | Yes      | The output table after scaling.                                                                                                                                                                                                                                                 | No default value                  |
| scaleCols            | Yes      | <p>The features that you want to scale.</p> <p>Sparse features are automatically filtered. You can select only the features of the numeric data types.</p>                                                                                                                      | No default value                  |

| Parameter    | Required | Description                                                                                                                                                                 | Default value    |
|--------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| labelCol     | No       | The label column.<br>If this parameter is specified, the $x$ - $y$ histogram that displays the relationship between the features and objective variables can be visualized. | No default value |
| categoryCols | No       | The selected fields that are processed as enumerated features. These fields do not support scaling.                                                                         | ""               |
| scaleMethod  | No       | The method used for scaling. Value values: <ul style="list-style-type: none"> <li>◦ log2</li> <li>◦ log10</li> <li>◦ ln</li> <li>◦ abs</li> <li>◦ sqrt</li> </ul>           | log2             |
| scaleTopN    | No       | If scaleCols is not specified, the system automatically selects the <i>top N</i> features that require scaling.                                                             | 10               |
| isSparse     | No       | Specifies whether features are sparse features in the <i>key-value</i> format.                                                                                              | Dense data       |
| itemSplitter | No       | The delimiter used between key-value pairs in the sparse format.                                                                                                            | ,                |
| kvSplitter   | No       | The delimiter used between keys and values in the sparse format.                                                                                                            | :                |
| lifecycle    | No       | The lifecycle of the output table.                                                                                                                                          | 7                |

| Parameter      | Required | Description                                                                                                                                          | Default value           |
|----------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| coreNum        | No       | The number of cores. The value of this parameter must be a positive integer. Valid values: $[1, 9999]$ . This parameter is used with memSizePerCore. | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Unit: MB. The value of this parameter must be a positive integer. Valid values: $[2048, 64 \times 1024]$ .             | Automatically allocated |

### Example

- Input data

```
create table if not exists pai_dense_10_1 as
select
  nr_employed
from bank_data limit 10;
```

- Parameters

On the Fields Setting tab, set Scaled Features to nr\_employed. Only the features of the numeric data types are supported. On the Parameters Setting tab, set **Scaling Function** to *log2*.

- Running result

| nr_employed        |
|--------------------|
| 12.352071021075528 |
| 12.34313018339218  |
| 12.285286613666395 |
| 12.316026916036957 |
| 12.309533196497519 |
| 12.352071021075528 |
| 12.316026916036957 |
| 12.316026916036957 |
| 12.309533196497519 |
| 12.316026916036957 |



## Feature Discretization

The Feature Discretization component supports the following types of discretization:

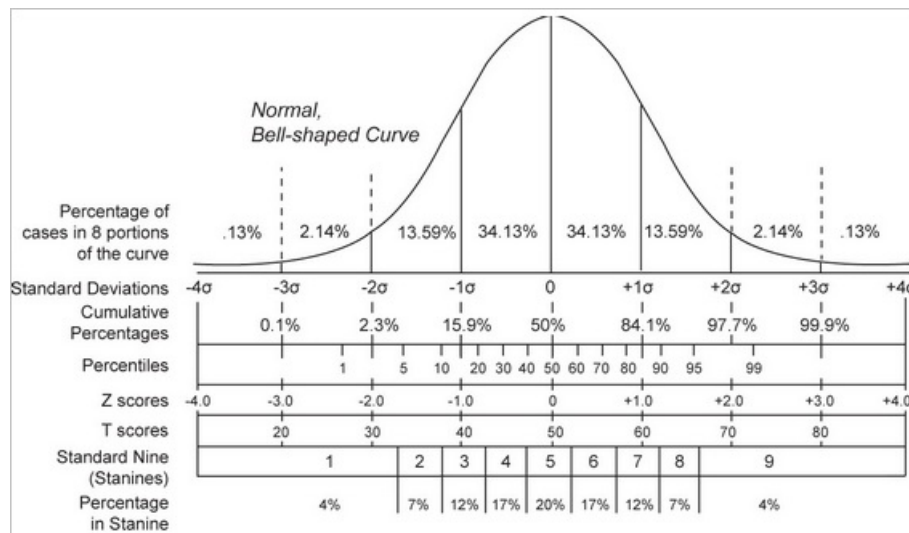
- Discretization of features of the numeric data types and in dense format
- Unsupervised discretization such as equal frequency discretization and equal width discretization

**Note** The default unsupervised discretization is equal width discretization.

- Supervised discretization such as Gini gain-based discretization and entropy gain-based discretization

**Note** The data type for label feature discretization must be ENUM, STRING, or BIGINT.

The Feature model prediction component can be used to predict feature discretization. The following figure shows the detailed directed acyclic graph (DAG) for modeling.



### **Note**

- You must use the same discrete model to predict feature discretization. This ensures that the same measurement is used during the prediction.
- Supervised discretization is used to search for segmentation points based on entropy gains by performing constant transversal. This type of discretization may take a long time to run. The number of bins obtained after segmentation is not limited by the value specified by maxBins.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab | Parameter         | Description                                                                                         |
|-----|-------------------|-----------------------------------------------------------------------------------------------------|
|     | Discrete Features | The features that require discretization. Sparse features are automatically filtered by the system. |

| Tab Fields Setting        | Parameter                      | Description                                                                                                                                                                                                                                                  |
|---------------------------|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                           | <b>Label Column</b>            | The label column. If this parameter is specified, the $x$ - $y$ histogram that displays the relationship between the features and objective variables can be visualized.                                                                                     |
| <b>Parameters Setting</b> | <b>Discretization Method</b>   | The method used for discretization. Valid values: <ul style="list-style-type: none"> <li>◦ Isometric Discretization</li> <li>◦ Isofrequency Discretization</li> <li>◦ Gini Gain-based Discretization</li> <li>◦ Entropy Gain-based Discretization</li> </ul> |
|                           | <b>Discretization Interval</b> | The interval for discretization. The value of this parameter must be a positive integer greater than 1.                                                                                                                                                      |
| <b>Tuning</b>             | <b>Cores</b>                   | The number of cores used in computing. The value of this parameter must be a positive integer.                                                                                                                                                               |
|                           | <b>Memory Size per Core</b>    | The memory size of each core.                                                                                                                                                                                                                                |

- PAI command

```
PAI -name fe_discrete_runner_1 -project algo_public
-DdiscreteMethod=SameFrequency
-Dlifecycle=28
-DmaxBins=5
-DinputTable=pai_dense_10_1
-DdiscreteCols=nr_employed
-DoutputTable=pai_temp_2262_20382_1
-DmodelTable=pai_temp_2262_20382_2;
```

| Parameter  | Required | Description                  | Default value    |
|------------|----------|------------------------------|------------------|
| inputTable | Yes      | The name of the input table. | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                              | Default value                     |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTablePartitions | No       | The partitions selected from the input table for training. Specify this parameter in the <b>Partition_name=value</b> format or the <b>name1=value1/name2=value2;</b> format (multi-level partitions). If you specify multiple partitions, separate them with commas (,). | All partitions in the input table |
| outputTable          | Yes      | The output table after discretization.                                                                                                                                                                                                                                   | No default value                  |
| discreteCols         | Yes      | The features that require discretization. Sparse features are automatically filtered by the system.                                                                                                                                                                      | ""                                |
| labelCol             | No       | The label column. If this field is specified, the <i>x-y</i> histogram that displays the relationship between the features and objective variables can be visualized.                                                                                                    | No default value                  |
| categoryCols         | No       | The selected fields that are processed as enumerated features. These fields do not support scaling.                                                                                                                                                                      | No default value                  |
| discreteMethod       | No       | The method used for discretization. Valid values: <ul style="list-style-type: none"> <li>◦ Isometric Discretization</li> <li>◦ Isofrequency Discretization</li> <li>◦ Gini Gain-based Discretization</li> <li>◦ Entropy Gain-based Discretization</li> </ul>             | <i>Isometric Discretization</i>   |

| Parameter      | Required | Description                                                                                                                                                                       | Default value           |
|----------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| discreteTopN   | No       | If discreteCols is not specified, the system automatically selects the <i>top N</i> features that require discretization. The value of this parameter must be a positive integer. | 10                      |
| maxBins        | No       | The interval for discretization. The value of this parameter must be a positive integer greater than 1.                                                                           | 100                     |
| isSparse       | No       | Specifies whether features are sparse features in the key-value format. Valid values:<br><ul style="list-style-type: none"> <li>◦ true</li> <li>◦ false</li> </ul>                | false                   |
| itemSplitter   | No       | The delimiter used between key-value pairs in the sparse format.                                                                                                                  | ,                       |
| kvSplitter     | No       | The delimiter used between keys and values in the sparse format.                                                                                                                  | :                       |
| lifecycle      | No       | The lifecycle of the output table. The value of this parameter must be a positive integer.                                                                                        | 7                       |
| coreNum        | No       | The number of cores. This parameter is used with <b>memSizePerCore</b> . The value of this parameter must be a positive integer.                                                  | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Unit: MB. The value of this parameter must be a positive integer.                                                                                   | Automatically allocated |

### Example

- Input data

```
create table if not exists pai_dense_10_1 as
select
  nr_employed
from bank_data limit 10;
```

- Parameters

The input data is *pai\_dense\_10\_1*. On the Fields Setting tab, set Discrete Features to *nr\_employed*. On the Parameters Setting tab, set Discretization Method to *Isometric Discretization* and Discrete Interval to *5*.

- Running result

| nr_employed |
|-------------|
| 4.0         |
| 3.0         |
| 1.0         |
| 3.0         |
| 2.0         |
| 4.0         |
| 3.0         |
| 3.0         |
| 2.0         |
| 3.0         |

## Feature Anomaly Smoothing

The Feature Anomaly Smoothing component can smooth anomalous features in input data to a specific interval. Data in both the sparse and dense formats are supported.

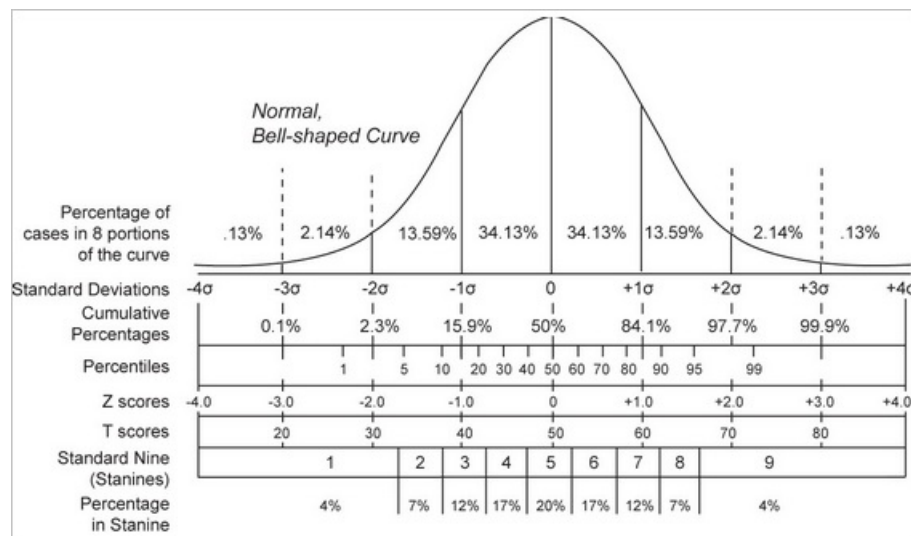
 **Note** The component corrects anomalous values but does not filter or delete records. The dimensions and number of input data records are not changed.

The following smoothing methods are supported:

- Z-Score

If a feature is in a normal distribution, the noise is distributed outside the range of  $-3\alpha$  to  $3\alpha$ . Z-Score smooths the noise to the range of  $[-3\alpha, 3\alpha]$ .

For example, a feature is in a normal distribution, the mean value is 0, and the standard deviation is 3. The feature value -10 is identified anomalous and corrected to -9 based on the smoothing rule of Z-Score. In the same way, the feature value 10 is corrected to 9.



- Percentile

Percentile smoothing is used to smooth the data distributed outside the range of  $[minPer, maxPer]$  to the minPer or maxPer quantile.

For example, the feature value of age is in the range of 0 to 200. Set minPer to 0 and maxPer to 50%. Feature values outside the range of 0 to 100 are corrected to 0 or 100.

- Threshold Smoothing

Threshold Smoothing is used to smooth the data distributed outside the range of  $[minThresh, maxThresh]$  to the minThresh or maxThresh data point.

For example, the feature value of age is in the range of 0 to 200. Set minThresh to 10 and maxThresh to 80. Feature values outside the range of 0 to 80 are corrected to 0 or 80.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter                | Description                                                                                                                                                        |
|----------------|--------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Smoothed Feature Columns | The feature columns that you want to smooth.                                                                                                                       |
|                | Label Column             | The label column. If this parameter is specified, the x-y histogram that displays the relationship between the features and objective variables can be visualized. |
|                |                          |                                                                                                                                                                    |

| Tab                | Parameter                  | Description                                                                                                                                                                                     |
|--------------------|----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | <b>Smoothing Method</b>    | The method used for smoothing. Valid values: <ul style="list-style-type: none"> <li>◦ Z-Score</li> <li>◦ Percentile</li> <li>◦ Threshold Smoothing</li> <li>◦ Box Plot</li> </ul>               |
|                    | <b>Confidence Interval</b> | The confidence level. This parameter is required when Smoothing Method is set to <b>Z-Score</b> .                                                                                               |
|                    | <b>Minimum Threshold</b>   | The minimum threshold. The default value is <i>-9999</i> , which indicates no minimum threshold.<br><br>This parameter is required when Smoothing Method is set to <b>Threshold Smoothing</b> . |
|                    | <b>Maximum Threshold</b>   | The maximum threshold. The default value is <i>-9999</i> , which indicates no maximum threshold.<br><br>This parameter is required when Smoothing Method is set to <b>Threshold Smoothing</b> . |
|                    | <b>Minimum Percentile</b>  | The minimum percentile.<br><br>This parameter is required when Smoothing Method is set to <b>Percentile</b> or <b>Box Plot</b> .                                                                |
|                    | <b>Maximum Percentile</b>  | The maximum percentile.<br><br>This parameter is required when Smoothing Method is set to <b>Percentile</b> or <b>Box Plot</b> .                                                                |

- PAI command

```
PAI -name fe_soften_runner -project algo_public
-DminThresh=5000
-Dlifecycle=28
-DsoftenMethod=min-max-thresh
-DsoftenCols=nr_employed
-DmaxThresh=6000
-DinputTable=pai_dense_10_1
-DoutputTable=pai_temp_2262_20381_1;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                              | Default value                     |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTable           | Yes      | The name of the input table.                                                                                                                                                                                                                                             | No default value                  |
| inputTablePartitions | No       | The partitions selected from the input table for training. Specify this parameter in the <b>Partition_name=value</b> format or the <b>name1=value1/name2=value2;</b> format (multi-level partitions). If you specify multiple partitions, separate them with commas (.). | All partitions in the input table |
| outputTable          | Yes      | The output table after smoothing.                                                                                                                                                                                                                                        | No default value                  |
| labelCol             | No       | The label column. If this parameter is specified, the x-y histogram that displays the relationship between the features and objective variables can be visualized.                                                                                                       | No default value                  |
| categoryCols         | No       | The selected fields that are processed as enumerated features.                                                                                                                                                                                                           | No default value                  |
| softenCols           | Yes      | The features that you want to smooth. Sparse features are automatically filtered by the system.                                                                                                                                                                          | No default value                  |



| Parameter    | Required | Description                                                                                                                                                                       | Default value  |
|--------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| softenMethod | No       | The method used for smoothing. Valid values: <ul style="list-style-type: none"> <li>◦ Z-Score</li> <li>◦ Percentile</li> <li>◦ Threshold Smoothing</li> <li>◦ Box Plot</li> </ul> | <i>Z-Score</i> |
| softenTopN   | No       | If softenCols is not specified, the system automatically selects the <i>top N</i> features that require smoothing. The value of this parameter must be a positive integer.        | <i>10</i>      |
| cl           | No       | The confidence level. This parameter is required when Smoothing Method is set to <b>Z-Score</b> .                                                                                 | <i>10</i>      |
| minPer       | No       | The minimum percentile. This parameter is required when Smoothing Method is set to <b>Percentile</b> or <b>Box Plot</b> .                                                         | <i>0.0</i>     |
| maxPer       | No       | The maximum percentile. This parameter is required when Smoothing Method is set to <b>Percentile</b> or <b>Box Plot</b> .                                                         | <i>1.0</i>     |
| minThresh    | No       | The minimum threshold. This parameter is required when Smoothing Method is set to <b>Threshold Smoothing</b> .                                                                    | <i>-9999</i>   |

| Parameter      | Required | Description                                                                                                                                                     | Default value           |
|----------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| maxThresh      | No       | The maximum threshold. This parameter is required when Smoothing Method is set to <b>Threshold Smoothing</b> .                                                  | <i>-9999</i>            |
| isSparse       | No       | Specifies whether features are sparse features in the key-value format. Valid values: <ul style="list-style-type: none"> <li>◦ true</li> <li>◦ false</li> </ul> | <i>false</i>            |
| itemSplitter   | No       | The delimiter used between key-value pairs in the sparse format.                                                                                                | <i>,</i>                |
| kvSplitter     | No       | The delimiter used between keys and values in the sparse format.                                                                                                | <i>:</i>                |
| lifecycle      | No       | The lifecycle of the output table. The value of this parameter must be a positive integer.                                                                      | <i>7</i>                |
| coreNum        | No       | The number of cores. This parameter is used with <b>memSizePerCore</b> . The value of this parameter must be a positive integer. Valid values: [1,9999].        | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Unit: MB. The value of this parameter must be a positive integer. Valid values: [2048,64 × 1024].                                 | Automatically allocated |

#### Example

- Input data

```
create table if not exists pai_dense_10_1 as
select
  nr_employed
from bank_data limit 10;
```

| nr_employed |
|-------------|
| 5228.1      |
| 5195.8      |
| 4991.6      |
| 5099.1      |
| 5076.2      |
| 5228.1      |
| 5099.1      |
| 5099.1      |
| 5076.2      |
| 5099.1      |

- Parameters

On the Fields Setting tab, set Smoothed Feature Columns to *nr\_employed*. On the **Parameters Setting** tab, set Smoothing Method to *Threshold Smoothing*, Minimum Threshold to *5000*, and Maximum Threshold to *6000*.

- Running result

| nr_employed |
|-------------|
| 5228.1      |
| 5195.8      |
| 5000.0      |
| 5099.1      |
| 5076.2      |
| 5228.1      |
| 5099.1      |
| 5099.1      |
| 5076.2      |

|             |
|-------------|
| nr_employed |
| 5099.1      |

## SVD

Singular-value Decomposition (SVD) is an important matrix decomposition in linear algebra. It is a generalization of the diagonalization of normal matrices in matrix analysis. SVD is widely used in fields such as signal processing and statistics.

Formula for singular value decomposition:  $X = U S V'$

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter                | Description                                                                                                                                                          |
|--------------------|--------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns          | The columns used to store key-value pairs. The keys and values are separated by colons (:), and multiple key-value pairs are separated by commas (,).                |
| Parameters Setting | Reserved Singular Values | The top N singular groups that you want to decompose. All singular groups are decomposed by default.                                                                 |
|                    | Precision Error          | The expected error precision that is allowed.                                                                                                                        |
| Tuning             | Memory Size per Node     | The memory size of each node. Unit: MB. This parameter is used with memSizePerCore. The value of this parameter must be a positive integer. Valid values: [1, 9999]. |
|                    | Cores                    | The value of this parameter must be a positive integer. Valid values: [1024, 64 × 1024].                                                                             |
|                    | Lifecycle                | The lifecycle of the output table.                                                                                                                                   |

- PAI command

```
PAI -name svd
-project algo_public
-DinputTableName=bank_data
-DselectedColNames=col0
-DenableSparse=true
-Dk=5
-DoutputUTableName=u_table
-DoutputVTableName=v_table
-DoutputSTableName=s_table;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                        | Default value    |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The input table used for training.                                                                                                                                                                                                                                                                 | No default value |
| selectedColNames     | No       | <p>The names of columns selected from the input table for training. Separate the names with commas (,).</p> <p>If a sparse matrix is used, the columns of the STRING data type are supported. If a data table is used, the columns of the INT and DOUBLE data types are supported.</p>             | All columns      |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in the <code>Partition_name=value</code> format or the <code>name1=value1/name2=value2;</code> format (multi-level partitions).</p> <p>If you specify multiple partitions, separate them with commas (,).</p> | All partitions   |

| Parameter        | Required | Description                                                                                                                                                                                                                 | Default value    |
|------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| outputUTableName | Yes      | The output table of the unitary matrix. The output table is generated from the $m * sgNum$ dimension. $m$ represents the number of rows of the data table, and $sgNum$ represents the number of calculated singular values. | No default value |
| outputSTableName | Yes      | The output table of the scattering matrix (S-matrix). The output table is generated from the $sgNum * sgNum$ dimension. $sgNum$ represents the number of calculated singular values.                                        | No default value |
| outputVTableName | Yes      | The output table of the V matrix. The output table is generated from the $n * sgNum$ dimension. $n$ represents the number of columns of the matrix, and $sgNum$ represents the number of calculated singular values.        | No default value |
| k                | Yes      | The number of expected singular values.<br><br>The number of generated singular values may be a positive integer less than the value specified by the k parameter.                                                          | No default value |
| tol              | No       | The convergence error.                                                                                                                                                                                                      | $1.0e-06$        |

| Parameter      | Required | Description                                                                                                                                                      | Default value           |
|----------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| enableSparse   | No       | Specifies whether data in the input table is in the sparse format.<br>Valid values:<br><ul style="list-style-type: none"> <li>◦ true</li> <li>◦ false</li> </ul> | false                   |
| itemDelimiter  | No       | The delimiter used between key-value pairs when data in the input table is in the sparse format.                                                                 | Space                   |
| kvDelimiter    | No       | The delimiter used between keys and values when data in the input table is in the sparse format.                                                                 | :                       |
| coreNum        | No       | The number of cores. This parameter is used with <b>memSizePerCore</b> . The value of this parameter must be a positive integer. Valid values: [1,9999].         | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Unit: MB. The value of this parameter must be a positive integer. Valid values: [1024,64 × 1024].                                  | Automatically allocated |
| lifecycle      | No       | The lifecycle of the output table. The value of this parameter must be a positive integer.                                                                       | No default value        |

#### Example

- Data generation

```
drop table if exists svd_test_input;
create table svd_test_input
as
select
*
from
(
select
'0:3.9079 2:0.0009 3:0.0416 4:0.17664 6:0.36460 8:0.091330' as col0
from dual
union all
select
'0:0.09229 2:0.4872172 5:0.5267 8:0.4544 9:0.23317' as col0
from dual
union all
select
'1:0.8312 3:0.9317 5:0.5680 7:0.5560 9:0.0508' as col0
from dual
union all
select
'2:0.767 5:0.01891 8:0.25235 ' as col0
from dual
union all
select
'0:0.29819 2:0.87598086 6:0.5315568 ' as col0
from dual
union all
select
'0:0.920260 2:0.5154311513 4:0.8104 5:0.188420 8:0.88' as col0
from dual
) a;
```

- PAI command

```
PAI -name svd
-project algo_public
-DinputTableName=svd_test_input
-DselectedColNames=col0
-DenableSparse=true
-Dk=5
-DoutputUTableName=u_table
-DoutputVTableName=v_table
-DoutputSTableName=s_table;
```

- Analysis scale: 100,000 columns

### 3.3.2. Feature importance evaluation

This topic describes the Random Forest Feature Importance Evaluation and Linear Model Feature Importance Evaluation components provided by Machine Learning Studio for feature importance evaluation.

#### Random Forest Feature Importance



You can use raw data and a random forest model to calculate feature importance. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter                | Description                                                                                                                                                                                                                                                        |
|--------------------|--------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns          | Optional. The feature columns selected from the input table for training. By default, all columns other than the label column are selected.                                                                                                                        |
|                    | Target Column            | Required.<br>Click the  icon. In the <b>Select Column</b> dialog box, enter a keyword to search for your desired column. Then, select the searched column and click <b>OK</b> . |
| Parameters Setting | Parallel Computing Cores | Optional. The number of cores used in parallel computing.                                                                                                                                                                                                          |
|                    | Memory Size per Core     | Optional. The memory size of each core. Unit: MB.                                                                                                                                                                                                                  |

- PAI command

```

pai -name feature_importance -project algo_public
-DinputTableName=pai_dense_10_10
-DmodelName=xlab_m_random_forests_1_20318_v0
-DoutputTableName=erkang_test_dev.pai_temp_2252_20319_1
-DlabelColName=y
-DfeatureColNames="pdays,previous,emp_var_rate,cons_price_idx,cons_conf_idx,euribor3m,nr_employed,age,campaign,poutcome"
-Dlifecycle=28 ;

```

| Parameter       | Required | Description                                      | Default value    |
|-----------------|----------|--------------------------------------------------|------------------|
| inputTableName  | Yes      | The name of the input table.                     | No default value |
| outputTableName | Yes      | The name of the output table.                    | No default value |
| labelColName    | Yes      | The name of the label column in the input table. | No default value |
| modelName       | Yes      | The name of the input model.                     | No default value |

| Parameter            | Required | Description                                                     | Default value                           |
|----------------------|----------|-----------------------------------------------------------------|-----------------------------------------|
| featureColNames      | No       | The feature columns selected from the input table for training. | All columns other than the label column |
| inputTablePartitions | No       | The partitions selected from the input table for training.      | Full table                              |
| lifecycle            | No       | The lifecycle of the output table.                              | No default value                        |
| coreNum              | No       | The number of cores.                                            | Automatically allocated                 |
| memSizePerCore       | No       | The memory size of each core.                                   | Automatically allocated                 |

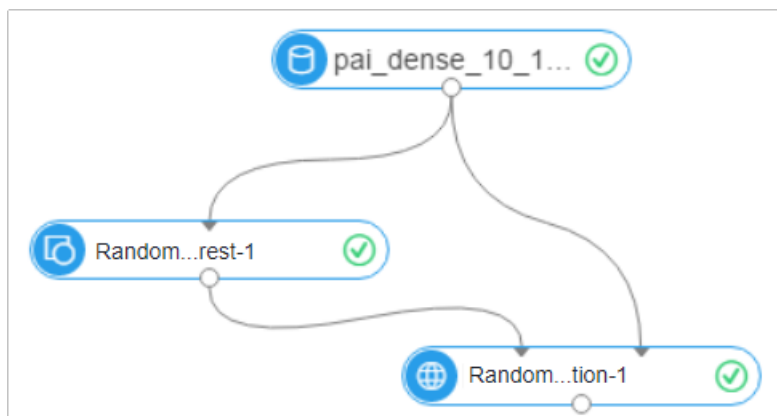
## Example

1. Execute the following SQL statement to generate training data:

```
drop table if exists pai_dense_10_10;
create table if not exists pai_dense_10_10 as
select
  age,campaign,pdays,previous,poutcome,emp_var_rate,cons_price_idx,cons_conf_idx,euribor3m,
  nr_employed,y
from bank_data limit 10;
```

2. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).

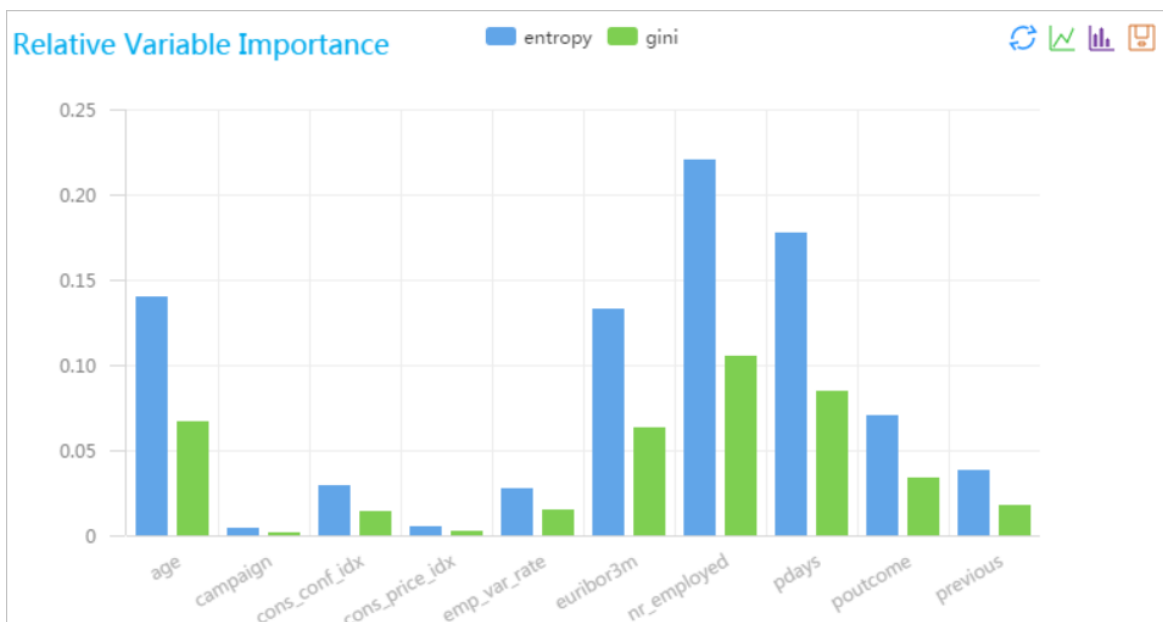
The data source is pai\_dense\_10\_10. y is the label column of the random forest model, and other columns are feature columns. Select age and campaign for **Forced Convert Columns**. This indicates that the two columns are processed as enumerated features, and default settings are retained for other columns.



3. Run the experiment and view the prediction result shown in the following figure.

| colname        | gini                  | entropy               |
|----------------|-----------------------|-----------------------|
| age            | 0.06625000000000003   | 0.13978726292803723   |
| campaign       | 0.0017500000000000003 | 0.004348515545596772  |
| cons_conf_idx  | 0.013999999999999999  | 0.02908409497018851   |
| cons_price_idx | 0.002                 | 0.0049804499913461255 |
| emp_var_rate   | 0.014700000000000003  | 0.026786360680260933  |
| euribor3m      | 0.06300000000000003   | 0.1321936348846039    |
| nr_employed    | 0.10499999999999998   | 0.2203227248076733    |
| pdays          | 0.0845                | 0.17750329234397513   |
| poutcome       | 0.03360000000000001   | 0.07050327193845542   |
| previous       | 0.017700000000000004  | 0.03810381005801592   |

4. After the running is complete, right-click Random Forest Feature Importance and select **View Analytics Report** to view the result.



## Linear Model Feature Importance

You can use the Linear Model Feature Importance component to calculate the feature importance for the linear model, such as linear regression and logistic regression for binary classification. Both the sparse and dense data formats are supported. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter                | Description                                                                                                                                                                                                                                                        |
|----------------|--------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns          | Optional. The feature columns selected from the input table for training. By default, all columns other than the label column are selected.                                                                                                                        |
|                | Target Column            | Required.<br>Click the  icon. In the <b>Select Column</b> dialog box, enter a keyword to search for your desired column. Then, select the searched column and click <b>OK</b> . |
|                | Input Sparse Format Data | Optional. Specifies whether data in the input table is in the sparse format.                                                                                                                                                                                       |
| Tuning         | Cores                    | Optional. The number of cores used in computing.                                                                                                                                                                                                                   |
|                | Memory Size per Core     | Optional. The memory size of each core. Unit: MB.                                                                                                                                                                                                                  |

- PAI command

```
PAI -name regression_feature_importance -project algo_public
-DmodelName=xlab_m_logisticregressi_20317_v0
-DoutputTableName=pai_temp_2252_20321_1
-DlabelColName=y
-DfeatureColNames=pdays,previous,emp_var_rate,cons_price_idx,cons_conf_idx,euribor3m,nr_employed,age,campaign
-DenableSparse=false -DinputTableName=pai_dense_10_9;
```

| Parameter       | Required | Description                                      | Default value    |
|-----------------|----------|--------------------------------------------------|------------------|
| inputTableName  | Yes      | The name of the input table.                     | No default value |
| outputTableName | Yes      | The name of the output table.                    | No default value |
| labelColName    | Yes      | The name of the label column in the input table. | No default value |
| modelName       | Yes      | The name of the input model.                     | No default value |

| Parameter            | Required | Description                                                                                      | Default value                           |
|----------------------|----------|--------------------------------------------------------------------------------------------------|-----------------------------------------|
| featureColNames      | No       | The feature columns selected from the input table for training.                                  | All columns other than the label column |
| inputTablePartitions | No       | The partitions selected from the input table for training.                                       | Full table                              |
| enableSparse         | No       | Specifies whether data in the input table is in the sparse format.                               | false                                   |
| itemDelimiter        | No       | The delimiter used between key-value pairs when data in the input table is in the sparse format. | Space                                   |
| kvDelimiter          | No       | The delimiter used between keys and values when data in the input table is in the sparse format. | :                                       |
| lifecycle            | No       | The lifecycle of the output table.                                                               | No default value                        |
| coreNum              | No       | The number of cores.                                                                             | Automatically allocated                 |
| memSizePerCore       | No       | The memory size of each core.                                                                    | Automatically allocated                 |

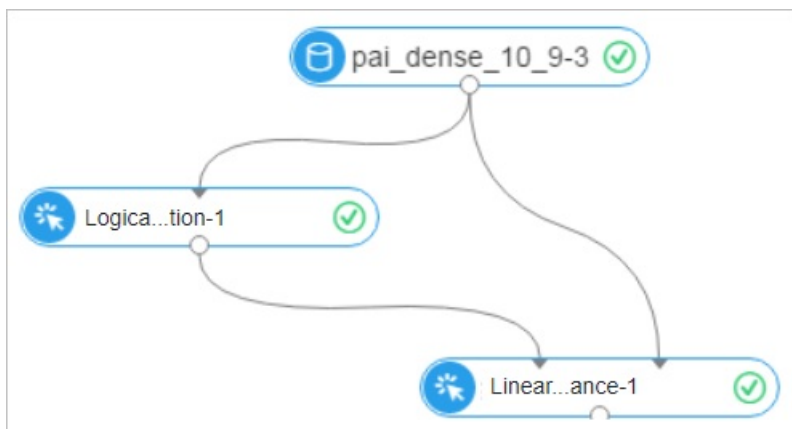
## Example

1. Execute the following SQL statement to generate training data:

```
create table if not exists pai_dense_10_9 as
select
  age,campaign,pdays,previous,emp_var_rate,cons_price_idx,cons_conf_idx,euribor3m,nr_employed,y
from bank_data limit 10;
```

2. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).

y in the preceding SQL statement is used as the label column of Logistic Regression for Multiclass Classification. Other fields in the statement are feature columns. Default values are retained for other parameters specified for the Logistic Regression for Multiclass Classification and Linear Model Feature Importance components.



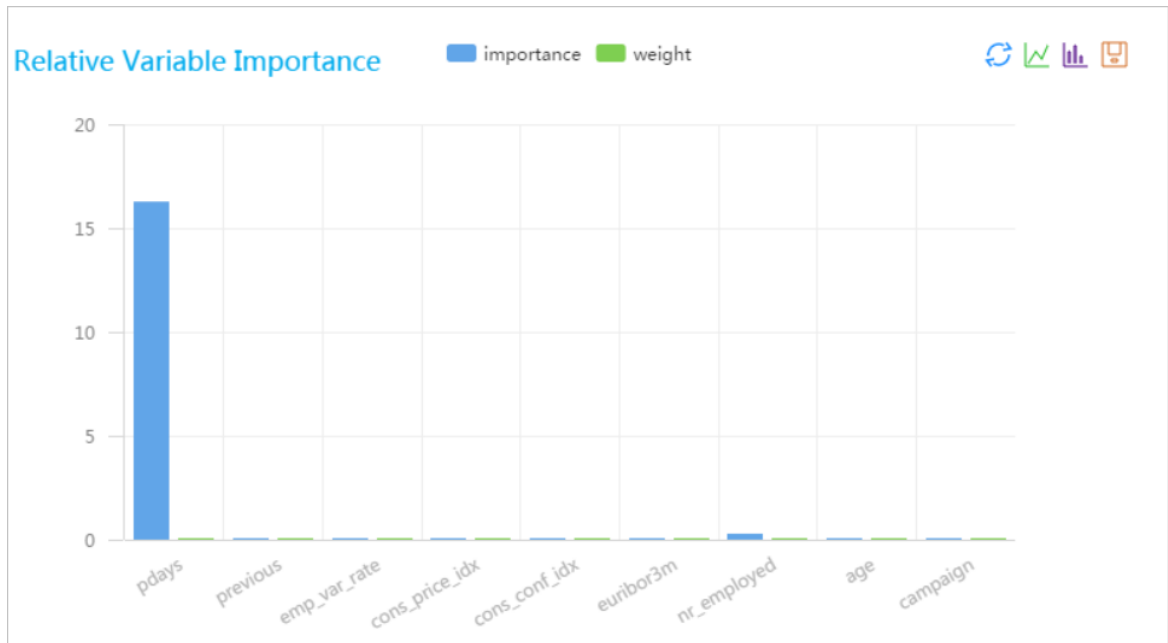
3. Run the experiment and view the prediction result shown in the following figure.

| colname        | weight                  | importance              |
|----------------|-------------------------|-------------------------|
| pdays          | 0.033942600256583334    | 16.31387797440866       |
| previous       | 0.00004248130342485344  | 0.000030038817725357177 |
| emp_var_rate   | 0.00006720242617694611  | 0.00010554561260753949  |
| cons_price_idx | 0.00012311047229142307  | 0.00006581255124425219  |
| cons_conf_idx  | 0.00017227965471819213  | 0.0008918770542818432   |
| euribor3m      | 0.00006113758212679113  | 0.00010427128177450988  |
| nr_employed    | 0.0034541377310490697   | 0.26048098230126043     |
| age            | 0.00009618162708080744  | 0.0009267659744232966   |
| campaign       | 0.000019142551785274455 | 0.000041793353660529855 |

The following table lists calculation formulas for metrics.

| Column     | Formula                               |
|------------|---------------------------------------|
| weight     | $\text{abs}(w\_)$                     |
| importance | $\text{abs}(w\_j) * \text{STD}(f\_i)$ |

4. After the running is complete, right-click Linear Model Feature Importance and select **View Analytics Report**. Then, view the result.



### 3.3.3. Feature Importance Filtering

The Feature Importance Filtering component provides the filtering feature for components including Linear Model Feature Importance, GBDT Feature Importance, and Random Forest Feature Importance. The Feature Importance Filtering component can be used to filter the top N features.

#### Configure the component

PAI command

```
PAI -name fe_filter_runner -project algo_public
-DselectedCols=pdays,previous,emp_var_rate,cons_price_idx,cons_conf_idx,euribor3m,nr_employed,age,campaign,poutcome
-DinputTable=pai_dense_10_10
-DweightTable=pai_temp_2252_20319_1
-DtopN=5
-DmodelTable=pai_temp_2252_20320_2
-DoutputTable=pai_temp_2252_20320_1;
```

| Parameter  | Description                  | Required |
|------------|------------------------------|----------|
| inputTable | The name of the input table. | Yes      |

| Parameter            | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Required |
|----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| inputTablePartitions | <p>By default, all partitions in the input table are selected. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>A single partition: <code>partition_name=value</code></li> <li>Multiple partitions: <code>name1=value1,name2=value2</code></li> </ul> <p> <b>Note</b> Multiple partitions are separated by commas (,).</p> <ul style="list-style-type: none"> <li>Multi-level partitions: <code>name1=value1/name2=value2</code></li> </ul> | No       |
| weightTable          | The weight tables of the feature importance. The weight tables are the output tables of Linear Model Feature Importance, GBDT Feature Importance, and Random Forest Feature Importance.                                                                                                                                                                                                                                                                                                                                                                                        | Yes      |
| outputTable          | The output table after the top N features are filtered.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Yes      |
| modelTable           | The model file generated by feature filtering.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Yes      |
| selectedCols         | By default, all the fields in the input table are selected.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | No       |
| topN                 | <p>The top N features that are filtered. Default value: 10.</p> <p> <b>Note</b> The value of this parameter must be a positive integer.</p>                                                                                                                                                                                                                                                                                                                                                 | No       |
| lifecycle            | <p>The lifecycle of the output table. Default value: 7.</p> <p> <b>Note</b> The value of this parameter must be a positive integer.</p>                                                                                                                                                                                                                                                                                                                                                     | No       |

### 3.3.4. Feature generation

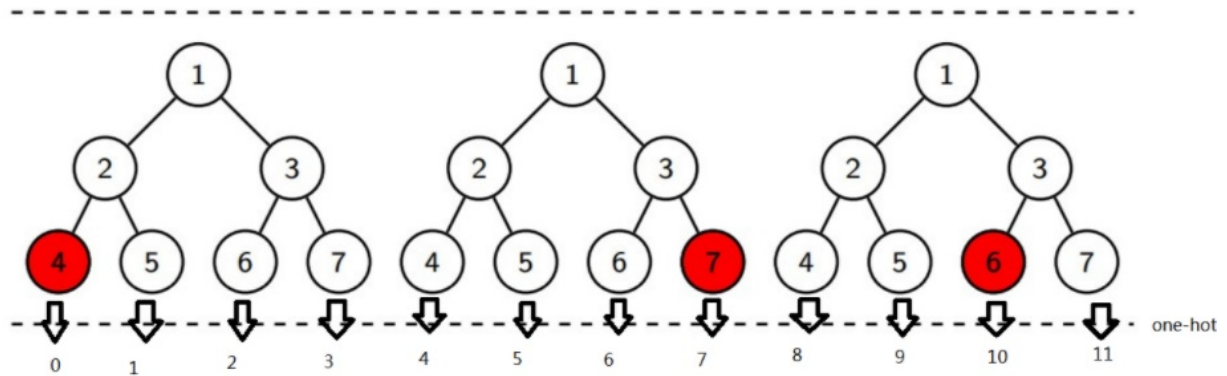
This topic describes the Feature Encoding and One Hot Encoding components provided by Machine Learning Studio for feature generation.

#### Feature Encoding



Feature Encoding is a technique that uses decision trees and ensemble methods to explore new features. The features are one-hot encoding results of the leaf nodes of decision trees. These nodes are composed of one or more features.

The following figure shows the feature encoding process. The three trees in the figure have 18 leaf nodes. The trees are encoded as features 0 to 11 successively. The first tree is encoded as features 0 to 3, the second tree is encoded as features 4 to 7, and the third tree is encoded as features 8 to 11. Feature Encoding can convert GBDT non-linear features to linear features.



You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter               | Description                                                                                                                                                                                                                                                                        |
|--------------------|-------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns         | The feature columns selected from the input table for training.                                                                                                                                                                                                                    |
|                    | Label Column            | Required. The label column.<br>Click the  icon. In the <b>Select Column</b> dialog box, enter the keywords of the column that you want to search for. Select the column and click <b>OK</b> . |
|                    | Appended Output Columns | Optional. The original features reserved in the output table.                                                                                                                                                                                                                      |
| Parameters Setting | Cores                   | The number of cores used in computing. The value of this parameter must be a positive integer.                                                                                                                                                                                     |
|                    | Memory size per core    | The memory size of each core. The value of this parameter must be a positive integer.                                                                                                                                                                                              |

- PAI command

```
PAI -name fe_encode_runner -project algo_public
-DinputTable="pai_temp_2159_19087_1"
-DencodeModel="xlab_m_GBDT_LR_1_19064"
-DselectedCols="pdays,previous,emp_var_rate,cons_price_idx,cons_conf_idx,euribor3m,nr_employed,age,campaign"
-DlabelCol="y"
-DoutputTable="pai_temp_2159_19061_1";
-DcoreNum=10
-DmemSizePerCore=1024
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                      | Default value                     |
|----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTable           | Yes      | The name of the input table.                                                                                                                                                                                                                                     | No default value                  |
| inputTablePartitions | No       | The partitions selected from the input table for training. Specify this parameter in the partition_name=value format or the name1=value1/name2=value2 format (multi-level partitions).<br><br>If you specify multiple partitions, separate them with commas (,). | All partitions in the input table |
| encodeModel          | Yes      | The encoded GBDT binary classification model that is imported.                                                                                                                                                                                                   | No default value                  |
| outputTable          | Yes      | The output table after scaling.                                                                                                                                                                                                                                  | No default value                  |
| selectedCols         | Yes      | The features that are encoded by using GBDT. These features are the training features of GBDT components.                                                                                                                                                        | No default value                  |
| labelCol             | Yes      | The label column.                                                                                                                                                                                                                                                | No default value                  |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                                               | 7                                 |

| Parameter      | Required | Description                                                                       | Default value                                                                                                   |
|----------------|----------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| coreNum        | No       | The number of cores. The value of this parameter must be of the BIGINT data type. | -1 (If you retain the default value, the system determines the number of cores based on the input data volume.) |
| memSizePerCore | No       | The memory size of each core.                                                     | -1 (If you retain the default value, the system determines the memory size based on the input data volume.)     |

## Example

1. Execute the following SQL statement to generate training data:

```

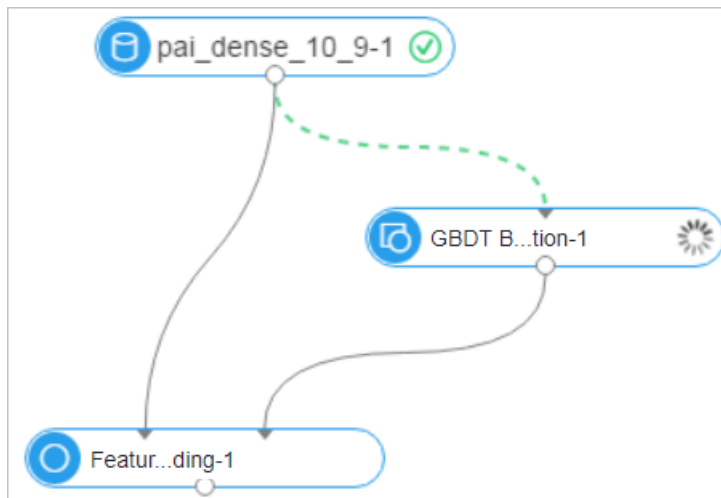
CREATE TABLE IF NOT EXISTS tdl_pai_bank_test1
(
  age      BIGINT COMMENT "",
  campaign BIGINT COMMENT "",
  pdays   BIGINT COMMENT "",
  previous BIGINT COMMENT "",
  emp_var_rate DOUBLE COMMENT "",
  cons_price_idx DOUBLE COMMENT "",
  cons_conf_idx DOUBLE COMMENT "",
  euribor3m DOUBLE COMMENT "",
  nr_employed DOUBLE COMMENT "",
  y        BIGINT COMMENT ""
)
LIFECYCLE 7;
insert overwrite table tdl_pai_bank_test1
select * from
(select 53 as age,1 as campaign,999 as pdays,0 as previous,-0.1 as emp_var_rate,
  93.2 as cons_price_idx,-42.0 as cons_conf_idx, 4.021 as euribor3m,5195.8 as nr_employed,0 as y
from dual
union all
select 28 as age,3 as campaign,6 as pdays,2 as previous,-1.7 as emp_var_rate,
  94.055 as cons_price_idx,-39.8 as cons_conf_idx, 0.729 as euribor3m,4991.6 as nr_employed,1 as y
from dual
union all
select 39 as age,2 as campaign,999 as pdays,0 as previous,-1.8 as emp_var_rate,
  93.075 as cons_price_idx,-47.1 as cons_conf_idx, 1.405 as euribor3m,5099.8 as nr_employed,0 as y
from dual
union all
select 55 as age,1 as campaign,3 as pdays,1 as previous,-2.9 as emp_var_rate,
  92.201 as cons_price_idx,-31.4 as cons_conf_idx, 0.869 as euribor3m,5076.2 as nr_employed,1 as y
from dual
union all
select 30 as age,8 as campaign,999 as pdays,0 as previous,1.4 as emp_var_rate,
  93.918 as cons_price_idx,-42.7 as cons_conf_idx, 4.961 as euribor3m,5228.2 as nr_employed,0 as y
from dual
union all
select 37 as age,1 as campaign,999 as pdays,0 as previous,-1.8 as emp_var_rate,
  92.893 as cons_price_idx,-46.2 as cons_conf_idx, 1.327 as euribor3m,5099.1 as nr_employed,0 as y
from dual
union all
select 39 as age,1 as campaign,999 as pdays,0 as previous,-1.8 as emp_var_rate,
  92.893 as cons_price_idx,-46.2 as cons_conf_idx, 1.313 as euribor3m,5099.1 as nr_employed,0 as y
from dual
union all
select 36 as age,1 as campaign,3 as pdays,1 as previous,-2.9 as emp_var_rate,
  92.963 as cons_price_idx,-40.8 as cons_conf_idx, 1.266 as euribor3m,5076.2 as nr_employed,1 as y
from dual
union all
select 27 as age,2 as campaign,999 as pdays,1 as previous,-1.8 as emp_var_rate,
  93.075 as cons_price_idx,-47.1 as cons_conf_idx, 1.41 as euribor3m,5099.1 as nr_employed,0 as y
from dual
) a

```

2. Create the experiment shown in the following figure. The GBDT Binary Classification component is

used in this experiment. For more information, see [Algorithm modeling](#).

Configure parameters for the GBDT Binary Classification component. Set Trees to 5, Maximum Tree Depth to 3, Label Column to y, and Feature Columns to other fields.



3. Run the experiment and view the prediction result.

| kv                                                                  | y   |
|---------------------------------------------------------------------|-----|
| 2:1,5:1,8:1,12:1,15:1,18:1,28:1,34:1,41:1,50:1,53:1,63:1,72:1       | 0.0 |
| 2:1,5:1,6:1,12:1,15:1,16:1,28:1,34:1,41:1,50:1,51:1,63:1,72:1       | 0.0 |
| 2:1,3:1,12:1,13:1,28:1,34:1,36:1,39:1,55:1,61:1                     | 1.0 |
| 2:1,3:1,12:1,13:1,20:1,21:1,22:1,42:1,43:1,46:1,63:1,64:1,67:1,68:1 | 0.0 |
| 0:1,10:1,28:1,29:1,32:1,36:1,37:1,55:1,56:1,59:1                    | 1.0 |

The generated result can be imported to the Logistic Regression for Binary Classification or Logistic Regression for Multiclass Classification component. This provides better performance than Linear Regression and GBDT Regression and can avoid overfitting of the generated result.

## One Hot Encoding

The One Hot Encoding component provides the training and prediction features.

- Training feature:
  - Input nodes: The first (left) input node of this component is the input of training data, and the second (right) input node does not need to be set during the training.
  - Output nodes: This component has two output nodes. The left one is the encoded output table and the right one is the output model table. The output model table is used to perform one-hot encoding for new data of the same type.
- Prediction feature:

The second (right) input node of the One Hot Encoding component is used to import a one-hot encoding model. An existing one-hot encoding model can be used to encode new data.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter                                  | Description                                                                                                                                                                                                                                                                                                                                               |
|--------------------|--------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | <b>Binarization Column</b>                 | Required. The fields that require binarization.                                                                                                                                                                                                                                                                                                           |
|                    | <b>Other Reserved Features</b>             | The features that are reserved and exported in the key-value format. The selected fields are exported as features in the key-value format. The fields must be of the DOUBLE data type. They are not subject to one-hot encoding and are encoded from 0.                                                                                                   |
|                    | <b>Appended Columns</b>                    | Optional. The columns appended to the output table.                                                                                                                                                                                                                                                                                                       |
| Parameters Setting | <b>Lifecycle</b>                           | The lifecycle of the output table. Default value: 7.                                                                                                                                                                                                                                                                                                      |
|                    | <b>Output table type</b>                   | The type of the output table. Valid values: KV and Table. If the number of the features that require discretization is large, we recommend that you set this parameter to KV. If you set this parameter to Table, the output table can contain a maximum of 1,024 columns. If the number of the exported columns exceeds the value, an error is reported. |
|                    | <b>Cores</b>                               | The number of cores.                                                                                                                                                                                                                                                                                                                                      |
|                    | <b>Memory Size per Node</b>                | The memory size of each core. Unit: MB.                                                                                                                                                                                                                                                                                                                   |
|                    | <b>Delete Encoding of Last Enumeration</b> | If this parameter is set to true, the linear independence of the encoded data is ensured.                                                                                                                                                                                                                                                                 |
|                    | <b>Ignore Empty Elements</b>               | If this parameter is set to true, empty elements are not encoded.                                                                                                                                                                                                                                                                                         |

- PAI command

```
PAI -name one_hot
-project algo_public
-DinputTable=one_hot_test
-DbinaryCols=f0,f1,f2
-DmodelTable=one_hot_model
-DoutputTable=one_hot_output
-Dlifecycle=28;
```

| Parameter            | Required | Description                                                                                                                                                                             | Default value                     |
|----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTable           | Yes      | The name of the input table.                                                                                                                                                            | No default value                  |
| inputTablePartitions | No       | The partitions selected from the input table for training.                                                                                                                              | All partitions in the input table |
| binaryCols           | Yes      | The fields that require one-hot encoding. These fields must be enumerated features and their data types are not limited.                                                                | No default value                  |
| reserveCols          | No       | The selected fields that are exported as features in the key-value format. The fields must be of the DOUBLE data type. They are not subject to one-hot encoding and are encoded from 0. | Empty string                      |
| appendCols           | No       | The selected fields that are exported to the output table the same as they are in the input table.                                                                                      | No default value                  |
| outputTable          | Yes      | The output table generated after one-hot encoding. The encoding result is stored in the key-value format.                                                                               | No default value                  |

| Parameter        | Required | Description                                                                                                                                                                                                                           | Default value |
|------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| inputModelTable  | No       | <p>The input model table for one-hot encoding.</p> <p> <b>Note</b> The value of inputModelTable or outputModelTable must be a non-empty string.</p>  | Empty string  |
| outputModelTable | No       | <p>The output model table for one-hot encoding.</p> <p> <b>Note</b> The value of inputModelTable or outputModelTable must be a non-empty string.</p> | Empty string  |
| lifecycle        | No       | The lifecycle of the output table.                                                                                                                                                                                                    | 7             |
| dropLast         | Yes      | Specifies whether to delete the encoding result of the last enumerator. If this parameter is set to true, the linear independence of the encoded data is ensured.                                                                     | <i>false</i>  |



| Parameter       | Required | Description                                                                                                                                                                                                                                                                                                                                               | Default value           |
|-----------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| outputTableType | Yes      | The type of the output table. Valid values: KV and Table. If the number of the features that require discretization is large, we recommend that you set this parameter to KV. If you set this parameter to Table, the output table can contain a maximum of 1,024 columns. If the number of the exported columns exceeds the value, an error is reported. | KV                      |
| ignoreNull      | Yes      | Specifies whether to ignore empty elements in the data that requires encoding. If this parameter is set to <i>true</i> , empty elements are not encoded.                                                                                                                                                                                                  | <i>false</i>            |
| coreNum         | No       | The number of cores.                                                                                                                                                                                                                                                                                                                                      | Automatically allocated |
| memSizePerCore  | No       | The memory size of each core. Unit: MB. Valid values: [2048, 64 × 1024].                                                                                                                                                                                                                                                                                  | Automatically allocated |

**Instructions:**

- The value of the inputModelTable or outputModelTable parameter must be a non-empty string. If the value of the inputModelTable parameter is a non-empty string, the table that is indicated by the parameter is a non-empty model table.
- For binaryCols, you can specify tens of millions of values for discretization.
- If you use the training model as the model for encoding next time, you cannot modify the values of the dropLast, ignoreNull, and reserveCols parameters. This is because the output results generated by specifying the three parameters are integrated into the model. If you want to modify the three parameters, you must train the model again.
- We recommend that you export the output table in the key-value format. If you use the table format, you can export a maximum of 1,024 columns. If the number of the exported columns exceed the value, an error is reported and encoding fails.
- By default, the output table in the key-value format generated after one-hot encoding is encoded from 0.

- If you use the trained model to encode new data and cannot find the discrete magnitude of the data in the model mapping table, the discrete magnitude is ignored. This indicates that the discrete magnitude is not encoded. If you need to encode the discrete magnitude, you must train the model mapping table again.

## Example

1. Execute the following SQL statement to generate training data:

```
PAI -project projectxlib4
-name one_hot
-DinputTable=one_hot_yh
-DbinaryCols=f0,f2,f4
-DoutputModelTable=one_hot_model_8
-DoutputTable=one_hot_in_table_1_output_8
-DdropLast=false
-DappendCols=f0,f2,f4
-DignoreNull=false
-DoutputTableType=table
-DreserveCols=f3
-DcoreNum=4
-DmemSizePerCore=2048;
```

2. Import the data listed in the following table for a test.

| f0   | f1      | f2                     | f3  | f4    |
|------|---------|------------------------|-----|-------|
| 12   | prefix1 | 1970-09-15<br>12:50:22 | 0.1 | true  |
| 12   | prefix3 | 1971-01-22<br>03:15:33 | 0.4 | false |
| NULL | prefix3 | 1970-01-01<br>08:00:00 | 0.2 | NULL  |
| 3    | NULL    | 1970-01-01<br>08:00:00 | 0.3 | false |
| 34   | NULL    | 1970-09-15<br>12:50:22 | 0.4 | NULL  |
| 3    | prefix1 | 1970-09-15<br>12:50:22 | 0.2 | true  |
| 3    | prefix1 | 1970-09-15<br>12:50:22 | 0.3 | false |
| 3    | prefix3 | 1970-01-01<br>08:00:00 | 0.2 | true  |
| 3    | prefix3 | 1971-01-22<br>03:15:33 | 0.1 | false |

| f0   | f1      | f2                     | f3  | f4    |
|------|---------|------------------------|-----|-------|
| NULL | prefix3 | 1970-01-01<br>08:00:00 | 0.3 | false |

In the preceding input table, the f0 column is of the BIGINT data type, the f1 column is of the STRING data type, the f2 column is of the DATETIME data type, the f3 column is of the DOUBLE data type, and the f4 column is of the BOOLEAN data type.

3. Obtain the following model mapping table in the test result.

| col_name  | col_value  | mapping |
|-----------|------------|---------|
| _reserve_ | f3         | 0       |
| f0        | 12         | 1       |
| f0        | 3          | 2       |
| f0        | 34         | 3       |
| f0        | null       | 4       |
| f2        | 2222222222 | 5       |
| f2        | 3333333333 | 6       |
| f2        | 4          | 7       |
| f4        | 0          | 8       |
| f4        | 1          | 9       |
| f4        | null       | 10      |

The top row in the model mapping table is the reserve row, and the column name is fixed to reserve. This row stores reserve information. The remaining rows correspond to the mapping information of the encoding.

o Encoded table in the table format

| f0 | f1      | f3  | f4   | _reserve_f3_0 | f0_1_2_1 | f0_3_2_2 | f0_3_4_3 | f0_nnull_4 | f2_2_22_22_22_2_5 | f2_3_33_33_33_3_6 | f2_4_7 | f4_0_8 | f4_1_9 | f4_nnull_10 |
|----|---------|-----|------|---------------|----------|----------|----------|------------|-------------------|-------------------|--------|--------|--------|-------------|
| 12 | prefix1 | 0.1 | true | 0.1           | 1.0      | 0.0      | 0.0      | 0.0        | 1.0               | 0.0               | 0.0    | 0.0    | 1.0    | 0.0         |

| f0   | f1      | f3  | f4    | _reserve_f3_0 | f0_1_2_1 | f0_3_2 | f0_3_4_3 | f0_null_4 | f2_2_22_22_22_2_5 | f2_3_33_33_33_3_6 | f2_4_7 | f4_0_8 | f4_1_9 | f4_null_1_0 |
|------|---------|-----|-------|---------------|----------|--------|----------|-----------|-------------------|-------------------|--------|--------|--------|-------------|
| 12   | prefix3 | 0.4 | false | 0.4           | 1.0      | 0.0    | 0.0      | 0.0       | 0.0               | 1.0               | 0.0    | 1.0    | 0.0    | 0.0         |
| NULL | prefix3 | 0.2 | NULL  | 0.2           | 0.0      | 0.0    | 0.0      | 1.0       | 0.0               | 0.0               | 1.0    | 0.0    | 0.0    | 1.0         |
| 3    | NULL    | 0.3 | false | 0.3           | 0.0      | 1.0    | 0.0      | 0.0       | 0.0               | 0.0               | 1.0    | 1.0    | 0.0    | 0.0         |
| 34   | NULL    | 0.4 | NULL  | 0.4           | 0.0      | 0.0    | 1.0      | 0.0       | 1.0               | 0.0               | 0.0    | 0.0    | 0.0    | 1.0         |
| 3    | prefix1 | 0.2 | true  | 0.2           | 0.0      | 1.0    | 0.0      | 0.0       | 1.0               | 0.0               | 0.0    | 0.0    | 1.0    | 0.0         |
| 3    | prefix1 | 0.3 | false | 0.3           | 0.0      | 1.0    | 0.0      | 0.0       | 1.0               | 0.0               | 0.0    | 1.0    | 0.0    | 0.0         |
| 3    | prefix3 | 0.2 | true  | 0.2           | 0.0      | 1.0    | 0.0      | 0.0       | 0.0               | 0.0               | 1.0    | 0.0    | 1.0    | 0.0         |
| 3    | prefix3 | 0.1 | false | 0.1           | 0.0      | 1.0    | 0.0      | 0.0       | 0.0               | 1.0               | 0.0    | 1.0    | 0.0    | 0.0         |
| NULL | prefix3 | 0.3 | false | 0.3           | 0.0      | 0.0    | 0.0      | 1.0       | 0.0               | 0.0               | 1.0    | 1.0    | 0.0    | 0.0         |

- Encoded table in the key-value format

| f0   | f1      | f3  | f4    | kv                 |
|------|---------|-----|-------|--------------------|
| 12   | prefix1 | 0.1 | true  | 0:0.1,1:1,5:1,9:1  |
| 12   | prefix3 | 0.4 | false | 0:0.4,1:1,6:1,8:1  |
| NULL | prefix3 | 0.2 | NULL  | 0:0.2,4:1,7:1,10:1 |
| 3    | NULL    | 0.3 | false | 0:0.3,2:1,7:1,8:1  |

| f0   | f1      | f3  | f4    | kv                 |
|------|---------|-----|-------|--------------------|
| 34   | NULL    | 0.4 | NULL  | 0:0.4,3:1,5:1,10:1 |
| 3    | prefix1 | 0.2 | true  | 0:0.2,2:1,5:1,9:1  |
| 3    | prefix1 | 0.3 | false | 0:0.3,2:1,5:1,8:1  |
| 3    | prefix3 | 0.2 | true  | 0:0.2,2:1,7:1,9:1  |
| 3    | prefix3 | 0.1 | false | 0:0.1,2:1,6:1,8:1  |
| NULL | prefix3 | 0.3 | false | 0:0.3,4:1,7:1,8:1  |

## Scalability test

Test data: The number of samples is 200 million, and the number of enumerators is 100,000. The test data is listed in the following table.

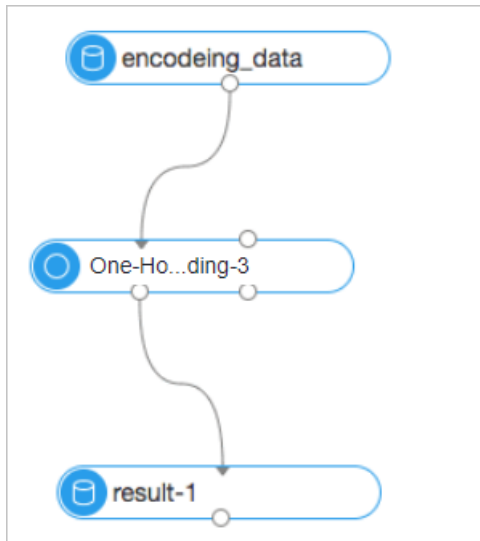
| f0   | f1         |
|------|------------|
| 94   | prefix3689 |
| 9664 | prefix5682 |
| 2062 | prefix5530 |
| 9075 | prefix9854 |
| 9836 | prefix1764 |
| 5140 | prefix1149 |
| 3455 | prefix7272 |
| 2508 | prefix7139 |
| 7993 | prefix1551 |
| 5602 | prefix4606 |
| 3132 | prefix5767 |

The test result is listed in the following table.

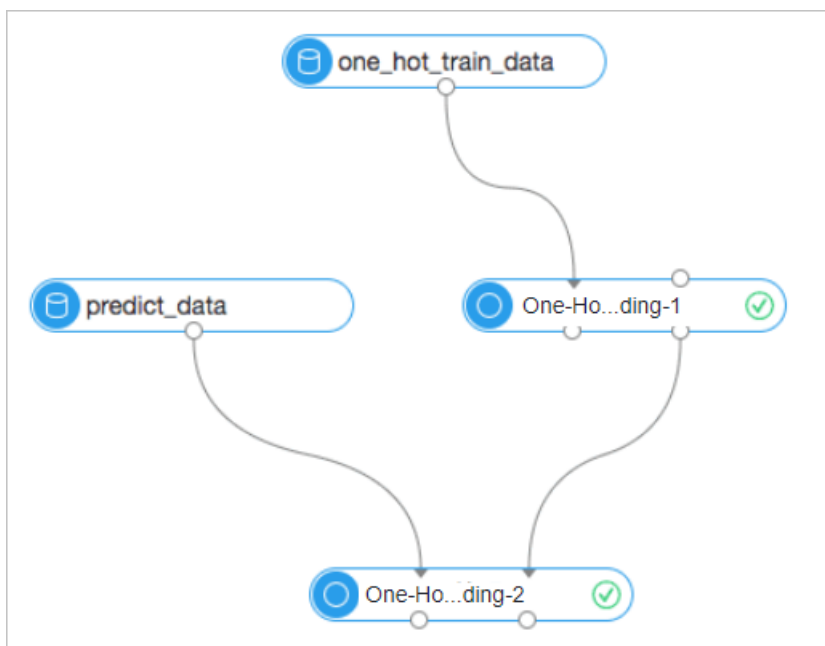
| core num | train time | predict time | Acceleration ratio |
|----------|------------|--------------|--------------------|
| 5        | 84s        | 181s         | 1/1                |
| 10       | 60s        | 93s          | 1.4/1.95           |
| 20       | 46s        | 56s          | 1.8/3.23           |

Usage notes:

- Use the One Hot Encoding component to encode data. The following figure shows the experiment process.



- Use the trained model of the One Hot Encoding component to encode data. The following figure shows the experiment process.



### 3.3.5. Anomaly Detection

The Anomaly Detection component is used to detect data with continuous or enumerated features. It helps you locate exceptions in the data.

#### Background information

Anomalous features in data can be detected by using the Box Plot or Attribute Value Frequency (AVF) method.

- Box Plot is used to detect data with continuous features. The detection is performed based on the

maximum and minimum values of the box plot.

- AVF is used to detect data with enumerated features. The detection is performed based on the frequency and threshold of enumerated features.

## Configure the component

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter                | Description                                                                                                                                              |
|----------------|--------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns          | The fields that you want to analyze.                                                                                                                     |
|                | Anomaly Detection Method | The method used to detect anomalous data. Box Plot is used to detect data with continuous features. AVF is used to detect data with continuous features. |

- PAI command

```
PAI -name fe_detect_runner -project algo_public
-DselectedCols="emp_var_rate,cons_price_rate,cons_conf_idx,euribor3m,nr_employed" \
-Dlifecycle="28"
-DdetectStrategy="boxPlot"
-DmodelTable="pai_temp_2458_23565_2"
-DinputTable="pai_bank_data"
-DoutputTable="pai_temp_2458_23565_1";
```

| Parameter            | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Required |
|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| inputTable           | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Yes      |
| inputTablePartitions | <p>The partitions selected from the input table for training. By default, all partitions in the input table are selected. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>◦ A single partition: <code>partition_name=value</code></li> <li>◦ Multiple partitions: <code>name1=value1,name2=value2</code></li> </ul> <div>  <b>Note</b> Multiple partitions are separated by commas (,).         </div> <ul style="list-style-type: none"> <li>◦ Multi-level partitions: <code>name1=value1/name2=value2</code></li> </ul> | No       |

| Parameter      | Description                                                                                                                                                                                                                                                              | Required |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| selectedCols   | The input features. The data types of the features are not limited.                                                                                                                                                                                                      | Yes      |
| detectStrategy | The detection method. Box Plot and AVF are supported. Box Plot is used to detect data with continuous features. AVF is used to detect data with enumerated features.                                                                                                     | Yes      |
| outputTable    | The output table that contains data with anomalous features.                                                                                                                                                                                                             | Yes      |
| modelTable     | The anomaly detection model.                                                                                                                                                                                                                                             | Yes      |
| lifecycle      | The lifecycle of the output table. Default value: 7.                                                                                                                                                                                                                     | No       |
| coreNum        | The number of cores. This parameter is used with memSizePerCore.<br><div>  <b>Note</b> The value of this parameter must be a positive integer. Valid values: [1,9999].         </div> | No       |
| memSizePerCore | The memory size of each core. Unit: MB. Valid values: [2048,64 × 1024].                                                                                                                                                                                                  | No       |

### 3.3.6. Use the Binning component to implement the discretization of continuous features

This topic describes how to use the Binning component to implement the discretization of continuous features.

#### Prerequisites

A project is created. For more information, see [Create a project](#).

#### Context

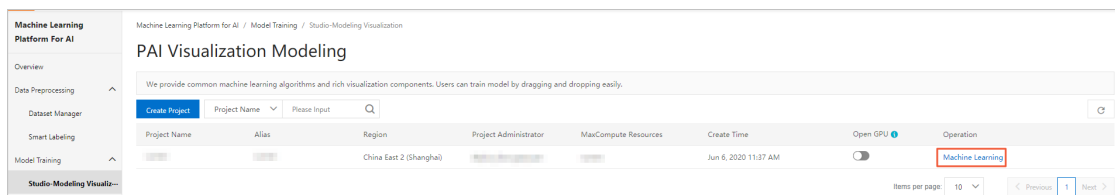
Feature discretization is a process of converting continuous data into multiple discrete intervals. To implement feature discretization, Machine Learning Platform for AI (PAI) provides the Binning component. This component supports the following binning modes: equal frequency binning, equal width binning, and automated binning.



In this topic, the **Read MaxCompute Table** component is used to read data from the `pai_online_project.iris_data` table. Then, the **Binning** component is used to put the data that is read into different bins. Finally, the **Data Conversion Module** component is used to convert the original continuous data in the bins into discrete data.

## Procedure

1. Go to the Machine Learning Studio console.
  - i. Log on to the **PAI console**.
  - ii. In the left-side navigation pane, choose **Model Training > Studio-Modeling Visualization**.
  - iii. On the **PAI Visualization Modeling** page, find the project in which you want to create an experiment and click **Machine Learning** in the Operation column.

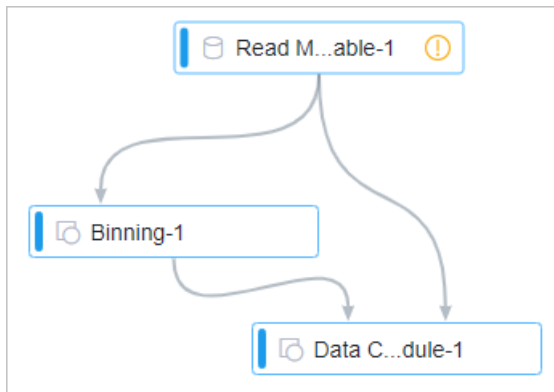


2. Create an experiment.
  - i. In the left-side navigation pane, click **Home**.
  - ii. Choose **New > New Experiment**.
  - iii. In the **New Experiment** dialog box, configure the parameters.

| Parameter          | Description                                                                                             |
|--------------------|---------------------------------------------------------------------------------------------------------|
| <b>Name</b>        | Enter Use the Binning component for the discretization of continuous features.                          |
| <b>Project</b>     | The name of the project to which the experiment belongs. You cannot change the value of this parameter. |
| <b>Description</b> | Enter Use the Binning component provided by PAI for the discretization of continuous features.          |
| <b>Save To</b>     | Click My Experiments.                                                                                   |

- iv. Click **OK**.
3. Configure the experiment.
    - i. In the left-side navigation pane, click **Components**.
    - ii. In the navigation tree, click **Data Source/Target**. Then, drag and drop the **Read MaxCompute Table** component onto the canvas.
    - iii. In the navigation tree, click **Financials**. Then, drag and drop the **Binning** and **Data Conversion Module** components onto the canvas.

iv. Connect the preceding components.



4. Configure component parameters.

- i. Click the **Read MaxCompute Table** component on the canvas. In the right-side pane, configure the following parameters.

| Tab                | Parameter            | Description                                                                                                                                                                                                                 |
|--------------------|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Table Selection    | Table Name           | Enter pai_online_project.iris_data.                                                                                                                                                                                         |
|                    | Partition            | The pai_online_project.iris_data table is not a partitioned table. Therefore, the <b>Partition</b> check box is dimmed.                                                                                                     |
| Fields Information | Source Table Columns | You do not need to manually specify this parameter. After you specify <b>Table Name</b> , the system synchronizes the information of columns in the table specified by Table Name to the <b>Source Table Columns</b> field. |

- ii. Click the **Binning** component on the canvas. In the right-side pane, configure the parameters listed in the following table and leave other parameters at their default values.

| Tab                | Parameter       | Description                                                                                                                                         |
|--------------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns | Select the f1, f2, f3, and f4 columns.                                                                                                              |
| Parameters Setting | Bins            | Set this parameter to 10. This value indicates that continuous features are converted into 10 discrete intervals.                                   |
|                    | Binning Mode    | Valid values: <b>Equal Frequency</b> , <b>Equal Width</b> , and <b>Automatic Binning</b> . In this topic, the value <b>Equal Frequency</b> is used. |

- iii. Click the **Data Conversion Module** component on the canvas. In the right-side pane, configure the parameters listed in the following table and leave other parameters at their default values.

| Tab            | Parameter                       | Description                                                                                 |
|----------------|---------------------------------|---------------------------------------------------------------------------------------------|
| Fields Setting | Columns without Data Conversion | Select the type column. Data in the output of this column is the same as that in the input. |
|                | Data Conversion Mode            | Select Index.                                                                               |

5. In the top toolbar of the canvas, click **Run**.
6. View the experiment results.
  - i. After the experiment is executed, right-click the **Data Conversion Module** component on the canvas and select **View Data**. Then, you can view the discretization results.

| type ▲      | f1 ▲ | f2 ▲ | f3 ▲ | f4 ▲ |
|-------------|------|------|------|------|
| Iris-setosa | 2    | 7    | 0    | 0    |
| Iris-setosa | 1    | 3    | 0    | 0    |
| Iris-setosa | 0    | 5    | 0    | 0    |
| Iris-setosa | 0    | 4    | 1    | 0    |
| Iris-setosa | 1    | 7    | 0    | 0    |
| Iris-setosa | 3    | 8    | 2    | 1    |
| Iris-setosa | 0    | 6    | 0    | 1    |
| Iris-setosa | 1    | 6    | 1    | 0    |
| Iris-setosa | 0    | 3    | 0    | 0    |
| Iris-setosa | 1    | 4    | 1    | 0    |
| Iris-setosa | 3    | 8    | 1    | 0    |
| Iris-setosa | 0    | 6    | 2    | 0    |
| Iris-setosa | 0    | 3    | 0    | 0    |
| Iris-setosa | 0    | 3    | 0    | 0    |
| Iris-setosa | 4    | 8    | 0    | 0    |
| Iris-setosa | 4    | 8    | 1    | 1    |
| Iris-setosa | 3    | 8    | 0    | 1    |

- ii. Right-click the **Binning** component on the canvas and select **Binning**.
- iii. Click **Details** in the **Action** column that corresponds to the feature you want to view. The f1 feature is used in this example.
- iv. Click the **Charts** tab to view the binning results.

## 3.4. Statistical analysis

## 3.4.1. Histogram (Multiple Columns)

A histogram is also known as a mass distribution profile. A histogram is a statistical report chart that consists of a series of vertical stripes or line segments with different heights to show the data distribution. The horizontal axis represents the data types, and the vertical axis represents the data distribution.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter            | Description                                                                                                                                                                                                                                       |
|--------------------|----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Select Column        | Select the columns to be analyzed. Only the DOUBLE and BIGINT types are supported. <div>  <b>Note</b> A maximum of 1,024 columns are supported.         </div> |
| Parameters Setting | Intervals            | The number of intervals into which the histogram is divided.                                                                                                                                                                                      |
| Tuning             | Cores                | The number of cores. The parameter value must be a positive integer.                                                                                                                                                                              |
|                    | Memory Size per Core | The memory size of each core, in MB. Valid values: 1 to 65536.                                                                                                                                                                                    |

- PAI command

```
PAI -name histogram
  -project algo_public
  -DinputTableName=maple_histogram_1to20_input
  -DoutputTableName=maple_histogram_1to20_output
  -DselectedColNames=col0,col1 -DintervalNum=20;
```

| Parameter      | Required | Description                  | Default value    |
|----------------|----------|------------------------------|------------------|
| inputTableName | Yes      | The name of the input table. | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                 | Default value           |
|----------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> Separate multiple partitions with commas (,).</p> | No default value        |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                               | No default value        |
| selectedColumnNames  | Yes      | <p>The names of the columns selected from the input table for training. Separate the names of multiple columns with commas (,). The INT and DOUBLE types are supported.</p> <p> <b>Note</b> A maximum of 1,024 columns are supported.</p>                                                                 | No default value        |
| intervalNum          | No       | The number of intervals into which the histogram is divided.                                                                                                                                                                                                                                                                                                                                | 100                     |
| lifecycle            | No       | The lifecycle of the table.                                                                                                                                                                                                                                                                                                                                                                 | No default value        |
| coreNum              | No       | The number of cores. The parameter value must be a positive integer. Valid values: [1,9999].                                                                                                                                                                                                                                                                                                | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core, in MB. Valid values: 1 to 65536.                                                                                                                                                                                                                                                                                                                              | Automatically allocated |

## Example

### • Input

| col0 | col1 |
|------|------|
| 1    | 1.0  |
| 2    | 2.0  |
| 3    | 3.0  |
| 4    | 4.0  |

| col0 | col1 |
|------|------|
| 5    | 5.0  |
| 6    | 6.0  |
| 7    | 7.0  |
| 8    | 8.0  |
| 9    | 9.0  |
| 10   | 10.0 |
| 11   | 11.0 |
| 12   | 12.0 |
| 13   | 13.0 |
| 14   | 14.0 |
| 15   | 15.0 |
| 16   | 16.0 |
| 17   | 17.0 |
| 18   | 18.0 |
| 19   | 19.0 |
| 20   | 20.0 |

- PAI command

```
PAI -name histogram
    -project algo_public
    -DinputTableName=maple_histogram_1to20_input
    -DoutputTableName=maple_histogram_1to20_output
    -DselectedColNames=col0,col1 -DintervalNum=20;
```

- Output

| colname | histogram                                                                                                                                                                                                                                                                                                |
|---------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| col0    | [1, 1.95):1;[1.95, 2.9):1;[2.9, 3.85):1;[3.85, 4.8):1;[4.8, 5.75):1;[5.75, 6.7):1;[6.7, 7.65):1;[7.65, 8.6):1;[8.6, 9.55):1;[9.55, 10.5):1;[10.5, 11.45):1;[11.45, 12.4):1;[12.4, 13.35):1;[13.35, 14.3):1;[14.3, 15.25):1;[15.25, 16.2):1;[16.2, 17.15):1;[17.15, 18.1):1;[18.1, 19.05):1;[19.05, 20]:1 |

| colname | histogram                                                                                                                                                                                                                                                                                                |
|---------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| col1    | [1, 1.95):1;[1.95, 2.9):1;[2.9, 3.85):1;[3.85, 4.8):1;[4.8, 5.75):1;[5.75, 6.7):1;[6.7, 7.65):1;[7.65, 8.6):1;[8.6, 9.55):1;[9.55, 10.5):1;[10.5, 11.45):1;[11.45, 12.4):1;[12.4, 13.35):1;[13.35, 14.3):1;[14.3, 15.25):1;[15.25, 16.2):1;[16.2, 17.15):1;[17.15, 18.1):1;[18.1, 19.05):1;[19.05, 20]:1 |

## 3.4.2. Pearson Coefficient

A Pearson coefficient is a linear correlation coefficient that measures the linear correlation between two variables. In Machine Learning Platform for AI, the Pearson Coefficient component is used to calculate the Pearson correlation coefficient of two numeric columns in an input table or partition. The calculation result is exported to an output table.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab            | Parameter      | Description                                                               |
|----------------|----------------|---------------------------------------------------------------------------|
| Fields Setting | Input Column 1 | The name of the column whose correlation coefficient is to be calculated. |
|                | Input Column 2 | The name of the column whose correlation coefficient is to be calculated. |

- PAI command

```
pai -name pearson
  -project algo_public
  -DinputTableName=wpbc
  -Dcol1Name=f1
  -Dcol2Name=f2
  -DoutputTableName=wpbc_pear;
```

| Parameter      | Description                  | Required |
|----------------|------------------------------|----------|
| inputTableName | The name of the input table. | Yes      |

| Parameter            | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Required |
|----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| inputTablePartitions | <p>The partitions in the input table. By default, all partitions are selected.</p> <ul style="list-style-type: none"> <li>Specify a single partition in the format of <code>partition_name=value</code>.</li> <li>Specify multiple partitions in the format of <code>name1=value1,name2=value2</code>.</li> </ul> <p> <b>Note</b> Separate multiple partitions with commas (,).</p> <ul style="list-style-type: none"> <li>Specify multi-level partitions in the format of <code>name1=value1/name2=value2</code>.</li> </ul> | No       |
| col1Name             | The name of input column 1.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Yes      |
| col2Name             | The name of input column 2.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Yes      |
| outputTableName      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Yes      |
| lifecycle            | <p>The lifecycle of the output table. By default, the output table has no lifecycle.</p> <p> <b>Note</b> The parameter value must be a positive integer.</p>                                                                                                                                                                                                                                                                                                                                                                | No       |

## Example

- Input table

```
create table pai_pearson_test_input as
select * from
(
select 1.0 as f0,0.11 as f1
union all
select 2.0 as f0,0.12 as f1
union all
select 3.0 as f0,0.13 as f1
union all
select 5.0 as f0,0.15 as f1
union all
select 8.0 as f0,0.18 as f1
)tmp;
```



- PAI command

```
pai -name pearson
  -project algo_public
  -DinputTableName=pai_pearson_test_input
  -Dcol1Name=f0
  -Dcol2Name=f1
  -DoutputTableName=pai_pearson_test_output;
```

- Output table

```
+-----+-----+-----+-----+-----+-----+
|src_table|src_parts|col1_name|col2_name|count_total|count_valid|pearson_coefficient|
+-----+-----+-----+-----+-----+-----+
|sre_mpi_algo_dev.pai_pearson_test_input|    |f0    |f1    |5    |5    |0.9999999999999973|
+-----+-----+-----+-----+-----+-----+
```

### 3.4.3. Percentile

Percentile is a measure used in statistics to calculate the percentile of data in the columns of a data table. When a set of data is ordered from the smallest to largest and is divided into 100 groups, the percentile indicates the value below which a given percentage of data falls.

#### Background information

- The system can calculate only the percentiles of data of the BIGINT, DOUBLE, or DATETIME type.
- Empty columns are skipped when the percentile is calculated. If all of the columns are empty, an error is returned.
- You can specify multiple columns of data in the colName parameter.

#### Configure the component

- Machine Learning Platform for AI console

| Tab                | Parameter            | Description                                  |
|--------------------|----------------------|----------------------------------------------|
| Parameters Setting | Input Columns        | Click Select Column to select input columns. |
| Tuning             | Number of Cores      | The number of cores.                         |
|                    | Memory Size per Core | The memory size of each core.                |

- PAI command

```
PAI -name Percentile
  -project algo_public
  -DinputTableName=maple_test_percentile_3col_input
  -DcolName=col0,col1,col2 -DoutputTableName=maple_test_percentile_3col_output;
```

| Parameter      | Description                  | Required |
|----------------|------------------------------|----------|
| inputTableName | The name of the input table. | Yes      |

| Parameter                   | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Required |
|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| outputTableName             | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Yes      |
| colName                     | <p>The names of columns to be calculated. By default, all columns are selected.</p> <p> <b>Note</b> Separate the names of multiple columns with commas (.).</p>                                                                                                                                                                                                                                                                                                                                                                 | No       |
| inputPartitions             | <p>The partitions in the input table. By default, all partitions are selected.</p> <ul style="list-style-type: none"> <li>Specify a single partition in the format of <code>partition_name=value</code>.</li> <li>Specify multiple partitions in the format of <code>name1=value1,name2=value2</code>.</li> </ul> <p> <b>Note</b> Separate multiple partitions with commas (.).</p> <ul style="list-style-type: none"> <li>Specify multi-level partitions in the format of <code>name1=value1/name2=value2</code>.</li> </ul> | No       |
| predictInputTableName       | The name of the prediction table. After you configure this parameter, the prediction result can be generated.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | No       |
| predictInputTablePartitions | The partitions in the input prediction table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | No       |
| predictSelectedColNames     | The names of the columns selected from the prediction table. By default, all the columns in the prediction table are selected. The column names must be the same as the column names in a training table.                                                                                                                                                                                                                                                                                                                                                                                                        | No       |

| Parameter                          | Description                                                                                                                                                                                                                                                   | Required |
|------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| predictSelectedOriginalColumnNames | The names of the columns whose data you want to retain. By default, all columns are selected. Separate the names of multiple columns with commas (,).                                                                                                         | No       |
| predictOutputTableName             | The name of the output prediction table. This parameter is used with the predictInputTableName parameter.                                                                                                                                                     | No       |
| lifecycle                          | The lifecycle of the output table. By default, the output table has no lifecycle.<br><br> <b>Note</b> The parameter value must be a positive integer.                        | No       |
| coreNum                            | The number of cores. Valid values: [1,9999]. This parameter is used with the memSizePerCore parameter.<br><br> <b>Note</b> The parameter value must be a positive integer. | No       |
| memSizePerCore                     | The memory size of each core, in MB. Valid values: [1024,64 × 1024].<br><br> <b>Note</b> The parameter value must be a positive integer.                                   | No       |

## Example

- Input table

| col0:double (1000 rows) | col1:bigint (100 rows) | col2:bigint (300 rows)       |
|-------------------------|------------------------|------------------------------|
| 962                     | 88                     | Tue Oct 15 00:26:40 CST 1974 |
| 218                     | 99                     | Thu Jan 04 20:53:20 CST 1973 |

| col0:double (1000 rows) | col1:bigint (100 rows) | col2:bigint (300 rows)       |
|-------------------------|------------------------|------------------------------|
| 565                     | 44                     | Sat Mar 09 02:40:00 CST 1974 |
| 314                     | 68                     | Mon Aug 11 22:40:00 CST 1975 |
| 583                     | 13                     | Sat Aug 23 12:26:40 CST 1975 |
| 615                     | 87                     | Tue May 25 14:13:20 CST 1971 |
| 70                      | 53                     | Fri Mar 23 09:20:00 CST 1979 |
| 929                     | 63                     | Mon Jul 03 16:26:40 CST 1972 |
| 249                     | 48                     | Thu Mar 15 07:33:20 CST 1973 |
| 428                     | 62                     | Wed Mar 17 03:33:20 CST 1971 |
| 119                     | 1                      | Thu Jun 26 15:33:20 CST 1975 |
| 756                     | 27                     | Mon Jan 30 17:20:00 CST 1978 |
| 490                     | 75                     | Wed Dec 11 21:20:00 CST 1974 |
| 957                     | 12                     | Sun Jul 05 12:26:40 CST 1970 |
| 80                      | 22                     | Wed Oct 04 06:40:00 CST 1972 |
| 681                     | 57                     | Wed Nov 03 15:06:40 CST 1971 |
| 13                      | 95                     | Sat Sep 12 23:06:40 CST 1970 |

- PAI command

```
PAI -name Percentile
-project algo_public
-DinputTableName=maple_test_percentile_3col_input
-DcolName=col0,col1,col2 -DoutputTableName=maple_test_percentile_3col_output;
```

- Output table

| quantile:bigint | col0:double | col1:bigint | col2:datetime                |
|-----------------|-------------|-------------|------------------------------|
| 0               | 0.0         | 0           | Thu Jan 01 08:00:00 CST 1970 |
| 1               | 9.0         | 0           | Sat Jan 24 11:33:20 CST 1970 |
| 2               | 19.0        | 1           | Sat Feb 28 04:53:20 CST 1970 |
| 3               | 29.0        | 2           | Fri Apr 03 22:13:20 CST 1970 |

| quantile:bigint | col0:double | col1:bigint | col2:datetime                |
|-----------------|-------------|-------------|------------------------------|
| 4               | 39.0        | 3           | Fri May 08 15:33:20 CST 1970 |
| 5               | 49.0        | 4           | Fri Jun 12 08:53:20 CST 1970 |
| 6               | 59.0        | 5           | Fri Jul 17 02:13:20 CST 1970 |
| 7               | 69.0        | 6           | Thu Aug 20 19:33:20 CST 1970 |
| 8               | 79.0        | 7           | Thu Sep 24 12:53:20 CST 1970 |
| 9               | 89.0        | 8           | Thu Oct 29 06:13:20 CST 1970 |
| 10              | 99.0        | 9           | Wed Dec 02 23:33:20 CST 1970 |
| 11              | 109.0       | 10          | Wed Jan 06 16:53:20 CST 1971 |
| 12              | 119.0       | 11          | Wed Feb 10 10:13:20 CST 1971 |
| 13              | 129.0       | 12          | Wed Mar 17 03:33:20 CST 1971 |
| 14              | 139.0       | 13          | Tue Apr 20 20:53:20 CST 1971 |
| 15              | 149.0       | 14          | Tue May 25 14:13:20 CST 1971 |
| 16              | 159.0       | 15          | Tue Jun 29 07:33:20 CST 1971 |
| ...             | ...         | ...         | ...                          |
| 84              | 839.0       | 83          | Thu Dec 15 10:13:20 CST 1977 |
| 85              | 849.0       | 84          | Thu Jan 19 03:33:20 CST 1978 |
| 86              | 859.0       | 85          | Wed Feb 22 20:53:20 CST 1978 |
| 87              | 869.0       | 86          | Wed Mar 29 14:13:20 CST 1978 |

| quantile:bigint | col0:double | col1:bigint | col2:datetime                |
|-----------------|-------------|-------------|------------------------------|
| 88              | 879.0       | 87          | Wed May 03 07:33:20 CST 1978 |
| 89              | 889.0       | 88          | Wed Jun 07 00:53:20 CST 1978 |
| 90              | 899.0       | 89          | Tue Jul 11 18:13:20 CST 1978 |
| 91              | 909.0       | 90          | Tue Aug 15 11:33:20 CST 1978 |
| 92              | 919.0       | 91          | Tue Sep 19 04:53:20 CST 1978 |
| 93              | 929.0       | 92          | Mon Oct 23 22:13:20 CST 1978 |
| 94              | 939.0       | 93          | Mon Nov 27 15:33:20 CST 1978 |
| 95              | 949.0       | 94          | Mon Jan 01 08:53:20 CST 1979 |
| 96              | 959.0       | 95          | Mon Feb 05 02:13:20 CST 1979 |
| 97              | 969.0       | 96          | Sun Mar 11 19:33:20 CST 1979 |
| 98              | 979.0       | 97          | Sun Apr 15 12:53:20 CST 1979 |
| 99              | 989.0       | 98          | Sun May 20 06:13:20 CST 1979 |
| 100             | 999.0       | 99          | Sat Jun 23 23:33:20 CST 1979 |

### 3.4.4. Whole Table Statistics

The Whole Table Statistics component is used to collect statistics of data in a table or only selected columns.

#### Configure the component

- Machine Learning Platform for AI console

| Tab | Parameter | Description |
|-----|-----------|-------------|
|-----|-----------|-------------|

| Tab            | Parameter     | Description                                                                        |
|----------------|---------------|------------------------------------------------------------------------------------|
| Fields Setting | Input Columns | Click Select Column to select input columns. By default, all columns are selected. |
| Tuning         | Cores         | The number of cores.                                                               |
|                | Memory Size   | The memory size of each core.                                                      |

- PAI command

```
PAI -name stat_summary
-project algo_public
-DinputTableName=test_data
-DoutputTableName=test_summary_out
-DinputTablePartitions="ds='20160101'"
-DselectColNames=col0,col1,col2
-Dlifecycle=1
```

| Parameter            | Description                                                                                                                                                                                                                                                   | Required |
|----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| inputTableName       | The name of the input table.                                                                                                                                                                                                                                  | Yes      |
| outputTableName      | The name of the output table.                                                                                                                                                                                                                                 | Yes      |
| inputTablePartitions | The partitions in the input table. By default, all partitions are selected.                                                                                                                                                                                   | No       |
| selectColNames       | The names of the columns that are selected for statistics.                                                                                                                                                                                                    | No       |
| lifecycle            | <p>The lifecycle of the output table. By default, the output table has no lifecycle.</p> <div>  <b>Note</b> The parameter value must be a positive integer.         </div> | No       |
| coreNum              | <p>The number of cores.</p> <div>  <b>Note</b> The parameter value must be a positive integer.         </div>                                                              | No       |

| Parameter      | Description                                                                                                                                                                                                                              | Required |
|----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| memSizePerCore | <div>The memory size of each core, in MB. Valid values: [1024,64 x 1024].</div> <div> <b>Note</b> The parameter value must be a positive integer.</div> | No       |

Output format

The following table describes all the columns in the output table.

| Column                | Description                             |
|-----------------------|-----------------------------------------|
| colname               | The name of the column.                 |
| datatype              | The data type.                          |
| totalcount            | The total number.                       |
| count                 | The number of non-NULL values.          |
| missingcount          | The number of NULL values.              |
| nancount              | The number of NaN values.               |
| positiveinfinitycount | The number of positive infinity values. |
| negativeinfinitycount | The number of negative infinity values. |
| min                   | The minimum value.                      |
| max                   | The maximum value.                      |
| mean                  | The average value.                      |
| variance              | The variance.                           |
| standarddeviation     | The standard deviation.                 |
| standarderror         | The standard error.                     |
| skewness              | The skewness.                           |
| kurtosis              | The kurtosis.                           |
| moment2               | The second moment.                      |
| moment3               | The third moment.                       |



| Column         | Description                   |
|----------------|-------------------------------|
| moment4        | The fourth moment.            |
| centralmoment2 | The second central moment.    |
| centralmoment3 | The third central moment.     |
| centralmoment4 | The fourth central moment.    |
| sum            | The sum.                      |
| sum2           | The sum of squares.           |
| sum3           | The sum of cubes.             |
| sum4           | The sum of the fourth powers. |

## Example

- Input data

```
drop table if exists summary_test_input;
create table summary_test_input as
select
  *
from
(
  select 'a' as col1, 1 as col2, 0.001 as col3 from dual
  union all
  select 'b' as col1, 2 as col2, 100.01 as col3 from dual
) tmp;
```

- PAI command

```
PAI -name stat_summary
    -project algo_public
    -DinputTableName=summary_test_input
    -DoutputTableName=summary_test_input_out
    -DselectColNames=col1,col2,col3
    -Dlifecycle=1;
```

- Output

```

| colname | datatype | totalcount | count | missingcount | nancount | positiveinfinitycount | negativ
einfinitycount | min | max | mean | variance | standarddeviation | standarderror | skewness | kur
tosis | moment2 | moment3 | moment4 | centralmoment2 | centralmoment3 | centralmoment4 | sum
| sum2 | sum3 | sum4 |
| col1 | string | 2 | 2 | 0 | 0 | 0 | 0 | NULL | NULL | NULL | NULL
| NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL | NULL |
NULL | NULL | NULL | NULL |
| col2 | bigint | 2 | 2 | 0 | 0 | 0 | 0 | 1 | 2 | 1.5 | 0.5 | 0.7071
067811865476 | 0.5 | 0 | -2 | 2.5 | 4.5 | 8.5 | 0.25 | 0 | 0.0625 | 3 | 5 | 9
| 17 |
| col3 | double | 2 | 2 | 0 | 0 | 0 | 0 | 0.001 | 100.01 | 50.0055 | 500
0.900040500001 | 70.71704207968544 | 50.00450000000001 | 2.327677906939552e-16 | -1.999999999999999
| 5001.000050500001 | 500150.0150005006 | 50020003.00020002 | 2500.45002025 | 2.91038304567337e-11 |
6252250.303768232 | 100.011 | 10002.000101 | 1000300.030001001 | 100040006.0004 |

```

### 3.4.5. Discrete Feature Analysis

This topic describes the Discrete Feature Analysis component provided by Machine Learning Studio.

The Discrete Feature Analysis component collects statistics on the distribution of discrete features by using the following metrics: gini, entropy, gini gain, information gain, and information gain ratio. The gini and entropy are calculated for each discrete value. The gini gain, information gain, and information gain ratio are calculated for each column.

- gini:

$$I_G(f) = \sum_{i=1}^m f_i(1 - f_i)$$

- entropy:

$$I_E(f) = - \sum_{i=1}^m f_i \log_2 f_i$$

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Parameter       | Description                                                                                       |
|-----------------|---------------------------------------------------------------------------------------------------|
| Feature Columns | The columns to represent the features of data in training samples.                                |
| Label Column    | The label column.                                                                                 |
| Sparse Matrix   | If data in an input table is in the sparse format, features must be in the key-value pair format. |

- PAI command

## PAI

```

-name enum_feature_selection
-project algo_public
-DinputTableName=enumfeautreselection_input
-DlabelColName=label
-DfeatureColNames=col0,col1
-DenableSparse=false
-DoutputCntTableName=enumfeautreselection_output_cntTable
-DoutputValueTableName=enumfeautreselection_output_valuetable
-DoutputEnumValueTableName=enumfeautreselection_output_enumvaluetable;

```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                  | Default value    |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                 | No default value |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> Separate multiple partitions with commas (,).</p> | Full table       |
| featureColNames      | No       | The names of the feature columns selected from the input table for training.                                                                                                                                                                                                                                                                                                                 | No default value |
| labelColName         | No       | The name of the label column in the input table.                                                                                                                                                                                                                                                                                                                                             | No default value |
| enableSparse         | No       | Specifies whether to configure input data in the sparse format. Valid values: {true,false}.                                                                                                                                                                                                                                                                                                  | false            |
| kvFeatureColNames    | No       | The names of the feature columns that are in the key-value pair format.                                                                                                                                                                                                                                                                                                                      | Full table       |
| kvDelimiter          | No       | The delimiter used to separate keys and values if data in an input table is in the sparse format.                                                                                                                                                                                                                                                                                            | Colon (:)        |
| itemDelimiter        | No       | The delimiter used to separate key-value pairs if data in an input table is in the sparse format.                                                                                                                                                                                                                                                                                            | Comma (,)        |
| outputCntTableName   | No       | The output distribution table that contains the enumerated values of discrete features.                                                                                                                                                                                                                                                                                                      | N/A              |

| Parameter                | Required | Description                                                                             | Default value           |
|--------------------------|----------|-----------------------------------------------------------------------------------------|-------------------------|
| outputValueTableName     | No       | The output table that contains gini and entropy values of discrete features.            | N/A                     |
| outputEnumValueTableName | No       | The output table that contains enumerated gini and entropy values of discrete features. | N/A                     |
| lifecycle                | No       | The lifecycle of the table.                                                             | No default value        |
| coreNum                  | No       | The number of cores. The parameter value must be a positive integer.                    | Automatically allocated |
| memSizePerCore           | No       | The memory size of each core, in MB. Valid values: 1 to 65536.                          | Automatically allocated |

## Example

Execute the following SQL statements to generate input data:

```
drop table if exists enum_feature_selection_test_input;
create table enum_feature_selection_test_input
as
select
  *
from
(
  select
    '00' as col_string,
    1 as col_bigint,
    0.0 as col_double
  from dual
  union all
  select
    cast(null as string) as col_string,
    0 as col_bigint,
    0.0 as col_double
  from dual
  union all
  select
    '01' as col_string,
    0 as col_bigint,
    1.0 as col_double
  from dual
  union all
  select
    '01' as col_string,
    1 as col_bigint,
    cast(null as double) as col_double
  from dual
  union all
  select
    '01' as col_string,
    1 as col_bigint,
    1.0 as col_double
  from dual
  union all
  select
    '00' as col_string,
    0 as col_bigint,
    0.0 as col_double
  from dual
) tmp;
```

Input data:

```

+-----+-----+-----+
| col_string | col_bigint | col_double |
+-----+-----+-----+
| 01      | 1      | 1.0      |
| 01      | 0      | 1.0      |
| 01      | 1      | NULL     |
| NULL    | 0      | 0.0      |
| 00      | 1      | 0.0      |
| 00      | 0      | 0.0      |
+-----+-----+-----+

```

- PAI command
  - Command

```

drop table if exists enum_feature_selection_test_input_enum_value_output;
drop table if exists enum_feature_selection_test_input_cnt_output;
drop table if exists enum_feature_selection_test_input_value_output;
PAI -name enum_feature_selection -project algo_public -DitemDelimiter=":" -Dlifecycle="28" -Doutput
ValueTableName="enum_feature_selection_test_input_value_output" -DkvDelimiter="," -DlabelColName="col_bigint" -DfeatureColNames="col_double,col_string" -DoutputEnumValueTableName="enum_
feature_selection_test_input_enum_value_output" -DenableSparse="false" -DinputTableName="enum_
feature_selection_test_input" -DoutputCntTableName="enum_feature_selection_test_input_cnt_
output";

```

- Command output
  - enum\_feature\_selection\_test\_input\_cnt\_output

```
+-----+-----+-----+-----+
|colname |colvalue |labelvalue|cnt  |
+-----+-----+-----+-----+
|col_double|NULL   |1      |1    |
|col_double|0      |0      |2    |
|col_double|0      |1      |1    |
|col_double|1      |0      |1    |
|col_double|1      |1      |1    |
|col_string|NULL   |0      |1    |
|col_string|00     |0      |1    |
|col_string|00     |1      |1    |
|col_string|01     |0      |1    |
|col_string|01     |1      |2    |
+-----+-----+-----+-----+
```

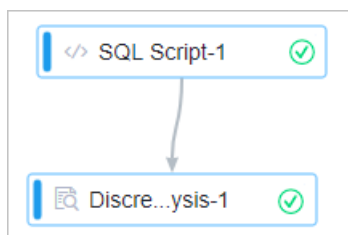
- enum\_feature\_selection\_test\_input\_value\_output

```
+-----+-----+-----+-----+-----+-----+
|colname |gini  |entropy |infogain |ginigain |infogainratio|
+-----+-----+-----+-----+-----+-----+
|col_double|0.3888888888888889|0.792481250360578|0.20751874963942196|0.1111111111111111|
|col_double|0.14221913160264427|
|col_string|0.38888888888888884|0.792481250360578|0.20751874963942196|0.1111111111111111|
|col_string|0.14221913160264427|
+-----+-----+-----+-----+-----+-----+
```

- enum\_feature\_selection\_test\_input\_enum\_value\_output

```
+-----+-----+-----+-----+
|colname |colvalue |gini  |entropy |
+-----+-----+-----+-----+
|col_double|NULL   |0.0   |0.0   |
|col_double|0      |0.22222222222222224|0.4591479170272448|
|col_double|1      |0.16666666666666666|0.3333333333333333|
|col_string|NULL   |0.0   |0.0   |
|col_string|00     |0.16666666666666666|0.3333333333333333|
|col_string|01     |0.22222222222222222|0.4591479170272448|
+-----+-----+-----+-----+
```

- Machine Learning Platform for AI console
  - Component interface



Parameter settings

Fields Setting

Feature Columns Required.

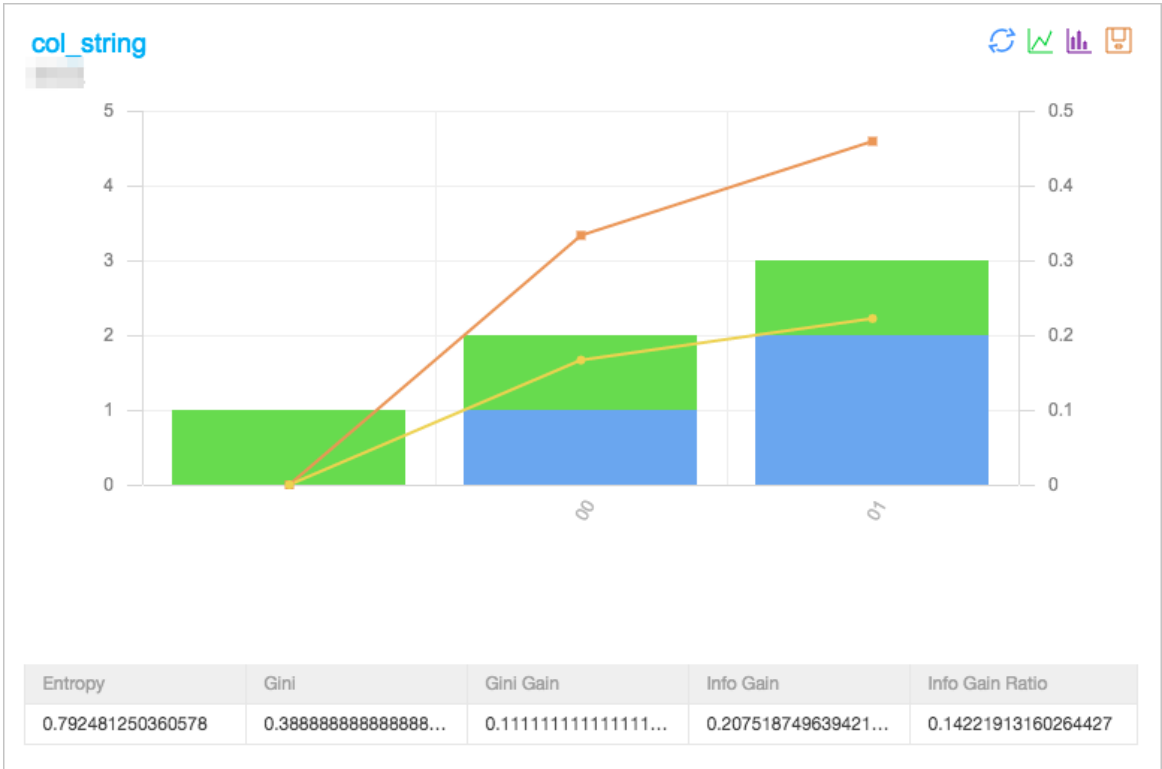
2 columns selected

Label Column Required.

col\_bigint

☐ Sparse Matrix

Results



3.4.6. One Sample T Test

This topic describes the One Sample T Test component provided by Machine Learning Studio.

The One Sample T Test component is used to determine whether a significant difference exists between the overall mean of a variable and a specific value. The sample on which you want to perform a T test must follow normal distribution. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter       | Description                        |
|----------------|-----------------|------------------------------------|
| Fields Setting | Sample 1 Column | The column that contains Sample 1. |



| Tab                       | Parameter                              | Description                                                                             |
|---------------------------|----------------------------------------|-----------------------------------------------------------------------------------------|
| <b>Parameters Setting</b> | Alternative Hypothesis Type            | The alternative hypothesis type.                                                        |
|                           | Confidence Level                       | The confidence level of the test results.                                               |
|                           | Hypothesized Mean                      | The hypothesized mean.                                                                  |
|                           | Variances of Two Populations Are Equal | Specifies whether the variances of two populations are equal. Valid values: Yes and No. |
|                           | Cores                                  | The number of cores. The value of this parameter must be a positive integer.            |
|                           | Memory Size per Core                   | The memory size of each core. Valid values: 1 to 65536. Unit: MB.                       |

- PAI command

```

pai -name t_test -project algo_public
-DxTableName=pai_t_test_all_type
-DxColName=col1_double
-DoutputTableName=pai_t_test_out
-DxTablePartitions=ds=2010/dt=1
-Dalternative=less
-Dmu=47
-DconfidenceLevel=0.95

```

| Parameter        | Required | Description                                                                  | Default value    |
|------------------|----------|------------------------------------------------------------------------------|------------------|
| xTableName       | Yes      | The name of the input table.                                                 | No default value |
| xColName         | Yes      | The column that you want to select from the input table for testing.         | No default value |
| outputTableName  | Yes      | The name of the output table.                                                | No default value |
| xTablePartitions | No       | The partitions that you want to select from the input table.                 | No default value |
| alternative      | No       | The alternative hypothesis type. Valid values: two.sided, less, and Greater. | two.sided        |

| Parameter       | Required | Description                                                                 | Default value |
|-----------------|----------|-----------------------------------------------------------------------------|---------------|
| mu              | No       | The hypothesized mean.                                                      | 0             |
| confidenceLevel | No       | The confidence level. Valid values: 0.8, 0.9, 0.95, 0.99, 0.995, and 0.999. | 0.95          |

### Output

The output table is in the JSON format and contains only one row and one column.

```
{
  "AlternativeHypothesis": "mean not equals to 0",
  "ConfidenceInterval": "(44.72234194006504, 46.27765805993496)",
  "ConfidenceLevel": 0.95,
  "alpha": 0.05,
  "df": 99,
  "mean": 45.5,
  "p": 0,
  "stdDeviation": 3.919647479510927,
  "t": 116.081867662439
}
```

## 3.4.7. Chi-square Goodness of Fit Test

This topic describes the Chi-square Goodness of Fit Test component provided by Machine Learning Studio.

### Configure the component

The Chi-square Goodness of Fit Test component is used in scenarios where categorical variables are used. This component is used to determine the difference between the observed frequency and expected frequency for each classification of a single multiclass categorical variable. The null hypothesis assumes that the observed frequency and expected frequency are the same. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Parameter         | Description                                                                                                                            |
|-------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| Input Column      | The column on which you want to perform a chi-square test.                                                                             |
| Class Probability | The class probability configuration. Specify this parameter in the <b>Class:Probability</b> format. The sum of all probabilities is 1. |

- PAI command

```
PAI -name chisq_test
-project algo_public
-DinputTableName=pai_chisq_test_input
-DcolName=f0
-DprobConfig=0:0.3,1:0.7
-DoutputTableName=pai_chisq_test_output0
-DoutputDetailTableName=pai_chisq_test_output0_detail
```

| Parameter             | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                       | Default value                                                                                   |
|-----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| inputTableName        | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                      | No default value                                                                                |
| colName               | Yes      | The name of the column.                                                                                                                                                                                                                                                                                                                                                                                                                                           | No default value                                                                                |
| outputTableName       | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                     | No default value                                                                                |
| outputDetailTableName | Yes      | The name of the output detail table.                                                                                                                                                                                                                                                                                                                                                                                                                              | No default value                                                                                |
| inputTablePartitions  | No       | <p>The partitions that you want to select from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>Multi-level partition: name1=value1/name2=value2</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (,).         </div> | No default value                                                                                |
| probConfig            | No       | <p>The class probability configuration. Specify this parameter in the <b>Class:Probability</b> format. The sum of all probabilities is 1.</p>                                                                                                                                                                                                                                                                                                                     | No default value (If this parameter is not specified, all the probability values are the same.) |

## Example

- Test data

```
create table pai_chisq_test_input as
select * from
(
  select '1' as f0,'2' as f1 from dual
  union all
  select '1' as f0,'3' as f1 from dual
  union all
  select '1' as f0,'4' as f1 from dual
  union all
  select '0' as f0,'3' as f1 from dual
  union all
  select '0' as f0,'4' as f1 from dual
)tmp;
```

- PAI command

```
PAI -name chisq_test
  -project algo_public
  -DinputTableName=pai_chisq_test_input
  -DcolName=f0
  -DprobConfig=0:0.3,1:0.7
  -DoutputTableName=pai_chisq_test_output0
  -DoutputDetailTableName=pai_chisq_test_output0_detail
```

- Output

- The output table that is specified by outputTableName is in the JSON format. It contains only one row and one column.

```
{
  "Chi-Square": {
    "comment": "Pearson's chi-square test",
    "df": 1,
    "p-value": 0.75,
    "value": 0.2380952380952381
  }
}
```

- The following table lists the columns in the output detail table that is specified by outputDetailTableName.

| column name | comment                                                                                                                                                                                                      |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| colName     | The data source class.                                                                                                                                                                                       |
| observed    | The observed frequency.                                                                                                                                                                                      |
| expected    | The expected frequency.                                                                                                                                                                                      |
| residuals   | The standard residuals, which are calculated by using the following expression: $(\text{Standard residuals} = (\text{Observed frequency} - \text{Expected frequency}) / \sqrt{\text{Expected frequency}})$ . |

- Generated data

| f0 |                                                                                   | observed | expected | residuals           |
|----|-----------------------------------------------------------------------------------|----------|----------|---------------------|
| 0  |  | 2.0      | 1.5      | 0.4082482904638631  |
| 1  |                                                                                   | 3.0      | 3.5      | -0.2672612419124244 |

## 3.4.8. Data Pivoting

The Data Pivoting component provided by Machine Learning Studio allows you to view the distributions of feature values, feature columns, and label columns. This facilitates future data analysis. This component supports both sparse and dense data formats. This topic describes how to configure the component and provides an example on how to use the component.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter                                   | Description                                                                                                     |
|--------------------|---------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns                             | The columns that indicate sample data features.                                                                 |
|                    | Target Column                               | The column that you want to use for training.                                                                   |
|                    | Enumeration Features                        | The features that you want to use as enumerated features.                                                       |
|                    | Sparse Format (K:V,K:V)                     | Specifies whether data in the sparse format is used.                                                            |
| Parameters Setting | Continuous Feature Discretization Intervals | The maximum number of intervals for the equal-distance division of continuous features.                         |
| Tuning             | Cores                                       | The number of cores that you want to use for computing. The value of this parameter must be a positive integer. |
|                    | Memory Size per Core                        | The memory size of each core. Valid values: 1 to 65536. Unit: MB.                                               |

- PAI command

## PAI

```
-name fe_meta_runner
-project algo_public
-DinputTable="pai_dense_10_10"
-DoutputTable="pai_temp_2263_20384_1"
-DmapTable="pai_temp_2263_20384_2"
-DselectedCols="pdays,previous,emp_var_rate,cons_price_idx,cons_conf_idx,euribor3m,nr_employed,age,campaign,poutcome"
-DlabelCol="y"
-DcategoryCols="previous"
-Dlifecycle="28"-DmaxBins="5" ;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                         | Default value    |
|----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTable           | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                        | No default value |
| inputTablePartitions | Yes      | <p>The partitions that you want to select from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>Multi-level partition: name1=value1/name2=value2</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | No default value |
| outputTable          | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                       | No default value |
| mapTable             | Yes      | The output mapping table. The Data Pivoting component maps STRING-type data to INT-type data for PAI to use for training.                                                                                                                                                                                                                                                                                                                           | No default value |
| selectedCols         | Yes      | The columns that you want to select from the input table.                                                                                                                                                                                                                                                                                                                                                                                           | No default value |
| categoryCols         | No       | The INT - or DOUBLE-type columns that you want to use as enumerated features.                                                                                                                                                                                                                                                                                                                                                                       | No default value |
| maxBins              | No       | The maximum number of intervals for the equal-distance division of continuous features.                                                                                                                                                                                                                                                                                                                                                             | 100              |
| isSparse             | No       | Specifies whether the input data is in the sparse format. Valid values: true and false.                                                                                                                                                                                                                                                                                                                                                             | false            |

| Parameter      | Required | Description                                                                                                                              | Default value           |
|----------------|----------|------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| itemSplitter   | No       | The delimiter that you want to use to separate key-value pairs when data in the input table is in the sparse format.                     | ,                       |
| kvSplitter     | No       | The delimiter that you want to use to separate keys and values when data in the input table is in the sparse format.                     | :                       |
| lifecycle      | No       | The lifecycle of the output table.                                                                                                       | 28                      |
| coreNum        | No       | The number of cores that you want to use for computing. The value of this parameter must be a positive integer. Valid values: 1 to 9999. | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Valid values: 1 to 65536. Unit: MB.                                                                        | Automatically allocated |

## 3.4.9. Covariance

This topic describes the Covariance component provided by Machine Learning Studio.

In probability theory and statistics, covariance is a measure of the joint variability of two random variables. Variance is a special case of covariance where the two measured variables are the same. If the expected values are  $E(X) = \mu$  and  $E(Y) = \nu$ , the covariance between real-number random variables  $X$  and  $Y$  is calculated by using the following expression:  $\text{cov}(X, Y) = E((X - \mu)(Y - \nu))$ .

You can configure the Covariance component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter     | Description                                                                                                                                           |
|----------------|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Input Columns | The input columns. You can select only BIGINT - or DOUBLE-type columns.                                                                               |
| Tuning         | Cores         | The number of cores that you want to use for computing. If you do not specify this parameter, the system automatically allocates the number of cores. |
|                | Memory Size   | The memory size of each core. If you do not specify this parameter, the system automatically allocates the memory size. Unit: MB.                     |

- PAI command

```
PAI -name cov
-project algo_public
-DinputTableName=maple_test_cov_basic12x10_input
-DoutputTableName=maple_test_cov_basic12x10_output
-DcoreNum=6
-DmemSizePerCore=110;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Default value                     |
|----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                    | No default value                  |
| inputTablePartitions | No       | <p>The partitions that you want to select from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>Multi-level partition: name1=value1/name2=value2</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (,).         </div> | All partitions of the input table |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                   | No default value                  |
| selectedColumnNames  | No       | The columns that you want to select from the input table.                                                                                                                                                                                                                                                                                                                                                                                                       | All columns                       |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                              | No default value                  |
| coreNum              | No       | The number of cores that you want to use for computing. The value of this parameter must be a positive integer. Valid values: 1 to 9999.                                                                                                                                                                                                                                                                                                                        | Automatically allocated           |
| memSizePerCore       | No       | The memory size of each core. Valid values: 1 to 65536. Unit: MB.                                                                                                                                                                                                                                                                                                                                                                                               | Automatically allocated           |

## 3.4.10. Empirical Probability Density Chart

This topic describes the Empirical Probability Density Chart component provided by Machine Learning Studio.

The Empirical Probability Density Chart component uses empirical distribution and kernel density estimation functions.

- Empirical distribution function



If accurate parametric distribution cannot be obtained, the empirical distribution function is used to estimate probability distribution based on data and generate non-parametric distribution. For more information, see [Empirical distribution function](#).

- Kernel density estimation function

The kernel density estimation function is used to estimate the probability density of sample data. Similar to a histogram, kernel distribution indicates the distribution of sample data. The difference is that kernel distribution is a smooth and continuous curve, whereas a histogram shows discrete data distribution. If the kernel density estimation function is used, the probability density of non-sample data points is not 0. Instead, the probability density is an overlay of weighted probability densities of all the sampling points in specific kernel distribution. The Empirical Probability Density Chart component uses Gaussian distribution as the kernel density estimation function. For more information, see [Kernel density estimation function](#).

## Configure the component

You can configure the Empirical Probability Density Chart component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter                       | Description                                                                                                                   |
|--------------------|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Input Columns                   | The input columns. You can select only BIGINT - or DOUBLE-type columns.                                                       |
|                    | Label Column                    | The label column.                                                                                                             |
| Parameters Setting | Number of Calculation Intervals | A large value indicates high accuracy. The value of this parameter is calculated based on the range of values in each column. |
| Tuning             | Cores                           | The number of cores that you want to use for computing. The value of this parameter must be a positive integer.               |
|                    | Memory Size                     | The memory size of each core. Valid values: 1 to 65536. Unit: MB.                                                             |

- PAI command

```
PAI -name empirical_pdf
-project algo_public
-DinputTableName="test_data"
-DoutputTableName="test_epdf_out"
-DfeatureColNames="col0,col1,col2"
-DinputTablePartitions="ds='20160101'"
-Dlifecycle=1
-DintervalNum=100
```

| Parameter      | Required | Description                  | Default value    |
|----------------|----------|------------------------------|------------------|
| inputTableName | Yes      | The name of the input table. | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Default value           |
|----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                   | No default value        |
| featureColNames      | Yes      | The names of the feature columns that you want to select from the input table for training.                                                                                                                                                                                                                                                                                                                                                                     | No default value        |
| labelColName         | No       | The name of the label column in the input table.                                                                                                                                                                                                                                                                                                                                                                                                                | No default value        |
| inputTablePartitions | No       | <p>The partitions that you want to select from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>Multi-level partition: name1=value1/name2=value2</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (,).         </div> | No default value        |
| intervalNum          | No       | The number of calculation intervals. A large number indicates a high accuracy. Valid values: [1,1E14).                                                                                                                                                                                                                                                                                                                                                          | No default value        |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                              | No default value        |
| coreNum              | No       | The number of cores that you want to use for computing. The value of this parameter must be a positive integer.                                                                                                                                                                                                                                                                                                                                                 | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core. Valid values: 1 to 65536. Unit: MB.                                                                                                                                                                                                                                                                                                                                                                                               | Automatically allocated |

## Example

Execute the following SQL statement to generate input data:

```
drop table if exists epdf_test;
create table epdf_test as
select
*
from
(
select 1.0 as col1 from dual
union all
select 2.0 as col1 from dual
union all
select 3.0 as col1 from dual
union all
select 4.0 as col1 from dual
union all
select 5.0 as col1 from dual
) tmp;
```

Run the following PAI command:

```
PAI -name empirical_pdf
-project algo_public
-DinputTableName=epdf_test
-DoutputTableName=epdf_test_out
-DfeatureColNames=col1;
```

- Input

You can select multiple columns to calculate. You can also select label columns and group these columns by label value. For example, the label columns contain the values 0 and 1. The columns are divided into two groups: label=0 and label=1. Then, the probability density of each group is provided.

 **Note** A maximum of 100 label columns can be specified.

- Output

A diagram and a result table are generated. The following table lists the columns that are contained in the result table. If no label columns are specified, NULL is generated for the label column in the output table.

| Column name | Data type |
|-------------|-----------|
| colName     | STRING    |
| label       | STRING    |
| x           | DOUBLE    |
| pdf         | DOUBLE    |

Output table

```
+-----+-----+-----+-----+
```

| colname | label | x                  | pdf                 |
|---------|-------|--------------------|---------------------|
| col1    | NULL  | 1.0                | 0.12775155176809325 |
| col1    | NULL  | 1.0404050505050506 | 0.1304256933829622  |
| col1    | NULL  | 1.0808101010101012 | 0.13306325897429525 |
| col1    | NULL  | 1.1212151515151518 | 0.1356613897616418  |
| col1    | NULL  | 1.1616202020202024 | 0.1382173796574596  |
| col1    | NULL  | 1.202025252525253  | 0.1407286844875733  |
| col1    | NULL  | 1.2424303030303037 | 0.14319293014274642 |
| col1    | NULL  | 1.2828353535353543 | 0.14560791960033242 |
| col1    | NULL  | 1.3232404040404049 | 0.14797163876379316 |
| col1    | NULL  | 1.3636454545454555 | 0.1502822610772349  |
| col1    | NULL  | 1.404050505050506  | 0.1525381508819247  |
| col1    | NULL  | 1.4444555555555567 | 0.1547378654919243  |
| col1    | NULL  | 1.4848606060606073 | 0.1568801559764068  |
| col1    | NULL  | 1.525265656565658  | 0.15896396664681753 |
| col1    | NULL  | 1.5656707070707085 | 0.16098843325768245 |
| col1    | NULL  | 1.6060757575757592 | 0.1629528799404685  |
| col1    | NULL  | 1.6464808080808098 | 0.16485681490034038 |
| col1    | NULL  | 1.6868858585858604 | 0.16669992491584543 |
| col1    | NULL  | 1.727290909090911  | 0.16848206869138338 |
| col1    | NULL  | 1.7676959595959616 | 0.17020326912168932 |
| col1    | NULL  | 1.8081010101010122 | 0.17186370453638117 |
| col1    | NULL  | 1.8485060606060628 | 0.17346369900080946 |
| col1    | NULL  | 1.8889111111111134 | 0.17500371175692428 |
| col1    | NULL  | 1.929316161616164  | 0.17648432589456017 |
| col1    | NULL  | 1.9697212121212146 | 0.17790623634938396 |
| col1    | NULL  | 2.0101262626262653 | 0.1792702373286898  |
| col1    | NULL  | 2.050531313131316  | 0.18057720927022053 |
| col1    | NULL  | 2.0909363636363665 | 0.18182810544221673 |
| col1    | NULL  | 2.131341414141417  | 0.18302393829491406 |
| col1    | NULL  | 2.1717464646464677 | 0.18416576567472337 |
| col1    | NULL  | 2.2121515151515183 | 0.1852546770123305  |
| col1    | NULL  | 2.252556565656569  | 0.18629177959496213 |
| col1    | NULL  | 2.2929616161616195 | 0.18727818503109434 |
| col1    | NULL  | 2.333366666666667  | 0.18821499601297229 |
| col1    | NULL  | 2.3737717171717208 | 0.18910329347850022 |
| col1    | NULL  | 2.4141767676767714 | 0.18994412426940221 |
| col1    | NULL  | 2.454581818181822  | 0.19073848937711185 |
| col1    | NULL  | 2.4949868686868726 | 0.19148733286168018 |
| col1    | NULL  | 2.535391919191923  | 0.1921915315221827  |
| col1    | NULL  | 2.575796969696974  | 0.19285188538972659 |
| col1    | NULL  | 2.6162020202020244 | 0.19346910910630113 |
| col1    | NULL  | 2.656607070707075  | 0.19404382424446043 |
| col1    | NULL  | 2.6970121212121256 | 0.1945765526142701  |
| col1    | NULL  | 2.7374171717171762 | 0.19506771059517916 |
| col1    | NULL  | 2.777822222222227  | 0.19551760452158667 |
| col1    | NULL  | 2.8182272727272775 | 0.19592642714194602 |
| col1    | NULL  | 2.858632323232328  | 0.1962942551623821  |
| col1    | NULL  | 2.8990373737373787 | 0.1966210478770638  |
| col1    | NULL  | 2.9394424242424293 | 0.1969066468790639  |
| col1    | NULL  | 2.979847474747478  | 0.19715077683721793 |
| col1    | NULL  | 3.0202525252525305 | 0.19735304731663747 |
| col1    | NULL  | 3.060657575757581  | 0.19751295561309964 |

|      |      |                    |                     |
|------|------|--------------------|---------------------|
| col1 | NULL | 3.1010626262626317 | 0.19762989056457925 |
| col1 | NULL | 3.1414676767676823 | 0.19770313729675995 |
| col1 | NULL | 3.181872727272733  | 0.19773188285349683 |
| col1 | NULL | 3.2222777777777836 | 0.19771522265793107 |
| col1 | NULL | 3.262682828282834  | 0.19765216774530828 |
| col1 | NULL | 3.303087878787885  | 0.19754165270453194 |
| col1 | NULL | 3.3434929292929354 | 0.19738254426210697 |
| col1 | NULL | 3.383897979797986  | 0.19717365043938664 |
| col1 | NULL | 3.4243030303030366 | 0.19691373021193162 |
| col1 | NULL | 3.4647080808080872 | 0.1966015035982942  |
| col1 | NULL | 3.505113131313138  | 0.19623566210464843 |
| col1 | NULL | 3.5455181818181885 | 0.19581487945135703 |
| col1 | NULL | 3.585923232323239  | 0.19533782250778076 |
| col1 | NULL | 3.6263282828282897 | 0.1948031623623475  |
| col1 | NULL | 3.6667333333333403 | 0.1942095854560816  |
| col1 | NULL | 3.707138383838391  | 0.19355580470939734 |
| col1 | NULL | 3.7475434343434415 | 0.19284057057394655 |
| col1 | NULL | 3.787948484848492  | 0.19206268194364004 |
| col1 | NULL | 3.8283535353535427 | 0.19122099686158253 |
| col1 | NULL | 3.8687585858585933 | 0.19031444296253852 |
| col1 | NULL | 3.909163636363644  | 0.1893420275936375  |
| col1 | NULL | 3.9495686868686946 | 0.18830284755928747 |
| col1 | NULL | 3.989973737373745  | 0.1871960984396676  |
| col1 | NULL | 4.030378787878796  | 0.18602108343567092 |
| col1 | NULL | 4.070783838383846  | 0.18477722169674377 |
| col1 | NULL | 4.111188888888897  | 0.1834640560916829  |
| col1 | NULL | 4.151593939393948  | 0.1820812603860928  |
| col1 | NULL | 4.191998989898998  | 0.18062864579383914 |
| col1 | NULL | 4.232404040404049  | 0.179106166873458   |
| col1 | NULL | 4.272809090909099  | 0.17751392674406796 |
| col1 | NULL | 4.31321414141415   | 0.17585218159888508 |
| col1 | NULL | 4.353619191919201  | 0.17412134449794325 |
| col1 | NULL | 4.394024242424251  | 0.1723219884250765  |
| col1 | NULL | 4.434429292929302  | 0.17045484859762067 |
| col1 | NULL | 4.4748343434343525 | 0.16852082402064342 |
| col1 | NULL | 4.515239393939403  | 0.1665209782808102  |
| col1 | NULL | 4.555644444444454  | 0.16445653957824907 |
| col1 | NULL | 4.596049494949504  | 0.16232889999798905 |
| col1 | NULL | 4.636454545454555  | 0.16013961402571825 |
| col1 | NULL | 4.6768595959596055 | 0.1578903963157465  |
| col1 | NULL | 4.717264646464656  | 0.15558311872216193 |
| col1 | NULL | 4.757669696969707  | 0.1532198066072439  |
| col1 | NULL | 4.798074747474757  | 0.1508026344442397  |
| col1 | NULL | 4.838479797979808  | 0.14833392073462115 |
| col1 | NULL | 4.878884848484859  | 0.14581612226291346 |
| col1 | NULL | 4.919289898989909  | 0.1432518277151203  |
| col1 | NULL | 4.95969494949496   | 0.1406437506896507  |
| col1 | NULL | 5.00010000000001   | 0.13799472213247665 |

+-----+-----+-----+-----+

## 3.4.11. Box Plot

This topic describes the Box Plot component provided by Machine Learning Studio.

A box plot shows the distribution of a set of data. It shows the distribution features of raw data. It can also be used to compare the distribution features of multiple sets of data.

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Parameter           | Description                              |
|---------------------|------------------------------------------|
| Continuous Features | Continuous features                      |
| Enumeration Feature | The enumeration feature                  |
| Stratified Samples  | The number of adopted stratified samples |

- PAI command

```
PAI -name box_plot -project algo_public
-DinputTable="boxplot"
-DcontinueCols="age"
-DcategoryCol="y"
-DoutputTable="pai_temp_6075_97181_1"
-DsampleSize="1000"
-Dlifecycle="7";
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                              | Default value    |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTable           | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                             | No default value |
| inputTablePartitions | Yes      | <p>The partitions selected from the input table for training. The system supports the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | No default value |
| outputTable          | Yes      | The output table that stores the box plot and samples.                                                                                                                                                                                                                                                                                                                                                                   | No default value |
| continueCols         | Yes      | Continuous feature columns.                                                                                                                                                                                                                                                                                                                                                                                              | No default value |
| categoryCol          | Yes      | The enumeration feature column.                                                                                                                                                                                                                                                                                                                                                                                          | No default value |
| sampleSize           | No       | The number of samples based on which the disturbance conditions of each feature are drawn.                                                                                                                                                                                                                                                                                                                               | 1000             |

| Parameter      | Required | Description                                                               | Default value           |
|----------------|----------|---------------------------------------------------------------------------|-------------------------|
| lifecycle      | No       | The lifecycle of the output table. Unit: days.                            | 28                      |
| coreNum        | No       | The number of cores. The value must be a positive integer.                | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. The value must range from 1 MB to 65536 MB. | Automatically allocated |

## Example

- Input data

```
create table boxplot as select age, y from bank_data limit 100;
```

| age | y |
|-----|---|
| 50  | 0 |
| 53  | 0 |
| 28  | 1 |
| 39  | 0 |
| 55  | 1 |
| 30  | 0 |
| 37  | 0 |
| 39  | 0 |
| 36  | 1 |
| 27  | 0 |
| 34  | 0 |
| 41  | 0 |
| 55  | 1 |
| 33  | 0 |
| 26  | 0 |
| 52  | 0 |
| 35  | 1 |

| age | y |
|-----|---|
| 27  | 1 |
| 28  | 0 |
| 26  | 0 |
| 41  | 0 |
| 35  | 0 |
| 40  | 0 |
| 32  | 0 |
| 41  | 0 |
| 34  | 0 |
| 49  | 0 |
| 37  | 0 |
| 35  | 0 |
| 38  | 0 |
| 47  | 0 |
| 46  | 0 |
| 27  | 0 |
| 29  | 1 |
| 32  | 0 |
| 36  | 0 |
| 29  | 0 |
| 47  | 0 |
| 44  | 0 |
| 54  | 0 |
| 36  | 0 |
| 42  | 0 |
| 44  | 0 |



| age | y |
|-----|---|
| 72  | 1 |
| 48  | 0 |
| 36  | 0 |
| 35  | 0 |
| 43  | 0 |
| 56  | 0 |
| 42  | 0 |
| 31  | 0 |
| 32  | 0 |
| 33  | 0 |
| 31  | 0 |
| 39  | 0 |
| 30  | 1 |
| 24  | 0 |
| 24  | 0 |
| 38  | 0 |
| 26  | 0 |
| 41  | 0 |
| 34  | 0 |
| 30  | 0 |
| 37  | 0 |
| 68  | 0 |
| 31  | 0 |
| 48  | 0 |
| 33  | 0 |
| 59  | 0 |

| age | y |
|-----|---|
| 44  | 0 |
| 28  | 0 |
| 50  | 0 |
| 33  | 0 |
| 45  | 0 |
| 40  | 0 |
| 45  | 0 |
| 43  | 0 |
| 54  | 0 |
| 53  | 0 |
| 35  | 0 |
| 30  | 0 |
| 25  | 0 |
| 35  | 0 |
| 54  | 1 |
| 30  | 0 |
| 38  | 0 |
| 35  | 0 |
| 47  | 0 |
| 32  | 0 |
| 27  | 0 |
| 40  | 1 |
| 31  | 0 |
| 42  | 0 |
| 40  | 0 |
| 31  | 0 |

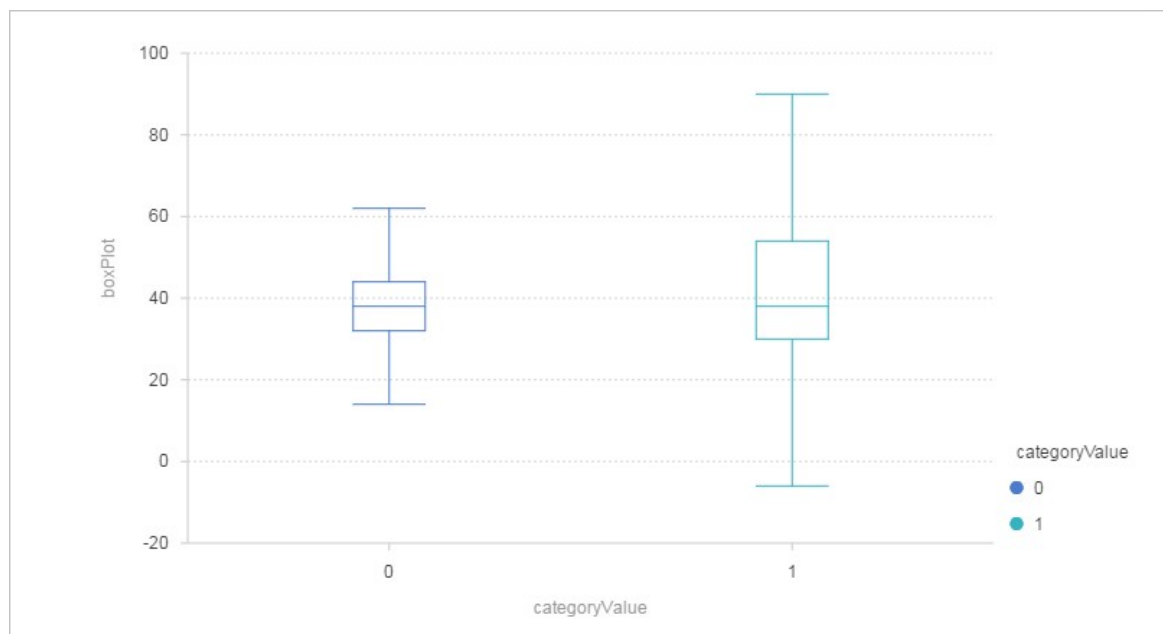
| age | y |
|-----|---|
| 57  | 0 |
| 38  | 1 |
| 39  | 0 |
| 37  | 0 |
| 44  | 0 |

- Parameter configuration

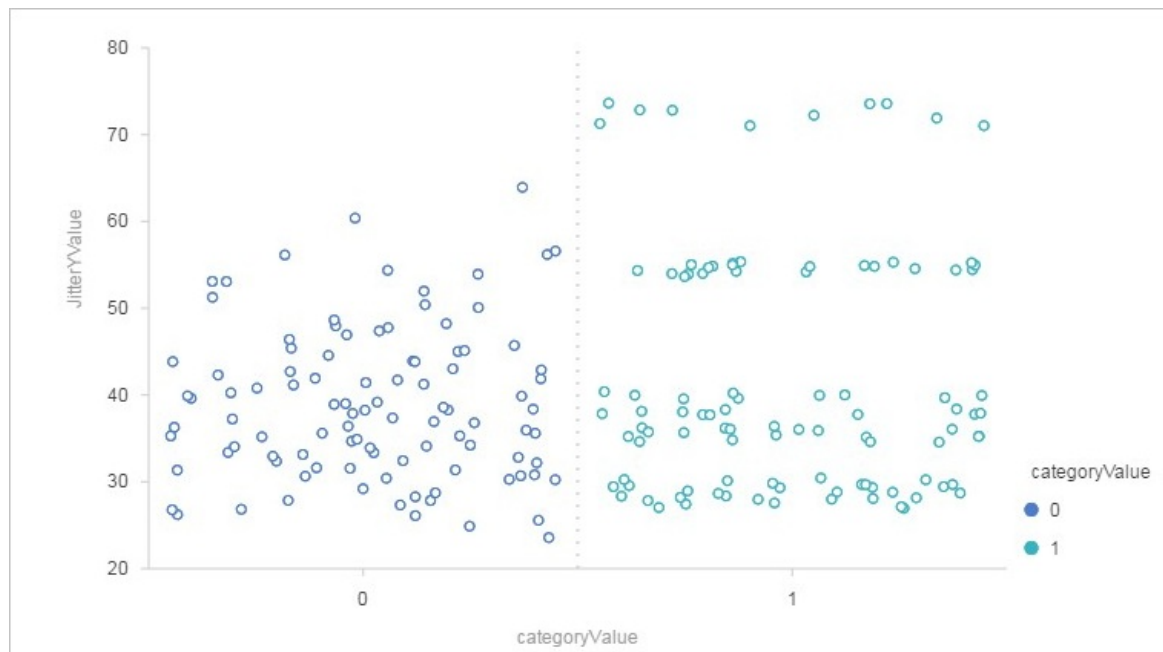
Specify the age column as the continuous feature column, and the y column as the enumeration feature column. Retain the default values of other parameters.

- Output

- The following figure shows a box plot.



- The following figure shows the distribution of disturbance points.



## 3.4.12. Scatter Plot

This topic describes the Scatter Plot component provided by Machine Learning Studio.

In regression analysis, a scatter plot shows the distribution of data points in a Cartesian coordinate system.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Parameter       | Description                                            |
|-----------------|--------------------------------------------------------|
| Feature Columns | Columns that are used to indicate sample data features |
| Label Column    | The label column                                       |
| Samples         | The number of samples                                  |

- PAI command

```
PAI -name scatter_diagram -project algo_public
-DselectedCols=emp_var_rate,cons_price_rate,cons_conf_idx,euribor3m
-DlabelCol=y
-DmapTable=pai_temp_2447_22859_2
-DinputTable=scatter_diagram
-DoutputTable=pai_temp_2447_22859_1;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                          | Default value    |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTable           | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                         | No default value |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. The system supports the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (,).         </div> | No default value |
| outputTable          | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                        | No default value |
| mapTable             | Yes      | The name of the output table that stores the maximum value, minimum value, and enumeration value of each feature.                                                                                                                                                                                                                                                                                                                    | No default value |
| selectedCols         | Yes      | The columns selected from the input table and used to draw a scatter plot. A maximum of five columns can be selected.                                                                                                                                                                                                                                                                                                                | No default value |
| labelCol             | Yes      | An INT or STRING field is used as an enumeration label.                                                                                                                                                                                                                                                                                                                                                                              | No default value |
| lifecycle            | Yes      | The lifecycle of the output table. Unit: days.                                                                                                                                                                                                                                                                                                                                                                                       | 28               |

## Example

- Input data

```
create table scatter_diagram as select emp_var_rate,cons_price_rate, cons_conf_idx,euribor3m,y from p
ai_bank_data limit 10
```

| emp_var_rate | cons_price_rate | cons_conf_idx | euribor3m | y |
|--------------|-----------------|---------------|-----------|---|
| 1.4          | 93.918          | -42.7         | 4.962     | 0 |
| -0.1         | 93.2            | -42.0         | 4.021     | 0 |
| -1.7         | 94.055          | -39.8         | 0.729     | 1 |
| -1.8         | 93.075          | -47.1         | 1.405     | 0 |
| -2.9         | 92.201          | 31.4          | 0.869     | 1 |

| emp_var_rate | cons_price_rate | cons_conf_idx | euribor3m | y |
|--------------|-----------------|---------------|-----------|---|
| 1.4          | 93.918          | -42.7         | 4.961     | 0 |
| -1.8         | 92.893          | -46.2         | 1.327     | 0 |
| -1.8         | 92.893          | 92.893        | 1.313     | 0 |
| -2.9         | 92.963          | -40.8         | 1.266     | 1 |
| -1.8         | 93.075          | -47.1         | 1.41      | 0 |
| 1.1          | 93.994          | -36.4         | 4.864     | 0 |
| 1.4          | 93.444          | -36.1         | 4.964     | 0 |
| 1.4          | 93.444          | -36.1         | 4.965     | 1 |
| -1.8         | 92.893          | -46.2         | 1.291     | 0 |
| 1.4          | 94.465          | -41.8         | 4.96      | 0 |
| 1.4          | 93.918          | -42.7         | 4.962     | 0 |
| -1.8         | 93.075          | -47.1         | 1.365     | 1 |
| -0.1         | 93.798          | -40.4         | 4.86      | 1 |
| 1.1          | 93.994          | -36.4         | 4.86      | 0 |
| 1.4          | 93.918          | -42.7         | 4.96      | 0 |
| -1.8         | 93.075          | -47.1         | 1.405     | 0 |
| 1.4          | 94.465          | -41.8         | 4.967     | 0 |
| 1.4          | 93.918          | -42.7         | 4.963     | 0 |
| 1.4          | 93.918          | -42.7         | 4.968     | 0 |
| 1.4          | 93.918          | -42.7         | 4.962     | 0 |
| -1.8         | 92.893          | -46.2         | 1.344     | 0 |
| -3.4         | 92.431          | -26.9         | 0.754     | 0 |
| -1.8         | 93.075          | -47.1         | 1.365     | 0 |
| -1.8         | 92.893          | -46.2         | 1.313     | 0 |
| 1.4          | 93.918          | -42.7         | 4.961     | 0 |
| 1.4          | 94.465          | -41.8         | 4.961     | 0 |

| emp_var_rate | cons_price_rate | cons_conf_idx | euribor3m | y |
|--------------|-----------------|---------------|-----------|---|
| -1.8         | 92.893          | -46.2         | 1.327     | 0 |
| -1.8         | 92.893          | -46.2         | 1.299     | 0 |
| -2.9         | 92.963          | -40.8         | 1.268     | 1 |
| 1.4          | 93.918          | -42.7         | 4.963     | 0 |
| -1.8         | 92.893          | -46.2         | 1.334     | 0 |
| 1.4          | 93.918          | -42.7         | 4.96      | 0 |
| -1.8         | 93.075          | -47.1         | 1.405     | 0 |
| 1.4          | 94.465          | -41.8         | 4.96      | 0 |
| 1.4          | 93.444          | -36.1         | 4.962     | 0 |
| 1.1          | 93.994          | -36.4         | 4.86      | 0 |
| 1.1          | 93.994          | -36.4         | 4.857     | 0 |
| 1.4          | 93.918          | -42.7         | 4.961     | 0 |
| -3.4         | 92.649          | -30.1         | 0.715     | 1 |
| 1.4          | 93.444          | -36.1         | 4.966     | 0 |
| -0.1         | 93.2            | -42.0         | 4.076     | 0 |
| 1.4          | 93.444          | -36.1         | 4.965     | 0 |
| -1.8         | 92.893          | -46.2         | 1.354     | 0 |
| 1.4          | 93.444          | -36.1         | 4.967     | 0 |
| 1.4          | 94.465          | -41.8         | 4.959     | 0 |
| -1.8         | 92.893          | -46.2         | 1.354     | 0 |
| 1.4          | 94.465          | -41.8         | 4.958     | 0 |
| -1.8         | 92.893          | -46.2         | 1.354     | 0 |
| 1.4          | 94.465          | -41.8         | 4.864     | 0 |
| 1.1          | 93.994          | -36.4         | 4.859     | 0 |
| 1.1          | 93.994          | -36.4         | 4.857     | 0 |
| -1.8         | 92.893          | -46.2         | 1.27      | 0 |

| emp_var_rate | cons_price_rate | cons_conf_idx | euribor3m | y |
|--------------|-----------------|---------------|-----------|---|
| 1.1          | 93.994          | -36.4         | 4.857     | 0 |
| 1.1          | 93.994          | -36.4         | 4.859     | 0 |
| 1.4          | 94.465          | -41.8         | 4.959     | 0 |
| 1.1          | 93.994          | -36.4         | 4.856     | 0 |
| -1.8         | 93.075          | -47.1         | 1.405     | 0 |
| -1.8         | 92.843          | -50.0         | 1.811     | 1 |
| -0.1         | 93.2            | -42.0         | 4.021     | 0 |
| -2.9         | 92.469          | -33.6         | 1.029     | 0 |
| 1.4          | 93.918          | -42.7         | 4.962     | 0 |
| -1.8         | 93.075          | -47.1         | 1.365     | 0 |
| 1.1          | 93.994          | -36.4         | 4.857     | 0 |
| -1.8         | 92.893          | -46.2         | 1.259     | 0 |
| 1.1          | 93.994          | -36.4         | 4.857     | 0 |
| 1.4          | 94.465          | -41.8         | 4.866     | 0 |
| -2.9         | 92.201          | -31.4         | 0.883     | 0 |
| -0.1         | 93.2            | -42.0         | 4.076     | 0 |
| 1.1          | 93.994          | -36.4         | 4.857     | 0 |
| 1.4          | 93.918          | -42.7         | 4.96      | 0 |
| 1.4          | 93.444          | -36.1         | 4.962     | 0 |
| 1.1          | 93.994          | -36.4         | 4.858     | 0 |
| 1.1          | 93.994          | -36.4         | 4.857     | 0 |
| 1.1          | 93.994          | -36.4         | 4.856     | 0 |
| 1.4          | 93.918          | -42.7         | 4.968     | 0 |
| 1.4          | 93.444          | -36.1         | 4.966     | 0 |
| 1.4          | 94.465          | -41.8         | 4.962     | 0 |
| 1.4          | 93.444          | -36.1         | 4.963     | 0 |



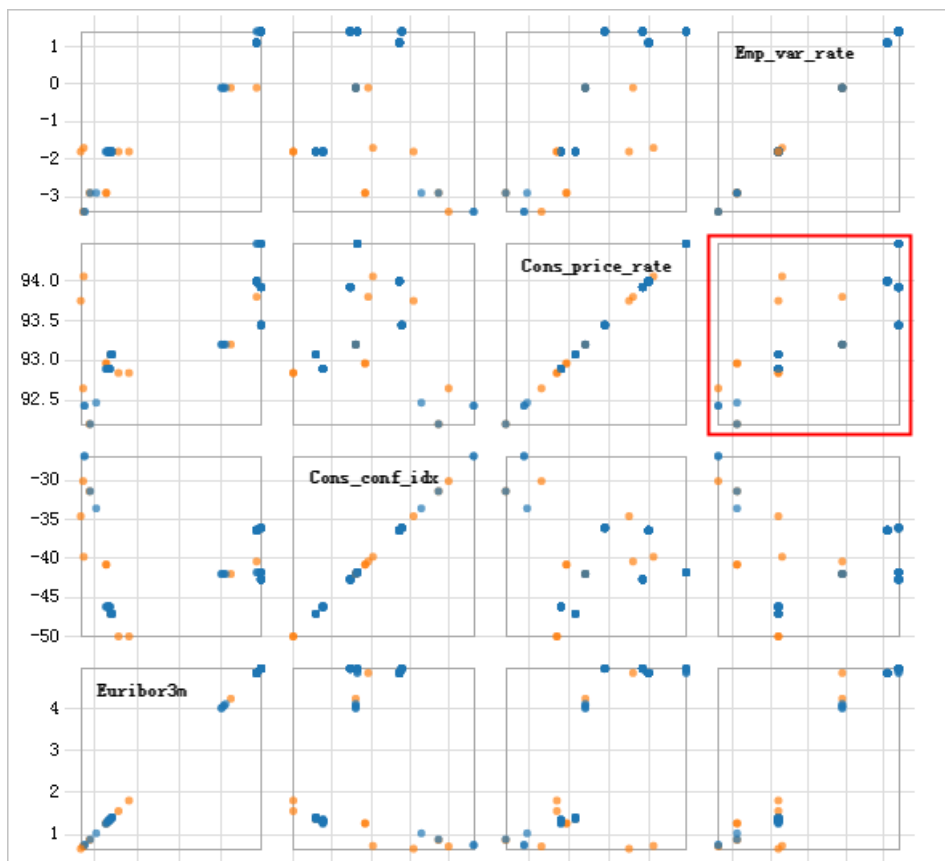
| emp_var_rate | cons_price_rate | cons_conf_idx | euribor3m | y |
|--------------|-----------------|---------------|-----------|---|
| -1.8         | 92.843          | -50.0         | 1.56      | 1 |
| 1.4          | 93.918          | -42.7         | 4.96      | 0 |
| 1.4          | 93.444          | -36.1         | 4.963     | 0 |
| -3.4         | 92.431          | -26.9         | 0.74      | 0 |
| 1.1          | 93.994          | -36.4         | 4.856     | 0 |
| 1.4          | 93.918          | -42.7         | 4.962     | 0 |
| 1.1          | 93.994          | -36.4         | 4.856     | 0 |
| -0.1         | 93.2            | -42.0         | 4.245     | 1 |
| 1.1          | 93.994          | -36.4         | 4.857     | 0 |
| -1.8         | 93.075          | -47.1         | 1.405     | 0 |
| -1.8         | 92.893          | -46.2         | 1.327     | 0 |
| -0.1         | 93.2            | -42.0         | 4.12      | 0 |
| 1.4          | 94.465          | -41.8         | 4.958     | 0 |
| -1.8         | 93.749          | -34.6         | 0.659     | 1 |
| 1.1          | 93.994          | -36.4         | 4.858     | 0 |
| 1.1          | 93.994          | -36.4         | 4.858     | 0 |
| 1.4          | 93.444          | -36.1         | 4.963     | 0 |

- Parameter configuration

Select the y column as the optional label column for the scatter plot. Select the select emp\_var\_rate, cons\_price\_rate, cons\_conf\_idx, and euribor3m columns as feature columns.

- Output

You can view the distribution of the objects specified by the label column for different features in the scatter plot.



### 3.4.13. Correlation Coefficient Matrix

This topic describes the Correlation Coefficient Matrix component provided by Machine Learning Studio.

The correlation coefficient indicates the correlation between columns in a matrix. The coefficient is in the range of  $[-1, 1]$ . The count parameter is measured when the value is the number of non-zero elements in two consecutive columns.

#### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab            | Parameter               | Description                              |
|----------------|-------------------------|------------------------------------------|
| Fields Setting | All Selected by Default | N/A                                      |
| Tuning         | Cores                   | This parameter is used with Memory Size. |
|                | Memory Size             | This parameter is used with Cores.       |

- PAI command

```
PAI -name corrcoeff
-project algo_public
-DinputTableName=maple_test_corrcoeff_basic12x10_input
-DoutputTableName=maple_test_corrcoeff_basic12x10_output
-DcoreNum=1
-DmemSizePerCore=110;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                          | Default value                        |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                         | No default value                     |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. The system supports the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (,).         </div> | No default value                     |
| outputTableName      | Yes      | The names of output tables.                                                                                                                                                                                                                                                                                                                                                                                                          | No default value                     |
| selectedColumnNames  | No       | The columns selected from the input table.                                                                                                                                                                                                                                                                                                                                                                                           | All columns are selected by default. |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                                                                                   | No default value                     |
| coreNum              | No       | This parameter is used with memSizePerCore. The value must be a positive integer. Valid values: [1, 9999].                                                                                                                                                                                                                                                                                                                           | Automatically calculated             |
| memSizePerCore       | No       | The memory size of each core. Unit: MB. A positive integer in the range of [1024, 64 × 1024]                                                                                                                                                                                                                                                                                                                                         | Automatically calculated             |

## Example

### • Input

| col0:double | col1:bigint | col2:double | col3:bigint | col4:double | col5:bigint | col6:double | col7:bigint | col8:double | col9:double |
|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 19          | 95          | 33          | 52          | 115         | 43          | 32          | 98          | 76          | 40          |

| col0:double | col1:bigint | col2:double | col3:bigint | col4:double | col5:bigint | col6:double | col7:bigint | col8:double | col9:double |
|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 114         | 26          | 101         | 69          | 56          | 59          | 116         | 23          | 109         | 105         |
| 103         | 89          | 7           | 9           | 65          | 118         | 73          | 50          | 55          | 81          |
| 79          | 20          | 63          | 71          | 5           | 24          | 77          | 31          | 21          | 75          |
| 87          | 16          | 66          | 47          | 25          | 14          | 42          | 99          | 108         | 57          |
| 11          | 104         | 38          | 37          | 106         | 51          | 3           | 91          | 80          | 97          |
| 84          | 30          | 70          | 46          | 8           | 6           | 94          | 22          | 45          | 48          |
| 35          | 17          | 107         | 64          | 10          | 112         | 53          | 34          | 90          | 96          |
| 13          | 61          | 39          | 1           | 29          | 117         | 112         | 2           | 82          | 28          |
| 62          | 4           | 102         | 88          | 100         | 36          | 67          | 54          | 12          | 85          |
| 49          | 27          | 44          | 93          | 68          | 110         | 60          | 72          | 86          | 58          |
| 92          | 119         | 0           | 113         | 41          | 15          | 74          | 83          | 18          | 111         |

- PAI command

```
PAI -name corrcoef
  -project algo_public
  -DinputTableName=maple_test_corrcoef_basic12x10_input
  -DoutputTableName=maple_test_corrcoef_basic12x10_output
  -DcoreNum=1
  -DmemSizePerCore=110;
```

- Output

| column names | col0                                | col1                                | col2                                | col3                                 | col4                                | col5                                 | col6                                | col7                                 | col8                                 | col9                            |
|--------------|-------------------------------------|-------------------------------------|-------------------------------------|--------------------------------------|-------------------------------------|--------------------------------------|-------------------------------------|--------------------------------------|--------------------------------------|---------------------------------|
| col0         | 1                                   | -<br>0.211<br>56572<br>51820<br>724 | 0.059<br>83062<br>59706<br>561      | 0.259<br>99035<br>70684<br>693       | -<br>0.348<br>32491<br>88225<br>586 | -<br>0.287<br>16254<br>39680<br>9926 | 0.478<br>80162<br>12743<br>5116     | -<br>0.136<br>46519<br>48421<br>3326 | -<br>0.195<br>00158<br>76468<br>0092 | 0.389<br>73902<br>40949<br>085  |
| col1         | -<br>0.211<br>56572<br>51820<br>724 | 1                                   | -<br>0.844<br>44773<br>77898<br>585 | -<br>0.175<br>07636<br>22159<br>4533 | 0.409<br>43384<br>15057<br>1377     | 0.091<br>35976<br>02610<br>1403      | -<br>0.301<br>85063<br>74626<br>574 | 0.407<br>33726<br>91280<br>8044      | -<br>0.118<br>27739<br>12459<br>0071 | 0.124<br>33851<br>38945<br>5183 |

| column names | col0                                 | col1                                 | col2                                 | col3                                  | col4                                 | col5                                 | col6                                  | col7                                 | col8                                 | col9                                 |
|--------------|--------------------------------------|--------------------------------------|--------------------------------------|---------------------------------------|--------------------------------------|--------------------------------------|---------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|
| col2         | 0.059<br>83062<br>59706<br>561       | -<br>0.844<br>44773<br>77898<br>585  | 1                                    | 0.185<br>18346<br>64729<br>3102       | -<br>0.209<br>34839<br>22805<br>7014 | -<br>0.189<br>64175<br>12389<br>659  | 0.179<br>93774<br>98863<br>213        | -<br>0.385<br>88856<br>76469<br>948  | 0.202<br>54569<br>20377<br>3892      | 0.134<br>76160<br>75375<br>6655      |
| col3         | 0.259<br>99035<br>70684<br>693       | -<br>0.175<br>07636<br>22159<br>4533 | 0.185<br>18346<br>64729<br>3102      | 1                                     | 0.039<br>88018<br>64985<br>4009      | -<br>0.437<br>37887<br>41832<br>9147 | -<br>0.053<br>81829<br>64252<br>67184 | 0.290<br>08564<br>41586<br>986       | -<br>0.360<br>75479<br>10075<br>688  | 0.491<br>20190<br>74930<br>449       |
| col4         | -<br>0.348<br>32491<br>88225<br>586  | 0.409<br>43384<br>15057<br>1377      | -<br>0.209<br>34839<br>22805<br>7014 | 0.039<br>88018<br>64985<br>4009       | 1                                    | 0.146<br>56052<br>09246<br>875       | -<br>0.501<br>60303<br>64347<br>955   | 0.549<br>60243<br>25711<br>117       | 0.013<br>74325<br>61153<br>94122     | 0.074<br>97231<br>55918<br>4887      |
| col5         | -<br>0.287<br>16254<br>39680<br>9926 | 0.091<br>35976<br>02610<br>1403      | -<br>0.189<br>64175<br>12389<br>659  | -<br>0.437<br>37887<br>41832<br>9147  | 0.146<br>56052<br>09246<br>875       | 1                                    | 0.167<br>29809<br>31087<br>3522       | -<br>0.298<br>90655<br>82879<br>6964 | 0.361<br>85181<br>01014<br>617       | -<br>0.171<br>39609<br>57286<br>885  |
| col6         | 0.478<br>80162<br>12743<br>5116      | -<br>0.301<br>85063<br>74626<br>574  | 0.179<br>93774<br>98863<br>213       | -<br>0.053<br>81829<br>64252<br>67184 | -<br>0.501<br>60303<br>64347<br>955  | 0.167<br>29809<br>31087<br>3522      | 1                                     | -<br>0.816<br>50198<br>80156<br>462  | -<br>0.111<br>73420<br>91872<br>1436 | -<br>0.103<br>63860<br>37834<br>7944 |
| col7         | -<br>0.136<br>46519<br>48421<br>3326 | 0.407<br>33726<br>91280<br>8044      | -<br>0.385<br>88856<br>76469<br>948  | 0.290<br>08564<br>41586<br>986        | 0.549<br>60243<br>25711<br>117       | -<br>0.298<br>90655<br>82879<br>6964 | -<br>0.816<br>50198<br>80156<br>462   | 1                                    | 0.074<br>35907<br>47154<br>4469      | 0.117<br>11976<br>05199<br>9162      |
| col8         | -<br>0.195<br>00158<br>76468<br>0092 | -<br>0.118<br>27739<br>12459<br>0071 | 0.202<br>54569<br>20377<br>3892      | -<br>0.360<br>75479<br>10075<br>688   | 0.013<br>74325<br>61153<br>94122     | 0.361<br>85181<br>01014<br>617       | -<br>0.111<br>73420<br>91872<br>1436  | 0.074<br>35907<br>47154<br>4469      | 1                                    | -<br>0.184<br>63012<br>54954<br>0175 |
| col9         | 0.389<br>73902<br>40949<br>085       | 0.124<br>33851<br>38945<br>5183      | 0.134<br>76160<br>75375<br>6655      | 0.491<br>20190<br>74930<br>449        | 0.074<br>97231<br>55918<br>4887      | -<br>0.171<br>39609<br>57286<br>885  | -<br>0.103<br>63860<br>37834<br>7944  | 0.117<br>11976<br>05199<br>9162      | -<br>0.184<br>63012<br>54954<br>0175 | 1                                    |

## 3.4.14. Normality Test

Normality test is a special goodness-of-fit hypothesis test in statistical determination. It determines whether the population follows normal distribution by using observations. This topic describes the Normality Test component provided by Machine Learning Studio.

The Normality Test component consists of **Anderson-Darling Test**, **Kolmogorov-Smirnov Test**, and **Q-Q Plot** tests. You can select one or more test methods based on your business requirements.

- The Anderson-Darling test compares the empirical distribution function of sample data with the expected normal distribution. If the difference is large, the test negates the hypothesis that the population has a normal distribution.
- The Kolmogorov-Smirnov test compares the distribution of two observations.
- A Q-Q plot tests the distribution of data by comparing the quantile of test sample data with the known distribution. If more than 1,000 samples are collected, the system uses these samples for calculation and generates a Q-Q plot. The data points in the plot do not necessarily cover all the samples.

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                       | Parameter                       | Description                                                                                                |
|---------------------------|---------------------------------|------------------------------------------------------------------------------------------------------------|
| <b>Fields Setting</b>     | Columns                         | N/A.                                                                                                       |
| <b>Parameters Setting</b> | Anderson-Darling Test           | Valid values:<br><ul style="list-style-type: none"> <li>◦ Yes</li> <li>◦ No</li> </ul> Default value: Yes. |
|                           | Kolmogorov-Smirnov Test         | Valid values:<br><ul style="list-style-type: none"> <li>◦ Yes</li> <li>◦ No</li> </ul> Default value: Yes. |
|                           | Use Q-Q Plot                    | Valid values:<br><ul style="list-style-type: none"> <li>◦ Yes</li> <li>◦ No</li> </ul> Default value: Yes. |
| <b>Tuning</b>             | Computing Cores                 | The number of cores used in computing. The value must be a positive integer.                               |
|                           | Memory Size per Core (Unit: MB) | The memory of each core.                                                                                   |

- PAI command

```
PAI -name normality_test
-project algo_public
-DinputTableName=test
-DoutputTableName=test_out
-DselectedColNames=col1,col2
-Dlifecycle=1;
```

| Parameter            | Required | Description                                                                                                                                                                        | Default value    |
|----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                       | No default value |
| outputTableName      | Yes      | The names of output tables.                                                                                                                                                        | No default value |
| selectedColNames     | No       | The columns selected from the input table. You can select multiple columns of the DOUBLE or BIGINT type.                                                                           | No default value |
| inputTablePartitions | No       | The name of the partition of the input table.                                                                                                                                      | ""               |
| enableQQplot         | No       | Specifies whether to use Q-Q plot testing. Valid values: true and false.                                                                                                           | true             |
| enableADtest         | No       | Optional. This parameter specifies whether to perform the Anderson-Darling test. Valid values: true and false.                                                                     | true             |
| enableKStest         | No       | Specifies whether to perform the Kolmogorov-Smirnov test. Valid values: true and false.                                                                                            | true             |
| lifecycle            | No       | The lifecycle of the output table. The value is an integer that is greater than or equal to -1.                                                                                    | -1               |
| coreNum              | No       | This parameter is used with memSizePerCore. The value must be a positive integer. The system calculates the number of instances based on the amount of input data.                 | -1               |
| memSizePerCore       | No       | The memory size of each core. Unit: MB. The value is a positive integer in the range of (100, 64 × 1024). The system calculates the memory size based on the amount of input data. | -1               |

## Example

- Input

```
drop table if exists normality_test_input;
create table normality_test_input as
select
*
from
(
select 1 as x from dual
union all
select 2 as x from dual
union all
select 3 as x from dual
union all
select 4 as x from dual
union all
select 5 as x from dual
union all
select 6 as x from dual
union all
select 7 as x from dual
union all
select 8 as x from dual
union all
select 9 as x from dual
union all
select 10 as x from dual
) tmp;
```

- PAI command

```
PAI -name normality_test
-project algo_public
-DinputTableName=normality_test_input
-DoutputTableName=normality_test_output
-DselectedColNames=x
-Dlifecycle=1;
```

- Input description

Input format: Select columns required for calculation. You can select multiple columns. The data type is DOUBLE or BIGINT.

- Output description

Output format: A diagram and a result table are provided. The following table lists fields in the result table. The result table has two partitions:

- The partition `p=Test` lists the results of a Anderson-Darling or Kolmogorov-Smirnov test. Data is provided if `enableADtest` or `enableKStest` is set to true.
- The partition `p=plot` lists the results of a Q-Q plot test. Data is provided if `enableQQplot` is set to true. The column `p=test` is reused. If the partition `p='plot'` is used, the `testvalue` column records the original observation (x-axis of the Q-Q plot), and the `pvalue` column records the expected data that is normally distributed (y-axis of the Q-Q plot).



| Column name | Data type | Description                              |
|-------------|-----------|------------------------------------------|
| colName     | STRING    | Column name                              |
| testname    | STRING    | Test name                                |
| testvalue   | DOUBLE    | Test value or the x-axis of the Q-Q plot |
| pvalue      | DOUBLE    | p value or the y-axis of the Q-Q plot    |
| p           | DOUBLE    | The partition name                       |

## Output

```

+-----+-----+-----+-----+-----+
| colname | testname | testvalue | pvalue | p |
+-----+-----+-----+-----+-----+
| x | NULL | 1.0 | 0.8173291742279805 | plot |
| x | NULL | 2.0 | 2.470864450785345 | plot |
| x | NULL | 3.0 | 3.5156067948020056 | plot |
| x | NULL | 4.0 | 4.3632330349313095 | plot |
| x | NULL | 5.0 | 5.128868067945126 | plot |
| x | NULL | 6.0 | 5.871131932054874 | plot |
| x | NULL | 7.0 | 6.6367669650686905 | plot |
| x | NULL | 8.0 | 7.4843932051979944 | plot |
| x | NULL | 9.0 | 8.529135549214654 | plot |
| x | NULL | 10.0 | 10.182670825772018 | plot |
| x | Anderson_Darling_Test | 0.1411092332197832 | 0.9566579606430077 | test |
| x | Kolmogorov_Smirnov_Test | 0.09551932503797644 | 0.9999888659426232 | test |
+-----+-----+-----+-----+-----+

```

### 3.4.15. Lorenz Curve

A Lorenz curve can be used to show the income distribution of a country or region. The slope of the curve indicates the degree of income inequality. The greater the slope, the more unequal the income distribution.

In a rectangle, the height represents the total wealth and is equally divided into N parts. The length represents the families arranged from least wealthy to most wealthy. The length is also equally divided into N parts. The first part indicates the least wealthy 1/N families. The points, each of which indicates a wealth proportion of 1/N families, are connected to form a Lorenz curve.

#### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab            | Parameter | Description |
|----------------|-----------|-------------|
| Fields Setting | Columns   | N/A.        |

| Tab                | Parameter                       | Description                                                                  |
|--------------------|---------------------------------|------------------------------------------------------------------------------|
| Parameters Setting | Quantile                        | Default value: 100.                                                          |
| Tuning             | Computing Cores                 | The number of cores used in computing. The value must be a positive integer. |
|                    | Memory Size per Core (Unit: MB) | The memory of each core.                                                     |

- PAI command

```
PAI -name LorenzCurve
-project algo_public
-DinputTableName=maple_test_lorenz_basic10_input
-DcolName=col0
-DoutputTableName=maple_test_lorenz_basic10_output -DcoreNum=20
-DmemSizePerCore=110;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                            | Default value            |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                           | No default value         |
| outputTableName      | Yes      | The names of output tables.                                                                                                                                                                                                                                                                                                                                                                                                            | No default value         |
| ColName              | No       | The columns selected from the input table. You can select multiple columns and separate them with commas (,).                                                                                                                                                                                                                                                                                                                          | No default value         |
| N                    | No       | Quantile.                                                                                                                                                                                                                                                                                                                                                                                                                              | 100                      |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. The system supports the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (,).         </div> | No default value         |
| lifecycle            | No       | The lifecycle of the output table. This value must be an integer. Unit: days.                                                                                                                                                                                                                                                                                                                                                          | 28                       |
| coreNum              | No       | This parameter is used with memSizePerCore. The value must be a positive integer. The system calculates the number of instances based on the amount of input data.                                                                                                                                                                                                                                                                     | Automatically calculated |

| Parameter      | Required | Description                                                                                                | Default value            |
|----------------|----------|------------------------------------------------------------------------------------------------------------|--------------------------|
| memSizePerCore | No       | The memory size of each core. Unit: MB. The value is a positive integer in the range of (1024, 64 × 1024). | Automatically calculated |

## Example

1. Generate the following test data.

| col0:double |
|-------------|
| 4           |
| 7           |
| 2           |
| 8           |
| 6           |
| 3           |
| 9           |
| 5           |
| 0           |
| 1           |
| 10          |

2. Run the following PAI command:

```
PAI -name LorenzCurve
    -project algo_public
    -DinputTableName=maple_test_lorenz_basic10_input
    -DcolName=col0
    -DoutputTableName=maple_test_lorenz_basic10_output
    -DcoreNum=20
    -DmemSizePerCore=110;
```

3. View the output as shown in the following table.

| quantile | col0                |
|----------|---------------------|
| 0        | 0                   |
| 1        | 0.01818181818181818 |
| 2        | 0.01818181818181818 |

| quantile | col0                |
|----------|---------------------|
| 3        | 0.01818181818181818 |
| 4        | 0.01818181818181818 |
| 5        | 0.01818181818181818 |
| 6        | 0.01818181818181818 |
| 7        | 0.01818181818181818 |
| 8        | 0.01818181818181818 |
| 9        | 0.01818181818181818 |
| 10       | 0.01818181818181818 |
| 11       | 0.05454545454545454 |
| 12       | 0.05454545454545454 |
| 13       | 0.05454545454545454 |
| 14       | 0.05454545454545454 |
| ...      | ...                 |
| 85       | 0.8181818181818182  |
| 86       | 0.8181818181818182  |
| 87       | 0.8181818181818182  |
| 88       | 0.8181818181818182  |
| 89       | 0.8181818181818182  |
| 90       | 1                   |
| 91       | 1                   |
| 92       | 1                   |
| 93       | 1                   |
| 94       | 1                   |
| 95       | 1                   |
| 96       | 1                   |
| 97       | 1                   |

| quantile | col0 |
|----------|------|
| 98       | 1    |
| 99       | 1    |
| 100      | 1    |

## 3.5. Machine Learning

### 3.5.1. Binary classification

This topic describes the binary classification algorithms provided by Machine Learning Studio. The binary classification algorithms include the Linear SVM, Logistic Regression for Binary Classification, GBDT Binary Classification, PS-SMART Binary Classification Training, and PS Logistic Regression for Binary Classification components.

#### Linear SVM

A support vector machine (SVM) is a machine learning model based on the statistical learning theory. It minimizes risks and improves the generalization capability of learning machines. This way, empirical risks and confidence intervals are minimized. The Linear SVM component is not implemented by using kernel functions. For more information about how to implement this component, see the "Trust region method for L2-SVM" section in [Trust Region Newton Method for Large-Scale Logistic Regression](#). This component can be used only in a binary classification scenario.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter               | Description                                                                                                                                                                                                                       |
|--------------------|-------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Column          | The input columns. The columns of the BIGINT and DOUBLE data types are supported.                                                                                                                                                 |
|                    | Label Column            | The label column. The columns of the BIGINT, DOUBLE, and STRING data types are supported.                                                                                                                                         |
| Parameters Setting | Positive Sample Label   | The objective reference value. A random label value is selected if you do not specify this parameter. If the difference between the positive example and negative example is large, we recommend that you specify this parameter. |
|                    | Positive Penalty Factor | The weight of the positive example. The default value is 1.0. Valid values: (0,+∞).                                                                                                                                               |
|                    | Negative Penalty Factor | The weight of the negative example. The default value is 1.0. Valid values: (0,+∞).                                                                                                                                               |
|                    | Convergence Coefficient | The convergence error. The default value is 0.001. Valid values: (0,1).                                                                                                                                                           |

| Tab    | Parameter            | Description                                                                                                                       |
|--------|----------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Tuning | Cores                | The number of cores used in computing. The system automatically allocates the number of cores if this parameter is not specified. |
|        | Memory Size per Core | The memory size of each core. The system automatically allocates the memory size if this parameter is not specified. Unit: MB.    |

- PAI command

```
PAI -name LinearSVM -project algo_public
-DinputTableName="bank_data"
-DmodelName="xlab_m_LinearSVM_6143"
-DfeatureColNames="pdays,emp_var_rate,cons_conf_idx"
-DlabelColName="y"
-DpositiveLabel="0";
-DpositiveCost="1.0"
-DnegativeCost="1.0"
-Depsilon="0.001"
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                           | Default value                                    |
|----------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                          | No default value                                 |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | All partitions in the input table                |
| modelName            | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                                                                                                                                         | No default value                                 |
| featureColNames      | Yes      | The feature columns selected from the input table for training.                                                                                                                                                                                                                                                                                                                                                                       | No default value                                 |
| labelColName         | Yes      | The name of the label column.                                                                                                                                                                                                                                                                                                                                                                                                         | No default value                                 |
| positiveLabel        | No       | The value of the positive example.                                                                                                                                                                                                                                                                                                                                                                                                    | A random value selected from the values of label |
| positiveCost         | No       | The weight of the positive example. It is also a positive penalty factor. Valid values: (0, +∞).                                                                                                                                                                                                                                                                                                                                      | 1.0                                              |

| Parameter      | Required | Description                                                                                          | Default value           |
|----------------|----------|------------------------------------------------------------------------------------------------------|-------------------------|
| negativeCost   | No       | The weight of the negative example. It is also a negative penalty factor. Valid values: (0, +∞).     | 1.0                     |
| epsilon        | No       | The convergence coefficient. Valid values: (0,1).                                                    | 0.001                   |
| enableSparse   | No       | Specifies whether the input data is in the sparse format. Valid values: true and false.              | false                   |
| itemDelimiter  | No       | The delimiter used to separate key-value pairs when data in the input table is in the sparse format. | ,                       |
| kvDelimiter    | No       | The delimiter used to separate keys and values when data in the input table is in the sparse format. | :                       |
| coreNum        | No       | The number of cores used in computing. The value of this parameter must be a positive integer.       | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Valid values: 1 to 65536. Unit: MB.                                    | Automatically allocated |

## Example

1. Use the following training data as the input.

| id | y  | f0         | f1        | f2        | f3        | f4        | f5         | f6        | f7        |
|----|----|------------|-----------|-----------|-----------|-----------|------------|-----------|-----------|
| 1  | -1 | -0.294118  | 0.487437  | 0.180328  | -0.292929 | -1        | 0.00149028 | -0.53117  | -0.033333 |
| 2  | +1 | -0.882353  | -0.145729 | 0.0819672 | -0.414141 | -1        | -0.207153  | -0.766866 | -0.666667 |
| 3  | -1 | -0.0588235 | 0.839196  | 0.0491803 | -1        | -1        | -0.305514  | -0.492741 | -0.633333 |
| 4  | +1 | -0.882353  | -0.105528 | 0.0819672 | -0.535354 | -0.777778 | -0.162444  | -0.923997 | -1        |

| id | y  | f0         | f1       | f2        | f3         | f4        | f5         | f6        | f7        |
|----|----|------------|----------|-----------|------------|-----------|------------|-----------|-----------|
| 5  | -1 | -1         | 0.376884 | -0.344262 | -0.292929  | -0.602837 | 0.28465    | 0.887276  | -0.6      |
| 6  | +1 | -0.411765  | 0.165829 | 0.213115  | -1         | -1        | -0.23696   | -0.894962 | -0.7      |
| 7  | -1 | -0.647059  | -0.21608 | -0.180328 | -0.353535  | -0.791962 | -0.0760059 | -0.854825 | -0.833333 |
| 8  | +1 | 0.176471   | 0.155779 | -1        | -1         | -1        | 0.052161   | -0.952178 | -0.733333 |
| 9  | -1 | -0.764706  | 0.979899 | 0.147541  | -0.0909091 | 0.283688  | -0.0909091 | -0.931682 | 0.0666667 |
| 10 | -1 | -0.0588235 | 0.256281 | 0.57377   | -1         | -1        | -1         | -0.868488 | 0.1       |

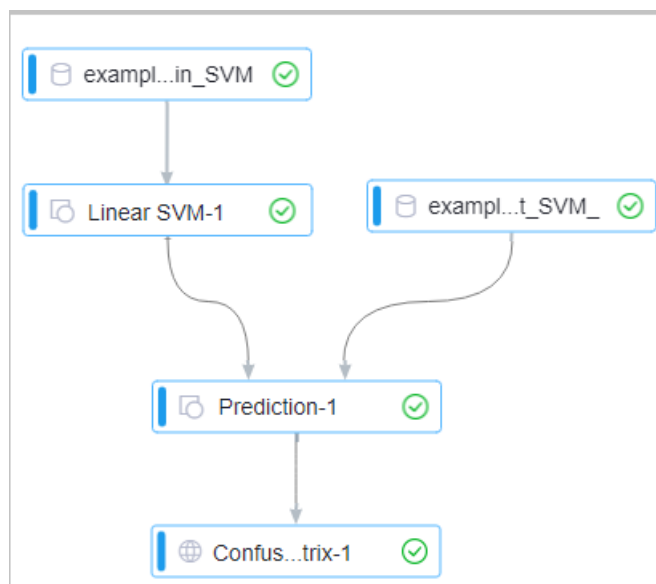
2. Use the following test data as the input.

| id | y  | f0        | f1         | f2         | f3        | f4        | f5        | f6        | f7        |
|----|----|-----------|------------|------------|-----------|-----------|-----------|-----------|-----------|
| 1  | +1 | -0.882353 | 0.0854271  | 0.442623   | -0.616162 | -1        | -0.19225  | -0.725021 | -0.9      |
| 2  | +1 | -0.294118 | -0.0351759 | -1         | -1        | -1        | -0.293592 | -0.904355 | -0.766667 |
| 3  | +1 | -0.882353 | 0.246231   | 0.213115   | -0.272727 | -1        | -0.171386 | -0.981213 | -0.7      |
| 4  | -1 | -0.176471 | 0.507538   | 0.278689   | -0.414141 | -0.702128 | 0.0491804 | -0.475662 | 0.1       |
| 5  | -1 | -0.529412 | 0.839196   | -1         | -1        | -1        | -0.153502 | -0.885568 | -0.5      |
| 6  | +1 | -0.882353 | 0.246231   | -0.0163934 | -0.353535 | -1        | 0.0670641 | -0.627669 | -1        |



| id | y  | f0        | f1         | f2        | f3        | f4        | f5        | f6         | f7        |
|----|----|-----------|------------|-----------|-----------|-----------|-----------|------------|-----------|
| 7  | -1 | -0.882353 | 0.819095   | 0.278689  | -0.151515 | -0.307329 | 0.19225   | 0.00768574 | -0.966667 |
| 8  | +1 | -0.882353 | -0.0753769 | 0.0163934 | -0.494949 | -0.903073 | -0.418778 | -0.654996  | -0.866667 |
| 9  | +1 | -1        | 0.527638   | 0.344262  | -0.212121 | -0.356974 | 0.23696   | -0.836038  | -0.8      |
| 10 | +1 | -0.882353 | 0.115578   | 0.0163934 | -0.737374 | -0.56974  | -0.28465  | -0.948762  | -0.933333 |

3. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).



4. Configure the parameters listed in the following table for the Linear SVM component. Retain the default values of the parameters that are not listed in the table.

| Tab            | Parameter      | Description                                            |
|----------------|----------------|--------------------------------------------------------|
| Fields Setting | Feature Column | Select the f0, f1, f2, f3, f4, f5, f6, and f7 columns. |
|                | Label Column   | Select the y column.                                   |

5. Run the experiment and view the prediction result.

| id ▲ | y ▲ | prediction_result ▲ | prediction_score ▲    | prediction_detail ▲                                     |
|------|-----|---------------------|-----------------------|---------------------------------------------------------|
| 1    | +1  | +1                  | 0.157594271402295...  | {"+1": 0.1575942714022959, "-1": -0.1575942714022959}   |
| 2    | +1  | +1                  | 0.6614698832250897    | {"+1": 0.6614698832250897, "-1": -0.6614698832250897}   |
| 3    | +1  | -1                  | -0.033364749587674... | {"+1": -0.03336474958767432, "-1": 0.03336474958767432} |
| 4    | -1  | -1                  | -0.7993936496277533   | {"+1": -0.7993936496277533, "-1": 0.7993936496277533}   |
| 5    | -1  | -1                  | -0.062511633256538... | {"+1": -0.06251163325653875, "-1": 0.06251163325653875} |
| 6    | +1  | +1                  | 0.099662488068409...  | {"+1": 0.09966248806840972, "-1": -0.09966248806840972} |
| 7    | -1  | -1                  | -0.7519810414882623   | {"+1": -0.7519810414882623, "-1": 0.7519810414882623}   |
| 8    | +1  | +1                  | 0.180293970573986...  | {"+1": 0.1802939705739862, "-1": -0.1802939705739862}   |
| 9    | +1  | -1                  | -0.1196222721567562   | {"+1": -0.1196222721567562, "-1": 0.1196222721567562}   |
| 10   | +1  | +1                  | 0.4110490612942082    | {"+1": 0.4110490612942082, "-1": -0.4110490612942082}   |

## Logistic Regression for Binary Classification

The Logistic Regression for Binary Classification component supports data in both the sparse and dense formats. You can configure the component by using one of the following methods:

- PAI console

| Tab                | Parameter                    | Description                                                                                                                                                                                       |
|--------------------|------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Training Feature Columns     | The feature columns selected from the input table for training. The columns of the DOUBLE and BIGINT data types are supported.<br><br><b>Note</b> A maximum of 20 million features are supported. |
|                    | Target Columns               | The objective columns in the input table.                                                                                                                                                         |
|                    | Positive Class Value         | None.                                                                                                                                                                                             |
|                    | Use Sparse Format            | Specifies whether data in the input table is in the sparse format.                                                                                                                                |
| Parameters Setting | Regularization Type          | The regularization type. Valid values: <b>None</b> , <b>L1</b> , and <b>L2</b> .                                                                                                                  |
|                    | Maximum Iterations           | The maximum number of iterations. Default value: 100.                                                                                                                                             |
|                    | Regularization Coefficient   | The regularization coefficient. This parameter is invalid when Regularization Type is set to None.                                                                                                |
|                    | Minimum Convergence Deviance | The minimum convergence error. Default value: 0.000001.                                                                                                                                           |
|                    | Cores                        | The number of cores. It is automatically allocated by the system.                                                                                                                                 |

| Tuning Tab | Parameter            | Description                                                                |
|------------|----------------------|----------------------------------------------------------------------------|
|            | Memory Size per Core | The memory size of each core. It is automatically allocated by the system. |

- PAI command

```
PAI -name logisticregression_binary
  -project algo_public
  -DmodelName="xlab_m_logistic_regression_6096"
  -DregularizedLevel="1"
  -DmaxIter="100"
  -DregularizedType="l1"
  -DEpsilon="0.000001"
  -DlabelColName="y"
  -DfeatureColNames="pdays,emp_var_rate"
  -DgoodValue="1"
  -DinputTableName="bank_data"
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                    | Default value                     |
|----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                   | No default value                  |
| featureColNames      | No       | The feature columns selected from the input table for training. <div>  <b>Note</b> A maximum of 20 million features are supported.           </div>                                                                                                                                                                                                         | All columns of numeric data types |
| labelColName         | Yes      | The name of the label column in the input table.                                                                                                                                                                                                                                                                                                                                                                                               | No default value                  |
| inputTablePartitions | No       | The partitions selected from the input table for training. Specify this parameter in one of the following formats: <ul style="list-style-type: none"> <li>partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (.).           </div> | Full table                        |
| modelName            | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                                                                                                                                                  | No default value                  |
| regularizedType      | No       | The regularization type. Valid values: l1, l2, and None.                                                                                                                                                                                                                                                                                                                                                                                       | l1                                |

| Parameter        | Required | Description                                                                                                                                                                                                                                                           | Default value           |
|------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| regularizedLevel | No       | The regularization coefficient. This parameter is invalid when regularizedType is set to None.                                                                                                                                                                        | 1.0                     |
| maxIter          | No       | The maximum number of iterations of the limited-memory BFGS (L-BFGS) algorithm.                                                                                                                                                                                       | 100                     |
| epsilon          | No       | The convergence error. This parameter specifies the conditions to terminate the L-BFGS algorithm. If the difference of log-likelihood between two iterations is less than the value specified by this parameter, the iteration of the L-BFGS algorithm is terminated. | 1.0e-06                 |
| goodValue        | No       | The objective reference value. This parameter specifies the value of label that corresponds to the training coefficient in binary classification. The system randomly allocates a value if this parameter is not specified.                                           | No default value        |
| enableSparse     | No       | Specifies whether data in the input table is in the sparse format. Valid values: true and false.                                                                                                                                                                      | false                   |
| itemDelimiter    | No       | The delimiter used to separate key-value pairs when data in the input table is in the sparse format.                                                                                                                                                                  | ,                       |
| kvDelimiter      | No       | The delimiter used to separate keys and values when data in the input table is in the sparse format.                                                                                                                                                                  | :                       |
| coreNum          | No       | The number of cores.                                                                                                                                                                                                                                                  | Automatically allocated |
| memSizePerCore   | No       | The memory size of each core. Unit: MB.                                                                                                                                                                                                                               | Automatically allocated |

In Machine Learning Studio, sparse data is presented in the key-value format. The following table lists some examples. `itemDelimiter` specifies the delimiter used to separate key-value pairs, and `kvDelimiter` specifies the delimiter used to separate keys and values.

| key_value         |
|-------------------|
| 1:100,4:200,5:300 |
| 1:10,2:20,3:30    |

 **Note** For the key-value format, keys are indicated from 0 by using indexes. If characters are used to indicate the values of keys, the system reports an error.

## Example

1. Execute the following SQL statement to generate training data:

```
drop table if exists lr_test_input;
create table lr_test_input
as
select
  *
from
(
  select
    cast(1 as double) as f0,
    cast(0 as double) as f1,
    cast(0 as double) as f2,
    cast(0 as double) as f3,
    cast(0 as bigint) as label
  from dual
  union all
  select
    cast(0 as double) as f0,
    cast(1 as double) as f1,
    cast(0 as double) as f2,
    cast(0 as double) as f3,
    cast(0 as bigint) as label
  from dual
  union all
  select
    cast(0 as double) as f0,
    cast(0 as double) as f1,
    cast(1 as double) as f2,
    cast(0 as double) as f3,
    cast(1 as bigint) as label
  from dual
  union all
  select
    cast(0 as double) as f0,
    cast(0 as double) as f1,
    cast(0 as double) as f2,
    cast(1 as double) as f3,
    cast(1 as bigint) as label
  from dual
  union all
  select
    cast(1 as double) as f0,
    cast(0 as double) as f1,
    cast(0 as double) as f2,
    cast(0 as double) as f3,
    cast(0 as bigint) as label
  from dual
)
```

```

union all
select
  cast(0 as double) as f0,
  cast(1 as double) as f1,
  cast(0 as double) as f2,
  cast(0 as double) as f3,
  cast(0 as bigint) as label
from dual
) a;

```

View the following training data table lr\_test\_input.

| f0  | f1  | f2  | f3  | label |
|-----|-----|-----|-----|-------|
| 1.0 | 0.0 | 0.0 | 0.0 | 0     |
| 0.0 | 0.0 | 1.0 | 0.0 | 1     |
| 0.0 | 0.0 | 0.0 | 1.0 | 1     |
| 0.0 | 1.0 | 0.0 | 0.0 | 0     |
| 1.0 | 0.0 | 0.0 | 0.0 | 0     |
| 0.0 | 1.0 | 0.0 | 0.0 | 0     |

- Run the following PAI command to submit the training parameters configured for the Logistic Regression for Binary Classification component:

```

drop offlinemodel if exists lr_test_model;
PAI -name logisticregression_binary
  -project algo_public
  -DmodelName="lr_test_model"
  -DitemDelimiter=","
  -DregularizedLevel="1"
  -DmaxIter="100"
  -DregularizedType="None"
  -DEpsilon="0.000001"
  -DkvDelimiter=":"
  -DlabelColName="label"
  -DfeatureColNames="f0,f1,f2,f3"
  -DenableSparse="false"
  -DgoodValue="1"
  -DinputTableName="lr_test_input";

```

- Run the following PAI command to submit the parameters configured for the Prediction component:

```
drop table if exists lr_test_prediction_result;
PAI -name prediction
  -project algo_public
  -DdetailColName="prediction_detail"
  -DmodelName="lr_test_model"
  -DitemDelimiter=","
  -DresultColName="prediction_result"
  -Dlifecycle="28"
  -DoutputTableName="lr_test_prediction_result"
  -DscoreColName="prediction_score"
  -DkvDelimiter=":"
  -DinputTableName="lr_test_input"
  -DenableSparse="false"
  -DappendColNames="label";
```

4. View the prediction result table lr\_test\_prediction\_result.

| label | prediction_result | prediction_score   | prediction_detail                                     |
|-------|-------------------|--------------------|-------------------------------------------------------|
| 0     | 0                 | 0.9999998793434426 | {"0": 0.9999998793434426, "1": 1.206565574533681e-07} |
| 1     | 1                 | 0.999999799574135  | {"0": 2.004258650156743e-07, "1": 0.999999799574135}  |
| 1     | 1                 | 0.999999799574135  | {"0": 2.004258650156743e-07, "1": 0.999999799574135}  |
| 0     | 0                 | 0.9999998793434426 | {"0": 0.9999998793434426, "1": 1.206565574533681e-07} |
| 0     | 0                 | 0.9999998793434426 | {"0": 0.9999998793434426, "1": 1.206565574533681e-07} |
| 0     | 0                 | 0.9999998793434426 | {"0": 0.9999998793434426, "1": 1.206565574533681e-07} |

## GBDT Binary Classification

The Gradient Boosting Decision Tree (GBDT) Binary Classification component is used to set a threshold. If a feature value is greater than the threshold, the feature is a positive example. Otherwise, the feature is a negative example.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab | Parameter | Description |
|-----|-----------|-------------|
|-----|-----------|-------------|

| Tab                | Parameter                     | Description                                                                                                                                                                                                                                                                                |
|--------------------|-------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns               | <p>The feature columns selected from the input table for training. The columns of the DOUBLE and BIGINT data types are supported.</p> <p> <b>Note</b> A maximum of 800 feature columns are supported.</p> |
|                    | Label Column                  | The label column. Only the column of the BIGINT data type is supported.                                                                                                                                                                                                                    |
|                    | Stratification Column         | The columns of the DOUBLE and BIGINT data types are supported. By default, the full table is a group.                                                                                                                                                                                      |
| Parameters Setting | Metric Type                   | The metric type. Valid values: <b>NDCG</b> and <b>DCG</b> .                                                                                                                                                                                                                                |
|                    | Trees                         | The number of trees. Valid values: 1 to 10000.                                                                                                                                                                                                                                             |
|                    | Learning Rate                 | The learning rate. Valid values: (0,1).                                                                                                                                                                                                                                                    |
|                    | Training Sample Fraction      | The ratio of samples selected for training. Valid values: (0,1).                                                                                                                                                                                                                           |
|                    | Training Feature Fraction     | The ratio of features selected for training. Valid values: (0,1).                                                                                                                                                                                                                          |
|                    | Maximum Leaves                | The maximum number of leaf nodes of each tree. Valid values: 1 to 1000.                                                                                                                                                                                                                    |
|                    | Testing Data Fraction         | The ratio of data selected for testing. Valid values: [0,1).                                                                                                                                                                                                                               |
|                    | Maximum Tree Depth            | The maximum depth of each tree. Valid values: 1 to 100.                                                                                                                                                                                                                                    |
|                    | Minimum Samples per Leaf Node | The minimum number of samples on each leaf node. Valid values: 1 to 1000.                                                                                                                                                                                                                  |
|                    | Random Seed                   | The random seed. Valid values:[0,10].                                                                                                                                                                                                                                                      |
| Tuning             | Maximum Splits per Feature    | The maximum number of splits of each feature. Valid values: 1 to 1000.                                                                                                                                                                                                                     |
|                    | Cores                         | The number of cores. The system automatically allocates the number of cores used for training based on the volume of input data.                                                                                                                                                           |
|                    | Memory Size                   | The memory size of each core. The system automatically allocates the memory size based on the volume of input data. Valid values: 1024 to 64 × 1024. Unit: MB.                                                                                                                             |



- PAI command

```
PAI -name gbd_t_lr
  -project algo_public
  -DfeatureSplitValueMaxSize="500"
  -DrandSeed="0"
  -Dshrinkage="0.5"
  -DmaxLeafCount="32"
  -DlabelColName="y"
  -DinputTableName="bank_data_partition"
  -DminLeafSampleCount="500"
  -DgroupIdColName="nr_employed"
  -DsampleRatio="0.6"
  -DmaxDepth="11"
  -DmodelName="xlab_m_GBDT_LR_21208"
  -DmetricType="2"
  -DfeatureRatio="0.6"
  -DinputTablePartitions="pt=20150501"
  -DtestRatio="0.0"
  -DfeatureColNames="age,previous,cons_conf_idx,euribor3m"
  -DtreeCount="500"
```

| Parameter                 | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                           | Default value                     |
|---------------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTableName            | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                          | No default value                  |
| featureColNames           | No       | The feature columns selected from the input table for training.                                                                                                                                                                                                                                                                                                                                                                       | All columns of numeric data types |
| labelColName              | Yes      | The name of the label column in the input table.                                                                                                                                                                                                                                                                                                                                                                                      | No default value                  |
| inputTablePartitions      | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | All partitions                    |
| modelName                 | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                                                                                                                                         | No default value                  |
| outputImportanceTableName | No       | The name of the output feature importance table.                                                                                                                                                                                                                                                                                                                                                                                      | No default value                  |
| groupIdColName            | No       | The name of the group column.                                                                                                                                                                                                                                                                                                                                                                                                         | Full table                        |

| Parameter          | Required | Description                                                                                                                                                                                                                                                                               | Default value |
|--------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| lossType           | No       | The loss function type. The following types of loss functions are supported: <ul style="list-style-type: none"> <li>0: GBRANK</li> <li>1: LAMBDAMART_DCG</li> <li>2: LAMBDAMART_NDCG</li> <li>3: LEAST_SQUARE</li> <li>4: LOG_LIKELIHOOD</li> </ul>                                       | 0             |
| metricType         | No       | The metric type. The following metric types are supported: <ul style="list-style-type: none"> <li>0: NDCG (Normalized Discounted Cumulative Gain)</li> <li>1: DCG (Discounted Cumulative Gain)</li> <li>2: AUC. This parameter is applicable only when label is set to 0 or 1.</li> </ul> | 0             |
| treeCount          | No       | The number of trees. Valid values: 1 to 10000.                                                                                                                                                                                                                                            | 500           |
| shrinkage          | No       | The learning rate. Valid values: (0,1).                                                                                                                                                                                                                                                   | 0.05          |
| maxLeafCount       | No       | The maximum number of leaf nodes of each tree. Valid values: 1 to 1000.                                                                                                                                                                                                                   | 32            |
| maxDepth           | No       | The maximum depth of each tree. Valid values: 1 to 100.                                                                                                                                                                                                                                   | 10            |
| minLeafSampleCount | No       | The minimum number of samples on each leaf node. Valid values: 1 to 1000.                                                                                                                                                                                                                 | 500           |
| sampleRatio        | No       | The ratio of samples selected for training. Valid values: (0,1).                                                                                                                                                                                                                          | 0.6           |
| featureRatio       | No       | The ratio of features selected for training. Valid values: (0,1).                                                                                                                                                                                                                         | 0.6           |
| tau                | No       | The Tau parameter for the GBRank loss function. Valid values: [0,1].                                                                                                                                                                                                                      | 0.6           |
| p                  | No       | The p parameter for the GBRank loss function. Valid values: [1,10].                                                                                                                                                                                                                       | 1             |
| randSeed           | No       | The random seed. Valid values: [0,10].                                                                                                                                                                                                                                                    | 0             |
| newtonStep         | No       | Specifies whether to use Newton's method. Valid values: 0 and 1.                                                                                                                                                                                                                          | 1             |

| Parameter                | Required | Description                                                            | Default value    |
|--------------------------|----------|------------------------------------------------------------------------|------------------|
| featureSplitValueMaxSize | No       | The maximum number of splits of each feature. Valid values: 1 to 1000. | 500              |
| lifecycle                | No       | The lifecycle of the output table.                                     | No default value |

#### Note

- By default, the loss function types of GBDT and those of GBDT\_LR are different. By default, GBDT uses regression loss: mean squared error loss as its loss function, and GBDT\_LR uses logistic regression loss as its loss function. Therefore, you do not need to configure a loss function for GBDT\_LR.
- The feature column, label column, and stratification column of GBDT must be of numeric data types.
- You must specify the objective reference value for the Prediction component to generate an ROC curve.

## Example

1. Execute the following SQL statement to generate training data:

```
drop table if exists gbdt_lr_test_input;
create table gbdt_lr_test_input
as
select
  *
from
(
  select
    cast(1 as double) as f0,
    cast(0 as double) as f1,
    cast(0 as double) as f2,
    cast(0 as double) as f3,
    cast(0 as bigint) as label
  from dual
  union all
  select
    cast(0 as double) as f0,
    cast(1 as double) as f1,
    cast(0 as double) as f2,
    cast(0 as double) as f3,
    cast(0 as bigint) as label
  from dual
  union all
  select
    cast(0 as double) as f0,
    cast(0 as double) as f1,
    cast(1 as double) as f2,
    cast(0 as double) as f3,
    cast(1 as bigint) as label
```

```
from dual
union all
select
  cast(0 as double) as f0,
  cast(0 as double) as f1,
  cast(0 as double) as f2,
  cast(1 as double) as f3,
  cast(1 as bigint) as label
from dual
union all
select
  cast(1 as double) as f0,
  cast(0 as double) as f1,
  cast(0 as double) as f2,
  cast(0 as double) as f3,
  cast(0 as bigint) as label
from dual
union all
select
  cast(0 as double) as f0,
  cast(1 as double) as f1,
  cast(0 as double) as f2,
  cast(0 as double) as f3,
  cast(0 as bigint) as label
from dual
) a;
```

The following training data table gbdt\_lr\_test\_input is generated.

| f0  | f1  | f2  | f3  | label |
|-----|-----|-----|-----|-------|
| 1.0 | 0.0 | 0.0 | 0.0 | 0     |
| 0.0 | 0.0 | 1.0 | 0.0 | 1     |
| 0.0 | 0.0 | 0.0 | 1.0 | 1     |
| 0.0 | 1.0 | 0.0 | 0.0 | 0     |
| 1.0 | 0.0 | 0.0 | 0.0 | 0     |
| 0.0 | 1.0 | 0.0 | 0.0 | 0     |

2. Run the following PAI command to submit the training parameters configured for the GBDT Binary Classification component:

```
drop offlinemodel if exists gbdt_lr_test_model;
PAI -name gbdt_lr
  -project algo_public
  -DfeatureSplitValueMaxSize="500"
  -DrandSeed="1"
  -Dshrinkage="1"
  -DmaxLeafCount="30"
  -DlabelColName="label"
  -DinputTableName="gbdt_lr_test_input"
  -DminLeafSampleCount="1"
  -DsampleRatio="1"
  -DmaxDepth="10"
  -DmodelName="gbdt_lr_test_model"
  -DmetricType="0"
  -DfeatureRatio="1"
  -DtestRatio="0"
  -DfeatureColNames="f0,f1,f2,f3"
  -DtreeCount="5"
```

- Run the following PAI command to submit the parameters configured for the Prediction component:

```
drop table if exists gbdt_lr_test_prediction_result;
PAI -name prediction
  -project algo_public
  -DdetailColName="prediction_detail"
  -DmodelName="gbdt_lr_test_model"
  -DitemDelimiter=","
  -DresultColName="prediction_result"
  -Dlifecycle="28"
  -DoutputTableName="gbdt_lr_test_prediction_result"
  -DscoreColName="prediction_score"
  -DkvDelimiter=":"
  -DinputTableName="gbdt_lr_test_input"
  -DenableSparse="false"
  -DappendColNames="label";
```

- View the prediction result table gbdt\_lr\_test\_prediction\_result.

| label | prediction_result | prediction_score   | prediction_detail                                    |
|-------|-------------------|--------------------|------------------------------------------------------|
| 0     | 0                 | 0.9984308925552831 | {"0": 0.9984308925552831, "1": 0.001569107444716943} |
| 0     | 0                 | 0.9984308925552831 | {"0": 0.9984308925552831, "1": 0.001569107444716943} |
| 1     | 1                 | 0.9982721832240973 | {"0": 0.001727816775902724, "1": 0.9982721832240973} |
| 1     | 1                 | 0.9982721832240973 | {"0": 0.001727816775902724, "1": 0.9982721832240973} |

| label | prediction_result | prediction_score   | prediction_detail                                    |
|-------|-------------------|--------------------|------------------------------------------------------|
| 0     | 0                 | 0.9984308925552831 | {"0": 0.9984308925552831, "1": 0.001569107444716943} |
| 0     | 0                 | 0.9984308925552831 | {"0": 0.9984308925552831, "1": 0.001569107444716943} |

## PS-SMART Binary Classification Training

Parameter servers are used to perform large-scale online and offline training tasks. Scalable Multiple Additive Regression Tree (SMART) is a parameter server-based iterative algorithm. It is developed based on Gradient Boosting Decision Tree (GBDT). The PS-SMART Binary Classification Training component supports training tasks for tens of billions of samples and hundreds of thousands of features. It can run the training tasks on thousands of nodes. This component also supports multiple data formats and optimization techniques such as approximation by using histograms.

You can configure the component by using one of the following methods:

- PAI console

| Tab            | Parameter                 | Description                                                                                                                                                                                                                                                                                                                                                                    |
|----------------|---------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Use Sparse Format         | Separate multiple key-value pairs in the sparse format with spaces and separate keys and values with colons (:). Example: 1:0.3 3:0.9.                                                                                                                                                                                                                                         |
|                | Feature Columns           | The feature columns selected from the input table for training. If data in the input table is in the dense format, only columns of the BIGINT or DOUBLE data types are supported. If data in the input table is key-value pairs in the sparse format and keys and values in the key-value pairs are of numeric data types, only columns of the STRING data type are supported. |
|                | Label Column              | The label column in the input table. The columns of the STRING data type and numeric data types are supported. However, only data of numeric data types can be stored in the columns. For example, column values can be 0 or 1 in binary classification.                                                                                                                       |
|                | Weight Column             | The weight column. This column contains the weight of each sample. The columns of numeric data types are supported.                                                                                                                                                                                                                                                            |
|                | Evaluation Indicator Type | The evaluation metric type of the training set. Valid values: <b>Negative Loglikelihood for Logistic Regression, Binary Classification Error, and AUC for Classification.</b>                                                                                                                                                                                                  |
|                | Trees                     | The number of trees. The volume of data used for training is in proportion to the value of <b>Trees</b> .                                                                                                                                                                                                                                                                      |
|                |                           |                                                                                                                                                                                                                                                                                                                                                                                |

| Tab                | Parameter                                 | Description                                                                                                                                                                                                                                                                                    |
|--------------------|-------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | <b>Maximum Tree Depth</b>                 | The maximum depth of each tree. The default value of this parameter is 5. This indicates that a maximum of 32 leaf nodes are supported.                                                                                                                                                        |
|                    | <b>Data Sampling Fraction</b>             | The data sampling ratio when trees are built. The sampled data is used to build a weak learner to accelerate training.                                                                                                                                                                         |
|                    | <b>Feature Sampling Fraction</b>          | The feature sampling ratio when trees are built. The sampled features are used to build a weak learner to accelerate training.                                                                                                                                                                 |
|                    | <b>L1 Penalty Coefficient</b>             | This parameter is used to control the size of each leaf node. The larger the value of this parameter is, the more evenly the leaf nodes are distributed. If overfitting occurs, increase the value of this parameter.                                                                          |
|                    | <b>L2 Penalty Coefficient</b>             | This parameter is used to control the size of each leaf node. The larger the value of this parameter is, the more evenly the leaf nodes are distributed. If overfitting occurs, increase the value of this parameter.                                                                          |
|                    | <b>Learning Rate</b>                      | The learning rate. Valid values: (0,1).                                                                                                                                                                                                                                                        |
|                    | <b>Sketch-based Approximate Precision</b> | The threshold for selecting quantiles when you build a sketch. The smaller the value of this parameter is, the more bins can be obtained. In most cases, the default value 0.03 is used.                                                                                                       |
|                    | <b>Minimum Split Loss Change</b>          | The minimum loss change required for splitting a node. The larger the value of this parameter is, the less likely the node is to split.                                                                                                                                                        |
|                    | <b>Features</b>                           | The number of features or the maximum feature ID. If this parameter is not specified during the assessment of resource usage, the system automatically runs an SQL task to calculate the number of features or the maximum feature ID.                                                         |
|                    | <b>Global Offset</b>                      | The initial prediction values of all samples.                                                                                                                                                                                                                                                  |
|                    | <b>Feature Importance Type</b>            | The feature importance type. Valid values: <b>Weight</b> , <b>Gain</b> , and <b>Cover</b> . Weight indicates the number of splits of the feature. Gain indicates the information gain provided by the feature. Cover indicates the number of samples covered by the feature on the split node. |
|                    | <b>Cores</b>                              | The number of cores. It is automatically allocated by the system.                                                                                                                                                                                                                              |

| Tuning Tab | Parameter            | Description                                                                            |
|------------|----------------------|----------------------------------------------------------------------------------------|
|            | Memory Size per Core | The memory size of each core. Unit: MB. This is automatically allocated by the system. |

- PAI command

```
#Training
PAI -name ps_smart
  -project algo_public
  -DinputTableName="smart_binary_input"
  -DmodelName="xlab_m_pai_ps_smart_bi_545859_v0"
  -DoutputTableName="pai_temp_24515_545859_2"
  -DoutputImportanceTableName="pai_temp_24515_545859_3"
  -DlabelColName="label"
  -DfeatureColNames="f0,f1,f2,f3,f4,f5"
  -DenableSparse="false"
  -Dobjective="binary:logistic"
  -Dmetric="error"
  -DfeatureImportanceType="gain"
  -DtreeCount="5";
  -DmaxDepth="5"
  -Dshrinkage="0.3"
  -DL2="1.0"
  -DL1="0"
  -Dlifecycle="3"
  -DsketchEps="0.03"
  -DsampleRatio="1.0"
  -DfeatureRatio="1.0"
  -DbaseScore="0.5"
  -DminSplitLoss="0"
#Prediction
PAI -name prediction
  -project algo_public
  -DinputTableName="smart_binary_input";
  -DmodelName="xlab_m_pai_ps_smart_bi_545859_v0"
  -DoutputTableName="pai_temp_24515_545860_1"
  -DfeatureColNames="f0,f1,f2,f3,f4,f5"
  -DappendColNames="label,qid,f0,f1,f2,f3,f4,f5"
  -DenableSparse="false"
  -Dlifecycle="28"
```

| Module | Parameter | Required | Description | Default value |
|--------|-----------|----------|-------------|---------------|
|--------|-----------|----------|-------------|---------------|



| Module          | Parameter                 | Required | Description                                                                                                                                                                                                                                                                                                                                                                   | Default value    |
|-----------------|---------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| Data parameters | featureColNames           | Yes      | The feature columns selected from the input table for training. If data in the input table is in the dense format, only columns of the BIGINT or DOUBLE data type are supported. If data in the input table is key-value pairs in the sparse format and keys and values in the key-value pairs are of numeric data types, only columns of the STRING data type are supported. | No default value |
|                 | labelColName              | Yes      | The label column in the input table. The columns of the STRING data type and numeric data types are supported. However, only data of numeric data types can be stored in the columns. For example, column values can be 0 or 1 in binary classification.                                                                                                                      | No default value |
|                 | weightCol                 | No       | The weight column. This column contains the weight of each sample. The columns of numeric data types are supported.                                                                                                                                                                                                                                                           | No default value |
|                 | enableSparse              | No       | Specifies whether data in the input table is in the sparse format. Valid values: true and false. Separate multiple key-value pairs in the sparse format with spaces and separate keys and values with colons (:). Example: 1:0.3 3:0.9.                                                                                                                                       | false            |
|                 | inputTableName            | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                  | No default value |
|                 | modelName                 | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                                                                                 | No default value |
|                 | outputImportanceTableName | No       | The name of the output feature importance table.                                                                                                                                                                                                                                                                                                                              | No default value |
|                 | inputTablePartitions      | No       | The partition selected from the input table for training. Specify this parameter in the ds=1/pt=1 format.                                                                                                                                                                                                                                                                     | No default value |
|                 | outputTableName           | No       | The MaxCompute table that is generated. The table is a binary file. It cannot be read and can only be obtained by using the prediction component of PS-SMART.                                                                                                                                                                                                                 | No default value |
|                 | lifecycle                 | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                            | 3                |
|                 | objective                 | Yes      | The objective function type. Set the value to binary:logistic if the training is performed by using binary classification components.                                                                                                                                                                                                                                         | No default value |

| Module               | Parameter    | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Default value    |
|----------------------|--------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| Algorithm parameters | metric       | No       | <p>The evaluation metric type of the training set, which is contained in stdout of the coordinator in a logview. The following types are supported:</p> <ul style="list-style-type: none"> <li>◦ logloss: corresponds to the <b>Negative Loglikelihood for Logistic Regression</b> type of Evaluation Indicator Type.</li> <li>◦ error: corresponds to the <b>Binary Classification Error</b> type of Evaluation Indicator Type.</li> <li>◦ auc: corresponds to the <b>AUC for Classification</b> type of Evaluation Indicator Type.</li> </ul> | No default value |
|                      | treeCount    | No       | The number of trees. The value of this parameter is in proportion to the training time.                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 1                |
|                      | maxDepth     | No       | The maximum depth of each tree. Valid values: 1 to 20.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 5                |
|                      | sampleRatio  | No       | The data sampling ratio. Valid values: (0,1]. If the value of this parameter is set to 1.0, no data is sampled.                                                                                                                                                                                                                                                                                                                                                                                                                                 | 1.0              |
|                      | featureRatio | No       | The feature sampling ratio. Valid values: (0,1]. If the value of this parameter is set to 1.0, no features are sampled.                                                                                                                                                                                                                                                                                                                                                                                                                         | 1.0              |
|                      | l1           | No       | L1 penalty coefficient. The larger the value of this parameter is, the more evenly the leaf nodes are distributed. If overfitting occurs, increase the value of this parameter.                                                                                                                                                                                                                                                                                                                                                                 | 0                |
|                      | l2           | No       | L2 penalty coefficient. The larger the value of this parameter is, the more evenly the leaf nodes are distributed. If overfitting occurs, increase the value of this parameter.                                                                                                                                                                                                                                                                                                                                                                 | 1.0              |
|                      | shrinkage    | No       | The learning rate. Valid values: (0,1).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 0.3              |
|                      | sketchEps    | No       | The threshold for selecting quantiles when you build a sketch. The number of bins is $O(1.0/sketchEps)$ . The smaller the value of this parameter is, the more bins can be obtained. In most cases, the default value is used. Valid values: (0,1).                                                                                                                                                                                                                                                                                             | 0.03             |

| Module            | Parameter             | Required | Description                                                                                                                                                                                                                                                                                                            | Default value           |
|-------------------|-----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
|                   | minSplitLoss          | No       | The minimum loss change required for splitting a node. The larger the value of this parameter is, the less likely the node is to split.                                                                                                                                                                                | 0                       |
|                   | featureNum            | No       | The number of features or the maximum feature ID. If this parameter is not specified during the assessment of resource usage, the system automatically runs an SQL task to calculate the number of features or the maximum feature ID.                                                                                 | No default value        |
|                   | baseScore             | No       | The initial prediction values of all samples.                                                                                                                                                                                                                                                                          | 0.5                     |
|                   | featureImportanceType | No       | The feature importance type. Valid values: <ul style="list-style-type: none"><li>◦ Weight: indicates the number of splits of the feature.</li><li>◦ Gain: indicates the information gain provided by the feature.</li><li>◦ Cover: indicates the number of samples covered by the feature on the split node.</li></ul> | gain                    |
| Tuning parameters | coreNum               | No       | The number of cores used in computing. The larger the value is, the faster the computing algorithm runs.                                                                                                                                                                                                               | Automatically allocated |
|                   | memSizePerCore        | No       | The memory size of each core. Unit: MB.                                                                                                                                                                                                                                                                                | Automatically allocated |

## Example

1. Execute the following SQL statement to generate training data. In this example, training data in the dense format is generated.

```

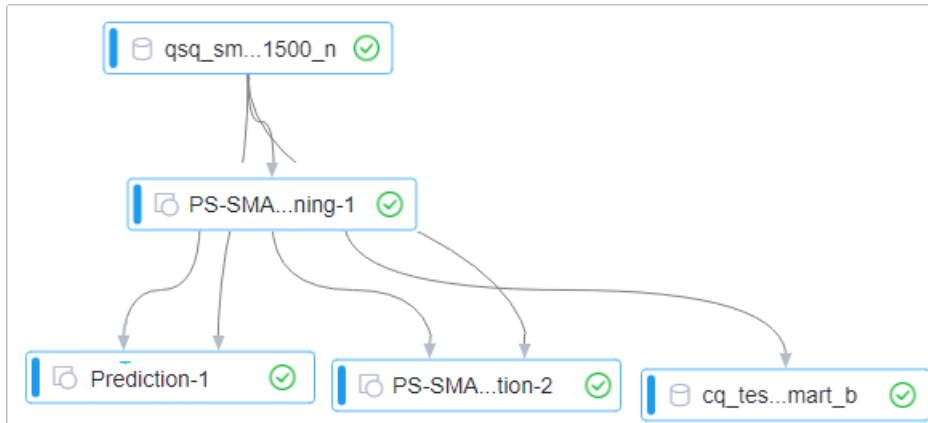
drop table if exists smart_binary_input;
create table smart_binary_input lifecycle 3 as
select
*
from
(
select 0.72 as f0, 0.42 as f1, 0.55 as f2, -0.09 as f3, 1.79 as f4, -1.2 as f5, 0 as label from dual
union all
select 1.23 as f0, -0.33 as f1, -1.55 as f2, 0.92 as f3, -0.04 as f4, -0.1 as f5, 1 as label from dual
union all
select -0.2 as f0, -0.55 as f1, -1.28 as f2, 0.48 as f3, -1.7 as f4, 1.13 as f5, 1 as label from dual
union all
select 1.24 as f0, -0.68 as f1, 1.82 as f2, 1.57 as f3, 1.18 as f4, 0.2 as f5, 0 as label from dual
union all
select -0.85 as f0, 0.19 as f1, -0.06 as f2, -0.55 as f3, 0.31 as f4, 0.08 as f5, 1 as label from dual
union all
select 0.58 as f0, -1.39 as f1, 0.05 as f2, 2.18 as f3, -0.02 as f4, 1.71 as f5, 0 as label from dual
union all
select -0.48 as f0, 0.79 as f1, 2.52 as f2, -1.19 as f3, 0.9 as f4, -1.04 as f5, 1 as label from dual
union all
select 1.02 as f0, -0.88 as f1, 0.82 as f2, 1.82 as f3, 1.55 as f4, 0.53 as f5, 0 as label from dual
union all
select 1.19 as f0, -1.18 as f1, -1.1 as f2, 2.26 as f3, 1.22 as f4, 0.92 as f5, 0 as label from dual
union all
select -2.78 as f0, 2.33 as f1, 1.18 as f2, -4.5 as f3, -1.31 as f4, -1.8 as f5, 1 as label from dual
) tmp;

```

The training data generated is listed in the following table.

|    | f0 ▲  | f1 ▲  | f2 ▲  | f3 ▲  | f4 ▲  | f5 ▲  | label ▲ |
|----|-------|-------|-------|-------|-------|-------|---------|
| 1  | 0.72  | 0.42  | 0.55  | -0.09 | 1.79  | -1.2  | 0       |
| 2  | 1.23  | -0.33 | -1.55 | 0.92  | -0.04 | -0.1  | 1       |
| 3  | -0.2  | -0.55 | -1.28 | 0.48  | -1.7  | 1.13  | 1       |
| 4  | 1.24  | -0.68 | 1.82  | 1.57  | 1.18  | 0.2   | 0       |
| 5  | -0.85 | 0.19  | -0.06 | -0.55 | 0.31  | 0.08  | 1       |
| 6  | 0.58  | -1.39 | 0.05  | 2.18  | -0.02 | 1.71  | 0       |
| 7  | -0.48 | 0.79  | 2.52  | -1.19 | 0.9   | -1.04 | 1       |
| 8  | 1.02  | -0.88 | 0.82  | 1.82  | 1.55  | 0.53  | 0       |
| 9  | 1.19  | -1.18 | -1.1  | 2.26  | 1.22  | 0.92  | 0       |
| 10 | -2.78 | 2.33  | 1.18  | -4.5  | -1.31 | -1.8  | 1       |

2. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).



- Configure the parameters listed in the following table for the PS-SMART Binary Classification component. Retain the default values of the parameters that are not listed in the table.

| Tab                | Parameter                 | Description                                                                  |
|--------------------|---------------------------|------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns           | The feature columns. Select the f0, f1, f2, f3, f4, and f5 columns.          |
|                    | Label Column              | The label column. Select the label column.                                   |
| Parameters Setting | Evaluation Indicator Type | The evaluation metric type. Set the value to <b>AUC for Classification</b> . |
|                    | Trees                     | Set the value to 5.                                                          |

- View the prediction result of the Prediction component.

|    | f0 ▲  | f1 ▲  | f2 ▲  | f3 ▲  | f4 ▲  | f5 ▲  | label ▲ | prediction_result ▲ | prediction_score ▲ | prediction_detail ▲                                  |
|----|-------|-------|-------|-------|-------|-------|---------|---------------------|--------------------|------------------------------------------------------|
| 1  | 0.72  | 0.42  | 0.55  | -0.09 | 1.79  | -1.2  | 0       | 0                   | 0.5563381910324097 | { "0": 0.5563382208347321, "1": 0.4436617791652679 } |
| 2  | 1.23  | -0.33 | -1.55 | 0.92  | -0.04 | -0.1  | 1       | 1                   | 0.6076213121414185 | { "0": 0.3923786878585815, "1": 0.6076213121414185 } |
| 3  | -0.2  | -0.55 | -1.28 | 0.48  | -1.7  | 1.13  | 1       | 1                   | 0.5768264532089233 | { "0": 0.4231735467910767, "1": 0.5768264532089233 } |
| 4  | 1.24  | -0.68 | 1.82  | 1.57  | 1.18  | 0.2   | 0       | 0                   | 0.7096900939941406 | { "0": 0.709690123796463, "1": 0.290309876203537 }   |
| 5  | -0.85 | 0.19  | -0.06 | -0.55 | 0.31  | 0.08  | 1       | 1                   | 0.7265769839286804 | { "0": 0.2734230160713196, "1": 0.7265769839286804 } |
| 6  | 0.58  | -1.39 | 0.05  | 2.18  | -0.02 | 1.71  | 0       | 0                   | 0.5573073625564575 | { "0": 0.5573073327541351, "1": 0.4426926672458649 } |
| 7  | -0.48 | 0.79  | 2.52  | -1.19 | 0.9   | -1.04 | 1       | 1                   | 0.5777847170829773 | { "0": 0.4222152829170227, "1": 0.5777847170829773 } |
| 8  | 1.02  | -0.88 | 0.82  | 1.82  | 1.55  | 0.53  | 0       | 0                   | 0.7096900939941406 | { "0": 0.709690123796463, "1": 0.290309876203537 }   |
| 9  | 1.19  | -1.18 | -1.1  | 2.26  | 1.22  | 0.92  | 0       | 0                   | 0.7096900939941406 | { "0": 0.709690123796463, "1": 0.290309876203537 }   |
| 10 | -2.78 | 2.33  | 1.18  | -4.5  | -1.31 | -1.8  | 1       | 1                   | 0.7265769839286804 | { "0": 0.2734230160713196, "1": 0.7265769839286804 } |

1 in the **prediction\_detail** column indicates a positive example, and 0 indicates a negative example.

- View the prediction result of the PS-SMART prediction component.

|    | original_label ▲ | prediction_score ▲ | leaf_index ▲ |
|----|------------------|--------------------|--------------|
| 1  | 0                | 0.4066115617752075 | 2 2 2 2 2    |
| 2  | 1                | 0.5709302425384521 | 2 2 1 1 2    |
| 3  | 1                | 0.6125051379203796 | 1 1 1 1 1    |
| 4  | 0                | 0.3217407166957855 | 2 1 2 2 1    |
| 5  | 1                | 0.6954338550567627 | 1 2 1 1 2    |
| 6  | 0                | 0.4794751703739166 | 2 1 1 1 1    |
| 7  | 1                | 0.5404139757156372 | 1 2 2 2 2    |
| 8  | 0                | 0.3217407166957855 | 2 1 2 2 1    |
| 9  | 0                | 0.3217407166957855 | 2 1 2 2 1    |
| 10 | 1                | 0.6954338550567627 | 1 2 1 1 2    |

- **prediction\_score** indicates the probability that the prediction result is a positive example. If the value of this parameter is greater than 0.5, the prediction result is a positive example. Otherwise, the prediction result is a negative example.
- **leaf\_index** indicates the IDs of leaf nodes for prediction. Each sample has N trees, and each tree corresponds to a number. The number indicates the ID of the leaf node on to which the sample falls.

 **Note** The PS-SMART prediction component needs to use a STRING-type column as the label column. The values of the column cannot be empty or NULL. You can use the Data Type Conversion component to convert the data type of a feature column into STRING.

6. Right-click the **PS-SMART Binary Classification Training** component and choose **View Data > View Output Port 3**. Then, view the feature importance table.

|   | id ▲ | value ▲             |
|---|------|---------------------|
| 1 | 0    | 0.5690338015556335  |
| 2 | 1    | 0.21714292466640472 |
| 3 | 4    | 0.21382322907447815 |

- **id** indicates the ID of a passed feature. In this example, the f0, f1, f2, f3, f4, and f5 features are passed. Therefore, in the **id** column, 0 represents the feature column of f0, and 4 represents the feature column of f4. If data in the input table is key-value pairs, the **id** column lists keys in the key-value pairs.
- **value** indicates the feature importance type. The default value of this parameter is Gain, which indicates the sum of the information gains provided by the feature for the model.
- The preceding feature importance table has only three features. This indicates that only the three features are used to split the tree. The feature importance of other features can be considered as 0.

Additional information:

- Only columns of numeric data types can be used by the PS-SMART Binary Classification Training component. 0 indicates a negative example, and 1 indicates a positive example. If the type of data in the MaxCompute table is STRING, the STRING data type must be converted. For example, you must convert Good/Bad to 1/0.
- If data is in the key-value format, feature IDs must be positive integers, and feature values must be real numbers. If the data type of feature IDs is STRING, you must use the serialization component to

serialize the data. If feature values are categorical strings, you must perform feature engineering such as feature discretization to process the features.

- The PS-SMART Binary Classification Training component supports hundreds of thousands of feature tasks. However, these tasks are resource-intensive and time-consuming. To address this issue, you can use GBDT algorithms to train the model. GBDT algorithms are suitable for scenarios where continuous features are used for training. You can perform one-hot encoding on categorical features to filter low-frequency features. However, we recommend that you do not perform feature discretization on continuous features of numeric data types.
- The PS-SMART algorithm may introduce randomness. For example, randomness may be introduced in the following scenarios: data and feature sampling based on `data_sample_ratio` and `fea_sample_ratio`, optimization of the PS-SMART algorithm by using histograms for approximation, and merge of a local sketch into a global sketch. The structures of trees are different when tasks run on multiple workers in distributed mode. However, the effect of the training model is theoretically the same. It is normal if you use the same data and parameters during training but obtain different results.
- If you want to accelerate training, you can set **Cores** to a larger value. The PS-SMART algorithm starts training tasks after the required resources are provided. Therefore, the more the resources are requested, the longer you must wait.

## PS Logistic Regression for Binary Classification

PS is developed to perform large-scale online and offline training tasks. Logistic Regression is a classical binary classification algorithm. It is widely used in advertising and searching scenarios. PS Logistic Regression supports training tasks of binary classification for hundreds of billions of samples and billions of features.

You can configure the component by using one of the following methods:

- PAI console

| Tab            | Parameter         | Description                                                                                                                                                                                                                                                                                                                                                              |
|----------------|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns   | The feature columns selected from the input table for training. If data in the input table is in the dense format, columns of the DOUBLE and BIGINT data types are supported. If data in the input table is key-value pairs in the sparse format, only columns of the STRING data type are supported.<br><br><b>Note</b> A maximum of 20 million features are supported. |
|                | Label Column      | The name of the label column in the input table.                                                                                                                                                                                                                                                                                                                         |
|                | Use Sparse Format | Specifies whether data in the input table is in the sparse format. If <b>Use Sparse Format</b> is set to true, the feature ID cannot be 0. We recommend that you number features from 1.                                                                                                                                                                                 |

| Tab                | Parameter                    | Description                                                                                                                                                                                                                                                                                                                              |
|--------------------|------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | L1 weight                    | L1 regularization coefficient. The larger the value of this parameter is, the fewer non-zero elements the model has. If overfitting occurs, increase the value of this parameter.                                                                                                                                                        |
|                    | L2 weight                    | L2 regularization coefficient. The larger the value of this parameter is, the smaller the absolute values of the model parameters are. If overfitting occurs, increase the value of this parameter.                                                                                                                                      |
|                    | Maximum Iterations           | The maximum number of iterations of the algorithm. If the value of this parameter is set to 0, the number of iterations is unlimited.                                                                                                                                                                                                    |
|                    | Minimum Convergence Deviance | The conditions to terminate the algorithm. The value of this parameter is the average value of the relative change rates of losses for 10 iterations. The smaller the value of this parameter is, the longer the algorithm runs.                                                                                                         |
|                    | Largest Feature ID           | The maximum feature ID or feature dimension. The value of this parameter can be greater than the actual value. The larger the value of this parameter is, the more the memory size will be used. If this parameter is not specified, the system automatically runs an SQL task to calculate the maximum feature ID or feature dimension. |
| Tuning             | Cores                        | The number of cores.                                                                                                                                                                                                                                                                                                                     |
|                    | Memory Size per Core         | The memory size of each core. Unit: MB.                                                                                                                                                                                                                                                                                                  |

- PAI commands



```
#Training
PAI -name ps_lr
  -project algo_public
  -DinputTableName="lm_test_input"
  -DmodelName="logistic_regression_model"
  -DlabelColName="label"
  -DfeatureColNames="f0,f1,f2,f3,f4,f5"
  -DL1Weight=1.0
  -DL2Weight=0.0
  -DmaxIter=100
  -Depsilon=1e-6
  -DenableSparse=false

#Prediction
drop table if exists logistic_regression_predict;
PAI -name prediction
  -DmodelName="logistic_regression_model"
  -DoutputTableName="logistic_regression_predict"
  -DinputTableName="lm_test_input"
  -DappendColNames=label
  -DfeatureColNames="f0,f1,f2,f3,f4,f5"
  -DenableSparse=false
```

| Parameter       | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | Default value    |
|-----------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName  | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                                      | No default value |
| featureColNames | Yes      | <p>The feature columns selected from the input table for training. If data in the input table is in the dense format, columns of the DOUBLE and BIGINT data types are supported. If data in the input table is key-value pairs in the sparse format, only columns of the STRING data type are supported.</p> <div>  <b>Note</b> A maximum of 20 million features are supported.         </div> | No default value |
| labelColName    | Yes      | The name of the label column. The columns of the DOUBLE and BIGINT data types are supported.                                                                                                                                                                                                                                                                                                                                                                                      | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                         | Default value    |
|----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | Full table       |
| modelName            | Yes      | The name of the output model. By default, the model is generated as an offline model. If enableModello is set to false, the model is generated in MaxCompute.                                                                                                                                                                                                                                                                       | No default value |
| enableModello        | No       | Specifies whether the model is generated as an offline model. Valid values: true and false. If enableModello is set to false, the model is generated in MaxCompute.                                                                                                                                                                                                                                                                 | true             |
| l1Weight             | No       | L1 regularization coefficient. The larger the value of this parameter is, the fewer non-zero elements the model has. If overfitting occurs, increase the value of this parameter.                                                                                                                                                                                                                                                   | 1.0              |
| l2Weight             | No       | L2 regularization coefficient. The larger the value of this parameter is, the smaller the absolute values of the model parameters are. If overfitting occurs, increase the value of this parameter.                                                                                                                                                                                                                                 | 0                |
| maxIter              | No       | The maximum number of iterations of the algorithm.                                                                                                                                                                                                                                                                                                                                                                                  | 100              |
| epsilon              | No       | The conditions to terminate the algorithm.                                                                                                                                                                                                                                                                                                                                                                                          | 1.0e-06          |
| modelSize            | No       | The maximum feature ID or feature dimension. The value of this parameter can be greater than the actual value. The larger the value of this parameter is, the more the memory size will be used. If this parameter is not specified, the system automatically runs an SQL task to calculate the maximum feature ID or feature dimension.                                                                                            | 0                |

| Parameter      | Required | Description                                                                                          | Default value           |
|----------------|----------|------------------------------------------------------------------------------------------------------|-------------------------|
| enableSparse   | No       | Specifies whether data in the input table is in the sparse format. Valid values: true and false.     | false                   |
| itemDelimiter  | No       | The delimiter used to separate key-value pairs when data in the input data is in the sparse format.  | ,                       |
| kvDelimiter    | No       | The delimiter used to separate keys and values when data in the input table is in the sparse format. | :                       |
| coreNum        | No       | The number of cores.                                                                                 | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Unit: MB.                                                              | Automatically allocated |

## Example

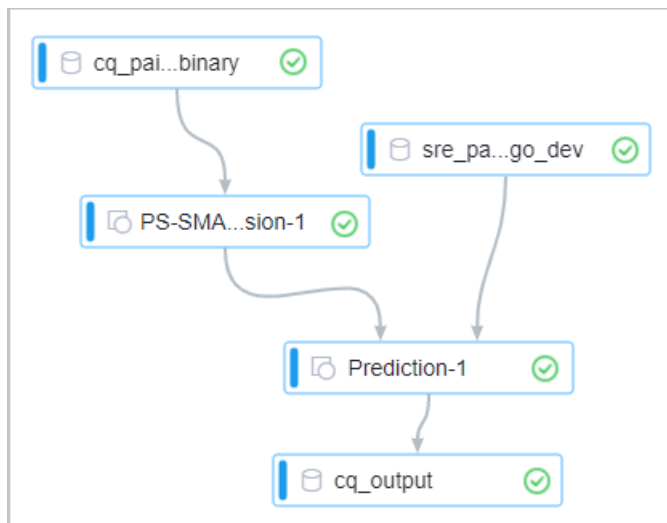
1. Execute the following SQL statement to generate training data. In this example, training data in the dense format is generated.

```
drop table if exists lm_test_input;
create table lm_test_input as
select
*
from
(
select 0.72 as f0, 0.42 as f1, 0.55 as f2, -0.09 as f3, 1.79 as f4, -1.2 as f5, 0 as label from dual
union all
select 1.23 as f0, -0.33 as f1, -1.55 as f2, 0.92 as f3, -0.04 as f4, -0.1 as f5, 1 as label from dual
union all
select -0.2 as f0, -0.55 as f1, -1.28 as f2, 0.48 as f3, -1.7 as f4, 1.13 as f5, 1 as label from dual
union all
select 1.24 as f0, -0.68 as f1, 1.82 as f2, 1.57 as f3, 1.18 as f4, 0.2 as f5, 0 as label from dual
union all
select -0.85 as f0, 0.19 as f1, -0.06 as f2, -0.55 as f3, 0.31 as f4, 0.08 as f5, 1 as label from dual
union all
select 0.58 as f0, -1.39 as f1, 0.05 as f2, 2.18 as f3, -0.02 as f4, 1.71 as f5, 0 as label from dual
union all
select -0.48 as f0, 0.79 as f1, 2.52 as f2, -1.19 as f3, 0.9 as f4, -1.04 as f5, 1 as label from dual
union all
select 1.02 as f0, -0.88 as f1, 0.82 as f2, 1.82 as f3, 1.55 as f4, 0.53 as f5, 0 as label from dual
union all
select 1.19 as f0, -1.18 as f1, -1.1 as f2, 2.26 as f3, 1.22 as f4, 0.92 as f5, 0 as label from dual
union all
select -2.78 as f0, 2.33 as f1, 1.18 as f2, -4.5 as f3, -1.31 as f4, -1.8 as f5, 1 as label from dual
) tmp;
```

The training data generated is listed in the following table.

|    | f0 ▲  | f1 ▲  | f2 ▲  | f3 ▲  | f4 ▲  | f5 ▲  | label ▲ |
|----|-------|-------|-------|-------|-------|-------|---------|
| 1  | 0.72  | 0.42  | 0.55  | -0.09 | 1.79  | -1.2  | 0       |
| 2  | 1.23  | -0.33 | -1.55 | 0.92  | -0.04 | -0.1  | 1       |
| 3  | -0.2  | -0.55 | -1.28 | 0.48  | -1.7  | 1.13  | 1       |
| 4  | 1.24  | -0.68 | 1.82  | 1.57  | 1.18  | 0.2   | 0       |
| 5  | -0.85 | 0.19  | -0.06 | -0.55 | 0.31  | 0.08  | 1       |
| 6  | 0.58  | -1.39 | 0.05  | 2.18  | -0.02 | 1.71  | 0       |
| 7  | -0.48 | 0.79  | 2.52  | -1.19 | 0.9   | -1.04 | 1       |
| 8  | 1.02  | -0.88 | 0.82  | 1.82  | 1.55  | 0.53  | 0       |
| 9  | 1.19  | -1.18 | -1.1  | 2.26  | 1.22  | 0.92  | 0       |
| 10 | -2.78 | 2.33  | 1.18  | -4.5  | -1.31 | -1.8  | 1       |

2. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).



3. Configure the parameters listed in the following table for the PS Logistic Regression for Binary Classification component. Retain the default values of the parameters that are not listed in the table.

| Tab            | Parameter            | Description                                    |
|----------------|----------------------|------------------------------------------------|
| Fields Setting | Feature Columns      | Select the f0, f1, f2, f3, f4, and f5 columns. |
|                | Label Column         | Select the label column.                       |
| Tuning         | Cores                | Set the value to 3.                            |
|                | Memory Size per Core | Set the value to 1024.                         |

4. Configure the parameters listed in the following table for the Prediction component. Retain the default values of the parameters that are not listed in the table.

| Tab | Parameter       | Description                                    |
|-----|-----------------|------------------------------------------------|
|     | Feature Columns | Select the f0, f1, f2, f3, f4, and f5 columns. |

| Fields Setting | Parameter        | Description                                                         |
|----------------|------------------|---------------------------------------------------------------------|
|                | Reserved Columns | Select the f0, f1, f2, f3, f4, and f5 columns and the label column. |

5. Run the experiment and view the prediction result.

|    | prediction_result ▲ | prediction_score ▲ | prediction_detail ▲                                    |
|----|---------------------|--------------------|--------------------------------------------------------|
| 1  | 0                   | 0.9083687005394571 | { "0": 0.9083687005394571, "1": 0.09163129946054296 }  |
| 2  | 0                   | 0.9018497676649229 | { "0": 0.9018497676649229, "1": 0.09815023233507712 }  |
| 3  | 1                   | 0.8688191866303958 | { "0": 0.1311808133696042, "1": 0.8688191866303958 }   |
| 4  | 0                   | 0.970607987507053  | { "0": 0.970607987507053, "1": 0.02939201249294704 }   |
| 5  | 1                   | 0.8705510193907658 | { "0": 0.1294489806092342, "1": 0.8705510193907658 }   |
| 6  | 0                   | 0.8426501566563311 | { "0": 0.8426501566563311, "1": 0.1573498433436689 }   |
| 7  | 1                   | 0.7559534361845501 | { "0": 0.2440465638154499, "1": 0.7559534361845501 }   |
| 8  | 0                   | 0.9722574685125349 | { "0": 0.9722574685125349, "1": 0.02774253148746508 }  |
| 9  | 0                   | 0.9788787481815239 | { "0": 0.9788787481815239, "1": 0.0211212518184761 }   |
| 10 | 1                   | 0.999816509540659  | { "0": 0.0001834904593409625, "1": 0.999816509540659 } |

1 in the prediction\_detail column indicates a positive example, and 0 indicates a negative example.

Additional information:

- Only columns of numeric data types can be used by the PS Logistic Regression for Binary Classification component. 0 indicates a negative example, and 1 indicates a positive example. If the type of data in the MaxCompute table is STRING, the STRING data type must be converted. For example, you must convert Good/Bad to 1/0.
- If data is in the key-value format, feature IDs must be positive integers, and feature values must be real numbers. If the data type of feature IDs is STRING, you must use the serialization component to serialize the data. If feature values are categorical strings, you must perform feature engineering such as feature discretization to process the features.

## 3.5.2. Multiclass classification

This topic describes the Multiclass Classification components provided by Machine Learning Studio. These components include PS-SMART Multiclass Classification, KNN, Logistic Regression for Multiclass Classification, Random Forest, and Naive Bayes.

### PS-SMART Multiclass Classification

A parameter server (PS) is used to process a large number of offline and online training tasks. SMART is short for scalable multiple additive regression tree. PS-SMART is an iteration algorithm that is implemented by using a PS-based gradient boosting decision tree (GBDT). PS-SMART Multiclass Classification supports training tasks with tens of billions of samples and hundreds of thousands of features. You can run these tasks on thousands of nodes. PS-SMART Multiclass Classification also supports a variety of data formats and optimization techniques, such as histogram-based approximation.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                 | Description                                                                                                                                                                                                                                                                                                                                   |
|--------------------|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Use Sparse Format         | Specifies whether data is in sparse format. Key-value pairs in sparse format are separated by spaces, and keys and values in key-value pairs are separated by colons (:), for example, 1:0.3 3:0.9.                                                                                                                                           |
|                    | Feature Columns           | The feature columns selected from the input table for training. If you do not select the Use Sparse Format check box, only columns of the BIGINT and DOUBLE types are supported. If you select the Use Sparse Format check box and keys and values in key-value pairs are of the numeric type, only columns of the STRING type are supported. |
|                    | Label Column              | The label columns selected from the input table. Columns of the STRING and numeric types are supported. Only numeric data can be stored in these columns, for example, 0 and 1 in binary classification.                                                                                                                                      |
|                    | Weight Column             | The columns selected to weigh samples in each row. Columns of the numeric type are supported.                                                                                                                                                                                                                                                 |
| Parameters Setting | Classes                   | The number of classes for multiclass classification. If you set the parameter value to n, the values of Label Column are {0,1,2,...,n-1}.                                                                                                                                                                                                     |
|                    | Evaluation Indicator Type | The parameter value can be set to <b>Multiclass Negative Log Likelihood</b> or <b>Multiclass Classification Error</b> .                                                                                                                                                                                                                       |
|                    | Trees                     | The number of trees. The amount of training data is proportional to the number of <b>trees</b> .                                                                                                                                                                                                                                              |
|                    | Maximum Tree Depth        | The default value is 5, which indicates that a maximum of 32 leaf nodes can be configured.                                                                                                                                                                                                                                                    |
|                    | Data Sampling Fraction    | The data sampling ratio when trees are built. The sample data is used to build a weak learner to accelerate training.                                                                                                                                                                                                                         |
|                    | Feature Sampling Fraction | The feature sampling ratio when trees are built. The sample features are used to build a weak learner to accelerate training.                                                                                                                                                                                                                 |
|                    | L1 Penalty Coefficient    | Controls the size of a leaf node. A larger value of this parameter indicates a more even distribution of leaf nodes. If overfitting occurs, increase the parameter value.                                                                                                                                                                     |
|                    | L2 Penalty Coefficient    | Controls the size of a leaf node. A larger value of this parameter indicates a more even distribution of leaf nodes. If overfitting occurs, increase the parameter value.                                                                                                                                                                     |

| Tab           | Parameter                                 | Description                                                                                                                                                                                                                                                                     |
|---------------|-------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|               | <b>Learning Rate</b>                      | Valid values: (0,1).                                                                                                                                                                                                                                                            |
|               | <b>Sketch-based Approximate Precision</b> | The quantile threshold for building a sketch. A smaller value indicates that more buckets can be obtained. In most cases, the default value 0.03 is used. Manual configurations are not required.                                                                               |
|               | <b>Minimum Split Loss Change</b>          | The minimum loss required for splitting a node. A larger value indicates that node splitting is less likely to occur.                                                                                                                                                           |
|               | <b>Features</b>                           | The number of features or the largest feature ID. Specify this parameter if you want to assess resource usage.                                                                                                                                                                  |
|               | <b>Global Offset</b>                      | The initial predicted values of all samples.                                                                                                                                                                                                                                    |
|               | <b>Feature Importance Type</b>            | The feature importance type, which can be set to Weight, Gain, or Cover. Weight indicates the number of splits of the feature. Gain indicates the information gain provided by the feature. Cover indicates the number of samples covered by the feature on the splitting node. |
| <b>Tuning</b> | <b>Cores</b>                              | By default, the number of cores is allocated by the system.                                                                                                                                                                                                                     |
|               | <b>Memory Size per Core (MB)</b>          | The memory size of each core, in MB. In most cases, the memory size is automatically allocated by the system.                                                                                                                                                                   |

- PAI command

```
# Training
PAI -name ps_smart
  -project algo_public
  -DinputTableName="smart_multiclass_input"
  -DmodelName="xlab_m_pai_ps_smart_bi_545859_v0"
  -DoutputTableName="pai_temp_24515_545859_2"
  -DoutputImportanceTableName="pai_temp_24515_545859_3"
  -DlabelColName="label"
  -DfeatureColNames="features"
  -DenableSparse="true"
  -Dobjective="multi:softprob"
  -Dmetric="mlogloss"
  -DfeatureImportanceType="gain"
  -DtreeCount="5";
  -DmaxDepth="5"
  -Dshrinkage="0.3"
  -DL2="1.0"
  -DL1="0"
  -Dlifecycle="3"
  -DsketchEps="0.03"
  -DsampleRatio="1.0"
  -DfeatureRatio="1.0"
  -DbaseScore="0.5"
  -DminSplitLoss="0"

# Prediction
PAI -name prediction
  -project algo_public
  -DinputTableName="smart_multiclass_input";
  -DmodelName="xlab_m_pai_ps_smart_bi_545859_v0"
  -DoutputTableName="pai_temp_24515_545860_1"
  -DfeatureColNames="features"
  -DappendColNames="label,features"
  -DenableSparse="true"
  -DkvDelimiter=":"
  -Dlifecycle="28"
```

| Category | Parameter       | Required | Description                                                                                                                                                                                                                                                                                               | Default value    |
|----------|-----------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
|          | featureColNames | Yes      | The feature columns selected from the input table for training. If enableSparse is set to false, only columns of the BIGINT and DOUBLE types are supported. If enableSparse is set to true and keys and values in key-value pairs are of the numeric type, only columns of the STRING type are supported. | No default value |



| Category        | Parameter                 | Required | Description                                                                                                                                                                                                                                             | Default value    |
|-----------------|---------------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| Data parameters | labelColName              | Yes      | The label columns selected from the input table. Columns of the STRING and numeric types are supported. Only numeric data can be stored in these columns, for example, {0,1,2,...,n-1} in multiclass classification. n indicates the number of classes. | No default value |
|                 | weightCol                 | No       | The columns selected to weigh samples in each row. Columns of the numeric type are supported.                                                                                                                                                           | No default value |
|                 | enableSparse              | No       | Specifies whether data is in sparse format. Valid values: {true,false}. Key-value pairs in sparse format are separated by spaces, and keys and values in key-value pairs are separated by colons (:), for example, 1:0.3 3:0.9.                         | false            |
|                 | inputTableName            | Yes      | The name of the input table.                                                                                                                                                                                                                            | No default value |
|                 | modelName                 | Yes      | The name of the output model.                                                                                                                                                                                                                           | No default value |
|                 | outputImportanceTableName | No       | The name of the table that provides feature importance.                                                                                                                                                                                                 | No default value |
|                 | inputTablePartitions      | No       | The partitions selected from the input table for training. Format: ds=1/pt=1.                                                                                                                                                                           | No default value |
|                 | outputTableName           | No       | The name of the output table in MaxCompute. The parameter value is in binary format and cannot be read. You can obtain the parameter value only from the PS-SMART prediction component.                                                                 | No default value |
|                 | lifecycle                 | No       | The lifecycle of the output table.                                                                                                                                                                                                                      | 3                |
|                 |                           |          |                                                                                                                                                                                                                                                         |                  |
|                 | classNum                  | Yes      | The number of classes in multiclass classification. If the parameter value is set to n, the values of labelColName are {0,1,2,...,n-1}.                                                                                                                 | No default value |
|                 | objective                 | Yes      | The type of the objective function. If you use multiclass classification for training, select the multi:softprob objective function.                                                                                                                    | No default value |
|                 |                           |          |                                                                                                                                                                                                                                                         |                  |

| Category             | Parameter    | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Default value    |
|----------------------|--------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| Algorithm parameters | metric       | No       | The type of evaluation indicator in the training set. You can obtain the parameter value from stdout of the Coordinator section in a Logview file. Valid values: <ul style="list-style-type: none"> <li>mlogloss: corresponds to <b>Multiclass Negative Log Likelihood</b> of Evaluation Indicator Type in the Machine Learning Platform for AI console.</li> <li>merror: corresponds to <b>Multiclass Classification Error</b> of Evaluation Indicator Type in the Machine Learning Platform for AI console.</li> </ul> | No default value |
|                      | treeCount    | No       | The number of trees, which is proportional to the training time.                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 1                |
|                      | maxDepth     | No       | The maximum depth of a tree. Valid values: 1 to 20.                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 5                |
|                      | sampleRatio  | No       | The data sampling ratio. Valid values: (0,1]. If the parameter value is set to 1.0, data sampling is not performed.                                                                                                                                                                                                                                                                                                                                                                                                      | 1.0              |
|                      | featureRatio | No       | The feature sampling ratio. Valid values: (0,1]. If the parameter value is set to 1.0, feature sampling is not performed.                                                                                                                                                                                                                                                                                                                                                                                                | 1.0              |
|                      | l1           | No       | The L1 penalty coefficient. A larger value of this parameter indicates a more even distribution of leaf nodes. If overfitting occurs, increase the parameter value.                                                                                                                                                                                                                                                                                                                                                      | 0                |
|                      | l2           | No       | The L2 penalty coefficient. A larger value of this parameter indicates a more even distribution of leaf nodes. If overfitting occurs, increase the parameter value.                                                                                                                                                                                                                                                                                                                                                      | 1.0              |
|                      | shrinkage    | No       | Valid values: (0,1).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 0.3              |
|                      | sketchEps    | No       | The quantile threshold for building a sketch. The number of buckets is equal to $O(1.0/sketchEps)$ . A smaller value indicates that more buckets can be obtained. In most cases, the default value is used. Manual configurations are not required. Valid values: (0,1).                                                                                                                                                                                                                                                 | 0.03             |
|                      | minSplitLoss | No       | The minimum loss required for splitting a node. A larger value indicates that node splitting is less likely to occur.                                                                                                                                                                                                                                                                                                                                                                                                    | 0                |

| Category                | Parameter             | Required | Description                                                                                                                                                                                                                                                                                                              | Default value           |
|-------------------------|-----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
|                         | featureNum            | No       | The number of features or the largest feature ID. Specify this parameter if you want to assess resource usage.                                                                                                                                                                                                           | No default value        |
|                         | baseScore             | No       | The initial predicted values of all samples.                                                                                                                                                                                                                                                                             | 0.5                     |
|                         | featureImportanceType | No       | The feature importance type. Valid values: <ul style="list-style-type: none"> <li>weight: indicates the number of splits of the feature.</li> <li>gain: indicates the information gain provided by the feature.</li> <li>cover: indicates the number of samples covered by the feature on the splitting node.</li> </ul> | gain                    |
| Optimization parameters | coreNum               | No       | The number of cores. A larger value of this parameter indicates faster running of PS-SMART Multiclass Classification.                                                                                                                                                                                                    | Automatically allocated |
|                         | memSizePerCore        | No       | The memory size of each core, in MB.                                                                                                                                                                                                                                                                                     | Automatically allocated |

## Example

1. Execute the following SQL statements to generate input data in key-value pair format:

```

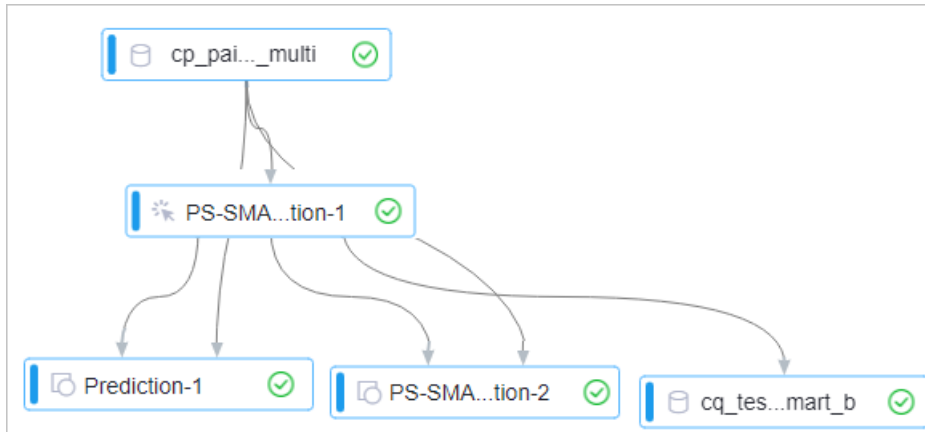
drop table if exists smart_multiclass_input;
create table smart_multiclass_input lifecycle 3 as
select
*
from
(
select 2 as label, '1:0.55 2:-0.15 3:0.82 4:-0.99 5:0.17' as features from dual
union all
select 1 as label, '1:-1.26 2:1.36 3:-0.13 4:-2.82 5:-0.41' as features from dual
union all
select 1 as label, '1:-0.77 2:0.91 3:-0.23 4:-4.46 5:0.91' as features from dual
union all
select 2 as label, '1:0.86 2:-0.22 3:-0.46 4:0.08 5:-0.60' as features from dual
union all
select 1 as label, '1:-0.76 2:0.89 3:1.02 4:-0.78 5:-0.86' as features from dual
union all
select 1 as label, '1:2.22 2:-0.46 3:0.49 4:0.31 5:-1.84' as features from dual
union all
select 0 as label, '1:-1.21 2:0.09 3:0.23 4:2.04 5:0.30' as features from dual
union all
select 1 as label, '1:2.17 2:-0.45 3:-1.22 4:-0.48 5:-1.41' as features from dual
union all
select 0 as label, '1:-0.40 2:0.63 3:0.56 4:0.74 5:-1.44' as features from dual
union all
select 1 as label, '1:0.17 2:0.49 3:-1.50 4:-2.20 5:-0.35' as features from dual
) tmp;

```

Generated data

|    | label ▲ | features ▲                             |
|----|---------|----------------------------------------|
| 1  | 2       | 1:0.55 2:-0.15 3:0.82 4:-0.99 5:0.17   |
| 2  | 1       | 1:-1.26 2:1.36 3:-0.13 4:-2.82 5:-0.41 |
| 3  | 1       | 1:-0.77 2:0.91 3:-0.23 4:-4.46 5:0.91  |
| 4  | 2       | 1:0.86 2:-0.22 3:-0.46 4:0.08 5:-0.60  |
| 5  | 1       | 1:-0.76 2:0.89 3:1.02 4:-0.78 5:-0.86  |
| 6  | 1       | 1:2.22 2:-0.46 3:0.49 4:0.31 5:-1.84   |
| 7  | 0       | 1:-1.21 2:0.09 3:0.23 4:2.04 5:0.30    |
| 8  | 1       | 1:2.17 2:-0.45 3:-1.22 4:-0.48 5:-1.41 |
| 9  | 0       | 1:-0.40 2:0.63 3:0.56 4:0.74 5:-1.44   |
| 10 | 1       | 1:0.17 2:0.49 3:-1.50 4:-2.20 5:-0.35  |

2. Create an experiment. For more information, see [Algorithm modeling](#).



3. Configure the parameters listed in the following table for the PS-SMART Multiclass Classification component. Parameters that are not listed in the following table use their default values.

| Tab                | Parameter                 | Description                                                               |
|--------------------|---------------------------|---------------------------------------------------------------------------|
| Fields Setting     | Feature Columns           | The feature columns you selected.                                         |
|                    | Label Column              | The label columns you selected.                                           |
|                    | Use Sparse Format         | Select the <b>Use Sparse Format</b> check box.                            |
| Parameters Setting | Classes                   | Enter 3 in the Classes field.                                             |
|                    | Evaluation Indicator Type | Select <b>Multiclass Negative Log Likelihood</b> from the drop-down list. |
|                    | Trees                     | Enter 5 in the Trees field.                                               |

4. Configure the parameters listed in the following table for the unified prediction component. Parameters that are not listed in the following table use their default values.

| Tab            | Parameter               | Description                                                                                      |
|----------------|-------------------------|--------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns         | By default, all feature columns are selected. Extra columns do not affect the prediction result. |
|                | Reserved Columns        | The label columns you selected.                                                                  |
|                | Sparse Matrix           | Select the <b>Sparse Matrix</b> check box.                                                       |
|                | Key and Value Delimiter | The key and value in a key-value pair are separated by colons (:).                               |
|                | KV Pair Delimiter       | Enter \u0020 in the KV Pair Delimiter field.                                                     |

5. Configure the parameters listed in the following table for the PS-SMART prediction component. Parameters that are not listed in the following table use their default values.

| Tab            | Parameter               | Description                                                                                      |
|----------------|-------------------------|--------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns         | By default, all feature columns are selected. Extra columns do not affect the prediction result. |
|                | Reserved Columns        | The label columns you selected.                                                                  |
|                | Sparse Matrix           | Select the <b>Sparse Matrix</b> check box.                                                       |
|                | Key and Value Delimiter | The key and value in a key-value pair are separated by colons (:).                               |
|                | KV Pair Delimiter       | Enter \u0020 in the KV Pair Delimiter field.                                                     |

6. Run the experiment and view the prediction result of the unified prediction component.

|    | label ▲ | features ▲     | prediction_result ▲ | prediction_score ▲  | prediction_detail ▲                                                           |
|----|---------|----------------|---------------------|---------------------|-------------------------------------------------------------------------------|
| 1  | 2       | 1.0.55 2-0.... | 1                   | 0.46681538224220276 | { "0": 0.1409690380096436, "1": 0.4668153822422028, "2": 0.3922155499458313 } |
| 2  | 1       | 1-1.26 2-1.... | 1                   | 0.7185402512550354  | { "0": 0.1393321454524994, "1": 0.7185402512550354, "2": 0.1421276330947876 } |
| 3  | 1       | 1-0.77 2-0.... | 1                   | 0.6726031303405762  | { "0": 0.1620725691318512, "1": 0.6726031303405762, "2": 0.165324330329895 }  |
| 4  | 2       | 1.0.86 2-0.... | 2                   | 0.4884149134159088  | { "0": 0.1755447387695312, "1": 0.3360402882099152, "2": 0.4884149134159088 } |
| 5  | 1       | 1-0.76 2-0.... | 1                   | 0.6707357168197632  | { "0": 0.1629969924688339, "1": 0.6707357168197632, "2": 0.1662672907114029 } |
| 6  | 1       | 1.2.22 2-0.... | 0                   | 0.4126577377319336  | { "0": 0.4126577377319336, "1": 0.2393952459096909, "2": 0.3479470014572144 } |
| 7  | 0       | 1-1.21 2-0.... | 0                   | 0.541054368019104   | { "0": 0.541054368019104, "1": 0.2916863262653351, "2": 0.1672592610120773 }  |
| 8  | 1       | 1.2.17 2-0.... | 1                   | 0.5768934488296509  | { "0": 0.1118654236197472, "1": 0.5768934488296509, "2": 0.311241090297699 }  |
| 9  | 0       | 1-0.40 2-0.... | 0                   | 0.5392904877662659  | { "0": 0.5392904877662659, "1": 0.2939955592155457, "2": 0.1667139828205109 } |
| 10 | 1       | 1.0.17 2-0.... | 1                   | 0.7185402512550354  | { "0": 0.1393321454524994, "1": 0.7185402512550354, "2": 0.1421276330947876 } |

The result contains the following columns:

- **prediction\_detail**: lists the classes used for multiclass classification. Valid values: 0, 1, and 2.
- **prediction\_result**: lists the classes of the prediction results.
- **prediction\_score**: lists the probabilities of classes in the **prediction\_result** column.

7. View the prediction result of the PS-SMART prediction component.

|    | original_label ▲ | score_class_0 ▲     | score_class_1 ▲     | score_class_2 ▲     | leaf_index ▲                  |
|----|------------------|---------------------|---------------------|---------------------|-------------------------------|
| 1  | 2                | 0.14096903800964355 | 0.46681538224220276 | 0.3922155499458313  | 1 1 2 1 1 2 1 1 2 1 2 2 1 2 2 |
| 2  | 1                | 0.1393321305513382  | 0.7185402512550354  | 0.1421276330947876  | 1 1 1 1 1 1 1 1 1 1 3 1 1 3 1 |
| 3  | 1                | 0.1620725691318512  | 0.6726031303405762  | 0.16532433032989502 | 1 1 1 1 1 1 1 1 1 1 3 1 1 2 1 |
| 4  | 2                | 0.17554473876953125 | 0.33604028820991516 | 0.4884149134159088  | 1 2 2 1 2 2 1 2 2 1 4 2 1 4 2 |
| 5  | 1                | 0.16299699246883392 | 0.6707357168197632  | 0.1662672907114029  | 1 1 1 1 1 1 1 1 1 1 2 1 1 3 1 |
| 6  | 1                | 0.4126577377319336  | 0.23939524590969086 | 0.34794700145721436 | 2 2 2 2 2 2 2 2 2 2 4 2 2 4 2 |
| 7  | 0                | 0.541054368019104   | 0.2916863262653351  | 0.16725926101207733 | 2 2 1 2 2 1 2 2 1 2 4 1 2 2 1 |
| 8  | 1                | 0.11186541616916656 | 0.5768935084342957  | 0.3112410604953766  | 1 1 2 1 1 2 1 1 2 1 3 2 1 3 2 |
| 9  | 0                | 0.5392904877662659  | 0.29399555921554565 | 0.16671398282051086 | 2 2 1 2 2 1 2 2 1 2 2 1 2 4 1 |
| 10 | 1                | 0.1393321305513382  | 0.7185402512550354  | 0.1421276330947876  | 1 1 1 1 1 1 1 1 1 1 3 1 1 3 1 |

The result contains the following columns:

- **score\_class\_k**: lists the probabilities that a predication falls into the Kth class.
- **leaf\_index**: lists the IDs of the predicted leaf nodes. If the number of trees is N and the number of classes is M, the value of the **leaf\_index** of each sample is N multiplied by M. For example, the value of **leaf\_index** in this example is 15 (5 x 3). Each tree maps to a number, which indicates the

ID of the leaf node to which the sample belongs on the tree.

- On the canvas, right-click **PS-SMART Multiclass Classification** and choose **View Data > View Output Port 3** to view the feature importance result.

|   | id ▲ | value ▲             |
|---|------|---------------------|
| 1 | 1    | 0.276059627532959   |
| 2 | 3    | 0.20854459702968597 |
| 3 | 4    | 0.31002077460289    |
| 4 | 5    | 0.20537501573562622 |

The feature importance result contains the following columns:

- id**: lists the sequence number of the input feature. In this example, the input data is in the key-value pair format. The values in the **id** column indicate the keys in the key-value pairs.
- value**: lists the feature importance type. The default value is gain, which indicates the sum of information gains that a feature provides for the model.

Instructions:

- Data in the destination columns for PS-SMART Multiclass Classification must be of the numeric type. If data in the MaxCompute table is of the STRING type, you must convert the data type. For example, if the classification object is a string, such as Good/Medium/Bad, you must convert the string to 0/1/2.
- If data is in key-value pair format, feature IDs must be positive integers and feature values must be real numbers. If feature IDs are of the STRING type, use the serialization component to serialize these IDs. If feature values are class-type strings, perform feature engineering, such as discretization.
- The PS-SMART Multiclass Classification component supports hundreds of thousands of feature tasks. However, this component is resource-intensive and time-consuming. You can use the GBDT classification algorithm for training instead. The GBDT classification algorithm is suitable for training with continuous features. We recommend that you do not discretize continuous numeric features. Discretization is required only when one-hot encoding is performed on categorical features to filter out infrequent features.
- The PS-SMART Multiclass Classification component introduces randomness. For example, data sampling, feature sampling, histogram-based approximation optimization, and merging of partial sketches into complete sketches are performed in random order. Data sampling is indicated by `data_sample_ratio` and feature sampling is indicated by `fea_sample_ratio`. When distributed workers are deployed for training, their tree structures are different. However, the modeling effect is theoretically approximate. This is a normal phenomenon if you use the same data and parameters during the training but obtain inconsistent results for multiple training tasks.
- To accelerate training, you can increase the number of **cores**. This is because the PS-SMART Multiclass Classification component starts training only after all servers obtain their resources. When workloads on a cluster are high, resource application is time-consuming.

## KNN

The KNN component selects the K-nearest records from a row in the prediction table for classification. The most common class of the K-nearest records is used as the class of the row.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                             | Description                                                                                                                                                             |
|--------------------|---------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns from Training Table   | The feature columns used for training.                                                                                                                                  |
|                    | Label Column from Training Table      | The destination columns for training.                                                                                                                                   |
|                    | Feature Columns from Prediction Table | If this parameter is not specified, the feature columns selected from the prediction table are the same as the feature columns in the training table.                   |
|                    | ID Column Appended to Output Table    | The ID columns that are used to obtain the predicted values of a column. By default, the feature columns selected from the prediction table are used as the ID columns. |
|                    | Input Sparse Format Data              | Data in the input table is in the key-value pair format.                                                                                                                |
|                    | KV Pair Delimiter                     | The default delimiter is a comma (,).                                                                                                                                   |
|                    | Key and Value Delimiter               | The default delimiter is a colon (:).                                                                                                                                   |
| Parameters Setting | Neighbors                             | The default value is 100.                                                                                                                                               |
| Tuning             | Cores                                 | Automatically allocated                                                                                                                                                 |
|                    | Memory Size                           | Automatically allocated                                                                                                                                                 |

- PAI command

```
PAI -name knn
-DtrainTableName=pai_knn_test_input
-DtrainFeatureColNames=f0,f1
-DtrainLabelColName=class
-DpredictTableName=pai_knn_test_input
-DpredictFeatureColNames=f0,f1
-DoutputTableName=pai_knn_test_output
-Dk=2;
```

| Parameter            | Required | Description                                                   | Default value    |
|----------------------|----------|---------------------------------------------------------------|------------------|
| trainTableName       | Yes      | The name of the training table.                               | No default value |
| trainFeatureColNames | Yes      | The names of the feature columns in the training table.       | No default value |
| trainLabelColName    | Yes      | The name of the label column in the training table.           | No default value |
| trainTablePartitions | No       | The partitions selected from the training table for training. | All partitions   |



| Parameter              | Required | Description                                                                                | Default value                               |
|------------------------|----------|--------------------------------------------------------------------------------------------|---------------------------------------------|
| predictTableName       | Yes      | The name of the prediction table.                                                          | No default value                            |
| outputTableName        | Yes      | The name of the output table.                                                              | No default value                            |
| predictFeatureColNames | No       | The names of the feature columns in the prediction table.                                  | Same as the value of trainFeatureColNames   |
| predictTablePartitions | No       | The partitions selected from the prediction table for prediction.                          | All partitions                              |
| appendColNames         | No       | The names of the columns appended to the output table.                                     | Same as the value of predictFeatureColNames |
| outputTablePartition   | No       | The partitions in the output table.                                                        | Full table                                  |
| k                      | No       | The number of K-nearest neighbors. Valid values: 1 to 1000.                                | 100                                         |
| enableSparse           | No       | Specifies whether data in the input table is in sparse format. Valid values: {true,false}. | false                                       |
| itemDelimiter          | No       | The delimiter between key-value pairs.                                                     | Comma (,)                                   |
| kvDelimiter            | No       | The delimiter between the key and value in a key-value pair.                               | Colon (:)                                   |
| coreNum                | No       | The number of cores. This parameter is used with memSizePerCore. Valid values: 1 to 20000. | Automatically allocated                     |
| memSizePerCore         | No       | The memory size of each core, in MB. Valid values: 1024 to 64 x 1024.                      | Automatically allocated                     |
| lifecycle              | No       | The lifecycle of the output table.                                                         | No default value                            |

## Example

1. Generate the training data.

```
create table pai_knn_test_input as
select * from
(
  select 1 as f0,2 as f1, 'good' as class from dual
  union all
  select 1 as f0,3 as f1, 'good' as class from dual
  union all
  select 1 as f0,4 as f1, 'bad' as class from dual
  union all
  select 0 as f0,3 as f1, 'good' as class from dual
  union all
  select 0 as f0,4 as f1, 'bad' as class from dual
)tmp;
```

- Run the following PAI command to submit the parameters of KNN:

```
pai -name knn
-DtrainTableName=pai_knn_test_input
-DtrainFeatureColNames=f0,f1
-DtrainLabelColName=class
-DpredictTableName=pai_knn_test_input
-DpredictFeatureColNames=f0,f1
-DoutputTableName=pai_knn_test_output
-Dk=2;
```

- View the training result.

| f0 | f1 | prediction_result | prediction_score | prediction_detail         |
|----|----|-------------------|------------------|---------------------------|
| 1  | 4  | bad               | 1.0              | {"bad": 1}                |
| 0  | 4  | bad               | 1.0              | {"bad": 1}                |
| 0  | 3  | bad               | 0.5              | {"bad": 0.5, "good": 0.5} |
| 1  | 3  | good              | 1.0              | {"good": 1}               |
| 1  | 2  | good              | 1.0              | {"good": 1}               |

The result contains the following columns:

- **f0** and **f1**: the appended columns.
- **prediction\_result**: lists the classification results.
- **prediction\_score**: lists the probabilities for the classification results.
- **prediction\_detail**: lists the classes of the K-nearest neighbors and their probabilities.

## Logistic Regression for Multiclass Classification

The Logistic Regression for Multiclass Classification component supports sparse and dense data formats. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab | Parameter | Description |
|-----|-----------|-------------|
|-----|-----------|-------------|

| Tab                | Parameter                    | Description                                                                                                                                                                              |
|--------------------|------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Training Feature Columns     | The feature columns selected from the data source for training. Columns of the DOUBLE and BIGINT types are supported.<br><br><b>Note</b> A maximum of 20 million features are supported. |
|                    | Target Columns               | The name of the destination columns in the data source.                                                                                                                                  |
|                    | Use Sparse Format            | Specifies whether data is in sparse format.                                                                                                                                              |
| Parameters Setting | Regularization Type          | Valid values: L1, L2, and None.                                                                                                                                                          |
|                    | Maximum Iterations           | Default value: 100.                                                                                                                                                                      |
|                    | Regularization Coefficient   | If Regularization Type is set to None, this parameter is invalid.                                                                                                                        |
|                    | Minimum Convergence Deviance | Default value: 0.000001.                                                                                                                                                                 |

- PAI command

```
PAI -name logisticregression_multi
  -project algo_public
  -DmodelName="xlab_m_logistic_regression_6096"
  -DregularizedLevel="1"
  -DmaxIter="100"
  -DregularizedType="l1"
  -Depsilon="0.000001"
  -DlabelColName="y"
  -DfeatureColNames="pdays,emp_var_rate"
  -DgoodValue="1"
  -DinputTableName="bank_data"
```

| Parameter       | Required | Description                                                                                                                        | Default value                   |
|-----------------|----------|------------------------------------------------------------------------------------------------------------------------------------|---------------------------------|
| inputTableName  | Yes      | The name of the input table.                                                                                                       | No default value                |
| featureColNames | No       | The feature columns selected from the input table for training.<br><br><b>Note</b> A maximum of 20 million features are supported. | All columns of the numeric type |
| labelColName    | Yes      | The name of the label column selected from the input table.                                                                        | No default value                |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                               | Default value           |
|----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Valid values:</p> <ul style="list-style-type: none"> <li>partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | Full table              |
| modelName            | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                                                                                             | No default value        |
| regularizedType      | No       | The regularization type. Valid values: {'l1','l2','None'}.                                                                                                                                                                                                                                                                                                                                | l1                      |
| regularizedLevel     | No       | The regularization coefficient. If regularizedType is set to None, this parameter is invalid.                                                                                                                                                                                                                                                                                             | 1.0                     |
| maxIter              | No       | The maximum number of iterations for the L-BFGS algorithm.                                                                                                                                                                                                                                                                                                                                | 100                     |
| epsilon              | No       | The convergence deviation. This parameter indicates the condition to terminate the iterations of the L-BFGS algorithm. If log-likelihood between two iterations is less than this value, the iterations are terminated.                                                                                                                                                                   | 1.0e-06                 |
| enableSparse         | No       | Specifies whether data is in sparse format. Valid values: {true,false}.                                                                                                                                                                                                                                                                                                                   | false                   |
| itemDelimiter        | No       | The delimiter between key-value pairs of the data in the input table.                                                                                                                                                                                                                                                                                                                     | Comma (,)               |
| kvDelimiter          | No       | The delimiter between the key and value in a key-value pair of data in the input table.                                                                                                                                                                                                                                                                                                   | Colon (:)               |
| coreNum              | No       | The number of cores.                                                                                                                                                                                                                                                                                                                                                                      | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core, in MB.                                                                                                                                                                                                                                                                                                                                                      | Automatically allocated |

## Example

1. Execute the following SQL statements to generate training data:

```
drop table if exists multi_lr_test_input;
create table multi_lr_test_input
as
select
*
from
(
select
cast(1 as double) as f0,
cast(0 as double) as f1,
cast(0 as double) as f2,
cast(0 as double) as f3,
cast(0 as bigint) as label
from dual
union all
select
cast(0 as double) as f0,
cast(1 as double) as f1,
cast(0 as double) as f2,
cast(0 as double) as f3,
cast(0 as bigint) as label
from dual
union all
select
cast(0 as double) as f0,
cast(0 as double) as f1,
cast(1 as double) as f2,
cast(0 as double) as f3,
cast(2 as bigint) as label
from dual
union all
select
cast(0 as double) as f0,
cast(0 as double) as f1,
cast(0 as double) as f2,
cast(1 as double) as f3,
cast(1 as bigint) as label
from dual
) a;
```

The following table provides the generated training data in the multi\_lr\_test\_input table.

| f0  | f1  | f2  | f3  | label |
|-----|-----|-----|-----|-------|
| 1.0 | 0.0 | 0.0 | 0.0 | 0     |
| 0.0 | 0.0 | 1.0 | 0.0 | 2     |
| 0.0 | 0.0 | 0.0 | 1.0 | 1     |
| 0.0 | 1.0 | 0.0 | 0.0 | 0     |

2. Run the following PAI command to submit the parameters of the Logistic Regression for Multiclass

Classification component:

```
drop offlinemodel if exists multi_lr_test_model;
PAI -name logisticregression_multi
  -project algo_public
  -DmodelName="multi_lr_test_model"
  -DitemDelimiter=","
  -DregularizedLevel="1"
  -DmaxIter="100"
  -DregularizedType="None"
  -Depsilon="0.000001"
  -DkvDelimiter=":"
  -DlabelColName="label"
  -DfeatureColNames="f0,f1,f2,f3"
  -DenableSparse="false"
  -DinputTableName="multi_lr_test_input";
```

3. Run the following PAI command to submit the parameters of the prediction component:

```
drop table if exists multi_lr_test_prediction_result;
PAI -name prediction
  -project algo_public
  -DdetailColName="prediction_detail"
  -DmodelName="multi_lr_test_model"
  -DitemDelimiter=","
  -DresultColName="prediction_result"
  -Dlifecycle="28"
  -DoutputTableName="multi_lr_test_prediction_result"
  -DscoreColName="prediction_score"
  -DkvDelimiter=":"
  -DinputTableName="multi_lr_test_input"
  -DenableSparse="false"
  -DappendColNames="label";
```

4. View the multi\_lr\_test\_prediction\_result table.

| label | prediction_result | prediction_score   | prediction_detail                                                                 |
|-------|-------------------|--------------------|-----------------------------------------------------------------------------------|
| 0     | 0                 | 0.9999997274902165 | {"0": 0.9999997274902165, "1": 2.324679066261573e-07, "2": 2.324679066261569e-07} |
| 0     | 0                 | 0.9999997274902165 | {"0": 0.9999997274902165, "1": 2.324679066261573e-07, "2": 2.324679066261569e-07} |
| 2     | 2                 | 0.9999999155958832 | {"0": 2.018833979850994e-07, "1": 2.324679066261573e-07, "2": 0.9999999155958832} |
| 1     | 1                 | 0.9999999155958832 | {"0": 2.018833979850994e-07, "1": 0.9999999155958832, "2": 2.324679066261569e-07} |

## Random Forest

The Random Forest component is a classifier that consists of multiple decision trees. The classification result is determined by the mode of output classes of individual trees. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                  | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|--------------------|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns            | By default, the columns except the label columns and weight columns are selected.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|                    | Excluded Columns           | The columns that are not used for training. These columns cannot be used as <b>feature columns</b> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|                    | Columns Forced to Convert  | Comply with the following rules to parse columns: <ul style="list-style-type: none"> <li>Parse the columns of the STRING, BOOLEAN, or DATETIME type to the columns of the discrete type.</li> <li>Parse the columns of the DOUBLE or BIGINT type to the columns of the continuous type.</li> </ul> <div> <p> <b>Note</b> To parse the columns of the BIGINT type to the columns of the categorical type, you must use the forceCategorical parameter to specify the type.</p> </div> |
|                    | Weight Columns             | The columns selected to weigh samples in each row. Columns of the numeric type are supported.                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|                    | Label Column               | The label columns selected from the input table. Columns of the STRING and numeric types are supported.                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Parameters Setting | Trees                      | Valid values: 1 to 1000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|                    | Single-tree Algorithm Type | <p>If a forest has N trees and the condition is algorithmTypes=[a,b]:</p> <ul style="list-style-type: none"> <li>[0,a) indicates the ID3 algorithm.</li> <li>[a,b) indicates the CART algorithm.</li> <li>[b,n] indicates the C4.5 algorithm.</li> </ul> <p>For example, if a forest has five trees and [2,4] indicates 0, 1 indicates the ID3 algorithm, 2,3 indicates the CART algorithm, and 4 indicates the C4.5 algorithm. If the value is None, tree algorithms are evenly allocated across the forest.</p>                                                      |
|                    | Random Features per Tree   | Valid values: [1,N], where N represents the number of features.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |

| Tab | Parameter                                            | Description                                          |
|-----|------------------------------------------------------|------------------------------------------------------|
|     | <b>Minimum Number of Leaf Nodes</b>                  | Valid values: positive integers. Default value: 2.   |
|     | <b>Minimum Fraction of Leaf Nodes to Parent Node</b> | Valid values: [0,1]. Default value: 0.               |
|     | <b>Maximum Tree Depth</b>                            | Valid values: [1,+∞). Default value: ∞.              |
|     | <b>Number of Random Data Entries Input per Tree</b>  | Valid values: (1000,1000000]. Default value: 100000. |

- PAI command

```
PAI -name randomforests
  -project algo_public
  -DinputTableName="pai_rf_test_input"
  -DmodelName="pai_rf_test_model"
  -DforceCategorical="f1"
  -DlabelColName="class"
  -DfeatureColNames="f0,f1"
  -DmaxRecordSize="100000"
  -DminNumPer="0"
  -DminNumObj="2"
  -DtreeNum="3";
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                        | Default value    |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                       | No default value |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. The following formats are supported:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | All partitions   |
| labelColName         | Yes      | The name of the label column selected from the input table.                                                                                                                                                                                                                                                                                                                                                        | No default value |
| modelName            | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                                                                                                                      | No default value |
| treeNum              | Yes      | The number of trees in the forest. Valid values: 1 to 1000.                                                                                                                                                                                                                                                                                                                                                        | 100              |



| Parameter        | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Default value                                                                                                |
|------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| excludedColNames | No       | The columns that are not used for training. The columns cannot be used as feature columns.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | N/A                                                                                                          |
| weightColName    | No       | The name of the weight column in the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | No default value                                                                                             |
| featureColNames  | No       | The feature columns selected from the input table for training.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | All columns except the label columns specified by labelColName and weight column specified by weightColName. |
| forceCategorical | No       | <p>Comply with the following rules to parse columns:</p> <ul style="list-style-type: none"> <li>Parse the columns of the STRING, BOOLEAN, or DATETIME type to the columns of the discrete type.</li> <li>Parse the columns of the DOUBLE or BIGINT type to the columns of the continuous type.</li> </ul> <div>  <b>Note</b> To parse the columns of the BIGINT type to the columns of the categorical type, you must use the forceCategorical parameter to specify the type.         </div> | INT is the continuous type.                                                                                  |
| algorithmTypes   | No       | <p>The location of a tree algorithm in the forest. If the forest has N trees and the condition is algorithmTypes=[a,b]:</p> <ul style="list-style-type: none"> <li>[0, a) indicates the ID3 algorithm.</li> <li>[a,b) indicates the CART algorithm.</li> <li>[b,n] indicates the C4.5 algorithm.</li> </ul> <p>For example, if a forest has five trees and [2,4] indicates 0, 1 indicates the ID3 algorithm, 2,3 indicates the CART algorithm, and 4 indicates the C4.5 algorithm. If the value is None, tree algorithms are evenly allocated across the forest.</p>            | Evenly allocated                                                                                             |
| randomColNum     | No       | The number of random features selected for each split when a single tree is generated. Valid values: [1,N], where N represents the number of features.                                                                                                                                                                                                                                                                                                                                                                                                                          | $\log_2 N$                                                                                                   |

| Parameter     | Required | Description                                                                               | Default value |
|---------------|----------|-------------------------------------------------------------------------------------------|---------------|
| minNumObj     | No       | The minimum amount of data on leaf nodes. The parameter value must be a positive integer. | 2             |
| minNumPer     | No       | The minimum ratio of data on leaf nodes to data on a parent node. Valid values: [0,1].    | 0.0           |
| maxTreeDeep   | No       | The maximum depth of a single tree. Valid values: [1,+∞).                                 | ∞             |
| maxRecordSize | No       | The number of random data inputs for a tree. Valid values: (1000,1000000].                | 100000        |

## Example

1. Execute the following SQL statements to generate training data:

```
create table pai_rf_test_input as
select * from
(
  select 1 as f0,2 as f1, "good" as class from dual
  union all
  select 1 as f0,3 as f1, "good" as class from dual
  union all
  select 1 as f0,4 as f1, "bad" as class from dual
  union all
  select 0 as f0,3 as f1, "good" as class from dual
  union all
  select 0 as f0,4 as f1, "bad" as class from dual
)tmp;
```

2. Run the following PAI command to submit the parameters of the Random Forest component:

```
PAI -name randomforests
-project algo_public
-DinputTableName="pai_rf_test_input"
-DmodelName="pai_rf_test_model"
-DforceCategorical="f1"
-DlabelColName="class"
-DfeatureColNames="f0,f1"
-DmaxRecordSize="100000"
-DminNumPer="0"
-DminNumObj="2"
-DtreeNum="3";
```

3. View the Predictive Model Markup Language (PMML) of the model.

```
<? xml version="1.0" encoding="utf-8"?
<PMML xmlns="http://www.dmg.org/PMML-4_2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="4.2" xsi:schemaLocation="http://www.dmg.org/PMML-4_2 http://www.dmg.org/v4-2/pmml-4-2.xsd"
  <Header copyright="Copyright (c) 2014, Alibaba Inc." description=""
    <Application name="ODPS/PMML" version="0.1.0"/>
```

```

<TimestampTue, 12 Jul 2016 07:04:48 GMT</Timestamp>
</Header>
<DataDictionary numberOfFields="2"
  <DataField name="f0" optype="continuous" dataType="integer"/>
  <DataField name="f1" optype="continuous" dataType="integer"/>
  <DataField name="class" optype="categorical" dataType="string"
    <Value value="bad"/>
    <Value value="good"/>
  </DataField>
</DataDictionary>
<MiningModel modelName="xlab_m_random_forests_1_75078_v0" functionName="classification" algorithmName="RandomForests"
  <MiningSchema
    <MiningField name="f0" usageType="active"/>
    <MiningField name="f1" usageType="active"/>
    <MiningField name="class" usageType="target"/>
  </MiningSchema>
  <Segmentation multipleModelMethod="majorityVote"
    <Segment id="0"
      <True/
      <TreeModel modelName="xlab_m_random_forests_1_75078_v0" functionName="classification" algorithmName="RandomForests"
        <MiningSchema
          <MiningField name="f0" usageType="active"/>
          <MiningField name="f1" usageType="active"/>
          <MiningField name="class" usageType="target"/>
        </MiningSchema>
        <Node id="1"
          <True/
          <ScoreDistribution value="bad" recordCount="2"/>
          <ScoreDistribution value="good" recordCount="3"/>
          <Node id="2" score="good"
            <SimplePredicate field="f1" operator="equal" value="2"/>
            <ScoreDistribution value="good" recordCount="1"/>
          </Node>
          <Node id="3" score="good"
            <SimplePredicate field="f1" operator="equal" value="3"/>
            <ScoreDistribution value="good" recordCount="2"/>
          </Node>
          <Node id="4" score="bad"
            <SimplePredicate field="f1" operator="equal" value="4"/>
            <ScoreDistribution value="bad" recordCount="2"/>
          </Node>
        </Node>
      </TreeModel>
    </Segment>
    <Segment id="1"
      <True/
      <TreeModel modelName="xlab_m_random_forests_1_75078_v0" functionName="classification" algorithmName="RandomForests"
        <MiningSchema
          <MiningField name="f0" usageType="active"/>
          <MiningField name="f1" usageType="active"/>
          <MiningField name="class" usageType="target"/>
        </MiningSchema>
      </TreeModel>
    </Segment>
  </Segmentation>
</MiningModel>

```

```

</MiningSchema
<Node id="1"
  <True/
  <ScoreDistribution value="bad" recordCount="2"/
  <ScoreDistribution value="good" recordCount="3"/
  <Node id="2" score="good"
    <SimpleSetPredicate field="f1" booleanOperator="isIn"
      <Array n="2" type="integer" 2 3</Array
    </SimpleSetPredicate
    <ScoreDistribution value="good" recordCount="3"/
  </Node
  <Node id="3" score="bad"
    <SimpleSetPredicate field="f1" booleanOperator="isNotIn"
      <Array n="2" type="integer" 2 3</Array
    </SimpleSetPredicate
    <ScoreDistribution value="bad" recordCount="2"/
  </Node
</Node
</TreeModel
</Segment
<Segment id="2"
  <True/
  <TreeModel modelName="xlab_m_random_forests_1_75078_v0" functionName="classification" algorithmName="RandomForests"
    <MiningSchema
      <MiningField name="f0" usageType="active"/
      <MiningField name="f1" usageType="active"/
      <MiningField name="class" usageType="target"/
    </MiningSchema
    <Node id="1"
      <True/
      <ScoreDistribution value="bad" recordCount="2"/
      <ScoreDistribution value="good" recordCount="3"/
      <Node id="2" score="bad"
        <SimplePredicate field="f0" operator="lessOrEqual" value="0.5"/
        <ScoreDistribution value="bad" recordCount="1"/
        <ScoreDistribution value="good" recordCount="1"/
      </Node
      <Node id="3" score="good"
        <SimplePredicate field="f0" operator="greaterThan" value="0.5"/
        <ScoreDistribution value="bad" recordCount="1"/
        <ScoreDistribution value="good" recordCount="2"/
      </Node
    </Node
  </TreeModel
</Segment
</Segmentation
</MiningModel
</PMML

```

4. View the visual output of the Random Forest component.



## Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayesian theorem with independent assumptions. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab            | Parameter                                               | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|----------------|---------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Column                                          | By default, all columns except the label columns are selected. Columns of the DOUBLE, STRING, and BIGINT types are supported.                                                                                                                                                                                                                                                                                                                                       |
|                | Excluded Columns                                        | The columns that are not used for training. These columns cannot be used as feature columns.                                                                                                                                                                                                                                                                                                                                                                        |
|                | Forced Conversion Column                                | Comply with the following rules to parse columns: <ul style="list-style-type: none"> <li>Parse the columns of the STRING, BOOLEAN, or DATETIME type to the columns of the discrete type.</li> <li>Parse the columns of the DOUBLE or BIGINT type to the columns of the continuous type.</li> </ul> <div> <p><b>Note</b> To parse the columns of the BIGINT type to the columns of the categorical type, you must use this parameter to specify the type.</p> </div> |
|                | Label Column                                            | The label column in the input table. The label column cannot be used as a feature column. Columns of the DOUBLE, STRING, and BIGINT types are supported.                                                                                                                                                                                                                                                                                                            |
|                | Input Sparse Format Data                                | If you select the Input Sparse Format Data check box, key-value pairs are used.                                                                                                                                                                                                                                                                                                                                                                                     |
|                | Separator between K:V when input is sparse              | The default delimiter is a comma (,).                                                                                                                                                                                                                                                                                                                                                                                                                               |
|                | The separator of key and value when the input is sparse | The default delimiter is a colon (:).                                                                                                                                                                                                                                                                                                                                                                                                                               |
|                | Number of cores                                         | Automatically allocated                                                                                                                                                                                                                                                                                                                                                                                                                                             |

| Training | Parameter               | Description             |
|----------|-------------------------|-------------------------|
|          | Memory Size of Core(MB) | Automatically allocated |

- PAI command

```
PAI -name NaiveBayes -project algo_public
-DinputTablePartitions="pt=20150501"
-DmodelName="xlab_m_NaiveBayes_23772"
-DlabelColName="poutcome"
-DfeatureColNames="age,previous,cons_conf_idx,euribor3m"
-DinputTableName="bank_data_partition";
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Default value                       |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | No default value                    |
| inputTablePartitions | No       | The partitions selected from the input table for training.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | All partitions                      |
| modelName            | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | No default value                    |
| labelColName         | Yes      | The name of the label column in the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | No default value                    |
| featureColNames      | No       | The feature columns selected from the input table for training.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | All columns except the label column |
| excludedColNames     | No       | All columns except the feature columns. The columns specified by this parameter cannot be used as the columns specified by featureColNames.                                                                                                                                                                                                                                                                                                                                                                                                                              | N/A                                 |
| forceCategorical     | No       | Comply with the following rules to parse columns: <ul style="list-style-type: none"> <li>Parse the columns of the STRING, BOOLEAN, or DATETIME type to the columns of the discrete type.</li> <li>Parse the columns of the DOUBLE or BIGINT type to the columns of the continuous type.</li> </ul> <div>  <b>Note</b> To parse the columns of the BIGINT type to the columns of the categorical type, you must use the forceCategorical parameter to specify the type.         </div> | INT is a continuous type.           |
| coreNum              | No       | The number of cores.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Automatically allocated             |

| Parameter      | Required | Description                                                    | Default value           |
|----------------|----------|----------------------------------------------------------------|-------------------------|
| memSizePerCore | No       | The memory size of each core, in MB. Valid values: 1 to 65536. | Automatically allocated |

## Example

1. Prepare the following training data.

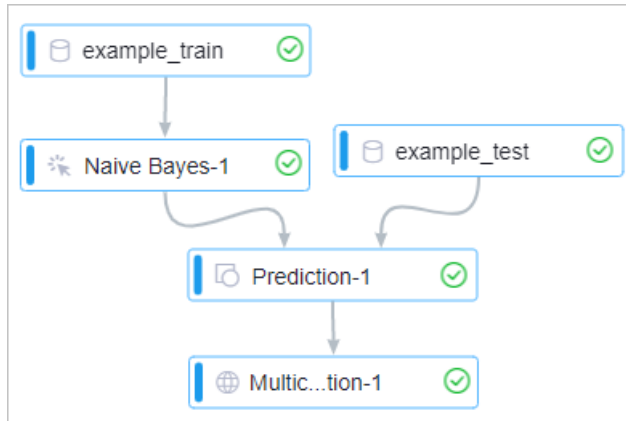
| id | y  | f0         | f1        | f2        | f3         | f4        | f5         | f6        | f7        |
|----|----|------------|-----------|-----------|------------|-----------|------------|-----------|-----------|
| 1  | -1 | -0.294118  | 0.487437  | 0.180328  | -0.292929  | -1        | 0.00149028 | -0.53117  | -0.033333 |
| 2  | +1 | -0.882353  | -0.145729 | 0.0819672 | -0.414141  | -1        | -0.207153  | -0.766866 | -0.666667 |
| 3  | -1 | -0.0588235 | 0.839196  | 0.0491803 | -1         | -1        | -0.305514  | -0.492741 | -0.633333 |
| 4  | +1 | -0.882353  | -0.105528 | 0.0819672 | -0.535354  | -0.777778 | -0.162444  | -0.923997 | -1        |
| 5  | -1 | -1         | 0.376884  | -0.344262 | -0.292929  | -0.602837 | 0.28465    | 0.887276  | -0.6      |
| 6  | +1 | -0.411765  | 0.165829  | 0.213115  | -1         | -1        | -0.23696   | -0.894962 | -0.7      |
| 7  | -1 | -0.647059  | -0.21608  | -0.180328 | -0.353535  | -0.791962 | -0.0760059 | -0.854825 | -0.833333 |
| 8  | +1 | 0.176471   | 0.155779  | -1        | -1         | -1        | 0.052161   | -0.952178 | -0.733333 |
| 9  | -1 | -0.764706  | 0.979899  | 0.147541  | -0.0909091 | 0.283688  | -0.0909091 | -0.931682 | 0.0666667 |
| 10 | -1 | -0.0588235 | 0.256281  | 0.57377   | -1         | -1        | -1         | -0.868488 | 0.1       |

2. Prepare the following test data.

| id | y  | f0        | f1         | f2         | f3        | f4        | f5        | f6         | f7        |
|----|----|-----------|------------|------------|-----------|-----------|-----------|------------|-----------|
| 1  | +1 | -0.882353 | 0.0854271  | 0.442623   | -0.616162 | -1        | -0.19225  | -0.725021  | -0.9      |
| 2  | +1 | -0.294118 | -0.0351759 | -1         | -1        | -1        | -0.293592 | -0.904355  | -0.766667 |
| 3  | +1 | -0.882353 | 0.246231   | 0.213115   | -0.272727 | -1        | -0.171386 | -0.981213  | -0.7      |
| 4  | -1 | -0.176471 | 0.507538   | 0.278689   | -0.414141 | -0.702128 | 0.0491804 | -0.475662  | 0.1       |
| 5  | -1 | -0.529412 | 0.839196   | -1         | -1        | -1        | -0.153502 | -0.885568  | -0.5      |
| 6  | +1 | -0.882353 | 0.246231   | -0.0163934 | -0.353535 | -1        | 0.0670641 | -0.627669  | -1        |
| 7  | -1 | -0.882353 | 0.819095   | 0.278689   | -0.151515 | -0.307329 | 0.19225   | 0.00768574 | -0.966667 |
| 8  | +1 | -0.882353 | -0.0753769 | 0.0163934  | -0.494949 | -0.903073 | -0.418778 | -0.654996  | -0.866667 |
| 9  | +1 | -1        | 0.527638   | 0.344262   | -0.212121 | -0.356974 | 0.23696   | -0.836038  | -0.8      |
| 10 | +1 | -0.882353 | 0.115578   | 0.0163934  | -0.737374 | -0.56974  | -0.28465  | -0.948762  | -0.933333 |

3. Create an experiment. For more information, see [Algorithm modeling](#).





4. Configure the parameters listed in the following table for the Naive Bayes component. Parameters that are not listed in the following table use their default values.

| Tab            | Parameter      | Description                                                                    |
|----------------|----------------|--------------------------------------------------------------------------------|
| Fields Setting | Feature Column | Select the f0, f1, f2, f3, f4, f5, f6, and f7 columns from the training table. |
|                | Label Column   | Select the y column from the training table.                                   |

5. Run the experiment and view the prediction results.

| id | y  | prediction_result | prediction_score    | prediction_detail                                    |
|----|----|-------------------|---------------------|------------------------------------------------------|
| 1  | +1 | +1                | 1.1253341740304466  | {"+1": 1.125334174030447, "-1": -4.116083356256129}  |
| 2  | +1 | +1                | 2.1096017126455773  | {"+1": 2.109601712645577, "-1": -9.431111061253313}  |
| 3  | +1 | +1                | 1.0119555555558313  | {"+1": 1.011955555555831, "-1": -3.065398632191728}  |
| 4  | -1 | -1                | -2.435966936637894  | {"+1": -33.18118539397123, "-1": -2.435966936637894} |
| 5  | -1 | -1                | -8.569186264886811  | {"+1": -10.87241674956984, "-1": -8.569186264886811} |
| 6  | +1 | -1                | -3.7653558619317407 | {"+1": -4.842928791659737, "-1": -3.765355861931741} |
| 7  | -1 | -1                | -4.895262140022778  | {"+1": -89.31454432308786, "-1": -4.895262140022778} |
| 8  | +1 | +1                | -2.701748842959141  | {"+1": -2.701748842959141, "-1": -3.683850999783979} |
| 9  | +1 | -1                | -4.321549531283424  | {"+1": -20.76204601810873, "-1": -4.321549531283424} |
| 10 | +1 | +1                | -3.0880514229489764 | {"+1": -3.088051422948976, "-1": -3.667255025859788} |

### 3.5.3. Clustering

This topic describes the Clustering components provided by Machine Learning Studio. The Clustering components include K-means Clustering and DBSCAN.

#### K-means Clustering

K-means Clustering randomly selects K objects as the initial centroids of each cluster, computes the distance between the remaining objects and the centroids, distributes the remaining objects to the nearest clusters, and then recalculates the centroids of each cluster. K-means Clustering assumes that clustering objects are spatial vectors. K-means Clustering minimizes the sum of the mean squared error (MSE) inside each cluster and constantly performs calculations and iterations until the criterion function converges.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                      | Description                                                                                                                                                    |
|--------------------|--------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns                | The feature columns. Columns of the DOUBLE and INT types are supported.                                                                                        |
|                    | Appended Columns               | The input columns that are appended to the clustering result table. Separate the column names with commas (,).                                                 |
|                    | Input Sparse Matrix            | The sparse format that uses key-value pairs.                                                                                                                   |
|                    | KV Pair Delimiter              | The default delimiter between key-value pairs is a comma (,).                                                                                                  |
|                    | KV Delimiter                   | The default delimiter between keys and values in key-value pairs is a colon (:).                                                                               |
| Parameters Setting | Clusters                       | Valid values: 1 to 1000.                                                                                                                                       |
|                    | Distance Measurement Method    | The method to measure distances. Valid values: <b>Euclidean</b> , <b>Cosine</b> , and <b>Cityblock</b> .                                                       |
|                    | Centroid Initialization Method | The method to initialize centroids. Valid values: <b>Random</b> , <b>First K</b> , <b>Uniform</b> , <b>K-means++</b> , and <b>Use Initial Centroid Table</b> . |
|                    | Maximum Iterations             | Valid values: 1 to 1000.                                                                                                                                       |
|                    | Convergence Criteria           | The threshold to terminate iterations.                                                                                                                         |
|                    | Initial Random Seed            | By default, the current time is used. If this parameter uses a fixed value, the clustering result is stable.                                                   |
| Tuning             | Cores                          | The number of cores. By default, it is allocated by the system.                                                                                                |
|                    | Memory Size per Core           | The memory size of each core. By default, it is allocated by the system.                                                                                       |

- PAI command

```

pai -name kmeans
  -project algo_public
  -DinputTableName=pai_kmeans_test_input
  -DselectedColNames=f0,f1
  -DappendColNames=f0,f1
  -DcenterCount=3
  -Dloop=10
  -Daccuracy=0.01
  -DdistanceType=euclidean
  -DinitCenterMethod=random
  -Dseed=1
  -DmodelName=pai_kmeans_test_output_model_
  -DidxTableName=pai_kmeans_test_output_idx
  -DclusterCountTableName=pai_kmeans_test_output_couter
  -DcenterTableName=pai_kmeans_test_output_center;
  
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                               | Default value    |
|----------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                                              | No default value |
| selectedColNames     | No       | The columns selected from the input table for training. The column names must be separated by commas (,). Columns of the INT and DOUBLE types are supported. If the input data is in sparse format, columns of the STRING type are supported.                                                                                                                                                                             | All columns      |
| inputTablePartitions | No       | The partitions selected from the input table for training. The following formats are supported: <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (,).         </div> | All partitions   |
| appendColNames       | No       | The input columns that are appended to the clustering result table. Separate the column names with commas (,).                                                                                                                                                                                                                                                                                                            | No default value |
| enableSparse         | No       | Specifies whether data in the input table is in sparse format. Valid values: {true,false}.                                                                                                                                                                                                                                                                                                                                | false            |
| itemDelimiter        | No       | The delimiter between key-value pairs.                                                                                                                                                                                                                                                                                                                                                                                    | Comma (,)        |
| kvDelimiter          | No       | The delimiter between keys and values in key-value pairs.                                                                                                                                                                                                                                                                                                                                                                 | Colon (:)        |
| centerCount          | Yes      | The number of clustering centroids. Valid values: 1 to 1000.                                                                                                                                                                                                                                                                                                                                                              | 10               |

| Parameter           | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | Default value       |
|---------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| distanceType        | No       | <p>The method to measure distances. Valid values:</p> <ul style="list-style-type: none"> <li>euclidean: the Euclidean distance that is calculated by using the following formula:<br/> <math display="block">d(x - c) = \sqrt{(x - c)(x - c)'}</math> </li> <li>cosine: the cosine that is calculated by using the following formula:<br/> <math display="block">d(x - c) = 0.5 - 0.5 * \frac{xc'}{\sqrt{xx'}\sqrt{cc'}}</math> </li> <li>cityblock: the city block distance, which is also called the Manhattan distance. It is calculated by using the following formula:<br/> <math display="block">d(x - c) =  x - c </math> </li> </ul>      | euclidean           |
| initCenterMethod    | No       | <p>The method to initialize centroids. Valid values:</p> <ul style="list-style-type: none"> <li>random: K initial centroids are randomly sampled from the input table. The initial random seed is specified by seed.</li> <li>topk: The first K rows in the input table are used as the initial centroids.</li> <li>uniform: K initial centroids are calculated from the minimum value to the maximum value. This ensures that these initial centroids are evenly distributed.</li> <li>kmpp: K initial centroids are obtained by using the k-means++ algorithm.</li> <li>external: the table that lists additional initial centroids.</li> </ul> | random              |
| initCenterTableName | No       | The name of the table that lists initial centroids. This parameter is valid only if initCenterMethod is set to external.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | No default value    |
| loop                | No       | The maximum number of iterations. Valid values: 1 to 1000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 100                 |
| accuracy            | No       | The condition for terminating an algorithm. The algorithm is terminated if the objective difference between two iterations is less than the value of this parameter.                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | 0.1                 |
| seed                | No       | The initial random seed.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Current system time |
| modelName           | No       | The name of the output model.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | No default value    |

| Parameter             | Required | Description                                                                                                                      | Default value           |
|-----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| idxTableName          | Yes      | The name of the clustering result table, which includes the ID of the cluster to which each record belongs after the clustering. | No default value        |
| idxTablePartition     | No       | The partition in the clustering result table.                                                                                    | No default value        |
| clusterCountTableName | No       | The clustering statistics table that counts the number of points included in each cluster.                                       | No default value        |
| centerTableName       | No       | The clustering centroid table.                                                                                                   | No default value        |
| coreNum               | No       | The number of cores. This parameter is used with memSizePerCore. Valid values: 1 to 9999.                                        | Automatically allocated |
| memSizePerCore        | No       | The memory size of each core, in MB. Valid values: 1024 to 64 x 1024.                                                            | Automatically allocated |
| lifecycle             | No       | The lifecycle of the output table.                                                                                               | No default value        |

## Output data

The output data of the K-means Clustering component includes the clustering result table, clustering statistics table, and clustering centroid table. Output format:

- Clustering result table

| Column         | Description                                                                  |
|----------------|------------------------------------------------------------------------------|
| appendColNames | The names of the appended columns.                                           |
| cluster_index  | The cluster to which each sample is assigned in the training table.          |
| distance       | The distance from each sample to the cluster centroid in the training table. |

- Clustering statistics table

| Column        | Description                            |
|---------------|----------------------------------------|
| cluster_index | The ID of the cluster.                 |
| cluster_count | The number of samples in each cluster. |

- Clustering centroid table

| Column        | Description            |
|---------------|------------------------|
| cluster_index | The ID of the cluster. |

| Column           | Description                                                |
|------------------|------------------------------------------------------------|
| selectedColNames | The columns selected from the training table for training. |

## Example

Input data in dense format:

1. You can generate test data by using one of the following methods:

- Using the initial centroid table

```
create table pai_kmeans_test_init_center as
select * from
(
  select 1 as f0,2 as f1 from dual
  union all
  select 1 as f0,3 as f1 from dual
  union all
  select 1 as f0,4 as f1 from dual
)tmp;
```

- Using other initial centroids

```
create table pai_kmeans_test_input as
select * from
(
  select 'id1' as id,1 as f0,2 as f1 from dual
  union all
  select 'id2' as id,1 as f0,3 as f1 from dual
  union all
  select 'id3' as id,1 as f0,4 as f1 from dual
  union all
  select 'id4' as id,0 as f0,3 as f1 from dual
  union all
  select 'id5' as id,0 as f0,4 as f1 from dual
)tmp;
```

2. Run PAI commands to submit the parameters of the K-means Clustering component.

- Using the initial centroid table

```
drop table if exists pai_kmeans_test_output_idx;
yes
drop table if exists pai_kmeans_test_output_couter;
yes
drop table if exists pai_kmeans_test_output_center;
yes
drop offlinemodel if exists pai_kmeans_test_output_model_;
yes
pai -name kmeans
  -project algo_public
  -DinputTableName=pai_kmeans_test_input
  -DinitCenterTableName=pai_kmeans_test_init_center
  -DselectedColNames=f0,f1
  -DappendColNames=f0,f1
  -DcenterCount=3
  -Dloop=10
  -Daccuracy=0.01
  -DdistanceType=euclidean
  -DinitCenterMethod=external
  -Dseed=1
  -DmodelName=pai_kmeans_test_output_model_
  -DidxTableName=pai_kmeans_test_output_idx
  -DclusterCountTableName=pai_kmeans_test_output_couter
  -DcenterTableName=pai_kmeans_test_output_center;
```

- o Using the initial centroids that are randomly selected

```
drop table if exists pai_kmeans_test_output_idx;
yes
drop table if exists pai_kmeans_test_output_couter;
yes
drop table if exists pai_kmeans_test_output_center;
yes
drop offlinemodel if exists pai_kmeans_test_output_model_;
yes
pai -name kmeans
  -project algo_public
  -DinputTableName=pai_kmeans_test_input
  -DselectedColNames=f0,f1
  -DappendColNames=f0,f1
  -DcenterCount=3
  -Dloop=10
  -Daccuracy=0.01
  -DdistanceType=euclidean
  -DinitCenterMethod=random
  -Dseed=1
  -DmodelName=pai_kmeans_test_output_model_
  -DidxTableName=pai_kmeans_test_output_idx
  -DclusterCountTableName=pai_kmeans_test_output_couter
  -DcenterTableName=pai_kmeans_test_output_center;
```

3. View the clustering result table, clustering statistics table, and clustering centroid table.

- o Clustering result table specified by idxTableName

| f0 | f1 | cluster_index | distance |
|----|----|---------------|----------|
| 1  | 2  | 0             | 0.0      |
| 1  | 3  | 1             | 0.5      |
| 1  | 4  | 2             | 0.5      |
| 0  | 3  | 1             | 0.5      |
| 0  | 4  | 2             | 0.5      |

- Clustering statistics table specified by clusterCountTableName

| cluster_index | cluster_count |
|---------------|---------------|
| 0             | 1             |
| 1             | 2             |
| 2             | 2             |

- Clustering centroid table specified by centerTableName

| cluster_index | f0  | f1  |
|---------------|-----|-----|
| 0             | 1.0 | 2.0 |
| 1             | 0.5 | 3.0 |
| 2             | 0.5 | 4.0 |

Input data in sparse format:

- Generate test data.

```
create table pai_kmeans_test_sparse_input as
select * from
(
  select 1 as id,"s1" as id_s,"0:0.1,1:0.2" as kvs0,"2:0.3,3:0.4" as kvs1 from dual
  union all
  select 2 as id,"s2" as id_s,"0:1.1,2:1.2" as kvs0,"4:1.3,5:1.4" as kvs1 from dual
  union all
  select 3 as id,"s3" as id_s,"0:2.1,3:2.2" as kvs0,"6:2.3,7:2.4" as kvs1 from dual
  union all
  select 4 as id,"s4" as id_s,"0:3.1,4:3.2" as kvs0,"8:3.3,9:3.4" as kvs1 from dual
  union all
  select 5 as id,"s5" as id_s,"0:5.1,5:5.2" as kvs0,"10:5.3,6:5.4" as kvs1 from dual
)tmp;
```

If input data is in sparse format, 0 is used to impute the columns with missing values. If multiple columns are used as an input, these columns are merged. For example, if kvs0 and kvs1 are used as an input, the first row contains the following data:

```
0:0.1,1:0.2,2:0.3,3:0.4,4:0,5:0,6:0,7:0,8:0,9:0,10:0
```

In this example, the sparse matrix is numbered from 0, and has 5 rows and 11 columns. If a column in



kvs contains 123456789:0.1 , the sparse matrix has 5 rows and 123456789 columns. This matrix consumes large amounts of CPU and memory resources. If kvs contains the columns that are incorrectly numbered, we recommend that you renumber the columns to reduce the size of the matrix.

- Run the following PAI command to submit the parameters of the K-means Clustering component:

```
pai -name kmeans
  -project algo_public
  -DinputTableName=pai_kmeans_test_sparse_input
  -DenableSparse=true
  -DselectedColNames=kvs0,kvs1
  -DappendColNames=id,id_s
  -DitemDelimiter=,
  -DkvDelimiter=:
  -DcenterCount=3
  -Dloop=100
  -Daccuracy=0.01
  -DdistanceType=euclidean
  -DinitCenterMethod=topk
  -Dseed=1
  -DmodelName=pai_kmeans_test_input_sparse_output_model
  -DidxTableName=pai_kmeans_test_sparse_output_idx
  -DclusterCountTableName=pai_kmeans_test_sparse_output_couter
  -DcenterTableName=pai_kmeans_test_sparse_output_center;
```

- View the clustering result table, clustering statistics table, and clustering centroid table.

- Clustering result table specified by idxTableName

| id | id_s | cluster_index | distance           |
|----|------|---------------|--------------------|
| 4  | s4   | 0             | 2.90215437218629   |
| 5  | s5   | 1             | 0.0                |
| 1  | s1   | 2             | 0.7088723439378913 |
| 2  | s2   | 2             | 1.1683321445547923 |
| 3  | s3   | 0             | 2.0548722588034516 |

- Clustering statistics table specified by clusterCountTableName

| cluster_index | cluster_count |
|---------------|---------------|
| 0             | 2             |
| 1             | 1             |
| 2             | 2             |

- Clustering centroid table specified by centerTableName

```

+-----+-----+-----+
| cluster_index | kvs0   | kvs1   |
+-----+-----+-----+
| 0             | 0:2.6,1:0,2:0,3:1.1,4:1.6,5:0 | 6:1.15,7:1.2,8:1.65,9:1.7,10:0 |
| 1             | 0:5.1,1:0,2:0,3:0,4:0,5:5.2 | 6:5.4,7:0,8:0,9:0,10:5.3 |
| 2             | 0:0.6,1:0.1,2:0.75,3:0.2,4:0.65,5:0.7 | 6:0,7:0,8:0,9:0,10:0 |
+-----+-----+-----+

```

**Notes:**

- If cosine is used, some clusters may be empty, that is, the number of clusters is less than K. K initial centroids may be parallel vectors. If the centroids are traversed in sequence, the sample is not distributed to the centroids that are parallel vectors. We recommend that you use the K centroids that you listed in the external centroid table.
- If the input table contains NULL or empty values, the system reports the following error: **Algo Job Failed-System Error-Null feature value found** . We recommend that you use the default values for imputation.
- If data in sparse format is used as an input and the largest column ID exceeds 2000000, the system reports the following error: **Algo Job Failed-System Error-Feature count can't be more than 2000000** . We recommend that you renumber the columns from 0 or 1.
- If a write operation fails due to a too large centroid model, the system reports the following error: **Algo Job Failed-System Error-kIOError:Write failed for message: comparison\_measure** . We recommend that you renumber the columns whose data is in sparse format from 0 or 1. If the value of `col*centerCount` is greater than 270000000, run commands to remove the modelName parameter, and then perform clustering again.
- If the name of a column in the input table contains SQL keywords, the system reports the following error: **FAILED: Failed Task createCenterTable:kOtherError:ODPS-0130161:[1,558] Parse exception - invalid token ',', expect ")"** .

## DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is an unsupervised clustering algorithm based on data density. In a cluster, DBSCAN uses specified radius and quantity thresholds to filter core points and neighbors in a region. DBSCAN uses density-reachable and density-connected theories to cluster data points.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab | Parameter              | Description                                                                         |
|-----|------------------------|-------------------------------------------------------------------------------------|
|     | <b>Input Data Type</b> | The type of the input data. Valid values: <b>Adjacency List</b> and <b>Vector</b> . |

| Tab                | Parameter                             | Description                                                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------|---------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Data Vector Dimension                 | <p>The vector dimension of the input data. This parameter is required only if <b>Input Data Type</b> is set to <b>Vector</b>.</p> <div> <p> <b>Note</b> If <b>Input Table Format</b> is set to <b>Multiple Columns</b>, the value of <b>Data Vector Dimension</b> must be the same as the number of data columns you selected.</p> </div> |
|                    | Neighborhood Point Distance Threshold | If the distance between two points is less than the threshold specified by this parameter, these points are neighbors of each other. This parameter is required only if <b>Input Data Type</b> is set to <b>Vector</b> .                                                                                                                                                                                                   |
|                    | Core Object Density Threshold         | If the number of points in the neighborhood of a point is greater than the threshold specified by this parameter, the point is a core object.                                                                                                                                                                                                                                                                              |
|                    | Input Table Format                    | <p>This parameter is required only if <b>Input Data Type</b> is set to <b>Vector</b>. Valid values:</p> <ul style="list-style-type: none"> <li>◦ <b>Multiple Columns</b>: Multiple columns are used to represent a vector.</li> <li>◦ <b>Two Columns</b>: A single column is used to represent a vector. Separate dimensions of the vector with commas (,).</li> </ul>                                                     |
| Fields Information | Data Columns (First Select ID Column) | This parameter is required only if <b>Input Table Format</b> is set to <b>Multiple Columns</b> .                                                                                                                                                                                                                                                                                                                           |
| Tuning             | Servers                               | The number of servers.                                                                                                                                                                                                                                                                                                                                                                                                     |
|                    | Workers (>1)                          | The number of workers.                                                                                                                                                                                                                                                                                                                                                                                                     |
|                    | CPUs per Server                       | The number of CPUs for a server.                                                                                                                                                                                                                                                                                                                                                                                           |
|                    | CPUs per Worker                       | The number of CPUs for a worker.                                                                                                                                                                                                                                                                                                                                                                                           |
|                    | Memory Size per Worker                | The memory size of each worker, in MB.                                                                                                                                                                                                                                                                                                                                                                                     |
|                    | Memory Size per Server                | The memory size of each server, in MB.                                                                                                                                                                                                                                                                                                                                                                                     |

- PAI command

DBSCAN allows you to use a neighbor table or a vector as an input, and to use two or multiple columns to represent a vector. Sample commands:

- Using a neighbor table as an input

```

pai -name ps_dbscan
-DinputTable=hxdb_neighbor_data_order
-DinputType="1"
-DoutputTable="hxtmp2"
-DminPoints="4"
-DserverNum="1"
-DserverCpu="300"
-DserverMemory="3000"
-DworkerNum="2"
-DworkerCpu="800"
-DworkerMemory="2000"

```

- Using a vector represented by multiple columns as an input

```

pai -name ps_dbscan
-DinputTable=hxdb_multicols_data
-DinputType="0"
-DoutputTable="hxtmp"
-DdataType="DenseMultiCols"
-DpointDim="12"
-Deps="4"
-DminPoints="20"
-DselectedColIds="all"
-DserverNum="2"
-DserverCpu="300"
-DserverMemory="3000"
-DworkerNum="10"
-DworkerCpu="800"
-DworkerMemory="2000"

```

- Using a vector represented by two columns as an input

```

pai -name ps_dbscan
-DinputTable="hxdb_sample_60w"
-DinputType="0"
-DoutputTable="hxtmp1"
-DdataType="Dense2Cols"
-DpointDim="2"
-Deps="0.01"
-DminPoints="10"
-DselectedColIds="all"
-DserverNum="2"
-DserverCpu="300"
-DserverMemory="3000"
-DworkerNum="10"
-DworkerCpu="800"
-DworkerMemory="2000"

```

| Parameter  | Required | Description                  | Default value    |
|------------|----------|------------------------------|------------------|
| inputTable | Yes      | The name of the input table. | No default value |

| Parameter   | Required | Description                                                                                                                                                                                                                                                                                                                                                                    | Default value    |
|-------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| outputTable | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                  | No default value |
| inputType   | No       | The type of the input data. Valid values: <ul style="list-style-type: none"> <li>0: A vector is used as an input.</li> <li>1: A neighbor table is used as an input.</li> </ul>                                                                                                                                                                                                 | 0                |
| pointDim    | No       | The vector dimension of the input data. This parameter is required only if inputType is set to 0. <div>  <b>Note</b> If dataType is set to DenseMultiCols, the value of pointDim must be the same as the number of the columns specified by selectedCols. </div>                              | 10               |
| eps         | No       | The threshold of the distance between two neighbors. If the distance between two points is less than the threshold, the points are neighbors of each other. This parameter is required only if inputType is set to 0.                                                                                                                                                          | 1.0              |
| minPoints   | No       | The density threshold for a core object. If the number of points in the neighborhood of a point is greater than the threshold specified by this parameter, this point is a core object.                                                                                                                                                                                        | 10               |
| dataType    | No       | The format of the input table. This parameter is required only if inputType is set to 0. Valid values: <ul style="list-style-type: none"> <li>DenseMultiCols: indicates that multiple columns are used to represent a vector.</li> <li>Dense2Cols: indicates that a single column is used to represent a vector. Separate dimensions of the vector with commas (,).</li> </ul> | Dense2Cols       |

| Parameter    | Required | Description                                                                                                                                                                                          | Default value |
|--------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| selectedCols | No       | <p>The column where the data is located. This parameter is required only if dataType is set to DenseMultiCols. The format is all or 0,1,3.</p> <p><b>Note</b> The ID column is the first column.</p> | all           |
| serverNum    | Yes      | The number of servers.                                                                                                                                                                               | 5             |
| workerNum    | Yes      | The number of workers.                                                                                                                                                                               | 30            |
| serverCpu    | Yes      | The number of CPUs for a server.                                                                                                                                                                     | 8             |
| workerCpu    | Yes      | The number of CPUs for a worker.                                                                                                                                                                     | 8             |
| workerMemory | Yes      | The memory size of each worker, in MB.                                                                                                                                                               | 10000         |
| serverMemory | Yes      | The memory size of each server, in MB.                                                                                                                                                               | 10000         |

## Input data

DBSCAN allows you to use a neighbor table or a vector as an input, and to use two or multiple columns to represent a vector. Example:

- Using a neighbor table

```
+-----+-----+
| mid(bigint) | f1(string) |
+-----+-----+
| 0 | 2,3,0 |
| 1 | 1,2,3,4 |
| 2 | 2,1,5 |
| 3 | 1,3 |
| 4 | 1,4 |
| 5 | 2,5,1,0 |
+-----+-----+
```

**Note** The neighbors of a point must include this point. For example, the neighbors of point 0 must include point 0.

- Using a two-dimensional vector that is represented by multiple columns

```
+-----+-----+-----+
| mid(bigint) | f1(double) | f2(double) |
+-----+-----+-----+
| 0 | 0.0 | 0.3 |
| 1 | 0.0 | 1.0 |
| 2 | 0.0 | 0.1 |
| 3 | 1.0 | 0.0 |
| 4 | 0.0 | 0.2 |
| 5 | 1.0 | 0.2 |
+-----+-----+-----+
```

The first column lists the sample IDs. The second and third columns list the values of each dimension of the vector.

- Using a two-dimensional vector that is represented by two columns

```
+-----+-----+
| mid(bigint) | f1(string) |
+-----+-----+
| 0 | 0.0,0.3 |
| 1 | 0.0,1.0 |
| 2 | 0.0,0.1 |
| 3 | 1.0,0.0 |
| 4 | 0.0,0.2 |
| 5 | 1.0,0.2 |
+-----+-----+
```

The first column lists the sample IDs. The second column lists the values of each dimension of the vector. Separate the values of different dimensions with commas (,).

 **Note** You can use the **Add ID Column** component under **Data Preprocessing** to add ID columns for each sample.

## 3.5.4. Regression

This topic describes the GBDT Regression, Linear Regression, PS-SMART Regression, and PS Linear Regression components provided by Machine Learning Studio for regressions.

### GBDT Regression

Gradient Boosting Decision Tree (GBDT) is an iteration decision tree algorithm that is suitable for linear and non-linear regression scenarios.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab | Parameter | Description |
|-----|-----------|-------------|
|-----|-----------|-------------|

| Tab                | Parameter                                   | Description                                                                                                                                                                                                                                                                                                |
|--------------------|---------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Input Columns                               | <p>The feature columns selected from the input table for training. The columns of the DOUBLE and BIGINT data types are supported.</p> <div>  <b>Note</b> A maximum of 800 feature columns can be selected.         </div> |
|                    | Label Column                                | The label column. The columns of the DOUBLE and BIGINT data types are supported.                                                                                                                                                                                                                           |
|                    | Stratification Column                       | The group column. The columns of the DOUBLE and BIGINT data types are supported. The full table is used as a group by default.                                                                                                                                                                             |
| Parameters Setting | Loss Function Type                          | The type of the loss function. Valid values: <b>GBrank Loss</b> , <b>Lambdamart DCG Loss</b> , <b>Lambdamart NDCG Loss</b> , and <b>Regression Loss</b> .                                                                                                                                                  |
|                    | Tau of GBrank Loss                          | This parameter is required only when Loss Function Type is set to GBrank Loss. Valid values: [0,1].                                                                                                                                                                                                        |
|                    | Exponent Base of GBrank and Regression Loss | This parameter is required only when Loss Function Type is set to GBrank Loss or Regression Loss. Valid values: [1,10].                                                                                                                                                                                    |
|                    | Metric Type                                 | The metric type. Valid values: <b>NDCG</b> and <b>DCG</b> .                                                                                                                                                                                                                                                |
|                    | Trees                                       | The number of trees. Valid values: 1 to 10000.                                                                                                                                                                                                                                                             |
|                    | Learning Rate                               | The learning rate. Valid values: (0,1).                                                                                                                                                                                                                                                                    |
|                    | Maximum Leaves                              | The maximum number of leaf nodes on each tree. Valid values: 1 to 1000.                                                                                                                                                                                                                                    |
|                    | Maximum Tree Depth                          | The maximum depth of each tree. Valid values: 1 to 100.                                                                                                                                                                                                                                                    |
|                    | Minimum Samples per Leaf Node               | The minimum number of samples on each leaf node. Valid values: 1 to 1000.                                                                                                                                                                                                                                  |
|                    | Sampling Fraction                           | The ratio of samples selected for training. Valid values: (0,1).                                                                                                                                                                                                                                           |
|                    | Fraction of Training Features               | The ratio of features selected for training. Valid values: (0,1).                                                                                                                                                                                                                                          |
|                    | Testing Sample Fraction                     | The ratio of samples selected for testing. Valid values: [0,1].                                                                                                                                                                                                                                            |
|                    | Random Seed                                 | The number of random seeds. Valid values: [0,10].                                                                                                                                                                                                                                                          |



| Tab    | Parameter                  | Description                                                                                                                   |
|--------|----------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| Tuning | Use Newton's Method        | Specifies whether to use Newton's method.                                                                                     |
|        | Maximum Splits per Feature | The maximum number of splits of each feature. Valid values: 1 to 1000.                                                        |
|        | Cores                      | The number of cores. The system automatically allocates the number of cores based on the volume of input data.                |
|        | Memory Size per Core       | The memory size of each core. The system automatically allocates the memory size based on the volume of input data. Unit: MB. |

- PAI command

```
PAI -name gbd
  -project algo_public
  -DfeatureSplitValueMaxSize="500"
  -DlossType="0"
  -DrandSeed="0"
  -DnewtonStep="0"
  -Dshrinkage="0.05"
  -DmaxLeafCount="32"
  -DlabelColName="campaign"
  -DinputTableName="bank_data_partition"
  -DminLeafSampleCount="500"
  -DsampleRatio="0.6"
  -DgroupIDColName="age"
  -DmaxDepth="11"
  -DmodelName="xlab_m_GBDT_83602"
  -DmetricType="2"
  -DfeatureRatio="0.6"
  -DinputTablePartitions="pt=20150501"
  -Dtau="0.6"
  -Dp="1"
  -DtestRatio="0.0"
  -DfeatureColNames="previous,cons_conf_idx,euribor3m"
  -DtreeCount="500"
```

| Parameter       | Required | Description                                                                                                                    | Default value                     |
|-----------------|----------|--------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTableName  | Yes      | The name of the input table.                                                                                                   | No default value                  |
| featureColNames | No       | The feature columns selected from the input table for training. The columns of the DOUBLE and BIGINT data types are supported. | All columns of numeric data types |

| Parameter                 | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                         | Default value    |
|---------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| labelColName              | Yes      | The label column in the input table. The columns of the DOUBLE and BIGINT data types are supported.                                                                                                                                                                                                                                                                                                                                 | No default value |
| inputTablePartitions      | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (,).</p> | All partitions   |
| modelName                 | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                                                                                                                                       | No default value |
| outputImportanceTableName | No       | The name of the feature importance table that is generated.                                                                                                                                                                                                                                                                                                                                                                         | No default value |
| groupIDColName            | No       | The name of the group column.                                                                                                                                                                                                                                                                                                                                                                                                       | Full table       |
| lossType                  | No       | <p>The type of the loss function. Valid values:</p> <ul style="list-style-type: none"> <li>0: GBRank</li> <li>1: LAMBDAMART_DCG</li> <li>2: LAMBDAMART_NDCG</li> <li>3: LEAST_SQUARE</li> <li>4: log-likelihood</li> </ul>                                                                                                                                                                                                          | 0                |
| metricType                | No       | <p>The metric type. Valid values:</p> <ul style="list-style-type: none"> <li>0: normalized discounted cumulative gain (NDCG).</li> <li>1: discounted cumulative gain (DCG).</li> <li>2: area under the curve (AUC). This metric type is suitable only for the scenario where the value of label is set to 0 or 1.</li> </ul>                                                                                                        | 0                |
| treeCount                 | No       | The number of trees. Valid values: 1 to 10000.                                                                                                                                                                                                                                                                                                                                                                                      | 500              |
| shrinkage                 | No       | The learning rate. Valid values: (0,1).                                                                                                                                                                                                                                                                                                                                                                                             | 0.05             |
| maxLeafCount              | No       | The maximum number of leaf nodes on each tree. Valid values: 1 to 1000.                                                                                                                                                                                                                                                                                                                                                             | 32               |

| Parameter                | Required | Description                                                               | Default value    |
|--------------------------|----------|---------------------------------------------------------------------------|------------------|
| maxDepth                 | No       | The maximum depth of each tree. Valid values: 1 to 100.                   | 10               |
| minLeafSampleCount       | No       | The minimum number of samples on each leaf node. Valid values: 1 to 1000. | 500              |
| sampleRatio              | No       | The ratio of samples selected for training. Valid values: (0,1).          | 0.6              |
| featureRatio             | No       | The ratio of samples selected for training. Valid values: (0,1).          | 0.6              |
| tau                      | No       | The Tau parameter for the GBRank loss function. Valid values: [0,1].      | 0.6              |
| p                        | No       | The p parameter for the GBRank loss function. Valid values: [1,10].       | 1                |
| randSeed                 | No       | The number of random seeds. Valid values: [0,10].                         | 0                |
| newtonStep               | No       | Specifies whether to use Newton's method. Valid values: 0 and 1.          | 1                |
| featureSplitValueMaxSize | No       | The maximum number of splits of each feature. Valid values: 1 to 1000.    | 500              |
| lifecycle                | No       | The lifecycle of the output table.                                        | No default value |

## Example

1. Execute the following SQL statement to generate training data:

```
drop table if exists gbdt_ls_test_input;
create table gbdt_ls_test_input
as
select
*
from
(
select
cast(1 as double) as f0,
cast(0 as double) as f1,
cast(0 as double) as f2,
cast(0 as double) as f3,
cast(0 as bigint) as label
from dual
union all
select
cast(0 as double) as f0,
cast(1 as double) as f1,
cast(0 as double) as f2,
```

```

    cast(0 as double) as f2,
    cast(0 as double) as f3,
    cast(0 as bigint) as label
from dual
union all
select
    cast(0 as double) as f0,
    cast(0 as double) as f1,
    cast(1 as double) as f2,
    cast(0 as double) as f3,
    cast(1 as bigint) as label
from dual
union all
select
    cast(0 as double) as f0,
    cast(0 as double) as f1,
    cast(0 as double) as f2,
    cast(1 as double) as f3,
    cast(1 as bigint) as label
from dual
union all
select
    cast(1 as double) as f0,
    cast(0 as double) as f1,
    cast(0 as double) as f2,
    cast(0 as double) as f3,
    cast(0 as bigint) as label
from dual
union all
select
    cast(0 as double) as f0,
    cast(1 as double) as f1,
    cast(0 as double) as f2,
    cast(0 as double) as f3,
    cast(0 as bigint) as label
from dual
) a;

```

The following test data table gbdt\_ls\_test\_input is generated.

| f0  | f1  | f2  | f3  | label |
|-----|-----|-----|-----|-------|
| 1.0 | 0.0 | 0.0 | 0.0 | 0     |
| 0.0 | 0.0 | 1.0 | 0.0 | 1     |
| 0.0 | 0.0 | 0.0 | 1.0 | 1     |
| 0.0 | 1.0 | 0.0 | 0.0 | 0     |
| 1.0 | 0.0 | 0.0 | 0.0 | 0     |
| 0.0 | 1.0 | 0.0 | 0.0 | 0     |

2. Run the following PAI command to submit the training parameters configured for the GBDT

Regression component :

```
drop offlinemodel if exists gbdt_ls_test_model;
PAI -name gbdt
  -project algo_public
  -DfeatureSplitValueMaxSize="500"
  -DlossType="3"
  -DrandSeed="0"
  -DnewtonStep="1"
  -Dshrinkage="0.5"
  -DmaxLeafCount="32"
  -DlabelColName="label"
  -DinputTableName="gbdt_ls_test_input"
  -DminLeafSampleCount="1"
  -DsampleRatio="1"
  -DmaxDepth="10"
  -DmetricType="0"
  -DmodelName="gbdt_ls_test_model"
  -DfeatureRatio="1"
  -Dp="1"
  -Dtau="0.6"
  -DtestRatio="0"
  -DfeatureColNames="f0,f1,f2,f3"
  -DtreeCount="10"
```

3. Run the following PAI command to submit the parameters configured for the Prediction component :

```
drop table if exists gbdt_ls_test_prediction_result;
PAI -name prediction
  -project algo_public
  -DdetailColName="prediction_detail"
  -DmodelName="gbdt_ls_test_model"
  -DitemDelimiter=","
  -DresultColName="prediction_result"
  -Dlifecycle="28"
  -DoutputTableName="gbdt_ls_test_prediction_result"
  -DscoreColName="prediction_score"
  -DkvDelimiter=":"
  -DinputTableName="gbdt_ls_test_input"
  -DenableSparse="false"
  -DappendColNames="label"
```

4. View the prediction result table gbdt\_ls\_test\_prediction\_result.

| label | prediction_result | prediction_score | prediction_detail       |
|-------|-------------------|------------------|-------------------------|
| 0     | NULL              | 0.0              | {"label": 0}            |
| 0     | NULL              | 0.0              | {"label": 0}            |
| 1     | NULL              | 0.9990234375     | {"label": 0.9990234375} |
| 1     | NULL              | 0.9990234375     | {"label": 0.9990234375} |

| label | prediction_result | prediction_score | prediction_detail |
|-------|-------------------|------------------|-------------------|
| 0     | NULL              | 0.0              | {"label": 0}      |
| 0     | NULL              | 0.0              | {"label": 0}      |

## Linear Regression

Linear Regression is a model used to analyze the linear relationship between a dependent variable and multiple independent variables. You can configure the component by using one of the following methods:

- PAI console

| Tab                | Parameter                         | Description                                                                                                                                                                                   |
|--------------------|-----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns                   | The feature columns selected from the input table for training.                                                                                                                               |
|                    | Label Column                      | The label column. The columns of the DOUBLE and BIGINT data types are supported.                                                                                                              |
|                    | Use Sparse Format                 | Specifies whether the input data is in the sparse format. Data in the sparse format is presented by using key-value pairs.                                                                    |
|                    | KV Pair Delimiter                 | The delimiter used to separate key-value pairs. Commas (,) are used by default.                                                                                                               |
|                    | KV Delimiter                      | The delimiter used to separate keys and values. Colons (:) are used by default.                                                                                                               |
| Parameters Setting | Maximum Iterations                | The maximum number of iterations performed by the algorithm.                                                                                                                                  |
|                    | Minimum Likelihood Deviance       | The algorithm is terminated if the difference of log-likelihood between two iterations is less than the value specified by this parameter.                                                    |
|                    | Regularization Type               | The regularization type. Valid values: L1, L2, and None.                                                                                                                                      |
|                    | Regularization Coefficient        | The regularization coefficient. This parameter is invalid when <b>Regularization Type</b> is set to <b>None</b> .                                                                             |
|                    | Generate Model Evaluation Table   | The metrics include R-Squared, adjusted R-Squared, AIC, degree of freedom, residual standard deviation, and residual deviation.                                                               |
|                    | Regression Coefficient Evaluation | The metrics include the t value and p value, and the confidence interval is [2.5%,97.5%]. This parameter is valid only when the <b>Generate Model Evaluation Table</b> check box is selected. |

| Tab    | Parameter            | Description                                                                |
|--------|----------------------|----------------------------------------------------------------------------|
| Tuning | Cores                | The number of cores. It is automatically allocated by the system.          |
|        | Memory Size per Core | The memory size of each core. It is automatically allocated by the system. |

- PAI command

```
PAI -name linearregression
  -project algo_public
  -DinputTableName=lm_test_input
  -DfeatureColNames=x
  -DlabelColName=y
  -DmodelName=lm_test_input_model_out;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                         | Default value    |
|----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                        | No default value |
| modelName            | Yes      | The name of the output model.                                                                                                                                                                                                                                                       | No default value |
| outputTableName      | No       | The name of the model evaluation table that is generated. This parameter is required when enableFitGoodness is set to true.                                                                                                                                                         | No default value |
| labelColName         | Yes      | The label column. This parameter specifies the dependent variable. The columns of the DOUBLE and BIGINT data types are supported. You can select only one column.                                                                                                                   | No default value |
| featureColNames      | Yes      | The feature columns. This parameter specifies the independent variables. If data in the input table is in the dense format, columns of the DOUBLE or BIGINT data type are supported. If the input data is in the sparse format, only columns of the STRING data type are supported. | No default value |
| inputTablePartitions | No       | The partitions selected from the input table for training.                                                                                                                                                                                                                          | No default value |
| enableSparse         | No       | Specifies whether the input data is in the sparse format. Valid values: true and false.                                                                                                                                                                                             | false            |

| Parameter                 | Required | Description                                                                                                                                                                                                                                      | Default value           |
|---------------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| itemDelimiter             | No       | The delimiter used to separate key-value pairs. This parameter is valid when enableSparse is set to true.                                                                                                                                        | ,                       |
| kvDelimiter               | No       | The delimiter used to separate keys and values. This parameter is valid when enableSparse is set to true.                                                                                                                                        | :                       |
| maxIter                   | No       | The maximum number of iterations performed by the algorithm.                                                                                                                                                                                     | 100                     |
| epsilon                   | No       | The minimum likelihood deviance. The algorithm is terminated if the difference of log-likelihood between two iterations is less than the value specified by this parameter.                                                                      | 0.000001                |
| regularizedType           | No       | The regularization type. Valid values: L1, L2, and None.                                                                                                                                                                                         | None                    |
| regularizedLevel          | No       | The regularization coefficient. This parameter is invalid when regularizedType is set to <b>None</b> .                                                                                                                                           | 1                       |
| enableFitGoodness         | No       | Specifies whether to generate the model evaluation table. The metrics include R-Squared, adjusted R-Squared, AIC, degree of freedom, residual standard deviation, and residual deviation. Valid values: true and false.                          | false                   |
| enableCoefficientEstimate | No       | Specifies whether to evaluate regularization coefficient. The metrics include the t value and p value, and the confidence interval is [2.5%,97.5%]. This parameter is valid when enableFitGoodness is set to true. Valid values: true and false. | false                   |
| lifecycle                 | No       | The lifecycle of the model evaluation table that is generated.                                                                                                                                                                                   | -1                      |
| coreNum                   | No       | The number of cores used in computing.                                                                                                                                                                                                           | Automatically allocated |



| Parameter      | Required | Description                                                                    | Default value           |
|----------------|----------|--------------------------------------------------------------------------------|-------------------------|
| memSizePerCore | No       | The memory size of each core.<br>Valid values: 1024 to 20 × 1024.<br>Unit: MB. | Automatically allocated |

## Example

1. Execute the following SQL statement to generate training data:

```
drop table if exists lm_test_input;
create table lm_test_input as
select
  *
from
(
  select 10 as y, 1.84 as x1, 1 as x2, '0:1.84 1:1' as sparsecol1 from dual
  union all
  select 20 as y, 2.13 as x1, 0 as x2, '0:2.13' as sparsecol1 from dual
  union all
  select 30 as y, 3.89 as x1, 0 as x2, '0:3.89' as sparsecol1 from dual
  union all
  select 40 as y, 4.19 as x1, 0 as x2, '0:4.19' as sparsecol1 from dual
  union all
  select 50 as y, 5.76 as x1, 0 as x2, '0:5.76' as sparsecol1 from dual
  union all
  select 60 as y, 6.68 as x1, 2 as x2, '0:6.68 1:2' as sparsecol1 from dual
  union all
  select 70 as y, 7.58 as x1, 0 as x2, '0:7.58' as sparsecol1 from dual
  union all
  select 80 as y, 8.01 as x1, 0 as x2, '0:8.01' as sparsecol1 from dual
  union all
  select 90 as y, 9.02 as x1, 3 as x2, '0:9.02 1:3' as sparsecol1 from dual
  union all
  select 100 as y, 10.56 as x1, 0 as x2, '0:10.56' as sparsecol1 from dual
) tmp;
```

2. Run the following PAI command to submit the parameters configured for the Linear Regression component:

```
PAI -name linearregression
  -project algo_public
  -DinputTableName=lm_test_input
  -DlabelColName=y
  -DfeatureColNames=x1,x2
  -DmodelName=lm_test_input_model_out
  -DoutputTableName=lm_test_input_conf_out
  -DenableCoefficientEstimate=true
  -DenableFitGoodness=true
  -Dlifecycle=1;
```

3. Run the following PAI command to submit the parameters configured for the Prediction component:

```

pai -name prediction
  -project algo_public
  -DmodelName=lm_test_input_model_out
  -DinputTableName=lm_test_input
  -DoutputTableName=lm_test_input_predict_out
  -DappendColNames=y;

```

4. View the model evaluation table `lm_test_input_conf_out` that is generated and indicated by the following code:

```

+-----+-----+-----+-----+-----+
| colname | value | tscore | pvalue | confidenceinterval | p |
+-----+-----+-----+-----+-----+
| Intercept | -6.42378496687763 | -2.2725755951390028 | 0.06 | {"2.5%": -11.964027, "97.5%": -0.883543} |
| coefficient |
| x1 | 10.260063429838898 | 23.270944360826963 | 0.0 | {"2.5%": 9.395908, "97.5%": 11.124219} | coefficient |
| x2 | 0.35374498323846265 | 0.2949247320997519 | 0.81 | {"2.5%": -1.997160, "97.5%": 2.704650} | coefficient |
| rsquared | 0.9879675667384592 | NULL | NULL | NULL | goodness |
| adjusted_rsquared | 0.9845297286637332 | NULL | NULL | NULL | goodness |
| aic | 59.331109494251805 | NULL | NULL | NULL | goodness |
| degree_of_freedom | 7.0 | NULL | NULL | NULL | goodness |
| standardErr_residual | 3.76577749448906 | NULL | NULL | NULL | goodness |
| deviance | 99.26757440771128 | NULL | NULL | NULL | goodness |
+-----+-----+-----+-----+-----+

```

5. View the prediction result table `lm_test_input_predict_out` indicated by the following code:

```

+-----+-----+-----+-----+
| y | prediction_result | prediction_score | prediction_detail |
+-----+-----+-----+-----+
| 10 | NULL | 12.808476727264404 | {"y": 12.8084767272644} |
| 20 | NULL | 15.43015013867922 | {"y": 15.43015013867922} |
| 30 | NULL | 33.48786177519568 | {"y": 33.48786177519568} |
| 40 | NULL | 36.565880804147355 | {"y": 36.56588080414735} |
| 50 | NULL | 52.674180388994415 | {"y": 52.67418038899442} |
| 60 | NULL | 62.82092871092313 | {"y": 62.82092871092313} |
| 70 | NULL | 71.34749583130122 | {"y": 71.34749583130122} |
| 80 | NULL | 75.75932310613193 | {"y": 75.75932310613193} |
| 90 | NULL | 87.1832221199846 | {"y": 87.18322211998461} |
| 100 | NULL | 101.92248485222113 | {"y": 101.9224848522211} |
+-----+-----+-----+-----+

```

## PS-SMART Regression

The parameter server (PS) is developed to serve large-scale online and offline training tasks. Scalable Multiple Additive Regression Tree (SMART) is an iterative algorithm achieved based on PS. PS-SMART supports training tasks for tens of billions of samples and hundreds of thousands of features. It can run the training tasks on thousands of nodes. PS-SMART also supports multiple data types and technologies such as the optimization of histogram-based approximation.

You can configure the component by using one of the following methods:

- PAI console

| Tab            | Parameter                  | Description                                                                                                                                                                                                                                                                                                                                                                                                    |
|----------------|----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Use Sparse Format          | If the input data is sparse data in the key-value format, separate key-value pairs with spaces, and separate keys and values with colons (:). Example: 1:0.3 3:0.9.                                                                                                                                                                                                                                            |
|                | Feature Columns            | The feature columns selected from the input table for training. If data in the input table is in the dense format, only columns of the BIGINT or DOUBLE data type are supported. If data in the input table is key-value pairs in the sparse format, and keys and values are of numeric data types, only columns of the STRING data type are supported.                                                        |
|                | Label Column               | The label column in the input table. The columns of the STRING data type or numeric data types are supported. However, only data of numeric data types can be stored in the columns. For example, in binary classification, column values can be 0 or 1.                                                                                                                                                       |
|                | Weight Column              | The column that contains the weight of each row of samples. The columns of numeric data types are supported.                                                                                                                                                                                                                                                                                                   |
|                | Objective Function Type    | The type of the objective function. Valid values: <b>Linear Regression</b> , <b>Logistic Regression</b> , <b>Poisson Regression</b> , <b>Gamma Regression</b> , and <b>Tweedie Regression</b> .                                                                                                                                                                                                                |
|                | Tweedie Distribution Index | The exponent for the relationship between the variance and average of Tweedie distribution.                                                                                                                                                                                                                                                                                                                    |
|                | Evaluation Indicator Type  | The type of the evaluation metric. Valid values: <b>Rooted Mean Square Error</b> , <b>Mean Absolute Error</b> , <b>Negative Loglikelihood for Logistic Regression</b> , <b>Negative Loglikelihood for Poisson Regression</b> , <b>Residual Deviance for Gamma Regression</b> , <b>Negative Log-likelihood for Gamma Regression</b> , <b>Negative Log-likelihood for Tweedie Regression</b> , and <b>None</b> . |
|                | Trees                      | The number of trees. The volume of training data is in proportion to the value of <b>Trees</b> .                                                                                                                                                                                                                                                                                                               |
|                | Maximum Tree Depth         | The maximum depth of each tree. The default value is 5. This indicates that each tree can have a maximum of 32 leaf nodes.                                                                                                                                                                                                                                                                                     |
|                | Data Sampling Fraction     | The data sampling rate when trees are built. The sample data is used to build a weak learner to accelerate training.                                                                                                                                                                                                                                                                                           |

| Tab                | Parameter                                 | Description                                                                                                                                                                                                                                                                                    |
|--------------------|-------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | <b>Feature Sampling Fraction</b>          | The feature sampling rate when trees are built. The sample features are used to build a weak learner to accelerate training.                                                                                                                                                                   |
|                    | <b>L1 Penalty Coefficient</b>             | This parameter is used to control the size of each leaf node. The larger the value of this parameter is, the more evenly the leaf nodes are distributed in size. If overfitting happens, increase the value of this parameter.                                                                 |
|                    | <b>L2 Penalty Coefficient</b>             | This parameter is used to control the size of each leaf node. The larger the value of this parameter is, the more evenly the leaf nodes are distributed in size. If overfitting happens, increase the value of this parameter.                                                                 |
|                    | <b>Learning Rate</b>                      | The learning rate. Valid values: (0,1).                                                                                                                                                                                                                                                        |
|                    | <b>Sketch-based Approximate Precision</b> | The threshold for selecting quantiles when you build a sketch. The smaller the value of this parameter is, the more bins can be obtained. In most cases, the default value 0.03 is used.                                                                                                       |
|                    | <b>Minimum Split Loss Change</b>          | The minimum loss change required for splitting a node. The larger the value of this parameter is, the less likely the node is to split.                                                                                                                                                        |
|                    | <b>Features</b>                           | The number of features or the maximum feature ID. If this parameter is not specified for the assessment of resource usage, the system automatically runs an SQL task to calculate the number of features or the maximum feature ID.                                                            |
|                    | <b>Global Offset</b>                      | The initial prediction values of all samples.                                                                                                                                                                                                                                                  |
|                    | <b>Feature Importance Type</b>            | The feature importance type. Valid values: <b>Weight</b> , <b>Gain</b> , and <b>Cover</b> . Weight indicates the number of splits of the feature. Gain indicates the information gain provided by the feature. Cover indicates the number of samples covered by the feature on the split node. |
| Tuning             | <b>Cores</b>                              | The number of cores. It is automatically allocated by the system.                                                                                                                                                                                                                              |
|                    | <b>Memory Size per Core</b>               | The memory size of each core. Unit: MB. In most cases, the system automatically allocates the memory size.                                                                                                                                                                                     |

- PAI command

```
#Training
PAI -name ps_smart
  -project algo_public
  -DinputTableName="smart_regression_input"
  -DmodelName="xlab_m_pai_ps_smart_bi_545859_v0"
  -DoutputTableName="pai_temp_24515_545859_2"
  -DoutputImportanceTableName="pai_temp_24515_545859_3"
  -DlabelColName="label"
  -DfeatureColNames="features"
  -DenableSparse="true"
  -Dobjective="reg:linear"
  -Dmetric="rmse"
  -DfeatureImportanceType="gain"
  -DtreeCount="5";
  -DmaxDepth="5"
  -Dshrinkage="0.3"
  -Dl2="1.0"
  -Dl1="0"
  -Dlifecycle="3"
  -DsketchEps="0.03"
  -DsampleRatio="1.0"
  -DfeatureRatio="1.0"
  -DbaseScore="0.5"
  -DminSplitLoss="0"

#Prediction
PAI -name prediction
  -project algo_public
  -DinputTableName="smart_regression_input";
  -DmodelName="xlab_m_pai_ps_smart_bi_545859_v0"
  -DoutputTableName="pai_temp_24515_545860_1"
  -DfeatureColNames="features"
  -DappendColNames="label,features"
  -DenableSparse="true"
  -Dlifecycle="28"
```

| Module | Parameter       | Required | Description                                                                                                                                                                                                                                                                                                                                   | Default value    |
|--------|-----------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
|        | featureColNames | Yes      | The feature columns selected from the input table for training. If data in the input table is dense data, only columns of the BIGINT or DOUBLE data type are supported. If data in the input table is sparse data in the key-value format, and keys and values are of numeric data types, only columns of the STRING data type are supported. | No default value |
|        | labelColName    | Yes      | The label column in the input table. The columns of the STRING data type or numeric data types are supported. However, only data of numeric data types can be stored in the columns. For example, in binary classification, column values can be 0 or 1.                                                                                      | No default value |

| Module          | Parameter                 | Required | Description                                                                                                                                                                                                                                                                                                  | Default value    |
|-----------------|---------------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| Data parameters | weightCol                 | No       | The column that contains the weight of each row of samples. The columns of numeric data types are supported.                                                                                                                                                                                                 | No default value |
|                 | enableSparse              | No       | Specifies whether the input data is in the sparse format. Valid values: true and false. If the input data is sparse data in the key-value format, separate key-value pairs with spaces, and separate keys and values with colons (:). Example: 1:0.3 3:0.9.                                                  | false            |
|                 | inputTableName            | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                 | No default value |
|                 | modelName                 | Yes      | The name of the output model.                                                                                                                                                                                                                                                                                | No default value |
|                 | outputImportanceTableName | No       | The name of the feature importance table that is generated.                                                                                                                                                                                                                                                  | No default value |
|                 | inputTablePartitions      | No       | The partition selected from the input table for training. Specify this parameter in the ds=1/pt=1 format.                                                                                                                                                                                                    | No default value |
|                 | outputTableName           | No       | The model is generated in a MaxCompute table. The table is a binary file. It cannot be read and can only be obtained by using the prediction component of SMART.                                                                                                                                             | No default value |
|                 | lifecycle                 | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                           | 3                |
|                 | objective                 | Yes      | <p>The type of the objective function. Valid values:</p> <ul style="list-style-type: none"> <li>reg:linear: Linear Regression</li> <li>reg:logistic: Logistic Regression</li> <li>count:poisson: Poisson Regression</li> <li>reg:gamma: Gamma Regression</li> <li>reg:tweedie: Tweedie Regression</li> </ul> | reg:linear       |
|                 |                           |          |                                                                                                                                                                                                                                                                                                              |                  |

| Module               | Parameter    | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Default value    |
|----------------------|--------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| Algorithm parameters | metric       | No       | <p>The evaluation metric type in the training set, which is contained in stdout of the coordinator in a logview. Valid values:</p> <ul style="list-style-type: none"> <li>◦ rmse: corresponds to the <b>Rooted Mean Square Error</b> value specified by the Evaluation Indicator Type parameter in the PAI console.</li> <li>◦ mae: corresponds to the <b>Mean Absolute Error</b> value specified by the Evaluation Indicator Type parameter in the PAI console.</li> <li>◦ logistic-nloglik: corresponds to the <b>Negative Loglikelihood for Logistic Regression</b> value specified by the Evaluation Indicator Type parameter in the PAI console.</li> <li>◦ poisson-nloglik: corresponds to the <b>Negative Loglikelihood for Poisson Regression</b> value specified by the Evaluation Indicator Type parameter in the PAI console.</li> <li>◦ gamma-deviance: corresponds to the <b>Residual Deviance for Gamma Regression</b> value specified by the Evaluation Indicator Type parameter in the PAI console.</li> <li>◦ gamma-nloglik: corresponds to the <b>Negative Log-likelihood for Gamma Regression</b> value specified by the Evaluation Indicator Type parameter in the PAI console.</li> <li>◦ tweedie-nloglik: corresponds to the <b>Negative Log-likelihood for Tweedie Regression</b> value specified by the Evaluation Indicator Type parameter in the PAI console.</li> </ul> | No default value |
|                      | treeCount    | No       | The number of trees. It is in proportion to the training time.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 1                |
|                      | maxDepth     | No       | The maximum depth of each tree. Valid values: 1 to 20.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 5                |
|                      | sampleRatio  | No       | The sampling ratio of data. Valid values: (0,1]. If this parameter is set to 1.0, no data is sampled.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | 1.0              |
|                      | featureRatio | No       | The sampling ratio of features. Valid values: (0,1]. If this parameter is set to 1.0, no features are sampled.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 1.0              |

| Module | Parameter             | Required | Description                                                                                                                                                                                                                                                                                                          | Default value           |
|--------|-----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
|        | l1                    | No       | L1 penalty coefficient. A large value indicates that the numbers of leaf nodes on all branches are almost the same. If overfitting happens, increase the value of this parameter.                                                                                                                                    | 0                       |
|        | l2                    | No       | L2 penalty coefficient. A large value indicates that the numbers of leaf nodes on all branches are almost the same. If overfitting happens, increase the value of this parameter.                                                                                                                                    | 1.0                     |
|        | shrinkage             | No       | The learning rate. Valid values: (0,1).                                                                                                                                                                                                                                                                              | 0.3                     |
|        | sketchEps             | No       | The threshold for selecting quantiles when you build a sketch. The number of buckets is $O(1.0/sketchEps)$ . The smaller the value of this parameter is, the more bins can be obtained. In most cases, the default value is used. Valid values: (0,1).                                                               | 0.03                    |
|        | minSplitLoss          | No       | The minimum loss change required for splitting a node. The larger the value of this parameter is, the less likely the node is to split.                                                                                                                                                                              | 0                       |
|        | featureNum            | No       | The number of features or the maximum feature ID. If this parameter is not specified for the assessment of resource usage, the system automatically runs an SQL task to calculate the number of features or the maximum feature ID.                                                                                  | No default value        |
|        | baseScore             | No       | The initial prediction values of all samples.                                                                                                                                                                                                                                                                        | 0.5                     |
|        | featureImportanceType | No       | The feature importance type. Valid values: <ul style="list-style-type: none"> <li>Weight: indicates the number of splits of the feature.</li> <li>Gain: indicates the information gain provided by the feature.</li> <li>Cover: indicates the number of samples covered by the feature on the split node.</li> </ul> | gain                    |
|        | tweedieVarPower       | No       | The exponent for the relationship between the variance and average of Tweedie distribution.                                                                                                                                                                                                                          | 1.5                     |
|        | coreNum               | No       | The number of cores. The larger the value of this parameter is, the more quickly the algorithm runs.                                                                                                                                                                                                                 | Automatically allocated |



| Tuning Module | Parameter      | Required | Description                             | Default value           |
|---------------|----------------|----------|-----------------------------------------|-------------------------|
|               | memSizePerCore | No       | The memory size of each core. Unit: MB. | Automatically allocated |

## Example

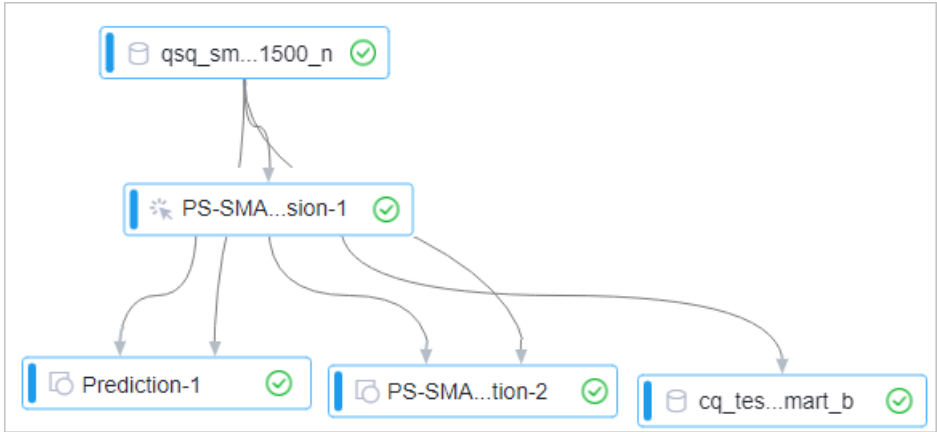
1. Execute the following SQL statement to generate input data. In this example, the input data in the key-value format is generated.

```
drop table if exists smart_regression_input;
create table smart_regression_input as
select
*
from
(
select 2.0 as label, '1:0.55 2:-0.15 3:0.82 4:-0.99 5:0.17' as features from dual
union all
select 1.0 as label, '1:-1.26 2:1.36 3:-0.13 4:-2.82 5:-0.41' as features from dual
union all
select 1.0 as label, '1:-0.77 2:0.91 3:-0.23 4:-4.46 5:0.91' as features from dual
union all
select 2.0 as label, '1:0.86 2:-0.22 3:-0.46 4:0.08 5:-0.60' as features from dual
union all
select 1.0 as label, '1:-0.76 2:0.89 3:1.02 4:-0.78 5:-0.86' as features from dual
union all
select 1.0 as label, '1:2.22 2:-0.46 3:0.49 4:0.31 5:-1.84' as features from dual
union all
select 0.0 as label, '1:-1.21 2:0.09 3:0.23 4:2.04 5:0.30' as features from dual
union all
select 1.0 as label, '1:2.17 2:-0.45 3:-1.22 4:-0.48 5:-1.41' as features from dual
union all
select 0.0 as label, '1:-0.40 2:0.63 3:0.56 4:0.74 5:-1.44' as features from dual
union all
select 1.0 as label, '1:0.17 2:0.49 3:-1.50 4:-2.20 5:-0.35' as features from dual
) tmp;
```

The data generated is shown in the following figure.

|    | label ▲ | features ▲                             |
|----|---------|----------------------------------------|
| 1  | 2.0     | 1:0.55 2:-0.15 3:0.82 4:-0.99 5:0.17   |
| 2  | 1.0     | 1:-1.26 2:1.36 3:-0.13 4:-2.82 5:-0.41 |
| 3  | 1.0     | 1:-0.77 2:0.91 3:-0.23 4:-4.46 5:0.91  |
| 4  | 2.0     | 1:0.86 2:-0.22 3:-0.46 4:0.08 5:-0.60  |
| 5  | 1.0     | 1:-0.76 2:0.89 3:1.02 4:-0.78 5:-0.86  |
| 6  | 1.0     | 1:2.22 2:-0.46 3:0.49 4:0.31 5:-1.84   |
| 7  | 0.0     | 1:-1.21 2:0.09 3:0.23 4:2.04 5:0.30    |
| 8  | 1.0     | 1:2.17 2:-0.45 3:-1.22 4:-0.48 5:-1.41 |
| 9  | 0.0     | 1:-0.40 2:0.63 3:0.56 4:0.74 5:-1.44   |
| 10 | 1.0     | 1:0.17 2:0.49 3:-1.50 4:-2.20 5:-0.35  |

2. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).



3. Configure the parameters listed in the following table for the PS-SMART Regression component. Retain the default values of the parameters that are not listed in the table.

| Tab                | Parameter                 | Description                                |
|--------------------|---------------------------|--------------------------------------------|
| Fields Setting     | Use Sparse Format         | Select the Use Sparse Format check box.    |
|                    | Feature Columns           | Select feature columns.                    |
|                    | Label Column              | Select the label column.                   |
| Parameters Setting | Objective Function Type   | Set the value to Linear Regression.        |
|                    | Evaluation Indicator Type | Set the value to Rooted Mean Square Error. |
|                    | Trees                     | Set the value to 5.                        |

4. Configure the parameters listed in the following table for the unified prediction component. Retain the default values of the parameters that are not listed in the table.

| Tab            | Parameter         | Description                                                                                                                                            |
|----------------|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns   | By default, all columns in the input table are selected. Some columns may not be used for training. These columns do not affect the prediction result. |
|                | Reserved Columns  | Select the label column.                                                                                                                               |
|                | Sparse Matrix     | Select the <b>Sparse Matrix</b> check box.                                                                                                             |
|                | KV Delimiter      | Use colons (:).                                                                                                                                        |
|                | KV Pair Delimiter | Use spaces or \u0020.                                                                                                                                  |

5. Configure the parameters listed in the following table for the PS-SMART prediction component. Retain the default values of the parameters that are not listed in the table.

| Tab            | Parameter         | Description                                                                                                                                            |
|----------------|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns   | By default, all columns in the input table are selected. Some columns may not be used for training. These columns do not affect the prediction result. |
|                | Reserved Columns  | Select the label column.                                                                                                                               |
|                | Use Sparse Format | Select the <b>Use Sparse Format</b> check box.                                                                                                         |
|                | KV Delimiter      | Use colons (:).                                                                                                                                        |
|                | KV Pair Delimiter | Use spaces or \u0020.                                                                                                                                  |

6. Run the experiment and view the prediction result of the Prediction component.

|    | label ▲ | features ▲      | prediction_result ▲ | prediction_score ▲  | prediction_detail ▲           |
|----|---------|-----------------|---------------------|---------------------|-------------------------------|
| 1  | 2.0     | 1:0.55 2:-0.... | 1.467519998550415   | 1.467519998550415   | {"label": 1.467519998550415}  |
| 2  | 1.0     | 1:-1.26 2:1.... | 0.8999134302139282  | 0.8999134302139282  | {"label": 0.8999134302139282} |
| 3  | 1.0     | 1:-0.77 2:0.... | 0.8999134302139282  | 0.8999134302139282  | {"label": 0.8999134302139282} |
| 4  | 2.0     | 1:0.86 2:-0.... | 1.467519998550415   | 1.467519998550415   | {"label": 1.467519998550415}  |
| 5  | 1.0     | 1:-0.76 2:0.... | 0.8999134302139282  | 0.8999134302139282  | {"label": 0.8999134302139282} |
| 6  | 1.0     | 1:2.22 2:-0.... | 0.8771200180053711  | 0.8771200180053711  | {"label": 0.8771200180053711} |
| 7  | 0.0     | 1:-1.21 2:0.... | 0.16383999586105347 | 0.16383999586105347 | {"label": 0.1638399958610535} |
| 8  | 1.0     | 1:2.17 2:-0.... | 0.8771200180053711  | 0.8771200180053711  | {"label": 0.8771200180053711} |
| 9  | 0.0     | 1:-0.40 2:0.... | 0.16383999586105347 | 0.16383999586105347 | {"label": 0.1638399958610535} |
| 10 | 1.0     | 1:0.17 2:0....  | 0.8999134302139282  | 0.8999134302139282  | {"label": 0.8999134302139282} |

7. View the prediction result of the PS-SMART component.

|    | original_label ▲ | prediction_score ▲  | leaf_index ▲ |
|----|------------------|---------------------|--------------|
| 1  | 2                | 1.467519998550415   | 15555        |
| 2  | 1                | 0.8999134302139282  | 13333        |
| 3  | 1                | 0.8999134302139282  | 13333        |
| 4  | 2                | 1.467519998550415   | 15555        |
| 5  | 1                | 0.8999134302139282  | 13333        |
| 6  | 1                | 0.8771200180053711  | 16666        |
| 7  | 0                | 0.16383999586105347 | 22222        |
| 8  | 1                | 0.8771200180053711  | 16666        |
| 9  | 0                | 0.16383999586105347 | 22222        |
| 10 | 1                | 0.8999134302139282  | 13333        |

The `prediction_score` column lists prediction values, and the `leaf_index` column lists the numbers of leaf nodes of the prediction.

8. Right-click **PS-SMART Regression** and choose **View Data > View Output Port 3**. Then, view the feature importance.

|   | id ▲ | value ▲             |
|---|------|---------------------|
| 1 | 1    | 0.14059734344482422 |
| 2 | 4    | 0.8594027161598206  |

The `id` column lists the numbers of input features. The input data in this example is key-value pairs. Therefore, the `id` column lists the keys in key-value pairs. The feature importance table has only two features. This indicates that only these two features are used in the split of the tree. The feature importance of other features can be considered as 0. The `value` column is the feature importance type. The default value of this parameter is Gain, which indicates the sum of information gains provided by the feature for the model.

Additional information:

- Only columns of numeric data types can be used by the PS-SMART Regression component. If the type of data in the MaxCompute table is STRING, a data type conversion must be performed for the data.
- If data is in the key-value format, feature IDs must be positive integers, and feature values must be real numbers. If the data type of feature IDs is STRING, you must use the serialization component to serialize the data. If feature values are categorical strings, you must perform feature engineering such as feature discretization to process the values.
- The PS-SMART Regression component can support hundreds of thousands of feature tasks. However, these tasks are resource-intensive and time-consuming. To address this issue, you can use GBDT algorithms to train the model. GBDT algorithms are suitable for the scenario where continuous features are used for training. You can perform one-hot encoding on categorical features to filter low-frequency features. However, we recommend that you do not perform feature discretization on continuous features of numeric data types.
- The PS-SMART algorithm may introduce randomness. For example, randomness may be introduced in the following scenarios: data and feature sampling based on `data_sample_ratio` and `fea_sample_ratio`, optimization of the PS-SMART algorithm by using histograms for approximation, and merge of a local sketch into a global sketch. The structures of trees are different when tasks run on multiple workers in distributed mode. However, the training effect of the model is theoretically the same. It is normal if you use the same data and parameters during training but obtain different results.
- If you want to accelerate training, you can set **Cores** to a larger value. The PS-SMART algorithm

starts training tasks after the required resources are provided. Therefore, the more resources that are requested, the longer the waiting time.

## PS Linear Regression

Linear regression is a model that is used to analyze the linear relationship between a dependent variable and multiple independent variables. The parameter server (PS) is developed to serve large-scale online and offline training tasks. PS Linear Regression can support large-scale linear training tasks for hundreds of billions of samples and billions of features.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter                    | Description                                                                                                                                                                                                          |
|--------------------|------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Feature Columns              | The feature columns selected from the input table for training.                                                                                                                                                      |
|                    | Label Column                 | The label column. The columns of the DOUBLE and BIGINT data types are supported.                                                                                                                                     |
|                    | Use Sparse Format            | Specifies whether the input data is in the sparse format. Data in the sparse format is presented by using key-value pairs.                                                                                           |
|                    | KV Pair Delimiter            | The delimiter used to separate key-value pairs. Spaces are used by default.                                                                                                                                          |
|                    | KV Delimiter                 | The delimiter used to separate keys and values. Colons (:) are used by default.                                                                                                                                      |
| Parameters Setting | L1 weight                    | L1 regularization coefficient. The larger the value of this parameter is, the few non-zero elements the model has. If overfitting happens, increase the value of this parameter.                                     |
|                    | L2 weight                    | L2 regularization coefficient. The larger the value of this parameter is, the smaller the absolute values of the model parameters are. If overfitting happens, increase the value of this parameter.                 |
|                    | Maximum Iterations           | The maximum number of iterations performed by the algorithm. If this parameter is set to 0, the number of iterations is unlimited.                                                                                   |
|                    | Minimum Convergence Deviance | The conditions for algorithm termination.                                                                                                                                                                            |
|                    | Largest Feature ID           | The largest feature ID or feature dimension. The value of this parameter can be greater than the actual value. The system automatically runs an SQL task to calculate the number if this parameter is not specified. |

| Tab    | Parameter            | Description                                                                |
|--------|----------------------|----------------------------------------------------------------------------|
| Tuning | Cores                | The number of cores. It is automatically allocated by the system.          |
|        | Memory Size per Core | The memory size of each core. It is automatically allocated by the system. |

- PAI commands

```
#Training
PAI -name ps_linearregression
  -project algo_public
  -DinputTableName="lm_test_input"
  -DmodelName="linear_regression_model"
  -DlabelColName="label"
  -DfeatureColNames="features"
  -DL1Weight=1.0
  -DL2Weight=0.0
  -DmaxIter=100
  -Depsilon=1e-6
  -DenableSparse=true
#Prediction
drop table if exists logistic_regression_predict;
PAI -name prediction
  -DmodelName="linear_regression_model"
  -DoutputTableName="linear_regression_predict"
  -DinputTableName="lm_test_input"
  -DappendColNames="label,features"
  -DfeatureColNames="features"
  -DenableSparse=true
```

| Parameter       | Required | Description                                                                                                                 | Default value    |
|-----------------|----------|-----------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName  | Yes      | The name of the input table.                                                                                                | No default value |
| modelName       | Yes      | The name of the output model.                                                                                               | No default value |
| outputTableName | No       | The name of the model evaluation table that is generated. This parameter is required when enableFitGoodness is set to true. | No default value |
| labelColName    | Yes      | The label column. The columns of the DOUBLE and BIGINT data types are supported.                                            | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                | Default value    |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| featureColNames      | Yes      | The feature columns selected from the input table for training. If data in the input table is in the dense format, columns of the DOUBLE or BIGINT data type are supported. If the input data is in the sparse format, only columns of the STRING data type are supported. | No default value |
| inputTablePartitions | No       | The partitions selected from the input table for training.                                                                                                                                                                                                                 | No default value |
| enableSparse         | No       | Specifies whether the input data is in the sparse format. Valid values: true and false.                                                                                                                                                                                    | false            |
| itemDelimiter        | No       | The delimiter used to separate key-value pairs. The parameter is valid when enableSparse is set to true.                                                                                                                                                                   | Space            |
| kvDelimiter          | No       | The delimiter used to separate keys and values. This parameter is valid when enableSparse is set to true.                                                                                                                                                                  | :                |
| enableModello        | No       | Specifies whether the model is generated as an offline model. If enableModello is set to false, the model is generated in a MaxCompute table. Valid values: true and false.                                                                                                | true             |
| maxIter              | No       | The maximum number of iterations performed by the algorithm. The value of this parameter must be a non-negative integer.                                                                                                                                                   | 100              |
| epsilon              | No       | The conditions for algorithm termination. Valid values: [0,1].                                                                                                                                                                                                             | 0.000001         |
| l1Weight             | No       | L1 regularization coefficient. The larger the value of this parameter is, the few non-zero elements the model has. If overfitting happens, increase the value of this parameter.                                                                                           | 1.0              |

| Parameter      | Required | Description                                                                                                                                                                                                                                                                      | Default value           |
|----------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| l2Weight       | No       | L2 regularization coefficient. The larger the value of this parameter is, the smaller the absolute values of the model parameters are. If overfitting happens, increase the value of this parameter.                                                                             | 0                       |
| modelSize      | No       | The largest feature ID or feature dimension. The value of this parameter can be greater than the actual value. The system automatically runs an SQL task to calculate the number if this parameter is not specified. The value of this parameter must be a non-negative integer. | 0                       |
| coreNum        | No       | The number of cores used in computing.                                                                                                                                                                                                                                           | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Unit: MB.                                                                                                                                                                                                                                          | Automatically allocated |

## Example

1. Execute the following SQL statement to generate input data. In this example, the input data in the key-value format is generated.



```

drop table if exists lm_test_input;
create table lm_test_input as
select
*
from
(
select 2 as label, '1:0.55 2:-0.15 3:0.82 4:-0.99 5:0.17' as features from dual
union all
select 1 as label, '1:-1.26 2:1.36 3:-0.13 4:-2.82 5:-0.41' as features from dual
union all
select 1 as label, '1:-0.77 2:0.91 3:-0.23 4:-4.46 5:0.91' as features from dual
union all
select 2 as label, '1:0.86 2:-0.22 3:-0.46 4:0.08 5:-0.60' as features from dual
union all
select 1 as label, '1:-0.76 2:0.89 3:1.02 4:-0.78 5:-0.86' as features from dual
union all
select 1 as label, '1:2.22 2:-0.46 3:0.49 4:0.31 5:-1.84' as features from dual
union all
select 0 as label, '1:-1.21 2:0.09 3:0.23 4:2.04 5:0.30' as features from dual
union all
select 1 as label, '1:2.17 2:-0.45 3:-1.22 4:-0.48 5:-1.41' as features from dual
union all
select 0 as label, '1:-0.40 2:0.63 3:0.56 4:0.74 5:-1.44' as features from dual
union all
select 1 as label, '1:0.17 2:0.49 3:-1.50 4:-2.20 5:-0.35' as features from dual
) tmp;

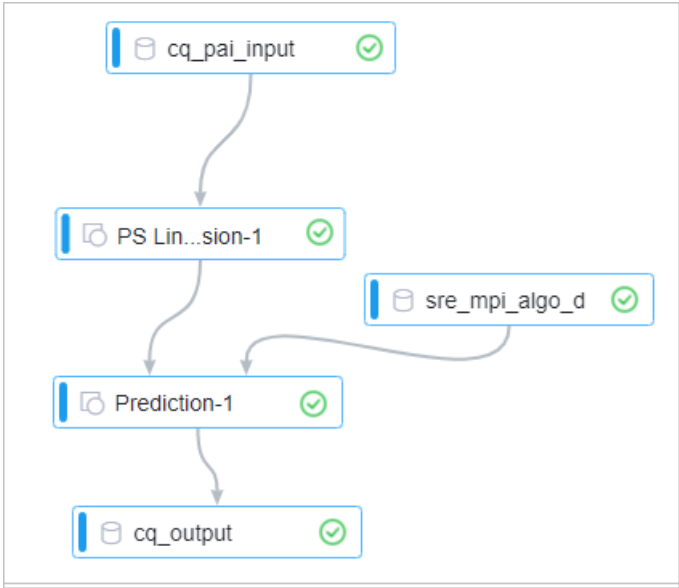
```

The data generated is shown in the following figure.

|    | label ▲ | features ▲                             |
|----|---------|----------------------------------------|
| 1  | 2       | 1:0.55 2:-0.15 3:0.82 4:-0.99 5:0.17   |
| 2  | 1       | 1:-1.26 2:1.36 3:-0.13 4:-2.82 5:-0.41 |
| 3  | 1       | 1:-0.77 2:0.91 3:-0.23 4:-4.46 5:0.91  |
| 4  | 2       | 1:0.86 2:-0.22 3:-0.46 4:0.08 5:-0.60  |
| 5  | 1       | 1:-0.76 2:0.89 3:1.02 4:-0.78 5:-0.86  |
| 6  | 1       | 1:2.22 2:-0.46 3:0.49 4:0.31 5:-1.84   |
| 7  | 0       | 1:-1.21 2:0.09 3:0.23 4:2.04 5:0.30    |
| 8  | 1       | 1:2.17 2:-0.45 3:-1.22 4:-0.48 5:-1.41 |
| 9  | 0       | 1:-0.40 2:0.63 3:0.56 4:0.74 5:-1.44   |
| 10 | 1       | 1:0.17 2:0.49 3:-1.50 4:-2.20 5:-0.35  |

 **Note** If data is in the key-value format, feature IDs must be positive integers, and feature values must be real numbers. If the data type of feature IDs is STRING, you must use the serialization component to serialize the data. If feature values are categorical strings, you must perform feature discretization to process the features.

2. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).



3. Configure the parameters listed in the following table for the PS Linear Regression component. Retain the default values of the parameters that are not listed in the table.

| Tab            | Parameter            | Description                      |
|----------------|----------------------|----------------------------------|
| Fields Setting | Use Sparse Format    | Set the value to yes.            |
|                | Feature Columns      | Select feature columns.          |
|                | Label Column         | Select the label column.         |
| Tuning         | Cores                | Set the value to 3.              |
|                | Memory Size per Core | Set the value to 1024. Unit: MB. |

4. Configure the parameters listed in the following table for the Prediction component. Retain the default values of the parameters that are not listed in the table.

| Tab            | Parameter         | Description                                                                                                                                            |
|----------------|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns   | By default, all columns in the input table are selected. Some columns may not be used for training. These columns do not affect the prediction result. |
|                | Reserved Columns  | Select the label column.                                                                                                                               |
|                | Sparse Matrix     | Select the <b>Sparse Matrix</b> check box.                                                                                                             |
|                | KV Delimiter      | Use colons (:).                                                                                                                                        |
|                | KV Pair Delimiter | Use \u0020.                                                                                                                                            |

5. Run the experiment to view the prediction result.

|    | label ▲ | features ▲                             | prediction_result ▲ | prediction_score ▲  | prediction_detail ▲       |
|----|---------|----------------------------------------|---------------------|---------------------|---------------------------|
| 1  | 2       | 1:0.55 2:-0.15 3:0.82 4:-0.99 5:0.17   | 0.9950045235424285  | 0.9950045235424285  | ("1": 0.9950045235424285) |
| 2  | 1       | 1:-1.28 2:1.38 3:-0.13 4:-2.82 5:-0.41 | 0.9022041049173464  | 0.9022041049173464  | ("1": 0.9022041049173464) |
| 3  | 1       | 1:-0.77 2:0.91 3:-0.23 4:-4.46 5:0.91  | 1.228147671448144   | 1.228147671448144   | ("1": 1.228147671448144)  |
| 4  | 2       | 1:0.86 2:-0.22 3:-0.46 4:0.08 5:-0.80  | 0.9043180917054381  | 0.9043180917054381  | ("1": 0.9043180917054381) |
| 5  | 1       | 1:-0.76 2:0.89 3:1.02 4:-0.78 5:-0.86  | 0.7116744886195954  | 0.7116744886195954  | ("1": 0.7116744886195954) |
| 6  | 1       | 1:2.22 2:-0.46 3:0.49 4:0.31 5:-1.84   | 1.135349294653747   | 1.135349294653747   | ("1": 1.135349294653747)  |
| 7  | 0       | 1:-1.21 2:0.09 3:0.23 4:2.04 5:0.30    | 0.22724956563861654 | 0.22724956563861654 | ("1": 0.2272495656386165) |
| 8  | 1       | 1:2.17 2:-0.45 3:-1.22 4:-0.48 5:-1.41 | 1.2369534428175184  | 1.2369534428175184  | ("1": 1.236953442817518)  |
| 9  | 0       | 1:-0.40 2:0.63 3:0.56 4:0.74 5:-1.44   | 0.5672805014883875  | 0.5672805014883875  | ("1": 0.5672805014883875) |
| 10 | 1       | 1:0.17 2:0.49 3:-1.50 4:-2.20 5:-0.35  | 1.0918540901263776  | 1.0918540901263776  | ("1": 1.091854090126378)  |

### 3.5.5. Evaluation

This topic describes the Binary Classification Evaluation, Regression Model Evaluation, Clustering Model Evaluation, Confusion Matrix, and Multiclass Classification Evaluation components provided by Machine Learning Studio for evaluation.

#### Binary Classification Evaluation

Binary Classification Evaluation is used to calculate AUC, KS, and F1 score metrics to generate Kolmogorov–Smirnov (KS) curves, precision-recall (P-R) curves, ROC curves, lift charts, and gain charts.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Parameter                                                                   | Description                                                                                                                                                                                                                   |
|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Original Label Column Name</b>                                           | The name of the objective column.                                                                                                                                                                                             |
| <b>Score Column</b>                                                         | The prediction score column. Default value: prediction_score.                                                                                                                                                                 |
| <b>Positive Sample Label</b>                                                | The value that indicates positive samples.                                                                                                                                                                                    |
| <b>Number of Bins with Equal Frequency for Indicators such as KS and PR</b> | The number of bins obtained by using the equal frequency binning method.                                                                                                                                                      |
| <b>Stratification Column</b>                                                | The group ID column. This parameter is used to calculate evaluation metrics for each group.                                                                                                                                   |
| <b>Advanced Options</b>                                                     | The <b>Prediction Result Detail Column</b> , <b>Prediction Targets Consistent With Evaluation Targets</b> , and <b>Save Performance Indicators</b> parameters are valid if the <b>Advanced Options</b> check box is selected. |
| <b>Prediction Result Detail Column</b>                                      | The name of the prediction result detail column.                                                                                                                                                                              |

| Parameter                                                    | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|--------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Prediction Targets Consistent with Evaluation Targets</b> | Specifies whether the prediction objective is consistent with the evaluation objective. For example, in a financial scenario, a program is trained to predict the probability that a customer is bad. The larger the probability is, the more likely the customer is bad. Related metrics such as lift evaluate the bad-customer detection rate. In this case, the prediction objective is consistent with the evaluation objective. In a credit scoring scenario, a program is trained to predict the probability that a customer is good. The larger the probability is, the more likely the customer is good. However, related metrics evaluate the bad-customer detection rate. In this case, the prediction objective is inconsistent with the evaluation objective. |
| <b>Save Performance Indicators</b>                           | Specifies whether to save performance metrics.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |

- PAI command

```
PAI -name=evaluate -project=algo_public
-DoutputMetricTableName=output_metric_table
-DoutputDetailTableName=output_detail_table
-DinputTableName=input_data_table
-DlabelColName=label
-DscoreColName=score
```

| Parameter             | Required | Description                                                                                                                  | Default value    |
|-----------------------|----------|------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName        | Yes      | The name of the input table.                                                                                                 | No default value |
| inputTablePartitions  | No       | The partitions selected from the input table for training.                                                                   | Full table       |
| labelColName          | Yes      | The name of the objective column.                                                                                            | No default value |
| scoreColName          | Yes      | The name of the score column.                                                                                                | No default value |
| groupColName          | No       | The name of the group column. This parameter is used for the evaluation of each group.                                       | No default value |
| binCount              | No       | The number of bins obtained by using the equal frequency binning method during the calculation of metrics such as KS and PR. | 1000             |
| outputMetricTableName | Yes      | The output metric table. The metrics include AUC, KS, and F1 score.                                                          | No default value |
| outputDetailTableName | No       | The detail data table that is generated.                                                                                     | No default value |
| positiveLabel         | No       | The value that indicates positive samples.                                                                                   | 1                |

| Parameter      | Required | Description                        | Default value           |
|----------------|----------|------------------------------------|-------------------------|
| lifecycle      | No       | The lifecycle of the output table. | No default value        |
| coreNum        | No       | The number of cores.               | Automatically allocated |
| memSizePerCore | No       | The memory size of each core.      | Automatically allocated |

## Regression Model Evaluation

Regression Model Evaluation is used to evaluate the advantages and disadvantages of the different models of regression algorithms based on prediction results and original results. Then, evaluation metrics and histograms of residuals are generated.

You can configure the component by using one of the following methods:

- PAI console

| Tab            | Parameter                  | Description                                                                                             |
|----------------|----------------------------|---------------------------------------------------------------------------------------------------------|
| Fields Setting | Original Regression Value  | The value of this parameter must be of a numeric data type.                                             |
|                | Predicted Regression Value | The value of this parameter must be of a numeric data type.                                             |
| Tuning         | Cores                      | The number of cores. Valid values: 1 to 9999. This parameter is used with <b>Memory Size per Node</b> . |
|                | Memory Size per Node       | The memory size of each core. Valid values: 1024 to 64 × 1024. Unit: MB.                                |

- PAI command

```
PAI -name regression_evaluation -project algo_public
-DinputTableName=input_table
-DyColName=y_col
-DpredictionColName=prediction_col
-DindexOutputTableName=index_output_table
-DresidualOutputTableName=residual_output_table;
```

| Parameter            | Required | Description                                                                                                                            | Default value    |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                           | No default value |
| inputTablePartitions | No       | The partitions selected from the input table for computing.                                                                            | Full table       |
| yColName             | Yes      | The name of the column that contains original dependent variables in the input table. The columns of numeric data types are supported. | No default value |

| Parameter               | Required | Description                                                                                                                         | Default value           |
|-------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| predictionColName       | Yes      | The name of the column that contains dependent variables in the prediction result. The columns of numeric data types are supported. | No default value        |
| indexOutputTableName    | Yes      | The name of the output table of regression metrics.                                                                                 | No default value        |
| residualOutputTableName | Yes      | The name of the output table of the histogram of residuals.                                                                         | No default value        |
| intervalNum             | No       | The number of intervals of the histogram.                                                                                           | 100                     |
| lifecycle               | No       | The lifecycle of the output table. The value of this parameter must be a positive integer.                                          | No default value        |
| coreNum                 | No       | The number of cores. Valid values: 1 to 9999.                                                                                       | Automatically allocated |
| memSizePerCore          | No       | The memory size of each core. Valid values: 1024 to 64 × 1024. Unit: MB.                                                            | Automatically allocated |

## Output of the Regression Model Evaluation component

The output table of regression metrics is generated in the JSON format and contains the following parameters.

| Parameter | Description                               |
|-----------|-------------------------------------------|
| SST       | The total sum of squares.                 |
| SSE       | The sum of squared errors.                |
| SSR       | The sum of squares due to regression.     |
| R2        | The coefficient of determination.         |
| R         | The coefficient of multiple correlations. |
| MSE       | The mean-square error.                    |
| RMSE      | The root-mean-square error.               |
| MAE       | The mean absolute error.                  |
| MAD       | The mean error.                           |
| MAPE      | The mean absolute percentage error.       |
| count     | The number of rows.                       |

| Parameter      | Description                               |
|----------------|-------------------------------------------|
| yMean          | The mean of original dependent variables. |
| predictionMean | The mean of prediction results.           |

## Clustering Model Evaluation

Clustering Model Evaluation is used to evaluate clustering models and generate evaluation metrics based on raw data and clustering results.

You can configure the component by using one of the following methods:

- PAI console

| Tab            | Parameter            | Description                                                                                                                                     |
|----------------|----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Evaluation Columns   | The columns selected from the input table for evaluation. The value of this parameter must be consistent with the feature columns in the model. |
|                | Input Sparse Format  | Specifies whether the input data is in the sparse format. Data in the sparse format is presented by using key-value pairs.                      |
|                | KV Pair Delimiter    | The delimiter used to separate key-value pairs. Commas (,) are used by default.                                                                 |
|                | KV Delimiter         | The delimiter used to separate keys and values. Colons (:) are used by default.                                                                 |
| Tuning         | Cores                | The number of cores. The value of this parameter must be a positive integer. This parameter is used with <b>Memory Size per Core</b> .          |
|                | Memory Size per Core | The memory size of each core. Unit: MB. This parameter is used with <b>Cores</b> .                                                              |

- PAI command

```
PAI -name cluster_evaluation
-project algo_public
-DinputTableName=pai_cluster_evaluation_test_input
-DselectedColNames=f0,f3
-DmodelName=pai_kmeans_test_model
-DoutputTableName=pai_ft_cluster_evaluation_out;
```

| Parameter      | Required | Description                  | Default value    |
|----------------|----------|------------------------------|------------------|
| inputTableName | Yes      | The name of the input table. | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                         | Default value    |
|----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| selectedColNames     | No       | The names of the columns selected from the input table for evaluation. Separate multiple columns with commas (.). The value of this parameter must be consistent with the feature columns in the model.                                                                                                                                                                                                                             | All columns      |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> If you specify multiple partitions, separate them with commas (.).</p> | Full table       |
| enableSparse         | No       | Specifies whether the input data is in the sparse format. Valid values: true and false.                                                                                                                                                                                                                                                                                                                                             | false            |
| itemDelimiter        | No       | The delimiter used to separate key-value pairs in the sparse format.                                                                                                                                                                                                                                                                                                                                                                | ,                |
| kvDelimiter          | No       | The delimiter used to separate keys and values in the sparse format.                                                                                                                                                                                                                                                                                                                                                                | :                |
| modelName            | Yes      | The name of the input clustering model.                                                                                                                                                                                                                                                                                                                                                                                             | No default value |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                       | No default value |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                                                                                  | No default value |

The Calinski-Harabasz index is also known as the variance ratio criterion (VRC). The following figure shows the formula used to calculate the VRC.

$$VRC_k = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$$

| Parameter | Description |
|-----------|-------------|
|-----------|-------------|



| Parameter | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $SS_B$    | <p>The variance between clusters. The following figure shows the formula used to calculate the variance between clusters.</p> $SS_B = \sum_{i=1}^k n_i \ m_i - m\ ^2$ <ul style="list-style-type: none"> <li>• <math>k</math>: indicates the number of cluster center points.</li> <li>• <math>m_i</math>: indicates the center point of the cluster.</li> <li>• <math>m</math>: indicates the mean of the input data.</li> </ul>                                                                    |
| $SS_W$    | <p>The variance within a cluster. The following figure shows the formula used to calculate the variance within a cluster.</p> $SS_W = \sum_{i=1}^k \sum_{x \in c_i} \ x - m_i\ ^2$ <ul style="list-style-type: none"> <li>• <math>k</math>: indicates the number of cluster center points.</li> <li>• <math>x</math>: indicates data points.</li> <li>• <math>c_i</math>: indicates the <math>i</math>th cluster.</li> <li>• <math>m_i</math>: indicates the center point of the cluster.</li> </ul> |
| $N$       | The total number of records.                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| $k$       | The number of cluster center points.                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |

## Example

1. Execute the following SQL statement to generate test data:

```
create table if not exists pai_cluster_evaluation_test_input as
select * from
(
  select 1 as id, 1 as f0, 2 as f3 from dual
  union all
  select 2 as id, 1 as f0, 3 as f3 from dual
  union all
  select 3 as id, 1 as f0, 4 as f3 from dual
  union all
  select 4 as id, 0 as f0, 3 as f3 from dual
  union all
  select 5 as id, 0 as f0, 4 as f3 from dual
)tmp;
```

2. Run the following PAI command to build a clustering model. A k-means clustering model is built in this example.

```

PAI -name kmeans
  -project algo_public
  -DinputTableName=pai_cluster_evaluation_test_input
  -DselectedColNames=f0,f3
  -DcenterCount=3
  -Dloop=10
  -Daccuracy=0.00001
  -DdistanceType=euclidean
  -DinitCenterMethod=random
  -Dseed=1
  -DmodelName=pai_kmeans_test_model
  -DidxTableName=pai_kmeans_test_idx

```

- Run the following PAI command to submit the parameters configured for the Clustering Model Evaluation component:

```

PAI -name cluster_evaluation
  -project algo_public
  -DinputTableName=pai_cluster_evaluation_test_input
  -DselectedColNames=f0,f3
  -DmodelName=pai_kmeans_test_model
  -DoutputTableName=pai_ft_cluster_evaluation_out;

```

- View the output evaluation table `pai_ft_cluster_evaluation_out`, the following table lists the mappings between the fields in the graph and those in the `pai_ft_cluster_evaluation_out` table.

| Table         | Graph                                          |
|---------------|------------------------------------------------|
| count         | The total number of records.                   |
| centerCount   | The number of cluster centers.                 |
| calinhara     | The VRC metric.                                |
| clusterCounts | The number of points included in each cluster. |

## Confusion Matrix

Confusion Matrix is suitable for supervised learning and corresponds to the matching matrix in unsupervised learning. In precision evaluation, Confusion Matrix is used to compare classification results with actual measured values and display the precision of classification results in a matrix.

You can configure the component by using one of the following methods:

- PAI console

| Parameter                      | Description                                                      |
|--------------------------------|------------------------------------------------------------------|
| Original Label Column          | The columns of numeric data types are supported.                 |
| Prediction Result Label Column | This parameter is required if <b>Threshold</b> is not specified. |

| Parameter                              | Description                                                                                                                                      |
|----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Threshold</b>                       | The threshold used to determine positive samples. Samples whose sample values are greater than the value of this parameter are positive samples. |
| <b>Prediction Result Detail Column</b> | You can specify only this parameter or <b>Prediction Result Label Column</b> . This parameter is required if <b>Threshold</b> is specified.      |
| <b>Positive Sample Label</b>           | This parameter is required if <b>Threshold</b> is specified.                                                                                     |

- PAI command
  - Threshold not specified

```

pai -name confusionmatrix -project algo_public
-DinputTableName=wpbc_pred
-DoutputTableName=wpbc_confu
-DlabelColName=label
-DpredictionColName=prediction_result;

```

- Threshold specified

```

pai -name confusionmatrix -project algo_public
-DinputTableName=wpbc_pred
-DoutputTableName=wpbc_confu
-DlabelColName=label
-DpredictionDetailColName=prediction_detail
-Dthreshold=0.8
-DgoodValue=N;

```

| Parameter                  | Required | Description                                                                                            | Default value    |
|----------------------------|----------|--------------------------------------------------------------------------------------------------------|------------------|
| inputTableName             | Yes      | The name of the input table. The value is also the name of the prediction output table.                | No default value |
| inputTablePartition        | No       | The partitions selected from the input table for training.                                             | Full table       |
| outputTableName            | Yes      | The name of the output table. The output table is used to store the confusion matrix.                  | No default value |
| labelColName               | Yes      | The name of the original label column.                                                                 | No default value |
| predictionColName          | No       | The name of the prediction result column. This parameter is required if Threshold is not specified.    | No default value |
| predictionDetailColumnName | No       | The name of the prediction result detail column. This parameter is required if Threshold is specified. | No default value |

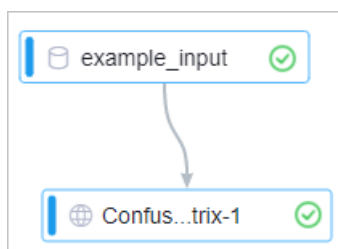
| Parameter      | Required | Description                                                                                                                               | Default value           |
|----------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| threshold      | No       | The threshold used to determine positive samples.                                                                                         | 0.5                     |
| goodValue      | No       | The label value that corresponds to the training coefficient in binary coefficient. This parameter is required if Threshold is specified. | No default value        |
| coreNum        | No       | The number of cores used in computing.                                                                                                    | Automatically allocated |
| memSizePerCore | No       | The memory size of each core. Unit: MB.                                                                                                   | Automatically allocated |
| lifecycle      | No       | The lifecycle of the output table.                                                                                                        | No default value        |

## Example

1. Use the following test data as the input.

| id | label | prediction_result |
|----|-------|-------------------|
| 0  | A     | A                 |
| 1  | A     | B                 |
| 2  | A     | A                 |
| 3  | A     | A                 |
| 4  | B     | B                 |
| 5  | B     | B                 |
| 6  | B     | A                 |
| 7  | B     | B                 |
| 8  | B     | A                 |
| 9  | A     | A                 |

2. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).



- Configure the parameters listed in the following table for Confusion Matrix. Retain the default values of the parameters that are not listed in the table.

| Parameter                      | Description                         |
|--------------------------------|-------------------------------------|
| Original Label Column          | Select the label column.            |
| Prediction Result Label Column | Set the value to prediction_result. |

- Run the experiment and view the output of the Confusion Matrix component.
  - Click the **Confusion Matrix** tab to view the output confusion matrix.
  - Click the **Proportion Matrix** tab to view the proportion matrix.
  - Click the **Statistics** tab to view the statistics of the model.

## Multiclass Classification Evaluation

Multiclass Classification Evaluation is used to evaluate the advantages and disadvantages of the models of multiclass classification algorithms based on the prediction results and original results of classification models. Then, evaluation metrics such as accuracy, kappa, and F1 score are generated.

You can configure the component by using one of the following methods:

- PAI console

| Tab            | Parameter                              | Description                                                                                                                                                                                                                                                          |
|----------------|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Original Classification Result Column  | The original label column can be selected for this parameter. A maximum of 1,000 classification results are supported.                                                                                                                                               |
|                | Predicted Classification Result Column | The predicted classification result column. In most cases, the default value prediction_result is used.                                                                                                                                                              |
|                | Advanced Options                       | <b>Prediction result Probability Column</b> is valid if the <b>Advanced Options</b> check box is selected.                                                                                                                                                           |
|                | Prediction result Probability Column   | The column that is used to calculate the log loss of the model. In most cases, this parameter is set to prediction_detail. This parameter is valid only for the random forest model. If you specify this parameter for other models, the system may report an error. |
| Tuning         | Cores                                  | The number of cores. It is automatically allocated by the system. This parameter is used with <b>Memory Size per Core</b> .                                                                                                                                          |
|                | Memory Size per Core                   | The memory size of each core. It is automatically allocated by the system. Unit: MB.                                                                                                                                                                                 |

- PAI command

```
PAI -name MultiClassEvaluation -project algo_public \
-DinputTableName="test_input" \
-DoutputTableName="test_output" \
-DlabelColName="label" \
-DpredictionColName="prediction_result" \
-Dlifecycle=30;
```

| Parameter                  | Required | Description                                                                                      | Default value           |
|----------------------------|----------|--------------------------------------------------------------------------------------------------|-------------------------|
| inputTableName             | Yes      | The name of the input table.                                                                     | No default value        |
| inputTablePartitions       | No       | The partitions selected from the input table for training.                                       | Full table              |
| outputTableName            | Yes      | The name of the output table.                                                                    | No default value        |
| labelColName               | Yes      | The name of the original label column in the input table.                                        | No default value        |
| predictionColName          | Yes      | The name of the label column of the prediction result.                                           | No default value        |
| predictionDetailColumnName | No       | The name of the probability column of the prediction result, such as {"A":0.2,"B":0.3,"C": 0.5}. | No default value        |
| lifecycle                  | No       | The lifecycle of the output table.                                                               | No default value        |
| coreNum                    | No       | The number of cores.                                                                             | Automatically allocated |
| memSizePerCore             | No       | The memory size of each core.                                                                    | Automatically allocated |

## Output description

The evaluation report generated by the Multiclass Classification Evaluation component contains the following tabs:

- Overview

On the Overview tab, the overall metrics are displayed. **MacroAveraged** is the average values of the metrics for each label.

- Confusion Matrix
- Proportion Matrix
- Statistics

On the Statistics tab, the metrics of each label are calculated by using the one-vs.-all method.

The following code shows the output table in the JSON format that is generated by the Multiclass Classification Evaluation component:

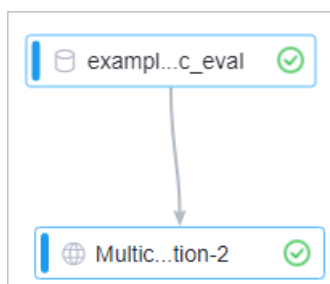
```
{
  "LabelNumber": 3,
  "LabelList": ["A", "B", "C"],
  "ConfusionMatrix": [ // Confusion matrix [actual][predict].
    [100, 10, 20],
    [30, 50, 9],
    [7, 40, 90] ],
  "ProportionMatrix": [ // Proportion matrix (based on the proportion of each row) [actual][predict].
    [0.6, 0.2, 0.2],
    [0.3, 0.6, 0.1],
    [0.1, 0.4, 0.5] ],
  "ActualLabelFrequencyList": [ // Actual data volume of each label.
    200, 300, 600],
  "ActualLabelProportionList": [ // Actual proportion of the data volume of each label.
    0.1, 0.2, 0.7],
  "PredictedLabelFrequencyList": [ // Predicted data volume of each label.
    300, 400, 400],
  "PredictedLabelProportionList": [ // Predicted proportion of the data volume of each label.
    0.2, 0.1, 0.7],
  "OverallMeasures": { // Overall metrics.
    "Accuracy": 0.70,
    "Kappa": 0.3,
    "MacroList": { // Average values of metrics of each label.
      "Sensitivity": 0.4,
      "Specificity": 0.3,
    },
    "MicroList": { // Calculate the metric based on the sum of TP, TN, FP, and FN of each label.
      "Sensitivity": 0.4,
      "Specificity": 0.3,
    },
    "LabelFrequencyBasedMicro": { // Calculate the weighted average values of metrics of each label.
      "Sensitivity": 0.4,
      "Specificity": 0.3,
    },
  },
  "LabelMeasuresList": [ // Metrics of each label.
    {
      "Accuracy": 0.6,
      "Sensitivity": 0.4,
      "Specificity": 0.3,
      "Kappa": 0.3
    },
    {
      "Accuracy": 0.6,
      "Sensitivity": 0.4,
      "Specificity": 0.3,
      "Kappa": 0.3
    },
  ]
}
```

## Example

1. Use the following test data as the input.

| id | label | prediction | detail                 |
|----|-------|------------|------------------------|
| 0  | A     | A          | {"A": 0.6, "B": 0.4}   |
| 1  | A     | B          | {"A": 0.45, "B": 0.55} |
| 2  | A     | A          | {"A": 0.7, "B": 0.3}   |
| 3  | A     | A          | {"A": 0.9, "B": 0.1}   |
| 4  | B     | B          | {"A": 0.2, "B": 0.8}   |
| 5  | B     | B          | {"A": 0.1, "B": 0.9}   |
| 6  | B     | A          | {"A": 0.52, "B": 0.48} |
| 7  | B     | B          | {"A": 0.4, "B": 0.6}   |
| 8  | B     | A          | {"A": 0.6, "B": 0.4}   |
| 9  | A     | A          | {"A": 0.75, "B": 0.25} |

2. Create the experiment shown in the following figure. For more information, see [Algorithm modeling](#).



3. Configure the parameters listed in the following table for the Multiclass Classification Evaluation component. Retain the default values of the parameters that are not listed in the table.

| Tab            | Parameter                              | Description                                   |
|----------------|----------------------------------------|-----------------------------------------------|
| Fields Setting | Original Classification Result Column  | Select the label column.                      |
|                | Predicted Classification Result Column | Set the value to prediction.                  |
|                | Advanced Options                       | Select the <b>Advanced Options</b> check box. |
|                | Prediction Result Probability Column   | Set the value to detail.                      |

4. Run the experiment and view the evaluation report generated by the Multiclass Classification



Evaluation component:

- Click the **Overview** tab to view the overview of the evaluation report.
- Click the **Confusion Matrix** tab to view the confusion matrix.
- Click the **Proportion Matrix** tab to view the proportion matrix.
- Click the **Statistics** tab to view the statistics of the model.

The following code shows the evaluation report in the JSON format:

```
{
  "ActualLabelFrequencyList": [5,
    5],
  "ActualLabelProportionList": [0.5,
    0.5],
  "ConfusionMatrix": [[4,
    1],
    [2,
    3]],
  "LabelList": ["A",
    "B"],
  "LabelMeasureList": [{
    "Accuracy": 0.7,
    "Auc": 0.9,
    "F1": 0.7272727272727273,
    "FalseDiscoveryRate": 0.3333333333333333,
    "FalseNegative": 1,
    "FalseNegativeRate": 0.2,
    "FalsePositive": 2,
    "FalsePositiveRate": 0.4,
    "Kappa": 0.3999999999999999,
    "NegativePredictiveValue": 0.75,
    "Precision": 0.6666666666666666,
    "Sensitivity": 0.8,
    "Specificity": 0.6,
    "TrueNegative": 3,
    "TruePositive": 4},
    {
    "Accuracy": 0.7,
    "Auc": 0.9,
    "F1": 0.6666666666666666,
    "FalseDiscoveryRate": 0.25,
    "FalseNegative": 2,
    "FalseNegativeRate": 0.4,
    "FalsePositive": 1,
    "FalsePositiveRate": 0.2,
    "Kappa": 0.3999999999999999,
    "NegativePredictiveValue": 0.6666666666666666,
    "Precision": 0.75,
    "Sensitivity": 0.6,
    "Specificity": 0.8,
    "TrueNegative": 4,
    "TruePositive": 3}],
  "LabelNumber": 2,
  "OverallMeasures": {
```

```

"Accuracy": 0.7,
"Kappa": 0.3999999999999999,
"LabelFrequencyBasedMicro": {
  "Accuracy": 0.7,
  "F1": 0.696969696969697,
  "FalseDiscoveryRate": 0.2916666666666666,
  "FalseNegative": 1.5,
  "FalseNegativeRate": 0.3,
  "FalsePositive": 1.5,
  "FalsePositiveRate": 0.3,
  "Kappa": 0.3999999999999999,
  "NegativePredictiveValue": 0.7083333333333333,
  "Precision": 0.7083333333333333,
  "Sensitivity": 0.7,
  "Specificity": 0.7,
  "TrueNegative": 3.5,
  "TruePositive": 3.5},
"LogLoss": 0.4548640449724484,
"MacroAveraged": {
  "Accuracy": 0.7,
  "F1": 0.696969696969697,
  "FalseDiscoveryRate": 0.2916666666666666,
  "FalseNegative": 1.5,
  "FalseNegativeRate": 0.3,
  "FalsePositive": 1.5,
  "FalsePositiveRate": 0.3,
  "Kappa": 0.3999999999999999,
  "NegativePredictiveValue": 0.7083333333333333,
  "Precision": 0.7083333333333333,
  "Sensitivity": 0.7,
  "Specificity": 0.7,
  "TrueNegative": 3.5,
  "TruePositive": 3.5},
"MicroAveraged": {
  "Accuracy": 0.7,
  "F1": 0.7,
  "FalseDiscoveryRate": 0.3,
  "FalseNegative": 3,
  "FalseNegativeRate": 0.3,
  "FalsePositive": 3,
  "FalsePositiveRate": 0.3,
  "Kappa": 0.3999999999999999,
  "NegativePredictiveValue": 0.7,
  "Precision": 0.7,
  "Sensitivity": 0.7,
  "Specificity": 0.7,
  "TrueNegative": 7,
  "TruePositive": 7}},
"PredictedLabelFrequencyList": [6,
4],
"PredictedLabelProportionList": [0.6,
0.4],
"ProportionMatrix": [[0.8,
0.2],
[0.4,
0.6]]

```

```
[0.4,  
0.6]]}
```

## 3.5.6. Recommendation

This topic describes the recommendation algorithm components provided by Machine Learning Studio. The recommendation algorithm components include Collaborative Filtering (etrec), FM Train, FM Prediction, and ALS Training.

### Collaborative Filtering (etrec)

etrec is a collaborative filtering algorithm based on item. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab                | Parameter             | Description                                                                                                                                                   |
|--------------------|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | User Column           | The name of the user column.                                                                                                                                  |
|                    | Item Column           | The name of the item column.                                                                                                                                  |
| Parameters Setting | Similarity Type       | The type of similarity. Valid values: <b>WbCosine</b> , <b>asymcosine</b> , and <b>Jaccard</b> .                                                              |
|                    | TopN                  | The maximum number of similar items that can be reserved in the output table.                                                                                 |
|                    | Calculation Method    | The method used to calculate the payload when an item of a user appears multiple times. Valid values: <b>Add</b> , <b>Mul</b> , <b>Min</b> , and <b>Max</b> . |
|                    | Minimum Item Value    | If the number of items of a user is less than the value of this parameter, the behavior of the user is ignored.                                               |
|                    | Maximum Items         | If the number of items of a user is larger than the value of this parameter, the behavior of the user is ignored.                                             |
|                    | Smoothing Factor      | This parameter is valid only when <b>Similarity Type</b> is set to <b>asymcosine</b> .                                                                        |
|                    | Weighting Coefficient | This parameter is valid only when <b>Similarity Type</b> is set to <b>asymcosine</b> .                                                                        |

- PAI command

```

PAI -name pai_etrec
  -project algo_public
  -DsimilarityType="wbcosine"
  -Dweight="1"
  -DminUserBehavior="2"
  -Dlifecycle="28"
  -DtopN="2000"
  -Dalpha="0.5"
  -DoutputTableName="etrec_test_result"
  -DmaxUserBehavior="500"
  -DinputTableName="etrec_test_input"
  -Doperator="add"
  -DuserColName="user"
  -DitemColName="item"

```

| Parameter            | Required | Description                                                               | Default value    |
|----------------------|----------|---------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                              | No default value |
| userColName          | Yes      | The name of the user column in the input table.                           | No default value |
| itemColName          | Yes      | The name of the item column in the input table.                           | No default value |
| inputTablePartitions | No       | The partitions selected from the input table for training.                | Full table       |
| outputTableName      | Yes      | The name of the output table.                                             | No default value |
| outputTablePartition | No       | The partitions in the output table.                                       | No default value |
| similarityType       | No       | The type of similarity. Valid values: wbcosine, asymlcosine, and jaccard. | wbcosine         |
| topN                 | No       | N items with the largest similarity. Valid values: 1 to 10000.            | 2000             |
| minUserBehavior      | No       | The minimum number of user behavior records.                              | 2                |
| maxUserBehavior      | No       | The maximum number of user behavior records.                              | 500              |
| itemDelimiter        | No       | The delimiter used between items in the output table.                     | Space            |
| kvDelimiter          | No       | The delimiter used between keys and values in the output table.           | :                |
| alpha                | No       | The smoothing factor when Similarity Type is set to asymlcosine.          | 0.5              |

| Parameter      | Required | Description                                                                                                                   | Default value           |
|----------------|----------|-------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| weight         | No       | The weight index when Similarity Type is set to asymcosine.                                                                   | 1.0                     |
| operator       | No       | The method used to calculate the payload when an item of a user appears multiple times. Valid values: add, mul, min, and max. | add                     |
| lifecycle      | No       | The lifecycle of the output table.                                                                                            | 1                       |
| coreNum        | No       | The number of cores.                                                                                                          | Automatically allocated |
| memSizePerCore | No       | The memory size of each core.                                                                                                 | Automatically allocated |

## Example

1. Execute the following SQL statement to generate training data:

```
drop table if exists etrec_test_input;
create table etrec_test_input
as
select
  *
from
(
  select
    cast(0 as string) as user,
    cast(0 as string) as item
  from dual
  union all
  select
    cast(0 as string) as user,
    cast(1 as string) as item
  from dual
  union all
  select
    cast(1 as string) as user,
    cast(0 as string) as item
  from dual
  union all
  select
    cast(1 as string) as user,
    cast(1 as string) as item
  from dual
) a;
```

The training data table etrec\_test\_input is generated.

| user | item |
|------|------|
| 0    | 0    |
| 0    | 1    |
| 1    | 0    |
| 1    | 1    |

2. Run the following PAI command to submit training parameters:

```
drop table if exists etrec_test_result;
PAI -name pai_etrec
  -project algo_public
  -DsimilarityType="wbcosine"
  -Dweight="1"
  -DminUserBehavior="2"
  -Dlifecycle="28"
  -DtopN="2000"
  -Dalpha="0.5"
  -DoutputTableName="etrec_test_result"
  -DmaxUserBehavior="500"
  -DinputTableName="etrec_test_input"
  -Doperator="add"
  -DuserColName="user"
  -DitemColName="item";
```

3. View the result output table etrec\_test\_result.

| itemid | similarity |
|--------|------------|
| 0      | 1:1        |
| 1      | 0:1        |

## FM algorithm

The Factorization Machine (FM) algorithm is a non-linear model that incorporates interactions among features. This algorithm is suitable for scenarios in which E-commerce, advertisements, and live video streaming are used to promote commodities. The FM algorithm template provided by Machine Learning Studio includes the FM Train and FM Prediction components. You can log on to the PAI console and switch back to the former version. Then, on the page that appears, click **Create** in the **FM-Embedding for Rec-System** section. In the New Experiment dialog box, configure the parameters to create an FM experiment.

You can configure the two components by using one of the following methods:

- PAI console

| Component | Tab | Parameter | Description |
|-----------|-----|-----------|-------------|
|-----------|-----|-----------|-------------|

| Component | Tab                | Parameter                    | Description                                                                                                                                                                  |
|-----------|--------------------|------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| FM Train  | Fields Setting     | Feature column               | The feature columns. Data in the feature columns must be in the key-value format. Separate multiple features with commas (,).                                                |
|           |                    | Label column                 | The label column. Only the columns of the DOUBLE data type are supported.                                                                                                    |
|           | Parameters Setting | Task Type                    | The task type. Valid values: <b>regression</b> and <b>binary_classification</b> .                                                                                            |
|           |                    | Iterations number            | None.                                                                                                                                                                        |
|           |                    | Regularization coefficient   | Three floating-point numbers separated by commas (,). The three numbers represent the regularization coefficients of the 0th order term, 1st order term, and 2nd order term. |
|           |                    | Learning rate                | The learning rate. If the training is diverged, set this parameter to a smaller value.                                                                                       |
|           |                    | Standard deviation           | None.                                                                                                                                                                        |
|           |                    | Dimension                    | Three positive integers separated by commas (,). The three positive integers represent the lengths of the 0th order term, 1st order term, and 2nd order term.                |
|           |                    | Block size                   | The performance parameter.                                                                                                                                                   |
|           |                    | Output table lifecycle       | None.                                                                                                                                                                        |
|           | Tuning             | cluster number               | The number of cores. This parameter is used with <b>Memory size of a single node</b> . Valid values: 1 to 9999.                                                              |
|           |                    | Memory size of a single node | The memory size of each core. This parameter is used with <b>cluster number</b> . Valid values: 1024 to 64 × 1024. Unit: MB.                                                 |
|           |                    | Prediction Result Column     | None.                                                                                                                                                                        |

| Component     | Tab                | Parameter                | Description                                                                                                         |
|---------------|--------------------|--------------------------|---------------------------------------------------------------------------------------------------------------------|
| FM Prediction | Parameters Setting | Prediction Score Column  | None.                                                                                                               |
|               |                    | Prediction Detail Column | None.                                                                                                               |
|               |                    | Reserved Columns         | The columns that you want to reserve in the output table.                                                           |
|               | Tuning             | Cores                    | The number of cores. This parameter is used with <b>Memory Size per Core</b> . Valid values: 1 to 9999.             |
|               |                    | Memory Size per Core     | The memory size of each core. This parameter is used with <b>Cores</b> . Valid values: 1024 to 64 × 1024. Unit: MB. |

- PAI command

| Component | Parameter     | Required | Description                                                                                                                                                                       | Default value    |
|-----------|---------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| FM Train  | tensorColName | Yes      | The name of the feature column. Data in the column must be in the key-value format. Separate multiple features with commas (,). Example: 1:1.0,3:1.0.                             | No default value |
|           | labelColName  | Yes      | The name of the label column. The columns of numeric data types are supported. If Task type is set to <code>binary_classification</code> , the value of the label must be 0 or 1. | No default value |
|           | task          | Yes      | The task type. Valid values: <code>regression</code> and <code>binary_classification</code> .                                                                                     | regression       |
|           | numEpochs     | No       | The number of iterations.                                                                                                                                                         | 10               |
|           | dim           | No       | Three positive integers separated by commas (,). The three positive integers represent the lengths of the 0th order term, 1st order term, and 2nd order term.                     | 1,1,10           |



| Component     | Parameter            | Required | Description                                                                                                                                                                            | Default value     |
|---------------|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
|               | learnRate            | No       | The learning rate.<br><br> <b>Note</b> If the training is diverged, set learnRate to a smaller value. | 0.01              |
|               | lambda               | No       | Three floating-point numbers separated by commas (.). The three numbers represent the regularization coefficients of the 0th order term, 1st order term, and 2nd order term.           | 0.01,0.01,0.01    |
|               | initStdev            | No       | The standard deviation of parameter initialization.                                                                                                                                    | 0.05              |
| FM Prediction | predResultColumnName | No       | The name of the prediction result column.                                                                                                                                              | prediction_result |
|               | predScoreColumnName  | No       | The name of the prediction score column.                                                                                                                                               | prediction_score  |
|               | predDetailColumnName | No       | The name of the prediction detail column.                                                                                                                                              | prediction_detail |
|               | keepColNames         | No       | The columns that you want to reserve in the output table.                                                                                                                              | All columns       |

If you use the following data as input for the FM algorithm template, the model AUC generated by the training is about 0.97.

|    | label ▲ | features ▲                                                           |
|----|---------|----------------------------------------------------------------------|
| 1  | 0       | 3:1,11:1,14:1,19:1,39:1,42:1,55:1,64:1,67:1,73:1,75:1,76:1,80:1,83:1 |
| 2  | 0       | 3:1,6:1,17:1,27:1,35:1,40:1,57:1,63:1,69:1,73:1,74:1,76:1,81:1,103:1 |
| 3  | 0       | 4:1,6:1,15:1,21:1,35:1,40:1,57:1,63:1,67:1,73:1,74:1,77:1,80:1,83:1  |
| 4  | 0       | 5:1,6:1,15:1,22:1,36:1,41:1,47:1,66:1,67:1,72:1,74:1,76:1,80:1,83:1  |
| 5  | 0       | 2:1,6:1,16:1,22:1,36:1,40:1,54:1,63:1,67:1,73:1,75:1,76:1,80:1,83:1  |
| 6  | 0       | 2:1,6:1,14:1,20:1,37:1,41:1,47:1,64:1,67:1,73:1,74:1,76:1,82:1,83:1  |
| 7  | 0       | 1:1,6:1,14:1,22:1,36:1,42:1,49:1,64:1,67:1,72:1,74:1,77:1,80:1,83:1  |
| 8  | 0       | 1:1,6:1,17:1,19:1,39:1,42:1,53:1,64:1,67:1,73:1,74:1,76:1,80:1,83:1  |
| 9  | 0       | 2:1,6:1,18:1,20:1,37:1,42:1,48:1,64:1,71:1,73:1,74:1,76:1,81:1,83:1  |
| 10 | 1       | 5:1,11:1,15:1,32:1,39:1,40:1,52:1,63:1,67:1,73:1,74:1,76:1,78:1,83:1 |

## ALS Training

Alternating Least Squares (ALS) is a model-based recommendation algorithm. It factorizes models by using sparse matrix factorization and predicts the values of missing entries to obtain a basic training model. ALS is also known as a hybrid collaborative filtering algorithm that combines users and items.

Machine Learning Studio provides an ALS algorithm template. You can log on to the Machine Learning Studio for AI console and switch back to the former version. On the page that appears, click **Create** in the **ALS music recommendations** section. Then, in the New Experiment dialog box, configure the parameters to create an ALS experiment. You can configure the component only in the PAI console.

| Tab                | Parameter                       | Description                                                                                                                                      |
|--------------------|---------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | User Column                     | The name of the user ID column in the input table. Data in the column must be of the BIGINT data type.                                           |
|                    | Item Column                     | The name of the item column in the input table. Data in the column must be of the BIGINT data type.                                              |
|                    | Score Column                    | The name of the column that contains the scores provided by users for items in the input table. The columns of numeric data types are supported. |
| Parameters Setting | Factors                         | The number of factors. Valid values: (0,+∞). Default value: 10.                                                                                  |
|                    | Iterations                      | The number of iterations. Valid values: (0,+∞). Default value: 10.                                                                               |
|                    | Regularization Coefficient      | The regularization coefficient. Valid values: (0,+∞). Default value: 0.1.                                                                        |
|                    | Implicit Preference Model       | Specifies whether to use the implicit preference model.                                                                                          |
|                    | Implicit Preference Coefficient | The implicit preference coefficient. Valid values: (0,+∞). Default value: 40.                                                                    |
| Tuning             | Cores                           | The number of cores. Valid values: 1 to 9999.                                                                                                    |
|                    | Node Memory                     | The memory size of each core. Valid values: 1024 to 64 × 1024. Unit: MB.                                                                         |

If you use the following data as input for the ALS algorithm template, you can obtain the following generated Matrix X and Matrix Y.

- Input data

| id ▲ | user ▲ | score ▲ | item ▲    |
|------|--------|---------|-----------|
| 5    | 3249   | 1       | 978245916 |
| 5    | 3176   | 2       | 978243085 |
| 5    | 1719   | 3       | 978244205 |
| 5    | 2806   | 2       | 978243085 |
| 5    | 2734   | 2       | 978242788 |
| 5    | 1649   | 4       | 978244667 |
| 5    | 321    | 3       | 978245863 |

- Matrix X

| user ▲ | factors ▲                                                                                                                                               |
|--------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1      | [0.5775652527809143,0.6737191677093506,0.9059759974479675,0.9866708517074585,0.15602371096611023,0.2735472023487091,0.4610620439052582,0.53126531839... |
| 2      | [0.8297491073608398,0.9742560982704163,0.20942062139511108,0.701496422290802,0.3298608958721161,0.2637876868247986,0.2463243007659912,1.072137832641... |
| 3      | [0.41283249855041504,0.05786990746855736,0.09279601275920868,0.17055073380470276,0.28811654448509216,0.520247757434845,0.6457393169403076,0.55301398... |
| 4      | [0.4568990468978882,0.706784725189209,0.27782294154167175,0.4249696731567383,0.2610655725002289,0.8090317845344543,0.38613465428352356,0.73447245359... |
| 5      | [0.27422088384628296,1.4948835372924805,0.7701016664505005,0.09342949837446213,0.0758112445473671,0.687049388885498,1.7983382940292358,0.56345194578... |
| 6      | [0.5954760313034058,0.3251837193965912,0.6079537868499756,0.24247819185256958,0.48808714747428894,0.29382362961769104,0.9361608624458313,0.706079840... |
| 7      | [-0.03147859871387482,0.15576674044132233,0.14959987998008728,0.6218625903129578,0.665594756603241,0.8321151733398438,1.3799917697906494,0.372453927... |
| 10     | [0.9043779969215393,1.0728607177734375,0.3242603838443756,0.3242107033729553,0.8776293992996216,0.4835788905620575,0.5912584662437439,0.761821210384... |
| 11     | [0.5306817889213562,0.25840917229652405,0.31859368085861206,0.4959831237792969,0.12244139611721039,0.9162191152572632,0.25332701206207275,0.74489533... |
| 14     | [0.5447552800178528,0.7779694199562073,0.523705005645752,0.8368392586708069,0.492429256439209,0.4384959638118744,0.4921191334724426,0.90973246097564... |
| 16     | [0.2452736794948578,0.882552045394897,0.5863954424858093,0.36698263835907,0.5939876437187195,0.3587069809436798,0.6400734186172485,1.3128618001937...   |
| 17     | [0.9096112847328186,0.7072085738182068,0.3837902545928955,0.21780475974082947,0.5543192028999329,0.6474085450172424,1.1136257648468018,0.26731985807... |
| 21     | [1.0358597040176392,0.40937551856040955,-0.1093253344297409,0.12899069488048553,1.085453987121582,1.1062564849853516,0.48638203740119934,0.351608008... |
| 24     | [0.5609134435653687,0.8832410573959351,0.8289874792098999,0.1667052060365677,-0.0539841428399086,0.2634010314941406,0.897288024255066,-0.06595267355... |
| 25     | [0.8964614868164062,1.0373890399932861,1.121972680091858,0.4297601282596588,0.9643515348434448,0.3594096601009369,0.8026714324951172,0.5237404108047... |
| 26     | [0.8279176950454712,0.06267868727445602,0.6535260081291199,0.7451098561286926,0.6139190793037415,0.9026572704315186,0.6037695407867432,0.82576447725... |
| 29     | [0.4705789387226105,0.8837192058563232,0.46882519125938416,0.2622719407081604,0.3242076635360718,-0.35604190826416016,1.17141854763031,0.35865238308... |

## Matrix Y

| item ▲    | factors ▲                                                                                                                                             |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| 978130429 | [0.2431642860174179,0.6019538044929504,0.40354019403345764,0.254305899143219,0.4056856632232666,0.46871861815452576,0.3701469600200653,0.37579229...  |
| 978130465 | [0.12887848913669586,0.5140372514724731,0.7569844126701355,0.43304914236068726,0.5548039078712463,0.1846630722284317,0.837582862190247,0.922644...    |
| 978130485 | [0.5187233090400696,0.01644902676343918,0.30217915773391724,0.7121358513832092,0.3559131622314453,0.6373451352119446,0.6356949210166931,0.7720878...  |
| 978130612 | [0.6717739701271057,0.07011739909648895,-0.14576716721057892,-0.5848092436790466,0.6463701128959656,1.4211229085922241,0.978522777557373,0.0279569... |
| 978130633 | [0.76789790391922,0.7038623094558716,0.5854156613349915,0.6497761607170105,0.5391698479652405,0.6755023002624512,0.7179682850837708,0.5982030034...   |
| 978130670 | [0.2403375208377838,0.5461633801460266,0.9200704097747803,1.064564824104309,0.6212608218193054,1.267152190208435,0.5108421444892883,0.4369839131...   |
| 978130703 | [0.28001174330711365,0.7113327383995056,0.5033919811248779,0.10984116047620773,0.7204229831695557,0.4389004111289978,0.7448101043701172,0.36075645... |
| 978130711 | [0.7834188342094421,0.6742879748344421,0.5627444386482239,0.5531031489372253,0.9128396511077881,0.5433675050735474,0.5437983870506287,0.662769436...  |
| 978130715 | [0.4402806758880615,0.5298744440078735,0.44332155585289,0.6304010152816772,0.45532888174057007,0.37801969051361084,0.8687032461166382,0.801231682...  |
| 978130729 | [-0.07635237276554108,0.7233363389968872,0.4513658583164215,0.687857985496521,0.7565320134162903,0.855117777824402,0.6841593980789185,0.27169799...   |
| 978130782 | [0.15669465065002441,0.30549323558807373,0.8395602107048035,0.5658870339393616,0.4259960353374481,0.19400911033153534,0.0137249743565917E-4,0.672...  |
| 978130855 | [0.4147326648235321,0.32279253005981445,0.5249242186546326,0.24316798150539398,0.5092757344245911,-0.018865084275603294,0.6107094883918762,0.4126...  |
| 978130895 | [0.4914812445640564,0.037220340222120285,0.43273890018463135,0.499946266412735,0.5319150686264038,0.39401403069496155,-0.05503401905298233,0.4671...  |
| 978130919 | [1.032468557357788,1.0190776586532593,0.4246324300765991,0.5157247185707092,0.28222957253456116,0.664626955986023,0.9744172096252441,0.8370118737...  |
| 978130991 | [0.6485525965690613,0.6814789772033691,0.3041754961013794,0.2130739688873291,0.9877475500106812,0.8997684711778345,0.6870101094245911,0.316646069...  |
| 978131048 | [0.6646682620048523,0.18477798998355865,0.9337624311447144,0.7535161375999451,0.6630390882492065,0.1786441206932068,0.76865154504776,0.4035094082...  |
| 978131094 | [0.45458555221557617,1.3108928203582764,0.04586312919855118,0.3885376453399658,0.6060879230499268,0.06554857641458511,0.5884337425231934,1.107482...  |

If you want to predict the score of the 994556636 item provided by User 1, multiply Vector X and Vector Y.

```
#Vector X
[-0.14220297,0.8327106,0.5352268,0.6336995,1.2326205,0.7112976,0.9794858,0.8489773,0.330319,0.7426911]
#Vector Y
[0.71699333,0.5847747,0.96564907,0.36637592,0.77271074,0.52454436,0.69028413,0.2341857,0.73444265,0.8352135]
```

## 3.5.7. Prediction

This topic describes the Prediction component provided by Machine Learning Studio.

### Configure the component

Traditional data mining algorithms can use the Prediction component to perform model prediction. The component uses the training model and prediction data as input and generates the prediction result.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter            | Description                                                                                                                            |
|----------------|----------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Column       | The feature columns selected from the input table for prediction. By default, all columns in the input table are selected.             |
|                | Reserved Columns     | The columns that you want to reserve in the output table. We recommend that you add a label column to facilitate evaluation.           |
|                | Output Result Column | The result column in the output table.                                                                                                 |
|                | Output Score Column  | The score column in the output table.                                                                                                  |
|                | Output Detail Column | The details column in the output table.                                                                                                |
|                | Sparse Matrix        | Specifies whether to indicate sparse data in the key-value format.                                                                     |
|                | KV Delimiter         | The delimiter used between keys and values. Colons (:) are used by default.                                                            |
|                | KV Pair Delimiter    | The delimiter used between key-value pairs. Commas (,) are used by default.                                                            |
| Tuning         | Cores                | The number of cores. This parameter is used with <b>Memory Size per Core</b> . The value of this parameter must be a positive integer. |
|                | Memory Size per Core | The memory size of each core. This parameter is used with <b>Cores</b> . Unit: MB.                                                     |

- PAI command

```

pai -name prediction
-DmodelName=nb_model
-DinputTableName=wpbc
-DoutputTableName=wpbc_pred
-DappendColNames=label;

```

| Parameter       | Required | Description                                                                                                  | Default value    |
|-----------------|----------|--------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName  | Yes      | The name of the input table.                                                                                 | No default value |
| featureColNames | No       | The feature columns selected from the input table for prediction. Separate multiple columns with commas (,). | All columns      |

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                                                       | Default value           |
|----------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| appendColNames       | No       | The prediction columns selected from the input table and appended to the output table.                                                                                                                                                                                                                                                                                                                                                            | No default value        |
| inputTablePartitions | No       | <p>The partitions selected from the input table for training. Specify this parameter in one of the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <div>  <b>Note</b> If you specify multiple partitions, separate them with commas (.).         </div> | Full table              |
| outputTablePartition | No       | The partitions whose results are contained in the output table.                                                                                                                                                                                                                                                                                                                                                                                   | No default value        |
| resultColName        | No       | The result column in the output table.                                                                                                                                                                                                                                                                                                                                                                                                            | prediction_result       |
| scoreColName         | No       | The score column in the output table.                                                                                                                                                                                                                                                                                                                                                                                                             | prediction_score        |
| detailColName        | No       | The details column in the output table.                                                                                                                                                                                                                                                                                                                                                                                                           | prediction_detail       |
| enableSparse         | No       | Specifies whether data in the input table is in the sparse format. Valid values: true and false.                                                                                                                                                                                                                                                                                                                                                  | false                   |
| itemDelimiter        | No       | The delimiter used between key-value pairs in the sparse format.                                                                                                                                                                                                                                                                                                                                                                                  | ,                       |
| kvDelimiter          | No       | The delimiter used between keys and values in the sparse format.                                                                                                                                                                                                                                                                                                                                                                                  | :                       |
| modelName            | Yes      | The name of the input clustering model.                                                                                                                                                                                                                                                                                                                                                                                                           | No default value        |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                     | No default value        |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                | No default value        |
| coreNum              | No       | The number of cores.                                                                                                                                                                                                                                                                                                                                                                                                                              | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core. Unit: MB.                                                                                                                                                                                                                                                                                                                                                                                                           | Automatically allocated |

## Example

1. Execute the following SQL statement to generate test data:

```
create table pai_rf_test_input as
select * from
(
select 1 as f0,2 as f1, "good" as class from dual
union all
select 1 as f0,3 as f1, "good" as class from dual
union all
select 1 as f0,4 as f1, "bad" as class from dual
union all
select 0 as f0,3 as f1, "good" as class from dual
union all
select 0 as f0,4 as f1, "bad" as class from dual
)tmp;
```

2. Run the following PAI command to build a model. The random forest algorithm is used in this example.

```
PAI -name randomforests
-project algo_public
-DinputTableName="pai_rf_test_input"
-DmodelName="pai_rf_test_model"
-DforceCategorical="f1"
-DlabelColName="class"
-DfeatureColNames="f0,f1"
-DmaxRecordSize="100000"
-DminNumPer="0"
-DminNumObj="2"
-DtreeNum="3";
```

3. Run the following PAI command to submit the parameters configured for the Prediction component:

```
PAI -name prediction
-project algo_public
-DinputTableName=pai_rf_test_input
-DmodelName=pai_rf_test_model
-DresultColName=prediction_result
-DscoreColName=prediction_score
-DdetailColName=prediction_detail
-DoutputTableName=pai_temp_2283_76333_1
```

4. View the output result table pai\_temp\_2283\_76333\_1 shown in the following figure.

| prediction_result ▲ | prediction_score ▲ | prediction_detail ▲                                     |
|---------------------|--------------------|---------------------------------------------------------|
| bad                 | 0.777777777777778  | { "bad": 0.777777777777778, "good": 0.222222222222222 } |
| bad                 | 0.833333333333334  | { "bad": 0.833333333333334, "good": 0.166666666666667 } |
| good                | 0.833333333333334  | { "bad": 0.166666666666667, "good": 0.833333333333334 } |
| good                | 0.888888888888888  | { "bad": 0.111111111111111, "good": 0.888888888888888 } |
| good                | 0.888888888888888  | { "bad": 0.111111111111111, "good": 0.888888888888888 } |

## 3.6. Time series

### 3.6.1. x13\_arima

X-13-ARIMA is a seasonal adjustment Autoregressive Integrated Moving Average Model (ARIMA) algorithm based on the open source X-13ARIMA-SEATS algorithm.

#### Background information

ARIMA was proposed by Box and Jenkins in the early 70s for time series forecasting. ARIMA is also known as Box-Jenkins Model. You can configure the component by using one of the following methods.

#### Machine Learning Platform for AI console

The following table describes the parameters of the x13\_arima component.

| Tab                | Parameter             | Description                                                                                                                                                                                                                                                                             |
|--------------------|-----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Time Series Column    | Required. It is used only to sort numeric values.                                                                                                                                                                                                                                       |
|                    | Value Column          | Required.                                                                                                                                                                                                                                                                               |
|                    | Stratification Column | Optional. You can separate multiple columns with commas (,), such as col0,col1. A time series is created for each group.                                                                                                                                                                |
| Parameters Setting | Format                | The supported format is p,d,q. p, d, and q are non-negative integers in the range of [0, 36]. <ul style="list-style-type: none"> <li>• p: autoregressive coefficient</li> <li>• d: difference</li> <li>• q: moving regression coefficient</li> </ul>                                    |
|                    | Start Date            | The supported format is year.seasonal. Example: 1986.1.                                                                                                                                                                                                                                 |
|                    | Series Frequency      | The value must be a positive integer in the range of [1,12].                                                                                                                                                                                                                            |
|                    | Format                | The supported format is sp,sd,sq. sp, sd, and sq are non-negative integers in the range of [0,36]. <ul style="list-style-type: none"> <li>• sp: seasonal autoregressive coefficient</li> <li>• sd: seasonal difference</li> <li>• sq: seasonal moving regression coefficient</li> </ul> |

| Tab    | Parameter                          | Description                                                                      |
|--------|------------------------------------|----------------------------------------------------------------------------------|
|        | <b>Seasonal Cycle</b>              | The value must be a number in the range of (0,12]. Default value: 12.            |
|        | <b>Prediction Entries</b>          | The value must be a positive integer in the range of (0,120]. Default value: 12. |
|        | <b>Prediction Confidence Level</b> | The value must be a number in the range of (0,1). Default value: 0.95.           |
| Tuning | <b>Cores</b>                       | The number of cores, which is automatically allocated.                           |
|        | <b>Memory Size</b>                 | The memory size of each core. Unit: MB.                                          |

## PAI command

```
PAI -name x13_arima
  -project algo_public
  -DinputTableName=pai_ft_x13_arima_input
  -DseqColName=id
  -DvalueColName=number
  -Dorder=3,1,1
  -Dstart=1949.1
  -Dfrequency=12
  -Dseasonal=0,1,1
  -Dperiod=12
  -DpredictStep=12
  -DoutputPredictTableName=pai_ft_x13_arima_out_predict
  -DoutputDetailTableName=pai_ft_x13_arima_out_detail
```

| Parameter            | Required | Description                                                                        | Default value    |
|----------------------|----------|------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                       | No default value |
| inputTablePartitions | No       | The names of feature columns selected from the input table for training.           | Full table       |
| seqColName           | Yes      | The series column. This parameter is used only to sort values in the value column. | No default value |
| valueColName         | Yes      | The value column.                                                                  | No default value |



| Parameter     | Required | Description                                                                                                                                                                                                           | Default value                                                                                                                                          |
|---------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| groupColNames | No       | The grouping columns. You can separate multiple columns with commas (,), such as col0,col1. A time series is created for each group.                                                                                  | No default value                                                                                                                                       |
| order         | Yes      | p indicates the autoregressive coefficient. d indicates the difference. q indicates the moving regression coefficient. The value must be a non-negative integer in the range of [0,36].                               | No default value                                                                                                                                       |
| start         | No       | The start date of a time series. A string in the year.seasonal format, such as 1986.1. For more information, see <a href="#">The time series format</a> .                                                             | 1.1                                                                                                                                                    |
| frequency     | No       | The time series frequency. The value must be a positive integer in the range of (0,12]. For more information, see <a href="#">The time series format</a> .                                                            | 12<br><div> <b>Note</b> 12 indicates 12 months or one year.</div> |
| seasonal      | No       | sp indicates the seasonal autoregressive coefficient. sd indicates the seasonal difference. sq indicates the seasonal moving regression coefficient. The value must be a non-negative integer in the range of [0,36]. | No seasonal                                                                                                                                            |
| period        | No       | The seasonal cycle. The value must be a number in the range of (0,100].                                                                                                                                               | frequency                                                                                                                                              |
| maxiter       | No       | The maximum number of iterations. The value must be a positive integer.                                                                                                                                               | 1500                                                                                                                                                   |

| Parameter              | Required | Description                                                                                                    | Default value                               |
|------------------------|----------|----------------------------------------------------------------------------------------------------------------|---------------------------------------------|
| tol                    | No       | The tolerance. The value is of the DOUBLE type.                                                                | 1e-5                                        |
| predictStep            | No       | The number of prediction entries. The value must be a positive integer in the range of (0,365].                | 12                                          |
| confidenceLevel        | No       | The confidence level. The value must be a number in the range of (0,1].                                        | 0.95                                        |
| outputPredictTableName | Yes      | The output table.                                                                                              | No default value                            |
| outputDetailTableName  | Yes      | The details table.                                                                                             | No default value                            |
| outputTablePartition   | No       | Specifies whether to export data to a partition.                                                               | Not exporting data to partitions by default |
| coreNum                | No       | The number of cores. The value must be a positive integer. This parameter is used with memSizePerCore.         | Automatically allocated                     |
| memSizePerCore         | No       | The memory size of each core. Unit: MB. The value must be a positive integer in the range of [1024,64 × 1024]. | Automatically allocated                     |
| lifecycle              | No       | The lifecycle of the output table.                                                                             | No default value                            |

## The time series format

start and frequency specify the time dimensions ts1 and ts2 for the value column.

- frequency indicates the data frequency within a unit period, which is equal to the frequency of ts2 in each ts1.
- The value of start is in the **n1.n2** format. This indicates that the start date is the n2th ts2 in the n1th ts1.

| Unit time | ts1 | ts2 | frequency | start |
|-----------|-----|-----|-----------|-------|
|-----------|-----|-----|-----------|-------|

| Unit time       | ts1           | ts2     | frequency | start                                                      |
|-----------------|---------------|---------|-----------|------------------------------------------------------------|
| 12 months/year  | Year          | Month   | 12        | 1949.2 indicates the second month of the year 1949.        |
| 4 quarters/year | Year          | Quarter | 4         | 1949.2 indicates the second quarter of the year 1949.      |
| 7 days/week     | Week          | Day     | 7         | 1949.2 indicates the second day of the 1949th week.        |
| 1               | Any time unit | 1       | 1         | 1949.1 indicates the year 1949, or the 1949th day or hour. |

Example: value=[1,2,3,5,6,7,8,9,10,11,12,13,14,15]

- **start=1949.3,frequency=12** indicates that the data is for 12 months and the prediction starts in June of the year 1950.

| year | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct |   |    |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|---|----|
| 1949 |     |     | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9 | 10 |
| 1950 | 11  | 12  | 13  | 14  | 15  |     |     |     |     |     |   |    |

- **start=1949.3,frequency=4** indicates that the data is for four quarters, and the prediction starts in the second quarter of the year 1953.

| year | Qtr1 | Qtr2 | Qtr3 | Qtr4 |
|------|------|------|------|------|
| 1949 |      |      | 1    | 2    |
| 1950 | 3    | 4    | 5    | 6    |
| 1951 | 7    | 8    | 9    | 10   |
| 1952 | 11   | 12   | 13   | 14   |
| 1953 | 14   | 15   |      |      |

- **start=1949.3,frequency=7** indicates that the data is for seven days, and the prediction starts on the fourth day of the 1951st week.

| week | Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|------|-----|-----|-----|-----|-----|-----|-----|
| 1949 |     |     | 1   | 2   | 3   | 4   | 5   |
| 1950 | 6   | 7   | 8   | 9   | 10  | 11  | 12  |
| 1951 | 13  | 14  | 15  |     |     |     |     |

- `start=1949.1,frequency=1` indicates that the prediction starts in 1963.00.

| cycle | p1 |
|-------|----|
| 1949  | 1  |
| 1950  | 2  |
| 1951  | 3  |
| 1952  | 4  |
| 1953  | 5  |
| 1954  | 6  |
| 1955  | 7  |
| 1956  | 8  |
| 1957  | 9  |
| 1958  | 10 |
| 1959  | 11 |
| 1960  | 12 |
| 1961  | 13 |
| 1962  | 14 |
| 1963  | 15 |

## Example

### Test data

The **AirPassengers** dataset is used.

| id | number |
|----|--------|
| 1  | 112    |
| 2  | 118    |

| id  | number |
|-----|--------|
| 3   | 132    |
| 4   | 129    |
| 5   | 121    |
| ... | ...    |

Use the command-line tool tunnel to upload data. Commands:

```
create table pai_ft_x13_arima_input(id bigint,number bigint);
tunnel upload data/airpassengers.csv pai_ft_x13_arima_input -h true;
```

### PAI command

```
PAI -name x13_arima
  -project algo_public
  -DinputTableName=pai_ft_x13_arima_input
  -DseqColName=id
  -DvalueColName=number
  -Dorder=3,1,1
  -Dseasonal=0,1,1
  -Dstart=1949.1
  -Dfrequency=12
  -Dperiod=12
  -DpredictStep=12
  -DoutputPredictTableName=pai_ft_x13_arima_out_predict
  -DoutputDetailTableName=pai_ft_x13_arima_out_detail
```

### Output

- Output table outputPredictTableName

| column name | comment                                                                                                                     |
|-------------|-----------------------------------------------------------------------------------------------------------------------------|
| pdate       | The date of the prediction.                                                                                                 |
| forecast    | The prediction conclusion.                                                                                                  |
| lower       | The lower threshold of the prediction results when the confidence level is specified. The default confidence level is 0.95. |
| upper       | The upper threshold of the prediction results when the confidence level is specified. The default confidence level is 0.95. |

The following figure describes the data.

|    | pdate  | forecast         | lower            | upper            |
|----|--------|------------------|------------------|------------------|
| 1  | 196101 | 444.445401291661 | 422.354385657496 | 466.536416925825 |
| 2  | 196102 | 420.971087699795 | 394.745551231805 | 447.196624167785 |
| 3  | 196103 | 453.432019696644 | 423.435661316528 | 483.428378076759 |
| 4  | 196104 | 490.922534668881 | 458.870488854835 | 522.974580482926 |
| 5  | 196105 | 503.877174753018 | 470.308964411549 | 537.445385094488 |
| 6  | 196106 | 566.536076521906 | 531.945171132091 | 601.126981911721 |
| 7  | 196107 | 652.606993945368 | 617.2596149799   | 687.954372910837 |
| 8  | 196108 | 639.841497141155 | 603.933582941945 | 675.749411340364 |
| 9  | 196109 | 542.341866147189 | 506.000630530659 | 578.683101763719 |
| 10 | 196110 | 494.745102803541 | 458.0614903363   | 531.428715270782 |
| 11 | 196111 | 426.635134341211 | 389.672783323704 | 463.597485358717 |
| 12 | 196112 | 468.722280768837 | 431.527372120416 | 505.917189417259 |

- Output table outputDetailTableName

The following table describes the parameters.

| column name | comment                                                                                                                                                                                               |
|-------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| key         | <ul style="list-style-type: none"> <li>◦ model: the model in use.</li> <li>◦ evaluation: the evaluation result.</li> <li>◦ parameters: training parameters.</li> <li>◦ log: training logs.</li> </ul> |
| summary     | The storage details.                                                                                                                                                                                  |

The following figure shows the data.

|   | key        | summary                                                       |
|---|------------|---------------------------------------------------------------|
| 1 | model      | { "comment": { "ma": "arma estimate", "mr": "regress...       |
| 2 | evaluation | { "comment": { "aic": "AIC", "aicc": "AICC (F-correcte...     |
| 3 | paramters  | { "arima": { "d": 1, "isSeasonal": true, "p": 3, "period":... |
| 4 | log        | 1 Log for X-13ARIMA-SEATS program (Version 1.1...             |

## PaiWeb display-model coefficient (key=model)

| operator | factor      | period | lag | estimate | standard error |
|----------|-------------|--------|-----|----------|----------------|
| AR       | Nonseasonal | 1      | 1   | 0.6135   | 0.0928         |
| AR       | Nonseasonal | 1      | 2   | 0.2403   | 0.1035         |
| AR       | Nonseasonal | 1      | 3   | -0.0732  | 0.0906         |
| MA       | Nonseasonal | 1      | 1   | 0.9737   | 0.0376         |
| MA       | Seasonal    | 12     | 12  | 0.1051   | 0.1031         |

## PaiWeb display-evaluation metrics (key=evaluation)

| Name                             | Indicator |
|----------------------------------|-----------|
| AIC                              | 1019.6973 |
| BIC                              | 1036.9485 |
| Hannan Quinn                     | 1026.7072 |
| Log likelihood                   | -503.8487 |
| Effective number of observations | 131       |
| Number of observations           | 144       |
| variance                         | 127.0384  |

## Algorithm scale

- Supported scale
  - Row: a maximum of 1,200 data records in a group
  - Column: one numeric column
- Resource calculation method
  - Default calculation method if groupColNames is not specified:
    - coreNum = 1
    - memSizePerCore = 4096
  - Default calculation method if groupColNames is specified:
    - coreNum = floor(Total number of rows/120,000)
    - memSizePerCore = 4096

## FAQ

Why are prediction results the same?

If an exception occurs during model training, the mean model is called. In this case, all prediction results are the mean of the training data.

Common exceptions include **instability after temporal-difference learning**, **training without convergence**, and **variance 0**. You can view the stderr file of individual nodes in Logview to obtain specific information about exceptions.

## 3.6.2. x13\_auto\_arima

The x13-auto-arima component uses the automatic selection program of the ARIMA model. The component is based on Gomez and Maravall (1998) programs, which are edited in and after TRAMO (1996).

## Background information

Selection process of x13\_auto\_arma:

- **default model estimation**

In the case of frequency = 1, the default model is (0,1,1).

In the case of frequency > 1, the default model is (0,1,1)(0,1,1).

- **identification of differencing orders**

If you specify diff and seasonalDiff, skip this step.

Use Unit root test (wiki) to determine the difference d and the seasonal difference D.

- **identification of ARMA model orders**

Select the most appropriate model based on BIC(wiki) criterion. The parameters maxOrder and maxSeasonalOrder take effect in this step.

- **comparison of identified model with default model**

Use Ljung-Box Q statistic(wiki) to compare the models. If both models are unacceptable, use the (3,d,1)(0,D,1) model.

- **final model checks**

For more information about ARIMA, see [wiki](#).

## Machine Learning Platform for AI console

The following table describes how to configure the parameters of the x13-auto-arma component by using the Machine Learning Platform for AI console.

| Tab            | Parameter                   | Description                                                                                                              |
|----------------|-----------------------------|--------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Time Series Column          | Required. It is used only to sort numeric values.                                                                        |
|                | Value Column                | Required.                                                                                                                |
|                | Stratification Column       | Optional. You can separate multiple columns with commas (,), such as col0,col1. A time series is created for each group. |
|                | Start Date                  | The supported format is year.seasonal. Example: 1986.1.                                                                  |
|                | Series Frequency            | The value must be a positive integer in the range of (0,12].                                                             |
|                | Maximum of p and q          | The value must be a positive integer in the range of (0,4].                                                              |
|                | Maximum of Seasonal p and q | The value must be a number in the range of (0,2].                                                                        |
|                | Maximum of Difference d     | The value must be a positive integer in the range of (0,2].                                                              |



| Tab                | Parameter                        | Description                                                                                                                                                                   |
|--------------------|----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Maximum of Seasonal Difference d | The value must be a positive integer in the range of (0,1].                                                                                                                   |
|                    | Difference d                     | The value must be a positive integer in the range of (0,2].<br>If both diff and maxDiff are configured, maxDiff does not take effect.<br>diff must be used with seasonalDiff. |
|                    | Seasonal Difference d            | The value must be a positive integer in the range of (0,1].<br>If both seasonalDiff and maxSeasonalDiff are configured, maxSeasonalDiff does not take effect.                 |
|                    | Prediction Entries               | The value must be a positive integer in the range of (0,120].                                                                                                                 |
|                    | Predicted Confidence Interval    | Default value: 0.95.                                                                                                                                                          |
|                    | Tolerance                        | Optional. Default value: 1e-5.                                                                                                                                                |
|                    | Maximum Iterations               | The value must be a positive integer. Default value: 1500.                                                                                                                    |
| Tuning             | Cores                            | The number of cores, which is automatically allocated.                                                                                                                        |
|                    | Memory Size                      | The memory size of each core.<br>Unit: MB.                                                                                                                                    |

## PAI command

```
PAI -name x13_auto_arima
  -project algo_public
  -DinputTableName=pai_ft_x13_arima_input
  -DseqColName=id
  -DvalueColName=number
  -Dstart=1949.1
  -Dfrequency=12
  -DpredictStep=12
  -DoutputPredictTableName=pai_ft_x13_arima_out_predict2
  -DoutputDetailTableName=pai_ft_x13_arima_out_detail2
```

| Parameter            | Required | Description                                                                                                                                                                                                | Default value                                                                                                                                          |
|----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                               | No default value                                                                                                                                       |
| inputTablePartitions | No       | The names of feature columns selected from the input table for training.                                                                                                                                   | Full table                                                                                                                                             |
| seqColName           | Yes      | The series column. This parameter is used only to sort values in the value column.                                                                                                                         | No default value                                                                                                                                       |
| valueColName         | Yes      | The value column.                                                                                                                                                                                          | No default value                                                                                                                                       |
| groupColNames        | No       | The grouping columns. You can separate multiple columns with commas (,), such as col0,col1. A time series is created for each group.                                                                       | No default value                                                                                                                                       |
| start                | No       | The start time of a time series. The value must be a string in the format of <code>year.seasonal</code> , such as <code>1986.1</code> . For more information, see <a href="#">The time series format</a> . | 1.1                                                                                                                                                    |
| frequency            | No       | The time series frequency. The value must be a positive integer in the range of (0,12]. For more information, see <a href="#">The time series format</a> .                                                 | 12<br><div> <b>Note</b> 12 indicates 12 months or one year.</div> |
| maxOrder             | No       | The maximum values of p and q. The values must be positive integers in the range of [0,4].                                                                                                                 | 2                                                                                                                                                      |
| maxSeasonalOrder     | No       | The maximum values of seasonal p and q. The values must be positive integers in the range of [0,2].                                                                                                        | 1                                                                                                                                                      |

| Parameter       | Required | Description                                                                                                                                                                                                    | Default value                                                                                                                                                                               |
|-----------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| maxDiff         | No       | The maximum value of the difference d. The value must be a positive integer in the range of [0,2].                                                                                                             | 2                                                                                                                                                                                           |
| maxSeasonalDiff | No       | The maximum value of the seasonal difference d. The value must be a positive integer in the range of [0,1].                                                                                                    | 1                                                                                                                                                                                           |
| diff            | No       | <p>The difference d. The value must be a positive integer in the range of [0,2].</p> <p>If both diff and maxDiff are configured, maxDiff does not take effect.</p> <p>diff must be used with seasonalDiff.</p> | <p>-1</p> <div>  <b>Note</b> The value -1 indicates that diff is not specified.         </div>           |
| seasonalDiff    | No       | <p>The seasonal difference d. The value must be a positive integer in the range of [0,1].</p> <p>If both seasonalDiff and maxSeasonalDiff are configured, maxSeasonalDiff does not take effect.</p>            | <p>-1</p> <div>  <b>Note</b> The value -1 indicates that seasonalDiff is not specified.         </div> |
| maxiter         | No       | The maximum number of iterations, which is a positive integer.                                                                                                                                                 | 1500                                                                                                                                                                                        |
| tol             | No       | The tolerance. The value is of the DOUBLE type.                                                                                                                                                                | 1e-5                                                                                                                                                                                        |
| predictStep     | No       | The number of prediction entries. The value must be a positive integer in the range of (0,365].                                                                                                                | 12                                                                                                                                                                                          |
| confidenceLevel | No       | The confidence level. The value must be a number in the range of (0,1).                                                                                                                                        | 0.95                                                                                                                                                                                        |

| Parameter              | Required | Description                                                                                                    | Default value                               |
|------------------------|----------|----------------------------------------------------------------------------------------------------------------|---------------------------------------------|
| outputPredictTableName | Yes      | The output table.                                                                                              | No default value                            |
| outputDetailTableName  | Yes      | The details table.                                                                                             | No default value                            |
| outputTablePartition   | No       | Specifies whether to export data to a partition.                                                               | Not exporting data to partitions by default |
| coreNum                | No       | The number of cores. The value must be a positive integer. This parameter is used with memSizePerCore.         | Automatically allocated                     |
| memSizePerCore         | No       | The memory size of each core. Unit: MB. The value must be a positive integer in the range of [1024,64 × 1024]. | Automatically allocated                     |
| lifecycle              | No       | The lifecycle of the output table.                                                                             | No default value                            |

## The time series format

start and frequency specify the time dimensions ts1 and ts2 for the value column.

- frequency indicates the data frequency within a unit period, which is equal to the frequency of ts2 in each ts1.
- The value of start is in the **n1.n2** format. This indicates that the start date is the n2th ts2 in the n1th ts1.

| Unit time       | ts1           | ts2     | frequency | start                                                      |
|-----------------|---------------|---------|-----------|------------------------------------------------------------|
| 12 months/year  | Year          | Month   | 12        | 1949.2 indicates the second month of the year 1949.        |
| 4 quarters/year | Year          | Quarter | 4         | 1949.2 indicates the second quarter of the year 1949.      |
| 7 days/week     | Week          | Day     | 7         | 1949.2 indicates the second day of the 1949th week.        |
| 1               | Any time unit | 1       | 1         | 1949.1 indicates the year 1949, or the 1949th day or hour. |

Example: value=[1,2,3,5,6,7,8,9,10,11,12,13,14,15]

- **start=1949.3,frequency=12** indicates that the data is for 12 months and the prediction starts in June of the year 1950.

| year | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1949 |     |     | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
| 1950 | 11  | 12  | 13  | 14  | 15  |     |     |     |     |     |     |     |

- **start=1949.3,frequency=4** indicates that the data is for four quarters, and the prediction starts in the second quarter of the year 1953.

| year | Qtr1 | Qtr2 | Qtr3 | Qtr4 |
|------|------|------|------|------|
| 1949 |      |      | 1    | 2    |
| 1950 | 3    | 4    | 5    | 6    |
| 1951 | 7    | 8    | 9    | 10   |
| 1952 | 11   | 12   | 13   | 14   |
| 1953 | 15   |      |      |      |

- **start=1949.3,frequency=7** indicates that the data is for seven days, and the prediction starts on the fourth day of the 1951st week.

| week | Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|------|-----|-----|-----|-----|-----|-----|-----|
| 1949 |     |     | 1   | 2   | 3   | 4   | 5   |
| 1950 | 6   | 7   | 8   | 9   | 10  | 11  | 12  |
| 1951 | 13  | 14  | 15  |     |     |     |     |

- **start=1949.1,frequency=1** indicates that the prediction starts in 1963.00.

| cycle | p1 |
|-------|----|
| 1949  | 1  |
| 1950  | 2  |
| 1951  | 3  |
| 1952  | 4  |
| 1953  | 5  |
| 1954  | 6  |

| cycle | p1 |
|-------|----|
| 1955  | 7  |
| 1956  | 8  |
| 1957  | 9  |
| 1958  | 10 |
| 1959  | 11 |
| 1960  | 12 |
| 1961  | 13 |
| 1962  | 14 |
| 1963  | 15 |

## Example

### Test data

The dataset [AirPassengers](#) is used, which records the number of international airline passengers each month from the year 1949 to the year 1960.

| id  | number |
|-----|--------|
| 1   | 112    |
| 2   | 118    |
| 3   | 132    |
| 4   | 129    |
| 5   | 121    |
| ... | ...    |

Use the command-line tool **tunnel** to upload data. Commands:

```
create table pai_ft_x13_arima_input(id bigint,number bigint);  
tunnel upload data/airpassengers.csv pai_ft_x13_arima_input -h true;
```

### PAI command

```
PAI -name x13_auto_arima
  -project algo_public
  -DinputTableName=pai_ft_x13_arima_input
  -DseqColName=id
  -DvalueColName=number
  -Dstart=1949.1
  -Dfrequency=12
  -DmaxOrder=4
  -DmaxSeasonalOrder=2
  -DmaxDiff=2
  -DmaxSeasonalDiff=1
  -DpredictStep=12
  -DoutputPredictTableName=pai_ft_x13_arima_auto_out_predict
  -DoutputDetailTableName=pai_ft_x13_arima_auto_out_detail
```

## Output

- Output table outputPredictTableName

| column name | comment                                                                                                                     |
|-------------|-----------------------------------------------------------------------------------------------------------------------------|
| pdate       | The date of the prediction.                                                                                                 |
| forecast    | The prediction conclusion.                                                                                                  |
| lower       | The lower threshold of the prediction results when the confidence level is specified. The default confidence level is 0.95. |
| upper       | The upper threshold of the prediction results when the confidence level is specified. The default confidence level is 0.95. |

The following figure provides the data.

|    | pdate  | forecast         | lower            | upper            |
|----|--------|------------------|------------------|------------------|
| 1  | 196101 | 444.445401291661 | 422.354385657496 | 466.536416925825 |
| 2  | 196102 | 420.971087699795 | 394.745551231805 | 447.196624167785 |
| 3  | 196103 | 453.432019696644 | 423.435661316528 | 483.428378076759 |
| 4  | 196104 | 490.922534668881 | 458.870488854835 | 522.974580482926 |
| 5  | 196105 | 503.877174753018 | 470.308964411549 | 537.445385094488 |
| 6  | 196106 | 566.536076521906 | 531.945171132091 | 601.126981911721 |
| 7  | 196107 | 652.606993945368 | 617.2596149799   | 687.954372910837 |
| 8  | 196108 | 639.841497141155 | 603.933582941945 | 675.749411340364 |
| 9  | 196109 | 542.341866147189 | 506.000630530659 | 578.683101763719 |
| 10 | 196110 | 494.745102803541 | 458.0614903363   | 531.428715270782 |
| 11 | 196111 | 426.635134341211 | 389.672783323704 | 463.597485358717 |
| 12 | 196112 | 468.722280768837 | 431.527372120416 | 505.917189417259 |

- Output table outputDetailTableName

The following table describes the parameters.

| column name | comment                                                                                                                                                                                               |
|-------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| key         | <ul style="list-style-type: none"> <li>◦ model: the model in use.</li> <li>◦ evaluation: the evaluation result.</li> <li>◦ parameters: training parameters.</li> <li>◦ log: training logs.</li> </ul> |
| summary     | The storage details.                                                                                                                                                                                  |

The following figure provides the data.

|   | key        | summary                                                       |
|---|------------|---------------------------------------------------------------|
| 1 | model      | { "comment": { "ma": "arma estimate", "mr": "regress...       |
| 2 | evaluation | { "comment": { "aic": "AIC", "aicc": "AICC (F-correcte...     |
| 3 | paramters  | { "arima": { "d": 1, "isSeasonal": true, "p": 3, "period":... |
| 4 | log        | 1 Log for X-13ARIMA-SEATS program (Version 1.1...             |



## PaiWeb display-model coefficient (key=model)

| operator | factor      | period | lag | estimate | standard error |
|----------|-------------|--------|-----|----------|----------------|
| AR       | Nonseasonal | 1      | 1   | 0.6135   | 0.0928         |
| AR       | Nonseasonal | 1      | 2   | 0.2403   | 0.1035         |
| AR       | Nonseasonal | 1      | 3   | -0.0732  | 0.0906         |
| MA       | Nonseasonal | 1      | 1   | 0.9737   | 0.0376         |
| MA       | Seasonal    | 12     | 12  | 0.1051   | 0.1031         |

## PaiWeb display-evaluation metrics (key=evaluation)

| Name                             | Indicator |
|----------------------------------|-----------|
| AIC                              | 1019.6973 |
| BIC                              | 1036.9485 |
| Hannan Quinn                     | 1026.7072 |
| Log likelihood                   | -503.8487 |
| Effective number of observations | 131       |
| Number of observations           | 144       |
| variance                         | 127.0384  |

## Algorithm scale

- Supported scale
  - Row: a maximum of 1,200 data records in a group
  - Column: one numeric column
- Resource calculation method
  - Default calculation method if groupColNames is not specified:  
coreNum = 1  
memSizePerCore = 4096
  - Default calculation method if groupColNames is specified:  
coreNum = floor(Total number of rows/120,000)  
memSizePerCore = 4096

## FAQ

Why are prediction results the same?

If an exception occurs during model training, the mean model is called. In this case, all prediction results are the mean of the training data.

Common exceptions include **instability after temporal-difference learning**, **training without convergence**, and **variance 0**. You can view the stderr file of individual nodes in Logview to obtain specific information about exceptions.

## 3.7. Text analysis

### 3.7.1. Word Splitting

This topic describes the Word Splitting component provided by Machine Learning Studio.

This component splits words in specific columns based on Alibaba Word Segmenter (AliWS). The words after splitting are separated by spaces. If you set the POS tagger or Semantic Tagger parameter, the system provides words after splitting, the Part-of-Speech (POS) tagging results, and the semantic tagging results. The POS tagging results are separated by forward slashes (/) and semantic tagging results are separated by vertical bars (|).

The tokenizer can be TAOBAO\_CHN or INTERNET\_CHN.

You can configure the component by using the Machine Learning Platform for AI console or a PAI command.

## Configure the component

- Machine Learning Platform for AI console

| Tab                | Parameter           | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|--------------------|---------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Settings    | Columns             | The columns used for word splitting.                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| Parameters Setting | Recognition Options | <p>The content types to be recognized. Valid values:</p> <ul style="list-style-type: none"> <li>Recognize Simple Entities</li> <li>Recognize Individual Names</li> <li>Recognize Organization Names</li> <li>Recognize Telephone Numbers</li> <li>Recognize Times</li> <li>Recognize Dates</li> <li>Recognize Alphanumeric Characters</li> </ul> <p>By default, Recognize Simple Entities, Recognize Telephone Numbers, Recognize Times, Recognize Dates, and Recognize Alphanumeric Characters are selected.</p> |
|                    | Merge Options       | <p>The types of content to be merged. Valid values:</p> <ul style="list-style-type: none"> <li>Merge Chinese Numbers</li> <li>Merge Arabic Numbers</li> <li>Merge Chinese Dates</li> <li>Merge Chinese Times</li> </ul> <p>Default value: Merge Arabic Numbers.</p>                                                                                                                                                                                                                                               |
|                    | Tokenizer           | The tokenizer. Valid values: TAOBAO_CHN and INTERNET_CHN. Default value: TAOBAO_CHN.                                                                                                                                                                                                                                                                                                                                                                                                                              |
|                    | POS Tagger          | Specifies whether to enable POS tagging. By default, POS tagging is enabled.                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|                    | Semantic Tagger     | Specifies whether to enable semantic tagging. By default, semantic tagging is disabled.                                                                                                                                                                                                                                                                                                                                                                                                                           |

| Tab    | Parameter                                                 | Description                                                                                                        |
|--------|-----------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|
|        | <b>Filter Out Words That Contain Only Numbers</b>         | Specifies whether to filter out words that contain only numbers. By default, such words are not filtered out.      |
|        | <b>Filter Out Words That Contain Only English Letters</b> | Specifies whether to filter out words that contain only letters. By default, such words are not filtered out.      |
|        | <b>Filter Out Words That Contain Only Punctuations</b>    | Specifies whether to filter out words that contain only punctuations. By default, such words are not filtered out. |
| Tuning | <b>Cores</b>                                              | Automatically allocated.                                                                                           |
|        | <b>Memory Size per Core</b>                               | Automatically allocated.                                                                                           |

- PAI command

```

pai -name split_word_model
  -project algo_public
  -DoutputModelName=aliws_model
  -DcolName=content
  -Dtokenizer=TAOBAO_CHN
  -DenableDfa=true
  -DenablePersonNameTagger=false
  -DenableOrgnizationTagger=false
  -DenablePosTagger=false
  -DenableTelephoneRetrievalUnit=true
  -DenableTimeRetrievalUnit=true
  -DenableDateRetrievalUnit=true
  -DenableNumberLetterRetrievalUnit=true
  -DenableChnNumMerge=false
  -DenableNumMerge=true
  -DenableChnTimeMerge=false
  -DenableChnDateMerge=false
  -DenableSemanticTagger=true

```

| Parameter      | Required | Description                  | Default value    |
|----------------|----------|------------------------------|------------------|
| inputTableName | Yes      | The name of the input table. | No default value |

| Parameter                       | Required | Description                                                                                                                                                                                                                                                                                                  | Default value    |
|---------------------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTablePartitions            | No       | The partitions selected from the input table for word splitting. This value must be in the <code>partition_name=value</code> format. To specify multiple partitions, use the following format: <code>name1=value1/name2=value2</code> . If multiple partitions are specified, separate them with commas (,). | Full table       |
| selectedColNames                | Yes      | The names of the columns selected from the input table for word splitting. If multiple partitions are specified, separate them with commas (,).                                                                                                                                                              | No default value |
| dictTableName                   | No       | Specifies whether to use a custom dictionary. The custom dictionary has only one column, and each row has only one word.                                                                                                                                                                                     | No default value |
| tokenizer                       | No       | The tokenizer. Valid values: TAOBAO_CHN and INTERNET_CHN.                                                                                                                                                                                                                                                    | TAOBAO_CHN       |
| enableDfa                       | No       | Specifies whether to recognize simple entities. Valid values: True and False.                                                                                                                                                                                                                                | True             |
| enablePersonNameTagger          | No       | Specifies whether to recognize person names. Valid values: True and False.                                                                                                                                                                                                                                   | False            |
| enableOrgnizationTagger         | No       | Specifies whether to recognize organizations. Valid values: True and False.                                                                                                                                                                                                                                  | False            |
| enablePosTagger                 | No       | Specifies whether to enable POS tagging. Valid values: True and False.                                                                                                                                                                                                                                       | False            |
| enableTelephoneRetrievalUnit    | No       | Specifies whether to recognize phone numbers. Valid values: True and False.                                                                                                                                                                                                                                  | True             |
| enableTimeRetrievalUnit         | No       | Specifies whether to recognize time. Valid values: True and False.                                                                                                                                                                                                                                           | True             |
| enableDateRetrievalUnit         | No       | Specifies whether to recognize dates. Valid values: True and False.                                                                                                                                                                                                                                          | True             |
| enableNumberLetterRetrievalUnit | No       | Specifies whether to recognize digits and letters. Valid values: True and False.                                                                                                                                                                                                                             | True             |
| enableChnNumMerge               | No       | Specifies whether to merge Chinese characters for numbers into retrieval units. Valid values: True and False.                                                                                                                                                                                                | False            |

| Parameter            | Required | Description                                                                                                                                             | Default value           |
|----------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| enableNumMerge       | No       | Specifies whether to merge Arabic numerals into retrieval units. Valid values: True and False.                                                          | True                    |
| enableChnTimeMerge   | No       | Specifies whether to merge Chinese characters for time into semantic units. Valid values: True and False.                                               | False                   |
| enableChnDateMerge   | No       | Specifies whether to merge Chinese characters for dates into semantic units. Valid values: True and False.                                              | False                   |
| enableSemanticTagger | No       | Specifies whether to enable semantic tagging. Valid values: True and False.                                                                             | False                   |
| outputTableName      | Yes      | The name of the output table.                                                                                                                           | No default value        |
| outputTablePartition | No       | The names of the partitions in the output table.                                                                                                        | No default value        |
| coreNum              | No       | The number of cores. This parameter takes effect only when memSizePerCore is configured. The value must be a positive integer in the range of [1,9999]. | Automatically allocated |
| memSizePerCore       | No       | The memory size for each core. Unit: MB. The value must be a positive integer in the range of [1024,64 × 1024].                                         | Automatically allocated |
| lifecycle            | No       | The lifecycle of the output table. The value must be a positive integer.                                                                                | No default value        |

If you use a regular table, we recommend that you do not set `coreNum` or `memSizePerCore`. The Word Splitting component automatically configures the parameters by default.

If your resources are limited, you can use the following code to calculate the number of cores and the memory size for each core.

```
def CalcCoreNumAndMem(row, col, kOneCoreDataSize=1024):
    """Calculates the number of cores and memory size for each core.
    Args:
        row: the number of rows in the input table.
        col: the number of columns in the input table.
        kOneCoreDataSize: the amount of data that can be computed by each core. Unit: MB. The value must
        be a positive integer. Default value: 1024.
    Return:
        coreNum, memSizePerCore
    Example:
        coreNum, memSizePerCore = CalcCoreNumAndMem(1000,99, 100, kOneCoreDataSize=2048)
    """
    kMBytes = 1024.0 * 1024.0
    #Number of cores involved in computing
    coreNum = max(1, int((row * col * 1000 / kMBytes / kOneCoreDataSize)))
    #Memory size per core = Data amount
    memSizePerCore = max(1024, int(kOneCoreDataSize*2))
    return coreNum, memSizePerCore
```

## Example

- Generated data

```
create table pai_alisws_test
as select
  1 as id,
  'Today is a good day. The weather is fine and sunny.' as content
from dual;
```

- PAI command

```
pai -name split_word
    -project algo_public
    -DinputTableName=pai_alisws_test
    -DselectedColNames=content
    -DoutputTableName=doc_test_split_word
```

- Input description

The input table consists of two columns: id and content.

```
+-----+-----+
| id   | content |
+-----+-----+
| 1    | Today is a good day. The weather is fine and sunny. |
```

- Output description

- The words in the tokenization column of the input table are split and then returned. The rest columns are returned without changes.
- When a custom dictionary is used, the system splits words based on the custom dictionary and context.

## 3.7.2. Convert Row, Column, and Value to KV Pair

This topic describes the Convert Row, Column, and Value to KV Pair component provided by Machine Learning Studio.

The component converts a triple table (row,col,value) to a key-value table (row,[col\_id:value]).

The data type of the triple table (row,col,value) is "XXD" or "XXL". "X" represents any data type, "D" represents the DOUBLE data type, and "L" represents the BIGINT data type. In the key-value table generated after conversion, the data types of row and value are the same as those of the original input data. The generated index table of col maps to the col\_id column. The data type of the col\_id column is BIGINT.

The following tables provide a conversion example.

- The following table is a triple table.

| id | word | count |
|----|------|-------|
| 01 | a    | 10    |
| 01 | b    | 20    |
| 01 | c    | 30    |

- The following table is the generated key-value table.

| id | key_value      |
|----|----------------|
| 01 | 1:10;2:20;3:30 |

 **Note** The delimiter between a key and a value, as well as between two key-value pairs in the key\_value column can be customized.

- The following table is the generated index table.

| key | key_id |
|-----|--------|
| a   | 1      |
| b   | 2      |
| c   | 3      |

You can configure the component by using the Machine Learning Platform for AI console or a PAI command.

### Configure the component

- Machine Learning Platform for AI console

| Tab             | Parameter                                   | Description                                                                                                                                         |
|-----------------|---------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Settings | Columns Reserved During KV Conversion       | The column name remains unchanged after a trituple table is converted to a key-value table.                                                         |
|                 | Output Keys                                 | The keys in the key-value table.                                                                                                                    |
|                 | Output Values                               | The values in the key-value table.                                                                                                                  |
|                 | Key Column in Input Index Table             | The name of the key column in the index table.                                                                                                      |
|                 | Index ID Column of Key in Input Index Table | The name of the index column in the index table.                                                                                                    |
|                 | KV Delimiter                                | The delimiter used to separate a key and a value in the key-value table. The default delimiter is a colon (:).                                      |
|                 | KV Pair Delimiter                           | The delimiter used to separate key-value pairs. The default delimiter is a comma (,).                                                               |
| Tuning          | Total Number of Instances                   | The total number of instances. The value must be a positive integer. By default, the system calculates the value based on the amount of input data. |
|                 | Memory Size (MB)                            | The total memory. The value must be a positive integer. By default, the system calculates the value based on the amount of input data.              |

- PAI command

```
PAI -name triple_to_kv
  -project algo_public
  -DinputTableName=test_data
  -DoutputTableName=test_kv_out
  -DindexOutputTableName=test_index_out
  -DidxColName=id
  -DkeyColName=word
  -DvalueColName=count
  -DinputTablePartitions=ds=test1
  -DindexInputTableName=test_index_input
  -DindexInputKeyColName=word
  -DindexInputKeyIdColName=word_id
  -DkvDelimiter=:
  -DpairDelimiter=;
  -Dlifecycle=3
```



| Parameter              | Required | Description                                                                                                       | Default value                                              |
|------------------------|----------|-------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------|
| inputTableName         | Yes      | The name of the input table.                                                                                      | No default value                                           |
| idColName              | Yes      | The column name remains unchanged after a trituple table is converted to a key-value table.                       | No default value                                           |
| keyColName             | Yes      | The name of the column that lists keys in the key-value table.                                                    | No default value                                           |
| valueColName           | Yes      | The name of the column that lists values in the key-value table.                                                  | No default value                                           |
| outputTableName        | Yes      | The name of the generated key-value table.                                                                        | No default value                                           |
| indexOutputTableName   | Yes      | The name of the generated index table.                                                                            | No default value                                           |
| indexInputTableName    | No       | The name of the existing index table. The table must contain data.                                                | No default value                                           |
| indexInputKeyColName   | No       | The name of the key column in the index table. This parameter is required if you configure indexInputTableName.   | No default value                                           |
| indexInputKeyIdColName | No       | The name of the index column in the index table. This parameter is required if you configure indexInputTableName. | No default value                                           |
| inputTablePartitions   | No       | The partition names of the input table. You can enter only one partition name.                                    | No default value                                           |
| kvDelimiter            | No       | The delimiter used to separate a key and a value in the key-value table.                                          | Colon ( : )                                                |
| pairDelimiter          | No       | The delimiter used to separate key-value pairs.                                                                   | Comma ( , )                                                |
| lifecycle              | No       | The lifecycle of the output table.                                                                                | No default value                                           |
| coreNum                | No       | The total number of instances. The value must be a positive integer.                                              | Automatically calculated based on the amount of input data |

| Parameter      | Required | Description                                             | Default value                                              |
|----------------|----------|---------------------------------------------------------|------------------------------------------------------------|
| memSizePerCore | No       | The total memory. The value must be a positive integer. | Automatically calculated based on the amount of input data |

## Example

- Input

```
drop table if exists triple2kv_test_input;
create table triple2kv_test_input as
select
*
from
(
select '01' as id, 'a' as word, 10 as count from dual
union all
select '01' as id, 'b' as word, 20 as count from dual
union all
select '01' as id, 'c' as word, 30 as count from dual
union all
select '02' as id, 'a' as word, 100 as count from dual
union all
select '02' as id, 'd' as word, 200 as count from dual
union all
select '02' as id, 'e' as word, 300 as count from dual
) tmp;
```

- PAI command

```
PAI -name triple_to_kv
    -project algo_public
    -DinputTableName=triple2kv_test_input
    -DoutputTableName=triple2kv_test_input_out
    -DindexOutputTableName=triple2kv_test_input_index_out
    -DidColName=id
    -DkeyColName=word
    -DvalueColName=count
    -Dlifecycle=1;
```

- Output

- Key-value table triple2kv\_test\_input\_out

```
+-----+-----+
| id   | key_value |
+-----+-----+
| 02   | 1:100;4:200;5:300 |
| 01   | 1:10;2:20;3:30 |
+-----+-----+
```

- Index table triple2kv\_test\_input\_index\_out

```
+-----+-----+
| key   | key_id  |
+-----+-----+
| a     | 1       |
| b     | 2       |
| c     | 3       |
| d     | 4       |
| e     | 5       |
+-----+-----+
```

### 3.7.3. String Similarity

This topic describes the String Similarity component provided by Machine Learning Studio.

String similarity calculation is a basic operation in machine learning. It is typically used in industries such as information retrieval, natural language processing, and bioinformatics. The component supports five calculation methods: Levenshtein (Levenshtein Distance), Longest Common SubString (LCS), String Subsequence Kernel (SSK), Cosine (Cosine), and SimHash\_Hamming. Input data may be distributed in two columns, and the value in one column can be used to calculate that in the other column.

- Levenshtein supports the calculation of distance and similarity.
  - Distance is indicated by the **levenshtein** parameter.
  - Similarity = 1 - Distance. Similarity is indicated by the **levenshtein\_sim** parameter.
- LCS supports the calculation of distance and similarity.
  - Distance is indicated by the **lcs** parameter.
  - Similarity = 1 - Distance. Similarity is indicated by the **lcs\_sim** parameter.
- SSK supports the calculation of similarity, which is indicated by the **ssk** parameter.
- Cosine supports the calculation of similarity, which is indicated by the **cosine** parameter.
- In the SimHash\_Hamming method, the SimHash algorithm maps the original text to a 64-bit binary fingerprint, and the Hamming Distance algorithm calculates the number of different characters of the binary fingerprint in the same position, the distance, and the similarity.
  - Distance is indicated by the **simhash\_hamming** parameter.
  - Similarity = 1 - Distance/64.0. Similarity is indicated by the **simhash\_hamming\_sim** parameter.

You can configure the component by using the Machine Learning Platform for AI console or a PAI command.

### Configure the component

- Machine Learning Platform for AI console

| Tab                | Parameter                                | Description                                                                                                                                                                                                                                                                                              |
|--------------------|------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Columns Appended to Output Table         | The columns appended to the specified output table.                                                                                                                                                                                                                                                      |
|                    | First Column for Similarity Calculation  | The default value is the first STRING column in the input table.                                                                                                                                                                                                                                         |
|                    | Second Column for Similarity Calculation | The default value is the second STRING column in the input table.                                                                                                                                                                                                                                        |
|                    | Similarity Columns in Output Table       | The similarity column in the specified output table.                                                                                                                                                                                                                                                     |
| Parameters Setting | Similarity Calculation Method            | The similarity calculation method. Valid values: <ul style="list-style-type: none"> <li>◦ levenshtein</li> <li>◦ levenshtein_sim</li> <li>◦ lcs</li> <li>◦ lcs_sim</li> <li>◦ ssk</li> <li>◦ cosine</li> <li>◦ simhash_hamming</li> <li>◦ simhash_hamming_sim</li> </ul> Default value: levenshtein_sim. |
|                    | Substring Length                         | This parameter is required only when <b>Similarity Calculation Method</b> is set to <b>ssk</b> , <b>cosine</b> , <b>simhash_hamming</b> , or <b>simhash_hamming_sim</b> . Valid values: (0,100). Default value: 2.                                                                                       |
|                    | Weight of Matching String                | This parameter is required only when <b>Similarity Calculation Method</b> is set to <b>ssk</b> . Valid values: (0,1). Default value: 0.5.                                                                                                                                                                |
| Tuning             | Computing Cores                          | Automatically allocated.                                                                                                                                                                                                                                                                                 |
|                    | Memory Size per Core (Unit: MB)          | Automatically allocated.                                                                                                                                                                                                                                                                                 |

- PAI command

#### PAI -name string\_similarity

```
-project algo_public
-DinputTableName="pai_test_string_similarity"
-DoutputTableName="pai_test_string_similarity_output"
-DinputSelectedColName1="col0"
-DinputSelectedColName2="col1";
```

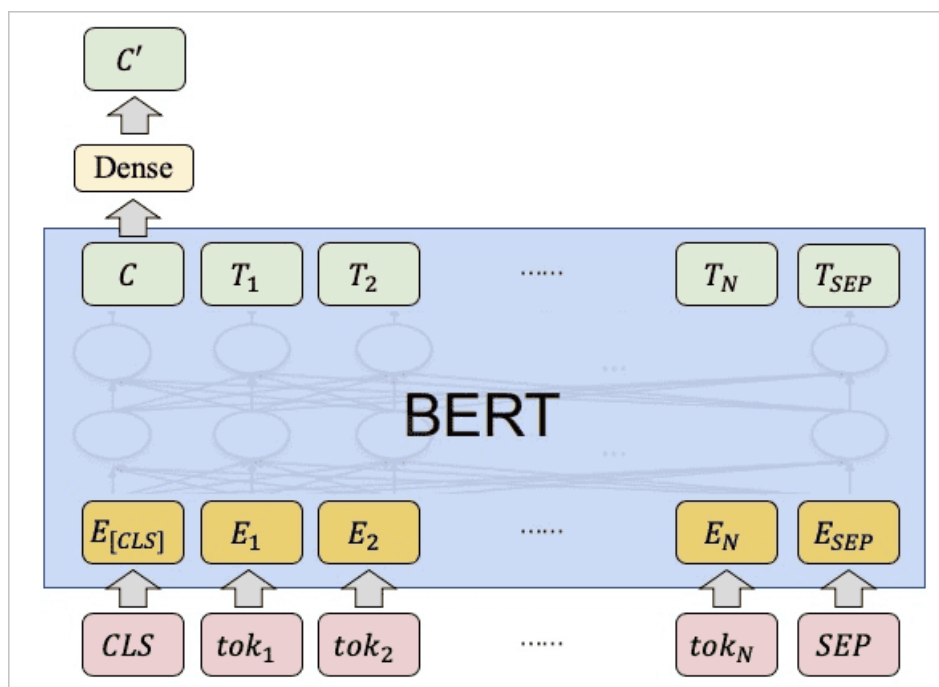
| Parameter             | Required | Description                                                                                                                                                                                                                                                              | Default value                               |
|-----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|
| inputTableName        | Yes      | The name of the input table.                                                                                                                                                                                                                                             | No default value                            |
| outputTableName       | Yes      | The name of the output table.                                                                                                                                                                                                                                            | No default value                            |
| inputSelectedColName1 | No       | The first column for similarity calculation.                                                                                                                                                                                                                             | The first STRING column in the input table  |
| inputSelectedColName2 | No       | The second column for similarity calculation.                                                                                                                                                                                                                            | The second STRING column in the input table |
| inputAppendColNames   | No       | The columns appended to the output table.                                                                                                                                                                                                                                | No default value                            |
| inputTablePartitions  | No       | The partitions in the input table.                                                                                                                                                                                                                                       | Full table                                  |
| outputColName         | No       | The name of the similarity column in the output table. A column name cannot contain special characters. It can contain only letters, digits, or underscores (_). A name must start with a letter and can be up to 128 bytes in length.                                   | output                                      |
| method                | No       | The similarity calculation method. Valid values: <ul style="list-style-type: none"> <li>◦ levenshtein</li> <li>◦ levenshtein_sim</li> <li>◦ lcs</li> <li>◦ lcs_sim</li> <li>◦ ssk</li> <li>◦ cosine</li> <li>◦ simhash_hamming</li> <li>◦ simhash_hamming_sim</li> </ul> | levenshtein_sim                             |

| Parameter      | Required | Description                                                                                                                                                               | Default value           |
|----------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| lambda         | No       | This parameter is required only when <b>Method</b> is set to <b>ssk</b> . Valid values: (0,1).                                                                            | 0.5                     |
| k              | No       | This parameter is required only when <b>Method</b> is set to <b>ssk</b> , <b>cosine</b> , <b>simhash_hamming</b> , or <b>simhash_hamming_sim</b> . Valid values: (0,100). | 2                       |
| lifecycle      | No       | The lifecycle of the output table. The value must be a positive integer.                                                                                                  | No default value        |
| coreNum        | No       | The number of cores involved in computing.                                                                                                                                | Automatically allocated |
| memSizePerCore | No       | The memory for each core.                                                                                                                                                 | Automatically allocated |

### 3.7.4. BERT Embedding

Bidirectional Encoder Representations from Transformers (BERT) embedding uses the original text as the input and provides a vector sequence after feature extraction by the system. After the CLS vector is processed at the Dense layer, the generated vector can be used as the vector of the whole sentence. This topic describes the BERT Embedding component provided by Machine Learning Studio.

The BERT Embedding component uses the original text as the input and provides the vector after BERT.



- **pool\_output** :  $C'$  in the figure, which is the vector after a sentence is encoded.

- `first_token_output` : C in the figure.
- `all_hidden_outputs` : [C, T1, T2, ..., TN, TSEP] in the figure.

This component has the following features:

- The command is simple and requires only four parameters in specific cases.
- E2E output is supported for MaxCompute tables. Vectors are provided after the original text is imported. You only need to specify the output table name.
- Parameters in the input table can be added to the output table.

## Billing

This component calls GPU computing resources. The billing rule for this component is different from that for other text analysis components. The price for other text analysis components is CNY 1.7 per computing hour. The price for this component is CNY 12 per computing hour in Beijing clusters (P100 GPU computing processor) and CNY 8.4 per computing hour in Shanghai clusters (M40 GPU computing processor).

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter              | Description                                                               |
|--------------------|------------------------|---------------------------------------------------------------------------|
| Fields Setting     | First Sequence Column  | The output field for the first column.                                    |
|                    | Second Sequence Column | The output field for the second column.                                   |
|                    | Appended Columns       | The output fields for the appended columns.                               |
| Parameters Setting | batchSize              | Default value: 256.                                                       |
|                    | sequenceLength         | Default value: 128.                                                       |
|                    | Output Schema          | Valid values: pool_output, first_token_output, and all_hidden_outputs.    |
|                    | Model                  | Valid values: pai-bert-base-zh, pai-bert-small-zh, and pai-bert-large-zh. |
| Tuning             | Workers                | Valid values: 1, 2, 3, and 4.                                             |
|                    | GPUs per Worker        | Valid values: 1 and 2.                                                    |
|                    | CPUs per Worker        | Default value: 1.                                                         |

- PAI command

PAI -name ez\_bert\_feat\_ext

-DinputTable=odps://{project}/tables/{Table name}

-DoutputTable=odps://{project}/tables/{Table name}

-DfirstSequence=col0

-DsecondSequence=""

-DappendCols=col1,col2,col0

-DoutputSchema=pool\_output,first\_token\_output,all\_hidden\_outputs

-DsequenceLength=128

-DmodelName=pai-bert-base-zh

-DbatchSize=100

-DworkerCount=1

-DworkerCPU=1

-DworkerGPU=1

-Dbuckets=oss://atp-modelzoo/tmp/? role\_arn=\${role\_arn}&host=oss-cn-hangzhou.aliyuncs.com

| Parameter      | Required | Description                                                                                                       | Default value                                                                                    |
|----------------|----------|-------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| inputTable     | Yes      | The name of the input table for feature extraction. The value must be a string in the format of project.table.    | No default value                                                                                 |
| outputTable    | Yes      | The name of the output feature table. The value must be a string is in the format of <code>project.table</code> . | No default value                                                                                 |
| firstSequence  | Yes      | The column that corresponds to the first text sequence in the input table. The value must be a string.            | No default value                                                                                 |
| secondSequence | No       | The column that corresponds to the second text sequence in the input table. The value must be a string.           | This parameter is empty by default.                                                              |
| appendCols     | No       | The name of the columns added to the output table from the input table. The value must be a string.               | This parameter is empty by default.                                                              |
| outputSchema   | No       | The features required in the output. The value must be a string.                                                  | 'pool_output', 'pool_output,first_token_output,all_hidden_outputs' (multiple features supported) |
| sequenceLength | No       | The maximum length of a sequence. The value must be an integer in the range of [1,512].                           | 128                                                                                              |
| modelName      | No       | The name of the pre-trained model. The value must be a string.                                                    | pai-bert-base-zh.                                                                                |
| batchSize      | No       | The batch size for feature extraction. The value must be a string.                                                | 256                                                                                              |



| Parameter   | Required | Description                                                       | Default value |
|-------------|----------|-------------------------------------------------------------------|---------------|
| workerCount | No       | The number of workers. The value must be an integer.              | 1             |
| workerGPU   | No       | The number of GPUs for each worker. The value must be an integer. | 1             |
| workerCPU   | No       | The number of CPUs for each worker. The value must be an integer. | 1             |

### 3.7.5. String Similarity - Top N

This topic describes the String Similarity - Top N component that is provided by Machine Learning Studio.

The String Similarity - Top N component is used to calculate the string similarity and obtain the Top N data records that best match the mapping table.

You can configure the component by using the Machine Learning Platform for AI console or running commands.

#### Configure the component

- Use the Machine Learning Platform for AI console

| Tab            | Parameter                                             | Description                                                                                                                                                                                                                                            |
|----------------|-------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Columns Appended to Output Table from Input Table     | The names of the columns to be appended to the output table from the input table.                                                                                                                                                                      |
|                | Columns Appended to Output Table from Mapping Table   | The names of the columns to be appended to the output table from the mapping table.                                                                                                                                                                    |
|                | Columns from Left Table for Similarity Calculation    | The names of the left-table columns that are used for similarity calculation.                                                                                                                                                                          |
|                | Columns from Mapping Table for Similarity Calculation | The names of the mapping-table columns that are used for similarity calculation. The similarities between the rows in the left table and all strings in the mapping table are calculated and the top N results are provided.                           |
|                | Similarity Columns in Output Table                    | The name of the similarity column in the output table. The value cannot contain special characters. It can contain only letters, digits, and underscores (_). It must start with a letter and can be up to 128 bytes in length. Default value: output. |

| Tab                | Parameter                                  | Description                                                                                                                                                                                                                                              |
|--------------------|--------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | <b>Number of Highest Similarity Scores</b> | The number of top N similarity values. The value must be a positive integer. Default value: 10.                                                                                                                                                          |
|                    | <b>Similarity Calculation Method</b>       | The method that is used for similarity calculation. Valid values: <ul style="list-style-type: none"> <li>◦ levenshtein_sim</li> <li>◦ lcs_sim</li> <li>◦ ssk</li> <li>◦ cosine</li> <li>◦ simhash_hamming_sim</li> </ul> Default value: levenshtein_sim. |
|                    | <b>Substring Length</b>                    | This parameter is required only when the <b>Similarity Calculation Method</b> parameter is set to ssk, cosine, or simhash_hamming_sim. Valid values: (0,100). Default value: 2.                                                                          |
|                    | <b>Weight of Matching String</b>           | This parameter is required only when the <b>Similarity Calculation Method</b> parameter is set to ssk. Valid values: (0,1). Default value: 0.5.                                                                                                          |
| Tuning             | <b>Computing Cores</b>                     | The number of cores that are used for calculation. By default, the cores are allocated by the system.                                                                                                                                                    |
|                    | <b>Memory Size per Core (Unit: MB)</b>     | The memory size of each core, in MB. By default, the memory is allocated by the system.                                                                                                                                                                  |

- Use commands

```
PAI -name string_similarity_topn
  -project algo_public
  -DinputTableName="pai_test_string_similarity_topn"
  -DoutputTableName="pai_test_string_similarity_topn_output"
  -DmapTableName="pai_test_string_similarity_map_topn"
  -DinputSelectedColName="col0"
  -DmapSelectedColName="col1";
```

| Parameter       | Required | Description                    | Default value |
|-----------------|----------|--------------------------------|---------------|
| inputTableName  | Yes      | The name of the input table.   | N/A           |
| mapTableName    | Yes      | The name of the mapping table. | N/A           |
| outputTableName | Yes      | The name of the output table.  | N/A           |

| Parameter                 | Required | Description                                                                                                                                                                                                                     | Default value                                                        |
|---------------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| inputSelectedColName1     | No       | The names of the left-table columns that are used for similarity calculation.                                                                                                                                                   | The name of the first column of the STRING type in the left table    |
| inputSelectedColName2     | No       | The names of the mapping-table columns that are used for similarity calculation.                                                                                                                                                | The name of the first column of the STRING type in the mapping table |
| inputAppendColNames       | No       | The names of the columns to be appended to the output table from the input table.                                                                                                                                               | N/A                                                                  |
| inputAppendRenameColNames | No       | The aliases of the columns to be appended to the output table from the input table.                                                                                                                                             | N/A                                                                  |
| mapSelectedColName        | Yes      | The names of the mapping-table columns that are used for similarity calculation.                                                                                                                                                | N/A                                                                  |
| mapAppendColNames         | No       | The names of the columns to be appended to the output table from the mapping table.                                                                                                                                             | N/A                                                                  |
| mapAppendRenameColNames   | No       | The aliases of the columns to be appended to the output table from the mapping table.                                                                                                                                           | N/A                                                                  |
| inputTablePartitions      | No       | The names of the partitions in the input table.                                                                                                                                                                                 | The names of all partitions                                          |
| mapTablePartitions        | No       | The names of the partitions in the mapping table.                                                                                                                                                                               | The names of all partitions                                          |
| outputColName             | No       | The name of the similarity column in the output table. The value cannot contain special characters. It can contain only letters, digits, and underscores (_). It must start with a letter and can be up to 128 bytes in length. | output                                                               |

| Parameter      | Required | Description                                                                                                                                                                                                    | Default value           |
|----------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| method         | No       | The method that is used for similarity calculation. Valid values: <ul style="list-style-type: none"> <li>levenshtein_sim</li> <li>lcs_sim</li> <li>ssk</li> <li>cosine</li> <li>simhash_hamming_sim</li> </ul> | levenshtein_sim         |
| lambda         | No       | This parameter is required only when the <b>Similarity Calculation Method</b> parameter is set to ssk. Valid values: (0,1).                                                                                    | 0.5                     |
| k              | No       | This parameter is required only when the <b>Similarity Calculation Method</b> parameter is set to ssk, cosine, or simhash_hamming_sim. Valid values: (0,100).                                                  | 2                       |
| lifecycle      | No       | The lifecycle of the output table. The value must be a positive integer.                                                                                                                                       | N/A                     |
| coreNum        | No       | The number of cores that are used for calculation.                                                                                                                                                             | Allocated by the system |
| memSizePerCore | No       | The memory size of each core.                                                                                                                                                                                  | Allocated by the system |

### 3.7.6. Deprecated Word Filter

This topic describes the Deprecated Word Filter component provided by Machine Learning Studio.

The Deprecated Word Filter component is a preprocessing method in text analysis. This component is used to filter noise (such as "of", "is", or "oops") in word tokenization results.

The input of the component includes an input table and deprecated word table. The data in the input table contains deprecated words. The deprecated word table has only one column. Each row has a deprecated word.

You can configure the component by using the Machine Learning Platform for AI console or a PAI command.

#### Configure the component

- Machine Learning Platform for AI console

| Tab | Parameter | Description |
|-----|-----------|-------------|
|-----|-----------|-------------|

| Tab            | Parameter            | Description                                                            |
|----------------|----------------------|------------------------------------------------------------------------|
| Fields Setting | Columns              | The columns to be filtered. Separate multiple columns with commas (,). |
| Tuning         | Cores                | Automatically allocated.                                               |
|                | Memory Size per Core | Automatically allocated.                                               |

- PAI command

```
PAI -name FilterNoise -project algo_public \
-DinputTableName="test_input" -DnoiseTableName="noise_input" \
-DoutputTableName="test_output" \
-DselectedColNames="words_seg1,words_seg2" \
-Dlifecycle=30
```

| Parameter            | Required | Description                                                              | Default value           |
|----------------------|----------|--------------------------------------------------------------------------|-------------------------|
| inputTableName       | Yes      | The name of the input table.                                             | No default value        |
| inputTablePartitions | No       | The names of the partitions in the input table.                          | Full table              |
| noiseTableName       | Yes      | The name of the deprecated word table.                                   | No default value        |
| noiseTablePartitions | No       | The names of the partitions in the deprecated word table.                | Full table              |
| outputTableName      | Yes      | The name of the output table.                                            | No default value        |
| selectedColNames     | Yes      | The columns to be filtered. Separate multiple columns with commas (,).   | No default value        |
| lifecycle            | No       | The lifecycle of the output table. The value must be a positive integer. | No default value        |
| coreNum              | No       | The number of cores involved in computing.                               | Automatically allocated |
| memSizePerCore       | No       | The memory for each core.                                                | Automatically allocated |

### 3.7.7. N-gram Counting

This topic describes the N-gram Counting component provided by Machine Learning Studio.

N-gram counting is a step in language model training. N-grams are generated based on words. The number of N-grams in all corpora is counted. The counting result is the number of N-grams in all documents rather than those in a single document. For more information, see [ngram-count](#).

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                                   | Description                                               |
|--------------------|---------------------------------------------|-----------------------------------------------------------|
| Fields Setting     | Column of Sentences in Input Table          | The column that contain the sentences in the input table. |
|                    | Column of Words in the Bag-of-Words         | The column that contain the words in the bag of words.    |
|                    | Words Column in Input Counting Result Table | The word column in the input counting result table.       |
|                    | Count Column in Input Counting Result Table | The count column in the input counting result table.      |
|                    | Sentence Weight Column                      | The column that contains weights of input sentences.      |
| Parameters Setting | Maximum N-gram Length                       | The maximum length of N-grams. Default value: 3.          |
| Tuning             | Optional. The number of cores.              | Automatically allocated.                                  |
|                    | Optional. Memory size per core.             | Automatically allocated.                                  |

- PAI command

```
PAI -name ngram_count
  -project algo_public
  -DinputTableName=pai_ngram_input
  -DoutputTableName=pai_ngram_output
  -DinputSelectedColNames=col0
  -DweightColName=weight
  -DcoreNum=2
  -DmemSizePerCore=1000;
```

| Parameter             | Required | Default value           | Description                            |
|-----------------------|----------|-------------------------|----------------------------------------|
| inputTableName        | Yes      | No default value        | The input table.                       |
| outputTableName       | Yes      | No default value        | The output table.                      |
| inputSelectedColNames | No       | The first STRING column | Columns selected from the input table. |

| Parameter            | Required | Default value           | Description                                                                                                    |
|----------------------|----------|-------------------------|----------------------------------------------------------------------------------------------------------------|
| weightColName        | No       | 1                       | The weight column.                                                                                             |
| inputTablePartitions | No       | Full table              | The partitions selected from the input table.                                                                  |
| countTableName       | No       | No default value        | The N-gram counting output table previously generated. The table is merged into the output result.             |
| countWordColName     | No       | The second column       | The word column in the counting table.                                                                         |
| countCountColName    | No       | The third column        | The name of the count column in the counting table.                                                            |
| countTablePartitions | No       | No default value        | The partitions in the counting table.                                                                          |
| vocabTableName       | No       | No default value        | The name of the bag-of-words table. The words that are not contained in the bag of words are marked as \<unk\. |
| vocabSelectedColName | No       | The first STRING column | The column that contains the words in the bag of words.                                                        |
| vocabTablePartitions | No       | No default value        | Partitions in the bag-of-words table.                                                                          |
| order                | No       | 3                       | The maximum length of N-grams.                                                                                 |
| lifecycle            | No       | No default value        | The lifecycle of the output table.                                                                             |
| coreNum              | No       | No default value        | The number of cores.                                                                                           |
| memSizePerCore       | No       | No default value        | The memory size for each core.                                                                                 |

## 3.7.8. Text Summarization

This topic describes the Text Summarization component provided by Machine Learning Studio.

The Text Summarization component can automatically generate abstracts. An abstract is a simple and coherent short text that accurately reflects the main idea of a document. The component allows computers to extract an abstract from a document.

The component uses a TextRank-based algorithm to extract sentences from a document to generate an abstract. For more information, see [TextRank: Bringing Order into Texts](#).

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                              | Description                                                                                                                                                                       |
|--------------------|----------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Settings    | Column of Marked Document IDs          | The name of the document ID column.                                                                                                                                               |
|                    | Sentence Column                        | You can specify a column.                                                                                                                                                         |
| Parameters Setting | Output First N Key Sentences           | Default value: 3.                                                                                                                                                                 |
|                    | Sentence Similarity Calculation Method | The method used to calculate sentence similarities. Valid values: <ul style="list-style-type: none"> <li>lcs_sim</li> <li>levenshtein_sim</li> <li>ssk</li> <li>cosine</li> </ul> |
|                    | Weight of Matching String              | This parameter takes effect only when <b>Sentence Similarity Calculation Method</b> is set to <b>ssk</b> . Default value: 0.5.                                                    |
|                    | Substring Length                       | This parameter takes effect only when <b>Sentence Similarity Calculation Method</b> is set to <b>ssk</b> or <b>cosine</b> . Default value: 2.                                     |
|                    | Damping coefficient                    | Default value: 0.85.                                                                                                                                                              |
|                    | Maximum Iterations                     | Default value: 100.                                                                                                                                                               |
|                    | Convergence Coefficient                | Default value: 0.000001.                                                                                                                                                          |
| Tuning             | Cores                                  | Automatically allocated.                                                                                                                                                          |
|                    | Memory Size per Core                   | Automatically allocated.                                                                                                                                                          |

- PAI command



#### PAI -name TextSummarization

```
-project algo_public
-DinputTableName="test_input"
-DoutputTableName="test_output"
-DdocIdCol="doc_id"
-DsentenceCol="sentence"
-DtopN=2
-Dlifecycle=30;
```

| Parameter            | Required | Description                                                                                                                                                                       | Default value    |
|----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                      | No default value |
| inputTablePartitions | No       | The partitions selected from the input table for computing.                                                                                                                       | Full table       |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                     | No default value |
| docIdCol             | Yes      | The name of the document ID column.                                                                                                                                               | No default value |
| sentenceCol          | Yes      | The sentence column. You can specify only one column.                                                                                                                             | No default value |
| topN                 | No       | The top N key sentences to be provided.                                                                                                                                           | 3                |
| similarityType       | No       | The method used to calculate sentence similarities. Valid values: <ul style="list-style-type: none"> <li>lcs_sim</li> <li>levenshtein_sim</li> <li>ssk</li> <li>cosine</li> </ul> | lcs_sim          |
| lambda               | No       | The weight of a matched string. This parameter takes effect only when similarityType is set to ssk.                                                                               | 0.5              |

| Parameter      | Required | Description                                                                                                             | Default value           |
|----------------|----------|-------------------------------------------------------------------------------------------------------------------------|-------------------------|
| k              | No       | The length of a substring. This parameter takes effect only when similarityType is set to <b>ssk</b> or <b>cosine</b> . | 2                       |
| dampingFactor  | No       | The damping coefficient.                                                                                                | 0.85                    |
| maxIter        | No       | The maximum number of iterations.                                                                                       | 100                     |
| epsilon        | No       | The convergence coefficient.                                                                                            | 0.000001                |
| lifecycle      | No       | The lifecycle of the input and output tables.                                                                           | No default value        |
| coreNum        | No       | The number of cores involved in computing.                                                                              | Automatically allocated |
| memSizePerCore | No       | The memory size for each core.                                                                                          | Automatically allocated |

## Example

The output table contains the doc\_id and abstract columns.

| doc_id  | abstract                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1000894 | In 2008, the Shanghai Stock Exchange published disclosure guidelines for the corporate social responsibility (CSR) of listed companies. Three types of companies were urged to disclose their CSR reports, and other qualified listed companies were encouraged to voluntarily disclose their CSR reports. In 2012, a total of 379 listed companies made up 40% of all listed companies disclosed CSR reports. Among those companies, 305 were mandated to disclose CSR reports and 74 voluntarily disclosed CSR reports. According to Hu Ruyin, the Shanghai Stock Exchange will explore how to expand the scope of CSR report disclosure, revise and refine the guidelines on disclosure of the CSR reports, and encourage more organizations to promote CSR product innovation. |

## 3.7.9. Keyword Extraction

This topic describes the Keyword Extraction component provided by Machine Learning Studio.

Keyword extraction is one of the important technologies in natural language processing. It is used to extract keywords from a document. The keyword extraction algorithm is based on TextRank, a variation of the PageRank algorithm. This keyword extraction algorithm uses the relationship between specific words to construct a network, calculate the importance of each word, and determine words with larger weights as keywords.

The keyword extraction process involves the following steps:

1. Raw corpora preparation
2. Tokenization
3. Word-based filtering
4. Keyword extraction

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                                       | Description                                     |
|--------------------|-------------------------------------------------|-------------------------------------------------|
| Fields Setting     | Column of Marked Document IDs                   | The document ID column.                         |
|                    | Word Splitting Result of Marked Documents       | The word splitting results of marked documents. |
| Parameters Setting | Output First N Keywords                         | The value must be an integer. Default value: 5. |
|                    | Window Size                                     | The value must be an integer. Default value: 2. |
|                    | Damping Coefficient                             | Default value: 0.85.                            |
|                    | Maximum Iterations                              | Default value: 100.                             |
|                    | Convergence Coefficient                         | Default value: 0.000001.                        |
| Tuning             | Cores. Auto-assigned by default.                | Allocated by the system by default.             |
|                    | Memory size per core. Auto-assigned by default. | Allocated by the system by default.             |

- PAI command

```
PAI -name KeywordsExtraction
-DinputTableName=maple_test_keywords_basic_input
-DdocIdCol=docid -DdocContent=word
-DoutputTableName=maple_test_keywords_basic_output
-DtopN=19;
```

| Parameter            | Required | Description                                                                                                                                                                                                                            | Default value    |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                           | No default value |
| inputTablePartitions | No       | The partitions selected from the input table for training, in the format of "Partition_name=value". To specify multiple partitions, use the following format: name1=value1/name2=value2. Separate multiple partitions with commas (,). | All partitions   |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                          | No default value |
| docIdCol             | Yes      | The document ID column. You can specify only one column.                                                                                                                                                                               | No default value |
| docContent           | Yes      | The word column. You can specify only one column.                                                                                                                                                                                      | No default value |
| topN                 | No       | The number of top N keywords to be provided. If the value of the parameter is greater than the total number of keywords, all keywords are provided.                                                                                    | 5                |
| windowSize           | No       | The window size of the TextRank algorithm.                                                                                                                                                                                             | 2                |
| dumpingFactor        | No       | The damping coefficient of the TextRank algorithm.                                                                                                                                                                                     | 0.85             |
| maxIter              | No       | The maximum number of iterations of the TextRank algorithm.                                                                                                                                                                            | 100              |
| epsilon              | No       | The convergence residual threshold of the TextRank algorithm.                                                                                                                                                                          | 0.000001         |

| Parameter      | Required | Description                             | Default value            |
|----------------|----------|-----------------------------------------|--------------------------|
| lifecycle      | No       | The lifecycle of the output table.      | No default value         |
| coreNum        | No       | The number of cores.                    | Automatically calculated |
| memSizePerCore | No       | The memory size of each core. Unit: MB. | Automatically calculated |

## Example

### 1. Input

Separate words in the input table with spaces, and filter out deprecated words (such as "of") and all punctuation marks.

| docid:string | word:string                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|--------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| doc0         | Blended Wing Body (BWB) aircraft new direction development aviation future Many research institutions home abroad have carried out research on BWB aircraft fully automatic shape optimization algorithm has become new research hotspot Based existing results home abroad methods characteristics commonly used modeling solution platforms analyzed compared Geometric modeling meshing flow field solution shape optimization BWB aircraft shape optimization designed optimization module compares advantages disadvantages different algorithms realizes shape optimization BWB aircraft conceptual design geometric modeling mesh generation module implements mesh generation algorithm based on over-limit interpolation modeling method based on splines flow field solving module includes finite difference solver finite element solver surface element method solver Among them finite difference solver mainly includes mathematical modeling potential flow based finite difference method variable step difference format derivation based Cartesian grid generation index algorithm Cartesian grid Neumann based Cartesian grid boundary condition expression form deduced calculation example two-dimensional airfoil aerodynamic parameters based finite difference solver realized finite element solver mainly includes theoretical modeling potential flow finite element based variational principle derivation two-dimensional finite element Kutta conditional expressions design velocity solving algorithm based on least squares two-dimensional airfoil space with wake based on Gmsh grid generator developed |

| docid:string | word:string<br>aerodynamic parameters based finite element                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              | <p>calculation example two-dimensional airfoil</p> <p>solver realized panel method solver mainly includes potential flow theory modeling based panel method automatic wake generation algorithm design surface-based development three-dimensional BWB flow field solver based element method design resistance estimation algorithm based Brasius plate solution transplantation solver Fortran language mixed compilation Python Fortran code parallelization based on OpenMP CUDA Accelerate design development algorithm realize calculation example aerodynamic parameters three-dimensional BWB based panel method solver shape optimization module implements mesh deformation algorithm based on free shape deformation genetic algorithm differential evolution algorithm aircraft surface area calculation algorithm aircraft volume calculation algorithm based on moment integral Developed data visualization format tool based on VTK</p> |

## 2. PAI command

```
PAI -name KeywordsExtraction
-DinputTableName=maple_test_keywords_basic_input
-DdocIdCol=docid -DdocContent=word
-DoutputTableName=maple_test_keywords_basic_output
-DtopN=19;
```

## 3. Output

| docid | keywords       | weight               |
|-------|----------------|----------------------|
| doc0  | based on       | 0.041306752223538405 |
| doc0  | algorithm      | 0.03089845626854151  |
| doc0  | modeling       | 0.021782865850562882 |
| doc0  | grid           | 0.020669749212693957 |
| doc0  | solver         | 0.020245609506360847 |
| doc0  | aircraft       | 0.019850761705313365 |
| doc0  | research       | 0.014193732541852615 |
| doc0  | finite element | 0.013831122054200538 |
| doc0  | solving        | 0.012924593244133104 |
| doc0  | module         | 0.01280216562287212  |

| docid | keywords        | weight               |
|-------|-----------------|----------------------|
| doc0  | derivation      | 0.011907588923852495 |
| doc0  | shape           | 0.011505456605632607 |
| doc0  | difference      | 0.011477831662367547 |
| doc0  | flow            | 0.010969269350293957 |
| doc0  | design          | 0.010830986516637251 |
| doc0  | implementation  | 0.010747536556701583 |
| doc0  | two-dimensional | 0.010695570768457084 |
| doc0  | development     | 0.010527342662670088 |
| doc0  | new             | 0.010096978306668461 |

## 3.7.10. Sentence Splitting

This topic describes the Sentence Splitting component provided by Machine Learning Studio.

Text in a document can be split by punctuation. This component is used to process text before text summarization. It splits the text into rows, each of which contains only one sentence.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab             | Parameter                      | Description                                                                      |
|-----------------|--------------------------------|----------------------------------------------------------------------------------|
| Fields Settings | Column of Marked Document IDs  | The name of the document ID column.                                              |
|                 | Marked Document Content Column | The name of the document column.                                                 |
|                 | Sentence Delimiter Set         | Default delimiters: periods (.), exclamation points (!), and question marks (?). |
| Tuning          | Cores                          | Automatically allocated.                                                         |
|                 | Memory Size per Core           | Automatically allocated.                                                         |

- PAI command

**PAI -name SplitSentences**

```
-project algo_public
-DinputTableName="test_input"
-DoutputTableName="test_output"
-DdocIdCol="doc_id"
-DdocContent="content"
-Dlifecycle=30
```

| Parameter            | Required | Description                                                               | Default value                                           |
|----------------------|----------|---------------------------------------------------------------------------|---------------------------------------------------------|
| inputTableName       | Yes      | The name of the input table.                                              | No default value                                        |
| inputTablePartitions | No       | The partitions selected from the input table for computing.               | Full table                                              |
| outputTableName      | Yes      | The name of the output table.                                             | No default value                                        |
| docIdCol             | Yes      | The name of the document ID column.                                       | No default value                                        |
| docContent           | Yes      | The name of the document content column. You can specify only one column. | No default value                                        |
| delimiter            | No       | The delimiter used to separate sentences.                                 | Period (.), exclamation point (!), or question mark (?) |
| lifecycle            | No       | The lifecycle of the input and output tables.                             | No default value                                        |
| coreNum              | No       | The number of cores involved in computing.                                | Automatically allocated                                 |
| memSizePerCore       | No       | The memory size for each core.                                            | Automatically allocated                                 |

## Example

The output table contains the doc\_id and sentence columns.

|        |          |
|--------|----------|
| doc_id | sentence |
|--------|----------|



| doc_id  | sentence                                                                                                                                                                                                                                                                                                   |
|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1000894 | In 2008, the Shanghai Stock Exchange published disclosure guidelines for the corporate social responsibility (CSR) of listed companies. Three types of companies were urged to disclose their CSR reports, and other qualified listed companies were encouraged to voluntarily disclose their CSR reports. |
| 1000894 | In 2012, a total of 379 listed companies made up 40% of all listed companies disclosed CSR reports. Among those companies, 305 were mandated to disclose CSR reports and 74 voluntarily disclosed CSR reports.                                                                                             |

### 3.7.11. Semantic Vector Distance

This topic describes the Semantic Vector Distance component provided by Machine Learning Studio.

You can calculate the extension words or sentences of the specified words or sentences based on the calculated semantic vectors, such as the word vectors calculated by the Word2Vec component. The extension words or sentences are a set of vectors that are closest to a certain vector. For example, you can generate a list of words that are most similar to a given word. This is based on the semantic vectors that are returned by the Word2Vec component.

#### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                           | Description                                                                                                                                                         |
|--------------------|-------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | ID Column                           | The ID of the column.                                                                                                                                               |
|                    | Vector Columns                      | The column that contain vector names. Example: f1 or f2.                                                                                                            |
| Parameters Setting | Number of Closest Vectors to Output | Default value: 5.                                                                                                                                                   |
|                    | Distance Calculation Mode           | The following calculation modes are supported: <ul style="list-style-type: none"><li>euclidean</li><li>cosine</li><li>manhattan</li></ul> Default value: euclidean. |
|                    | Distance Threshold                  | When the distance between two vectors is less than this value, the distance is provided. Default value: $+\infty$ .                                                 |

| Tab    | Parameter                       | Description                                                                     |
|--------|---------------------------------|---------------------------------------------------------------------------------|
| Tuning | Computing Cores                 | The number of cores used for calculation. The value is automatically allocated. |
|        | Memory Size per Core (Unit: MB) | The size of memory required by each core. The value is automatically allocated. |

- PAI command

```
PAI -name SemanticVectorDistance
-project algo_public
-DinputTableName="test_input"
-DoutputTableName="test_output"
-DidColName="word"
-DvectorColNames="f0,f1,f2,f3,f4,f5"
-Dlifecycle=30
```

| Parameter            | Required | Description                                                                                                                                                                                                                                    | Default value                     |
|----------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                   | No default value                  |
| inputTablePartitions | No       | The partitions that are selected from the input table for calculation.                                                                                                                                                                         | All partitions of the input table |
| outputTableName      | Yes      | The name of the output table.                                                                                                                                                                                                                  | No default value                  |
| idTableName          | No       | The name of the vector ID table for vector calculation. The table contains only one column, and each row stores a vector ID. This parameter is empty by default, which indicates that all vectors in the input table are used for calculation. | No default value                  |
| idTablePartitions    | No       | The partitions that are selected from the ID table for calculation. By default, all partitions are selected for calculation.                                                                                                                   | No default value                  |
| idColName            | Yes      | The name of the ID column.                                                                                                                                                                                                                     | 3                                 |

| Parameter         | Required | Description                                                                                                                                                        | Default value            |
|-------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| vectorColNames    | No       | A list of vector column names in the f1,f2 format.                                                                                                                 | No default value         |
| topN              | No       | The number of the closest vectors in the output. Valid values: [1,+∞].                                                                                             | 5                        |
| distanceType      | No       | The method that is used to calculate the distance between vectors.                                                                                                 | euclidean                |
| distanceThreshold | No       | The threshold for the distance between vectors. The threshold is provided when the distance between the two vectors is less than this value. Valid values: (0,+∞). | +∞                       |
| lifecycle         | No       | The lifecycle of the input table. Valid values: any non-zero positive integer.                                                                                     | No default value         |
| coreNum           | No       | The number of cores used for calculation. Valid values: any non-zero positive integer.                                                                             | Automatically calculated |
| memSizePerCore    | No       | The size of memory required by each core. Valid values: any non-zero positive integer.                                                                             | Automatically calculated |

## Example

The output table contains the following four columns: original\_id, near\_id, distance, and rank.

| original_id | near_id | distance | rank |
|-------------|---------|----------|------|
| hello       | hi      | 0.2      | 1    |
| hello       | xxx     | xx       | 2    |
| Man         | Woman   | 0.3      | 1    |
| Man         | xx      | xx       | 2    |

## 3.7.12. Word Frequency Statistics

During word frequency calculation, a program is used to calculate the total number of words in strings and the number of times each word appears in the strings. The strings can be manually entered or read from a specified file. The total number of words is the number of distinct words. This topic describes the Word Frequency Statistics component provided by Machine Learning Studio.

Word frequency refers to the number of times that a word appears in a corpus. Based on the word segmentation results, this component generates the words in their original order and calculates the number of times each word appears in the documents (docContent) specified by the document ID column.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform For AI console

| Tab                  | Parameter               | Description                                                      |
|----------------------|-------------------------|------------------------------------------------------------------|
| Fields Setting       | Document ID Column      | The column that contains the IDs of the specified documents.     |
|                      | Document Content Column | The column that contains the content of the specified documents. |
|                      | Tuning                  | Cores                                                            |
| Memory Size per Core |                         | The size of memory required by each core. Unit: MB.              |

- PAI command

```
pai -name doc_word_stat
  -project algo_public
  -DinputTableName=tdl_doc_test_split_word
  -DdocId=docid
  -DdocContent=content
  -DoutputTableNameMulti=doc_test_stat_multi
  -DoutputTableNameTriple=doc_test_stat_triple
  -Dlifecycle=7
```

| Parameter      | Required | Description                                                                | Default value    |
|----------------|----------|----------------------------------------------------------------------------|------------------|
| inputTableName | Yes      | The name of the input table.                                               | No default value |
| docId          | Yes      | The name of the document ID column. Only one column can be specified.      | No default value |
| docContent     | Yes      | The name of the document content column. Only one column can be specified. | No default value |

| Parameter             | Required | Description                                                                                                                                                                                                                                                                                                                                                                                                          | Default value    |
|-----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| outputTableNameMulti  | Yes      | The name of the output table that lists the words in their original order after word segmentation.                                                                                                                                                                                                                                                                                                                   | No default value |
| outputTableNameTriple | No       | The name of the output table that lists the number of times each word appears in the documents.                                                                                                                                                                                                                                                                                                                      | No default value |
| inputTablePartitions  | No       | <p>The partitions that are selected from the input table for training. The system supports the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <div>  <b>Note</b> Separate multiple partitions with commas.         </div> | Full table       |
| lifecycle             | No       | The lifecycle of the output table. Valid values: any non-zero positive integer.                                                                                                                                                                                                                                                                                                                                      | -1               |

## FAQ

- The outputTableNameMulti parameter specifies the output table that lists words in their original order in the documents after word segmentation is performed based on docId and docContent.
- The outputTableNameTriple parameter specifies the output table that lists the number of times each word appears in the documents after word segmentation is performed based on docId and docContent.

## 3.7.13. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a commonly used weighting technique for information retrieval and text mining. TF-IDF is used by search engines as a tool in scoring and ranking the relevance of a document for a given search query.

Term frequency (TF) refers to the number of times a given word appears in a document. Based on Inverse Document Frequency (IDF), a larger number of times a word appears across documents indicates a higher IDF score of the word and a stronger capability of distinguishing the word.

TF-IDF is a statistical measure that is used to evaluate the importance of a word or document.

Examples:

- The importance of a word increases proportionally when the number of times it appears in the document increases.
- The importance of a word decreases when the number of times it appears in the corpus increases.

The TF-IDF component is used to calculate the TF-IDF value of each word that appears in a collection of documents based on word frequency statistics.

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform For AI console

| Tab            | Parameter            | Description                                                                     |
|----------------|----------------------|---------------------------------------------------------------------------------|
| Fields Setting | Document ID Column   | The column that contains document IDs.                                          |
|                | Word Column          | The column that contains words.                                                 |
|                | Word Counting Column | The column that contains the number of times each word appears.                 |
| Tuning         | Cores                | The number of cores used for calculation. The value is automatically allocated. |
|                | Memory Size per Core | The size of the memory required by each core. Unit: MB.                         |

- PAI command

```
PAI -name tfidf
  -project algo_public
  -DinputTableName=rgdoc_split_triple_out
  -DdocIdCol=id
  -DwordCol=word
  -DcountCol=count
  -DoutputTableName=rg_tfidf_out;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                      | Default value    |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                     | No default value |
| inputTablePartitions | No       | The partitions that are selected from the input table for training.<br><br>This value must be in the <b>partition_name=value</b> format. If you want to specify more than one partition, use the following format: <b>name1=value1/name2=value2</b> . Separate these partitions with commas (,). | Full table       |
| docIdCol             | Yes      | The name of the document ID column. Only one column can be specified.                                                                                                                                                                                                                            | No default value |
| wordCol              | Yes      | The name of the word column. Only one column can be specified.                                                                                                                                                                                                                                   | No default value |

| Parameter       | Required | Description                                                                                                                               | Default value            |
|-----------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| countCol        | Yes      | The number of the word counting column. Only one column can be specified.                                                                 | No default value         |
| outputTableName | Yes      | The name of the input table.                                                                                                              | No default value         |
| lifecycle       | No       | The lifecycle of the output table. Valid values: any non-zero positive integer. Unit: days.                                               | No default value         |
| coreNum         | No       | The number of cores used for calculation. This parameter and the memSizePerCore parameter take effect only when they are both configured. | Automatically calculated |
| memSizePerCore  | No       | The size of the memory required by each core. This parameter and the coreNum parameter take effect only when they are both configured.    | Automatically calculated |

## 3.7.14. PLDA

The topic model is a type of statistical model that is used to discover abstract topics from a collection of documents. In the Machine Learning Platform for AI, you can set the Topics parameter for the PLDA component to abstract different topics for each document.

Latent Dirichlet allocation (LDA) is a topic model that provides topics of each document based on probability distribution. LDA is an unsupervised learning algorithm. You need only to specify the number of topics in a document set by using K. You do not need to manually annotate training sets. K is the Topics parameter of the PLDA component.

LDA is a technique that is developed by David M. Blei, Andrew Y. Ng, and Michael I. Jordan in 2003. It is used to recognize texts, classify texts, and calculate the similarity between texts in the text mining field.

### Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

#### Parameters

| Tab            | Parameter       | Description                                              |
|----------------|-----------------|----------------------------------------------------------|
| Fields Setting | Feature Columns | The feature columns that are used for training.          |
|                | Topics          | The number of topics generated by LDA.                   |
|                | Alpha           | The prior Dirichlet distribution parameter of $P(z/d)$ . |

| Tab                | Parameter          | Description                                                                                                                         |
|--------------------|--------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | beta               | The prior Dirichlet distribution parameter of $P(w/z)$ .                                                                            |
|                    | Burn-in Iterations | The number of burn-in iterations. The value of this parameter must be less than the total number of iterations. Default value: 100. |
|                    | Total Iterations   | The total number of iterations. Valid values: any non-zero positive integer. Default value: 150. This parameter is optional.        |

- PAI command

```
pai -name PLDA
  -project algo_public
  -DinputTableName=lda_input
  -DtopicNum=10
  -topicWordTableName=lda_output;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                                                                                                              | Default value                  |
|----------------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------|
| inputTableName       | Yes      | The name of the input table.                                                                                                                                                                                                                                                                                                                                                                             | No default value               |
| inputTablePartitions | No       | <p>The partitions that are selected from the input table for training. The system supports the following formats:</p> <ul style="list-style-type: none"> <li>Partition_name=value</li> <li>name1=value1/name2=value2: multi-level partitions</li> </ul> <p> <b>Note</b> Separate multiple partitions with commas.</p> | Full table                     |
| selectedColumnNames  | No       | The names of the columns selected from the input table for LDA.                                                                                                                                                                                                                                                                                                                                          | All columns of the input table |
| topicNum             | Yes      | The number of topics. Valid values: [2, 500].                                                                                                                                                                                                                                                                                                                                                            | No default value               |
| kvDelimiter          | No       | <p>The delimiter used to separate keys and values. Valid values:</p> <ul style="list-style-type: none"> <li>Space</li> <li>Comma (,)</li> <li>Colon (:)</li> </ul>                                                                                                                                                                                                                                       | Colon (:)                      |



| Parameter          | Required | Description                                                                                                                                                                                                                                         | Default value                        |
|--------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------|
| itemDelimiter      | No       | The delimiter used to separate keys.<br>Valid values: <ul style="list-style-type: none"> <li>Space</li> <li>Comma (,)</li> <li>Colon (:)</li> </ul>                                                                                                 | Space                                |
| alpha              | No       | The prior Dirichlet distribution parameter of $P(z/d)$ . Valid values: (0, $\infty$ ).                                                                                                                                                              | 0.1                                  |
| beta               | No       | The prior Dirichlet distribution parameter of $P(w/z)$ . Valid values: (0, $\infty$ ).                                                                                                                                                              | 0.01                                 |
| topicWordTableName | Yes      | The name of the topic-word frequency contribution table.                                                                                                                                                                                            | No default value                     |
| pwzTableName       | No       | The name of the $P(w/z)$ output table.                                                                                                                                                                                                              | The $P(w/z)$ table is not generated. |
| pzwTableName       | No       | The name of the $P(z/w)$ output table.                                                                                                                                                                                                              | The $P(z/w)$ table is not generated. |
| pdzTableName       | Yes      | The name of the $P(d/z)$ table.                                                                                                                                                                                                                     | The $P(d/z)$ table is not generated. |
| pzdTableName       | No       | The name of the $P(z/d)$ output table.                                                                                                                                                                                                              | The $P(z/d)$ table is not generated. |
| pzTableName        | No       | The name of the $P(z)$ output table.                                                                                                                                                                                                                | The $P(z)$ table is not generated.   |
| burnIterations     | No       | The number of burn-in iterations. Valid values: any non-zero positive integer. The value must be less than the value of totalIterations.                                                                                                            | 100                                  |
| totalIterations    | No       | The total number of iterations. Valid values: any non-zero positive integer.<br><div>  <b>Note</b> z indicates the topic, w the word, and d the document. </div> | 150                                  |

| Parameter      | Required | Description                                                                                                                                                              | Default value            |
|----------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| enableSparse   | No       | Specifies whether the data in the input table is key-value pairs. The data can be key-value pairs or word segmentation results.                                          | Key-value pairs          |
| coreNum        | No       | The parameter is used together with memSizePerCore. Valid values: any non-zero positive integer. The system calculates the number of instances based on the data volume. | Automatically calculated |
| memSizePerCore | No       | The size of memory required by each core. Unit: MB. Valid values: (1024, 65536). The value must be an integer.                                                           | Automatically calculated |

## Input and output settings

- Input:

The data must be in the format of a sparse matrix. You can use the Convert Row, Column, and Value to KV Pair component to convert the data.

**Input format** shows the input format.

Input format

| id | features                                                                                                                                                                                                                                                                                |
|----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2  | 38:3.0,39:1.0,40:3.0,41:1.0,42:1.0,43:2.0,44:1.0,45:1.0,46:1.0,47:1.0,48:1.0,49:2.0,50:1.0,51:1.0,52:1.0,53:1.0,54:1.0,55:1.0,56:1.0,57:1.0,58:1.0,59:1.0,60:1.0,61:1.0,62:1.0,63:1.0,64:1.0,65:1.0,66:1.0,67:1.0,68:1.0,69:1.0,70:1.0,71:1.0,72:1.0,73:1.0,74:1.0,75:1.0,76:1.0,77:2.0 |
| 1  | 0:1.0,1:2.0,3:1.0,4:1.0,5:1.0,6:1.0,7:1.0,8:1.0,9:1.0,10:1.0,11:1.0,12:1.0,13:1.0,14:2.0,15:1.0,16:1.0,17:1.0,18:1.0,19:1.0,20:1.0,21:1.0,22:1.0,23:1.0,24:1.0,25:1.0,26:1.0,27:1.0,28:1.0,29:1.0,30:1.0,31:1.0,32:1.0,33:1.0,34:1.0,35:1.0,36:2.0,39:2.0,77:3.0                        |

- Column 1: the ID of a document.
- Column 2: key-value data of words and word frequencies.

- Output:

The following tables are generated in sequence: topic-word frequency contribution table,  $P(w/z)$  table,  $P(z/w)$  table,  $P(d/z)$  table,  $P(z/d)$  table, and  $P(z)$  table.

**Output format** shows the output format of the topic-word frequency contribution table.

Output format

| wordid | topic_0 | topic_1 |
|--------|---------|---------|
| 0      | 1       | 0       |
| 1      | 2       | 0       |
| 2      | 0       | 0       |
| 3      | 1       | 0       |
| 4      | 1       | 0       |
| 5      | 1       | 0       |
| 6      | 1       | 0       |
| 7      | 1       | 0       |
| 8      | 0       | 1       |
| 9      | 1       | 0       |
| 10     | 1       | 0       |
| 11     | 1       | 0       |

## 3.7.15. Word Splitting (Generate Models)

This topic describes the Word Splitting (Generate Models) component provided by Machine Learning Studio.

The Word Splitting (Generate Models) component is based on Alibaba Word Segmenter (AliWS). The component is used to generate a word segmentation model based on parameters and custom dictionaries.

The component supports only Chinese Taobao word segmentation and Internet word segmentation.

The Word Splitting (Generate Models) component differs from the Word Splitting component in the following ways:

- The Word Splitting component splits texts into words.
- The Word Splitting (Generate Models) component generates a word segmentation model. To split texts, you must deploy a model and make a prediction or call an API operation.

## Configure the component

- Machine Learning Platform for AI console

| Tab                | Parameter           | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|--------------------|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Columns             | The column that is used to generate a model.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Parameters Setting | Recognition Options | <p>The types of content for recognition. Valid values:</p> <ul style="list-style-type: none"><li>◦ Recognize Simple Entities</li><li>◦ Recognize Individual Names</li><li>◦ Recognize Organization Names</li><li>◦ Recognize Telephone Numbers</li><li>◦ Recognize Times</li><li>◦ Recognize Dates</li><li>◦ Recognize Alphanumeric Characters</li></ul> <p>By default, the following options are selected: Recognize Simple Entities, Recognize Telephone Numbers, Recognize Times, Recognize Dates, and Recognize Alphanumeric Characters.</p> |
|                    | Merge Options       | <p>The types of content for merging. Valid values:</p> <ul style="list-style-type: none"><li>◦ Merge Chinese Numbers</li><li>◦ Merge Arabic Numerals</li><li>◦ Merge Chinese Dates</li><li>◦ Merge Chinese Times</li></ul> <p>Default value: Merge Arabic Numbers.</p>                                                                                                                                                                                                                                                                           |
|                    | Tokenizer           | The type of the filter. Valid values: TAOBAO_CHN and INTERNET_CHN. Default value: TAOBAO_CHN.                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|                    | Pos Tagger          | Specifies whether to perform part-of-speech tagging. By default, part-of-speech tagging is not performed.                                                                                                                                                                                                                                                                                                                                                                                                                                        |

| Tab    | Parameter                                                      | Description                                                                                                                      |
|--------|----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
|        | <b>Semantic Tagger</b>                                         | Specifies whether to perform semantic role labeling. By default, semantic role labeling is not performed.                        |
|        | <b>Filter Out Words That Contain Only Numbers</b>              | Specifies whether to filter out words whose word segmentation results are numbers. By default, this option is cleared.           |
|        | <b>Filter Out Words That Contain Only English Letters</b>      | Specifies whether to filter out words whose word segmentation results are English letters. By default, this option is cleared.   |
|        | <b>Filter Out Words That Contain Only English Punctuations</b> | Specifies whether to filter out words whose word segmentation results are punctuation marks. By default, this option is cleared. |
| Tuning | <b>Cores</b>                                                   | The number of cores used for calculation. The value is automatically allocated.                                                  |
|        | <b>Memory Size per Core</b>                                    | The size of memory required by each core. The value is automatically allocated.                                                  |

- PAI command

```

pai -name split_word_model
  -project algo_public
  -DoutputModelName=aliws_model
  -DcolName=content
  -Dtokenizer=TAOBAO_CHN
  -DenableDfa=true
  -DenablePersonNameTagger=false
  -DenableOrgnizationTagger=false
  -DenablePosTagger=false
  -DenableTelephoneRetrievalUnit=true
  -DenableTimeRetrievalUnit=true
  -DenableDateRetrievalUnit=true
  -DenableNumberLetterRetrievalUnit=true
  -DenableChnNumMerge=false
  -DenableNumMerge=true
  -DenableChnTimeMerge=false
  -DenableChnDateMerge=false
  -DenableSemanticTagger=true

```

| Parameter | Required | Description | Default value |
|-----------|----------|-------------|---------------|
|-----------|----------|-------------|---------------|

| Parameter                       | Required | Description                                                                                                                 | Default value    |
|---------------------------------|----------|-----------------------------------------------------------------------------------------------------------------------------|------------------|
| userDictTableName               | No       | Specifies whether to use a custom dictionary. A custom dictionary has only one column, and each row contains only one word. | No default value |
| outputModelName                 | Yes      | The name of the output model.                                                                                               | No default value |
| colName                         | No       | The column name of the prediction text.                                                                                     | context          |
| dictTableName                   | No       | Specifies whether to use a custom dictionary. A custom dictionary has only one column, and each row contains only one word. | No default value |
| tokenizer                       | No       | The type of the filter. Valid values: TAOBAO_CHN and INTERNET_CHN.                                                          | TAOBAO_CHN       |
| enableDfa                       | No       | Specifies whether to recognize simple entities. Valid values: True and False.                                               | True             |
| enablePersonNameTagger          | No       | Specifies whether to recognize individual names. Valid values: True and False.                                              | False            |
| enableOrgnizationTagger         | No       | Specifies whether to recognize organization names. Valid values: True and False.                                            | False            |
| enablePosTagger                 | No       | Specifies whether to perform part-of-speech tagging. Valid values: True and False.                                          | False            |
| enableTelephoneRetrievalUnit    | No       | Specifies whether to recognize telephone numbers. Valid values: True and False.                                             | True             |
| enableTimeRetrievalUnit         | No       | Specifies whether to recognize time expressions. Valid values: True and False.                                              | True             |
| enableDateRetrievalUnit         | No       | Specifies whether to recognize date expressions. Valid values: True or False.                                               | True             |
| enableNumberLetterRetrievalUnit | No       | Specifies whether to recognize alphanumeric characters. Valid values: True and False.                                       | True             |
| enableChnNumMerge               | No       | Specifies whether to merge Chinese numbers into a retrieval unit. Valid values: True and False.                             | False            |
| enableNumMerge                  | No       | Specifies whether to merge Arabic numerals into a retrieval unit. Valid values: True and False.                             | True             |

| Parameter            | Required | Description                                                                                             | Default value |
|----------------------|----------|---------------------------------------------------------------------------------------------------------|---------------|
| enableChnTimeMerge   | No       | Specifies whether to merge Chinese time expressions into a semantic unit. Valid values: True and False. | False         |
| enableChnDateMerge   | No       | Specifies whether to merge Chinese date expressions into a semantic unit. Valid values: True and False. | False         |
| enableSemanticTagger | No       | Specifies whether to perform semantic role labeling. Valid values: True and False.                      | False         |

## Examples

- PAI command

```
pai -name split_word_model
    -project algo_public
    -DoutputModelName=aliws_model
```

- Model deployment

```
create onlinemodel ning_test_aliws_model_2 -offlinemodelName ning_test_aliws_model -instanceNum 1
    -cpu 100 -memory 4096;
```

- Online word segmentation

```
KVJsonRequest request = new KVJsonRequest();
Map<String, JsonFeatureValue> row = request.addRow();
row.put(col_name, new JsonFeatureValue("The big data algorithm platform is new"));
KVJsonResponse res = predictClient.syncPredict(new JsonPredictRequest(project_name, model_name, request));
List<ResponseItem> ri = res.getOutputs();
for (ResponseItem item : ri) {
    System.out.println(item.getOutputLabel());
}
```

- Offline word segmentation

```
pai -name prediction
    -DmodelName=ning_test_aliws_model
    -DinputTableName=ning_test_aliws
    -DoutputTableName=ning_test_aliws_offline_predict;
```

## 3.7.16. Word2Vec

This topic describes the Word2Vec component provided by Machine Learning Studio.

The Word2Vec component uses a neural network to map words to vectors in the K-dimensional space based on extensive training. The component supports operations on the vectors to show the semantics of the vectors. The input is a word column or a vocabulary, and the output is a vector table and a vocabulary.

## Configure the component

- Machine Learning Platform For AI console

| Tab                | Parameter                  | Description                                                                                                              |
|--------------------|----------------------------|--------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Word Column                | The word column used for training.                                                                                       |
| Parameters Setting | Word Feature Dimension     | The number of dimensions of the word feature. Valid values: 0 to 1000. Default value: 100.                               |
|                    | Language Model             | The language model used for training. Valid values: <b>Skip-gram</b> and <b>Cbow</b> . Default value: <b>Skip-gram</b> . |
|                    | Window Size of Words       | The window size of words. Valid values: any non-zero positive integer. Default value: 5.                                 |
|                    | Random Window              | Specifies whether to use a random window. Random Window is selected by default.                                          |
|                    | Minimum Frequency of Words | Valid values: any non-zero positive integer. Default value: 5.                                                           |
|                    | Hierarchical Softmax       | Specifies whether to use hierarchical softmax. Hierarchical Softmax is selected by default.                              |
|                    | Negative Sampling          | The window size of negative sampling. Default value: 0.                                                                  |
|                    | Downsampling Threshold     | The threshold for downsampling. Default value: 0.                                                                        |
|                    | Initial Learning Rate      | The value is greater than 0. The default value is 0.025.                                                                 |
|                    | Iterations                 | The value is greater than or equal to 1. The default value is 1.                                                         |
| Tuning             | Cores                      | The number of cores used for calculation. The value is automatically allocated.                                          |
|                    | Memory Size per Core       | The size of memory required by each core. The value is automatically allocated.                                          |

- PAI command

```

pai -name Word2Vec
  -project algo_public
  -DinputTableName=w2v_input
  -DwordColName=word
  -DoutputTableName=w2v_output;

```

| Parameter              | Required | Description                                                                                                                                                                                                                                                                                     | Default value                                                        |
|------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| inputTableName         | Yes      | The name of the input vocabulary.                                                                                                                                                                                                                                                               | No default value                                                     |
| inputTablePartitions   | No       | The name of the partitions used for word segmentation. The value must be in the <code>partition_name=value</code> format. To specify multiple partitions, use the following format: <code>name1=value1/name2=value2</code> . If you specify multiple partitions, separate them with commas (,). | No default value                                                     |
| wordColumnName         | Yes      | The name of the word column. Each row in the word column contains only a single word. The <code>&lt;/s&gt;</code> tag indicates a line feed.                                                                                                                                                    | No default value                                                     |
| inVocabularyTableName  | No       | The output of the wordcount operation on the input vocabulary.                                                                                                                                                                                                                                  | The wordcount operation that the system performs on the output table |
| inVocabularyPartitions | No       | The names of the partitions in the output after a wordcount operation is performed on the input vocabulary.                                                                                                                                                                                     | All partitions in the output of inVocabularyTableName                |
| layerSize              | No       | The number of dimensions of the word feature. Valid values: 0 to 1000.                                                                                                                                                                                                                          | 100                                                                  |
| cbow                   | No       | The language model used for training. Valid values: 0 and 1. The value 0 indicates the skip-gram model, and the value 1 indicates the CBOW model.                                                                                                                                               | 0                                                                    |
| window                 | No       | The window size of words. Valid values: any non-zero positive integer.                                                                                                                                                                                                                          | 5                                                                    |
| minCount               | No       | The minimum frequency of words for truncation. Valid values: any non-zero positive integer.                                                                                                                                                                                                     | 5                                                                    |



| Parameter              | Required | Description                                                                                                                                                                                                         | Default value           |
|------------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| hs                     | No       | Specifies whether to use hierarchical softmax. Valid values: 0 and 1. The value 0 indicates that hierarchical softmax is not used, and the value 1 indicates that hierarchical softmax is used.                     | 1                       |
| negative               | No       | The window size of negative sampling. Valid values: any non-zero positive integer.                                                                                                                                  | 0                       |
| sample                 | No       | The threshold for downsampling. Valid values: 1e-3 to 1e-5.                                                                                                                                                         | 0                       |
| alpha                  | No       | The value is greater than 0.                                                                                                                                                                                        | 0.025                   |
| iterTrain              | No       | The value is greater than or equal to 1.                                                                                                                                                                            | 1                       |
| randomWindow           | No       | Specifies the mode that is used to display the word window. Valid values: 0 and 1. The value 1 indicates a random value from 1 to 5, and the value 0 indicates that the value is specified by the window parameter. | 1                       |
| outVocabularyTableName | No       | The name of the output vocabulary.                                                                                                                                                                                  | No default value        |
| outputTableName        | Yes      | The name of the output vector table.                                                                                                                                                                                | No default value        |
| lifecycle              | No       | The lifecycle of the output table. Valid values: any non-zero positive integer.                                                                                                                                     | No default value        |
| coreNum                | No       | The number of cores. This parameter and the memSizePerCore parameter take effect only when they are both configured. Valid values: any non-zero positive integer.                                                   | Automatically allocated |
| memSizePerCore         | No       | The size of the memory required by each core. This parameter and the coreNum parameter take effect only when they are both configured. Valid values: any non-zero positive integer.                                 | Automatically allocated |

## FAQ

The error message "Vocab size is zero! vocab\_size: 0" is reported. This indicates that the dictionary is empty. To solve the issue, set minCount to a smaller value.

## 3.7.17. Doc2Vec

This topic describes the Doc2Vec component provided by Machine Learning Studio.

You can use the Doc2Vec component to map articles to vectors. The input is a vocabulary. The output is a document vector table, a word vector table, or a vocabulary.

You can configure the component by using one of the following methods:

## Machine Learning Platform for AI console

| Tab                | Parameter                   | Description                                                                                                                                |
|--------------------|-----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Document ID Column          | The name of the document column that is used for training.                                                                                 |
|                    | Document Content            | The words used for training. Separate these words by spaces.                                                                               |
| Parameters Setting | Dimensions of Word Features | The number of dimensions of the word feature. Valid values: 0 to 1000. Default value: 100.                                                 |
|                    | Language Model              | The language model used for training. Valid values: <b>Skip-gram Model</b> and <b>CBOW Model</b> . Default value: <b>Skip-gram Model</b> . |
|                    | Window Size of Words        | The window size of words. Valid values: any non-zero positive integer. Default value: 5.                                                   |
|                    | Minimum Frequency of Words  | Valid values: any non-zero positive integer. Default value: 5.                                                                             |
|                    | Hierarchical Softmax        | Specifies whether hierarchical softmax is used. Hierarchical softmax is used by default.                                                   |
|                    | Negative Sampling           | The window size of negative sampling. Valid values: any non-zero positive integer. Default value: 5.                                       |
|                    | Downsampling Threshold      | The threshold for downsampling. Valid values: 1e-3 to 1e-5. Default value: 1e-3.                                                           |
|                    | Initial Learning Rate       | The value is greater than 0. Default value: 0.025.                                                                                         |
|                    | Training Iterations         | The value is greater than or equal to 1. Default value: 1.                                                                                 |

| Tab    | Parameter                       | Description                                                                                                                                                                                                    |
|--------|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | Use Random Window               | Specifies the mode that is used to display the word window. Valid values: <b>A Random Value Between 1 to 5 and Specified by the Window Parameter.</b> Default value: <b>Specified by the Window Parameter.</b> |
| Tuning | Computing Cores                 | The number of cores used for calculation. The value is automatically allocated.                                                                                                                                |
|        | Memory Size per Core (Unit: MB) | The size of memory required by each core. The value is automatically allocated.                                                                                                                                |

## PAI command

```
PAI -name pai_doc2vec
  -project algo_public
  -DinputTableName=d2v_input
  -DdocIdColName=docid
  -DdocColName=text_seg
  -DoutputWordTableName=d2v_word_output
  -DoutputDocTableName=d2v_doc_output;
```

| Parameter            | Required | Description                                                                                                                                                                                                                                                                                                              | Default value    |
|----------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputTableName       | Yes      | The name of the input vocabulary.                                                                                                                                                                                                                                                                                        | No default value |
| inputTablePartitions | No       | The names of the partitions in the input vocabulary, which are used for word segmentation. This value must be in the <b>partition_name=value</b> format. To specify multiple partitions, use the following format: <b>name1=value1/name2=value2</b> . If you specify multiple partitions, separate them with commas (,). | No default value |
| docIdColName         | Yes      | The name of the document column used for training.                                                                                                                                                                                                                                                                       | No default value |

| Parameter  | Required | Description                                                                                                                                                                                 | Default value    |
|------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| docColName | Yes      | The words used for training. Separate these words with spaces.                                                                                                                              | No default value |
| layerSize  | No       | The number of dimensions of the word feature. Valid values: 0 to 1000.                                                                                                                      | 100              |
| cbow       | No       | The language model used for training. Valid values: 0 and 1. The value 0 indicates the skip-gram model, and the value 1 indicates the CBOW model.                                           | 0                |
| window     | No       | The window size of words. Valid values: any non-zero positive integer.                                                                                                                      | 5                |
| minCount   | No       | The minimum frequency of words for truncation. Valid values: any non-zero positive integer.                                                                                                 | 5                |
| hs         | No       | Specifies whether to use hierarchical softmax. Valid values: 0 and 1. The value 0 indicates that hierarchical softmax is not used. The value 1 indicates that hierarchical softmax is used. | 1                |
| negative   | No       | The window size of negative sampling. Valid values: any non-zero positive integer.                                                                                                          | 5                |
| sample     | No       | The threshold for downsampling. Valid values: 1e-3 to 1e-5. Default value: 1e-3.                                                                                                            | 1e-3             |
| alpha      | No       | The value is greater than 0.                                                                                                                                                                | 0.025            |
| iterTrain  | No       | The value is greater than or equal to 1.                                                                                                                                                    | 1                |

| Parameter              | Required | Description                                                                                                                                                                                                         | Default value           |
|------------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| randomWindow           | No       | Specifies the mode that is used to display the word window. Valid values: 0 and 1. The value 1 indicates a random value from 1 to 5, and the value 0 indicates that the value is specified by the window parameter. | 1                       |
| outVocabularyTableName | No       | The name of the output vocabulary.                                                                                                                                                                                  | No default value        |
| outputWordTableName    | Yes      | The name of the output word vector table.                                                                                                                                                                           | No default value        |
| outputDocTableName     | Yes      | The name of the output document vector table.                                                                                                                                                                       | No default value        |
| lifecycle              | No       | The lifecycle of the output table. Valid values: any non-zero positive integer.                                                                                                                                     | No default value        |
| coreNum                | No       | The number of cores. This parameter and the memSizePerCore parameter take effect only when they are both configured. Valid values: any non-zero positive integer.                                                   | Automatically allocated |
| memSizePerCore         | No       | The size of the memory required by each core. This parameter and the coreNum parameter take effect only when they are both configured. Valid values: any non-zero positive integer.                                 | Automatically allocated |

## 3.7.18. Conditional Random Field

This topic describes the Conditional Random Field component provided by Machine Learning Studio.

A conditional random field (CRF) is a conditional probability distribution model of a group of output random variables based on a group of input random variables. This model presumes that the output random variables constitute a Markov random field (MRF). CRFs can be used in different prediction scenarios. The linear chain CRF is mostly used, especially in annotation scenarios. For more information, see [Wikipedia](#).

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                                  | Description                                                                                                                                                                                                                             |
|--------------------|--------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | ID Columns                                 | The column that contains the ID of each sample. Samples are stored in n-tuples.                                                                                                                                                         |
|                    | Feature Columns                            | The word to be annotated and their features (if any).                                                                                                                                                                                   |
|                    | Target Columns                             | The column that you want to select.                                                                                                                                                                                                     |
| Parameters Setting | Feature Generation Template                | Default value:<br><pre>[-2:0],[-1:0],[0:0],[1:0],[2:0],[-1:0]/[0:0], [0:0]/[1:0],[-2:1],[-1:1],[0:1], [1:1],[2:1], [-2:1]/[-1:1],[-1:1]/[0:1], [0:1]/[1:1],[1:1]/[2:1], [-2:1]/[-1:1]/[0:1],[-1:1]/[0:1]/[1:1], [0:1]/[1:1]/[2:1]</pre> |
|                    | Infrequently Used Word Filtering Threshold | Default value: 1.                                                                                                                                                                                                                       |
|                    | L 1 Regularization Coefficient             | Default value: 1.                                                                                                                                                                                                                       |
|                    | L 2 Regularization Coefficient             | Default value: 0.                                                                                                                                                                                                                       |
|                    | Maximum Iterations                         | Default value: 100.                                                                                                                                                                                                                     |

| Tab    | Parameter                    | Description                                                                         |
|--------|------------------------------|-------------------------------------------------------------------------------------|
|        | <b>Convergence Threshold</b> | Default value: 0.00001.                                                             |
| Tuning | <b>Cores</b>                 | The number of cores. The value is automatically allocated.                          |
|        | <b>Memory Size per Core</b>  | The size of the memory required by each core. The value is automatically allocated. |

- PAI command

```
PAI -name=linearcrf
  -project=algo_public
  -DinputTableName=crf_input_table
  -DidColName=sentence_id
  -DfeatureColNames=word,f1
  -DlabelColName=label
  -DoutputTableName=crf_model
  -Dlifecycle=28
  -DcoreNum=10
```

| Parameter             | Required | Description                                                              | Default value                           |
|-----------------------|----------|--------------------------------------------------------------------------|-----------------------------------------|
| inputTableName        | Yes      | The table that contains the input features.                              | No default value                        |
| inputTablePartitions  | No       | The partitions selected from the table that contains the input features. | Full table                              |
| featureColNames       | No       | The feature columns that are selected from the input table.              | All columns, excluding the label column |
| labelColName          | Yes      | The column that you want to select.                                      | No default value                        |
| idColName             | Yes      | The column that contains sample labels.                                  | No default value                        |
| outputTableName       | Yes      | The table that contains output models.                                   | No default value                        |
| outputTablePartitions | No       | The partitions that are selected from the output model table.            | Full table                              |

| Parameter  | Required       | Description                                                                                                 | Default value                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|------------|----------------|-------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| template   | No             | The template that is used to generate features.                                                             | <ul style="list-style-type: none"> <li>Definition <pre>&lt;template . =. &lt;template_item, &lt;template_item, ..., &lt;template_item &lt;template_item . =. [row_offset:col_index]/[row_offset:col_index]/.../[row_offset:col_index] row_offset . =. integer col_index . =. integer</pre> </li> <li>Default value <pre>[-2:0],[-1:0],[0:0], [1:0],[2:0],[-1:0]/[0:0], [0:0]/[1:0],[-2:1], [-1:1],[0:1],[1:1], [2:1],[-2:1]/[-1:1], [-1:1]/[0:1], [0:1]/[1:1], [1:1]/[2:1],[-2:1]/[-1:1]/[0:1],[-1:1]/[0:1]/[1:1], [0:1]/[1:1]/[2:1]</pre> </li> </ul> |
| freq       | No             | The parameter for filtering features. It saves only feature values greater than or equal to the freq value. | 1                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| iterations | No             | The maximum number of iterations of optimizations.                                                          | 100                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| l1Weight   | All partitions | The parameter weight of L1 regularization.                                                                  | 1.0                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| l2Weight   | No             | The parameter weight of L2 regularization.                                                                  | 1.0                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |



| Parameter      | Required | Description                                                                                                                                                                   | Default value           |
|----------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| epsilon        | No       | The convergence deviation. This parameter specifies the requirement to finish the L-BFGS process, which is the deviation between the log-likelihood values in two iterations. | 0.0001                  |
| lbfgsStep      | No       | The historical step size for optimization that is performed by using the L-BFGS algorithm. Only the L-BFGS algorithm supports this parameter.                                 | 10                      |
| threadNum      | No       | The number of parallel threads launched for model training.                                                                                                                   | 3                       |
| lifecycle      | No       | The lifecycle of the output table.                                                                                                                                            | No default value        |
| coreNum        | No       | The number of cores.                                                                                                                                                          | Automatically allocated |
| memSizePerCore | No       | The memory size of each core.                                                                                                                                                 | Automatically allocated |

## Examples

- Input data

| sentence_id | word          | f1  | label |
|-------------|---------------|-----|-------|
| 1           | Rockwell      | NNP | B-NP  |
| 1           | International | NNP | I-NP  |
| 1           | Corp          | NNP | I-NP  |
| 1           | 's            | POS | B-NP  |
| ...         | ...           | ... | ...   |
| 823         | Ohio          | NNP | B-NP  |
| 823         | grew          | VBD | B-VP  |
| 823         | 3.8           | CD  | B-NP  |

| sentence_id | word | f1 | label |
|-------------|------|----|-------|
| 823         | %    | NN | I-NP  |
| 823         | .    | .  | O     |

- Prediction algorithm PAI command

```
PAI -name=crf_predict
  -project=algo_public
  -DinputTableName=crf_test_input_table
  -DmodelTableName=crf_model
  -DIdColName=sentence_id
  -DfeatureColNames=word,f1
  -DlabelColName=label
  -DoutputTableName=crf_predict_result
  -DdetailColName=prediction_detail
  -Dlifecycle=28
  -DcoreNum=10
```

| Parameter             | Required | Description                                                      | Default value                           |
|-----------------------|----------|------------------------------------------------------------------|-----------------------------------------|
| inputTableName        | Yes      | The table that contains the input features.                      | No default value                        |
| inputTablePartitions  | No       | The partitions selected from the input table.                    | Full table                              |
| featureColNames       | No       | The feature columns selected from the input table.               | All columns, excluding the label column |
| labelColName          | No       | The column that you want to select.                              | No default value                        |
| IdColName             | Yes      | The sample label column.                                         | No default value                        |
| resultColName         | No       | The result column in the output table.                           | prediction_result                       |
| scoreColName          | No       | The score column in the output table.                            | prediction_score                        |
| detailColName         | No       | The detail column in the output table.                           | No default value                        |
| outputTableName       | Yes      | The output prediction result table.                              | No default value                        |
| outputTablePartitions | No       | The partitions selected from the output prediction result table. | Full table                              |

| Parameter            | Required | Description                                             | Default value           |
|----------------------|----------|---------------------------------------------------------|-------------------------|
| modelTableName       | Yes      | The algorithm model table.                              | No default value        |
| modelTablePartitions | No       | The partitions selected from the algorithm model table. | Full table              |
| lifecycle            | No       | The lifecycle of the output table.                      | No default value        |
| coreNum              | No       | The number of cores.                                    | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core.                           | Automatically allocated |

- Output data

| sentence_id | word       | f1  | label |
|-------------|------------|-----|-------|
| 1           | Confidence | NN  | B-NP  |
| 1           | in         | IN  | B-PP  |
| 1           | the        | DT  | B-NP  |
| 1           | pound      | NN  | I-NP  |
| ...         | ...        | ... | ...   |
| 77          | have       | VBP | B-VP  |
| 77          | announced  | VCN | I-VP  |
| 77          | similar    | JJ  | B-NP  |
| 77          | increases  | NNS | I-NP  |
| 77          | .          | .   | O     |

 **Note** The label column is optional.

## 3.7.19. Document Similarity

This topic describes the Document Similarity component provided by Machine Learning Studio.

Document similarity is based on string similarity and calculates the similarity between articles or between sentences. Documents or sentences are separated by spaces. The document similarity is calculated the same way as the string similarity. The document similarity supports the following calculation methods: Levenshtein Distance (Levenshtein), Longest Common SubString (LCS), String Subsequence Kernel (SSK), Cosine, and Simhash\_Hamming.

- The Levenshtein method supports distance and similarity calculation.
  - The distance is expressed as the **levenshtein** parameter.
  - The similarity is calculated based on the following formula:  $\text{Similarity} = 1 - \text{Distance}$ . The similarity is expressed as the **levenshtein\_sim** parameter.
- The LCS method supports distance and similarity calculation.
  - The distance is expressed as the **lcs** parameter.
  - The similarity is calculated based on the following formula:  $\text{Similarity} = 1 - \text{Distance}$ . The similarity is expressed as the **lcs\_sim** parameter.
- The SSK method supports similarity calculation and is expressed as the **ssk** parameter.
- The Cosine method supports similarity calculation and is expressed as the **cosine** parameter.
- In Simhash\_Hamming, the SimHash algorithm is to map the original documents to 64-bit binary fingerprints. The Hamming distance is used to calculate the number of characters of binary fingerprints on the same position. The Simhash\_Hamming method supports distance and similarity calculation.
  - The distance is expressed as the **simhash\_hamming** parameter.
  - The similarity is calculated based on the following formula:  $\text{Similarity} = 1 - \text{Distance}/64.0$ . The similarity is expressed as the **simhash\_hamming\_sim** parameter.

 **Note**

- For more information about SimHash, see [Similarity Estimation Techniques from Rounding Algorithms](#).
- For more information about the Hamming distance, see [Wikipedia](#).

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform For AI console

| Tab            | Parameter                                       | Description                                                             |
|----------------|-------------------------------------------------|-------------------------------------------------------------------------|
| Fields Setting | <b>First Column for Similarity Calculation</b>  | The default value is the name of the first string column in the table.  |
|                | <b>Second Column for Similarity Calculation</b> | The default value is the name of the second string column in the table. |
|                | <b>Columns Appended Output Table</b>            | The names of columns appended to the output table.                      |

| Tab                       | Parameter                                                      | Description                                                                                                                                                                                                                                                                                                                                      |
|---------------------------|----------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                           | <b>Similarity Column in Output Table</b>                       | <p>The name of the similarity column in the output table. Default value: output.</p> <div>  <b>Note</b> The column name can be up to 128 characters in length and can contain letters, digits, and underscores (_). It must start with a letter.         </div> |
| <b>Parameters Setting</b> | <b>Similarity Calculation Method</b>                           | <p>The similarity calculation method. Valid values:</p> <ul style="list-style-type: none"> <li>◦ levenshtein</li> <li>◦ levenshtein_sim</li> <li>◦ lcs</li> <li>◦ lcs_sim</li> <li>◦ ssk</li> <li>◦ cosine</li> <li>◦ simhash_hamming</li> <li>◦ simhash_hamming_sim</li> </ul> <p>Default value: levenshtein_sim.</p>                           |
|                           | <b>Substring Length (Available in SSK and Cosine)</b>          | <p>This parameter appears only when the <b>Similarity Calculation Method</b> parameter is set to <b>levenshtein</b>, <b>ssk</b>, or <b>cosine</b>. Valid values: (0,100). Default value: 2.</p>                                                                                                                                                  |
|                           | <b>Matching Word Pair Weight (Available in SSK and Cosine)</b> | <p>This parameter appears only when the <b>Similarity Calculation Method</b> parameter is set to <b>ssk</b>. Valid values: (0,1). Default value: 0.5.</p>                                                                                                                                                                                        |
| <b>Tuning</b>             | <b>Computing Cores</b>                                         | <p>The number of the cores used for calculation. The value is automatically allocated.</p>                                                                                                                                                                                                                                                       |
|                           | <b>Memory Size per Core (Unit: MB)</b>                         | <p>The size of the memory required by each core. The value is automatically allocated.</p>                                                                                                                                                                                                                                                       |

• PAI command

```
PAI -name doc_similarity
-project algo_public
-DinputTableName="pai_test_doc_similarity"
-DoutputTableName="pai_test_doc_similarity_output"
-DinputSelectedColName1="col0"
-DinputSelectedColName2="col1"
```

| Parameter      | Required | Description                  | Default value    |
|----------------|----------|------------------------------|------------------|
| inputTableName | Yes      | The name of the input table. | No default value |

| Parameter              | Required | Description                                                                                                                                                                                                                                                                                                                 | Default value                                     |
|------------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| outputTableName        | Yes      | The name of the output table.                                                                                                                                                                                                                                                                                               | No default value                                  |
| inputSelectedColName 1 | No       | The name of the first column for similarity calculation.                                                                                                                                                                                                                                                                    | The name of the first string column in the table  |
| inputSelectedColName 2 | No       | The name of the second column for similarity calculation.                                                                                                                                                                                                                                                                   | The name of the second string column in the table |
| inputAppendColNames    | No       | The name of the column appended to the output table.                                                                                                                                                                                                                                                                        | No column appended by default                     |
| inputTablePartitions   | No       | The partitions that are selected from the input table.                                                                                                                                                                                                                                                                      | Full table                                        |
| outputColName          | No       | <p>The name of the similarity column in the output table.</p> <div>  <b>Note</b> The column name can be up to 128 characters in length and can contain letters, digits, and underscores (_). It must start with a letter.         </div> | output                                            |
| method                 | No       | <p>The similarity calculation method. Valid values:</p> <ul style="list-style-type: none"> <li>◦ levenshtein</li> <li>◦ levenshtein_sim</li> <li>◦ lcs</li> <li>◦ lcs_sim</li> <li>◦ ssk</li> <li>◦ cosine</li> <li>◦ simhash_hamming</li> <li>◦ simhash_hamming_sim</li> </ul>                                             | levenshtein_sim                                   |

| Parameter      | Required | Description                                                                                            | Default value           |
|----------------|----------|--------------------------------------------------------------------------------------------------------|-------------------------|
| lambda         | No       | The weight of a matched word pair. The SSK method supports this parameter. Valid values: (0,1).        | 0.5                     |
| k              | No       | The length of the substring. The SSK and Cosine methods support this parameter. Valid values: (0,100). | 2                       |
| lifecycle      | No       | The lifecycle of the output table.                                                                     | No default value        |
| coreNum        | No       | The number of the cores used for calculation.                                                          | Automatically allocated |
| memSizePerCore | No       | The size of the memory required by each core. Unit: MB.                                                | Automatically allocated |

## Examples

- Generate data.

```
drop table if exists pai_doc_similarity_input;
create table pai_doc_similarity_input as
select * from
(
select 0 as id, "Beijing Shanghai" as col0, "Beijing Shanghai" as col1 from dual
union all
select 1 as id, "Beijing Shanghai" as col0, "Beijing Shanghai Hong Kong (China)" as col1 from dual
)tmp
```

The following table is the input table pai\_doc\_similarity\_input.

| id | col0             | col1                               |
|----|------------------|------------------------------------|
| 1  | Beijing Shanghai | Beijing Shanghai Hong Kong (China) |
| 0  | Beijing Shanghai | Beijing Shanghai                   |

- Run the PAI command.

```
drop table if exists pai_doc_similarity_output;
PAI -name doc_similarity
  -project algo_public
  -DinputTableName=pai_doc_similarity_input
  -DoutputTableName=pai_doc_similarity_output
  -DinputSelectedColName1=col0
  -DinputSelectedColName2=col1
  -Dmethod=levenshtein_sim
  -DinputAppendColNames=id,col0,col1;
```

- Generate the output.

The following table is the output table pai\_doc\_similarity\_output.

| id | col0             | col1                               | output             |
|----|------------------|------------------------------------|--------------------|
| 1  | Beijing Shanghai | Beijing Shanghai Hong Kong (China) | 0.6666666666666667 |
| 0  | Beijing Shanghai | Beijing Shanghai                   | 1.0                |

## FAQ

- Similarity calculation is based on the result of word segmentation. Words are separated by spaces, and each word serves as a unit of similarity calculation. If the input is a string as a whole, use the string similarity method.
- In the method parameter, levenshtein, lcs, and simhash\_hamming are used to calculate the distance. levenshtein\_sim, lcs\_sim, ssk, cosine, and simhash\_hamming\_sim are used to calculate the similarity. The distance is calculated based on the following formula: Distance = 1.0 - Similarity.
- When the method parameter is set to cosine or ssk, the k parameter is available, indicating that k words are used as a combination for similarity calculation. If the value of k is greater than the number of words, two strings are the same. The similarity output is 0. In this case, you need to change the value of k to a value less than or equal to the minimum number of words.

## 3.7.20. PMI

This topic describes the Pointwise Mutual Information (PMI) component provided by Machine Learning Studio.

Mutual information (MI) is a measure of information in the information theory. It can be regarded as the amount of information that is contained in a random variable about another variable, or the reduction in uncertainty of a random variable due to the known random variable.

This algorithm is used to count the co-occurrence of all words in several documents and calculate the pointwise mutual information (PMI). PMI definition:  $PMI(x,y) = \ln(p(x,y)/(p(x)p(y))) = \ln(\#(x,y)D/(\#x\#y))$ . In the definition,  $\#(x,y)$  indicates the number of pairs  $(x,y)$ . D indicates the total number of pairs. If x and y appear in the same window, the output is  $\#x+=1$ ;  $\#y+=1$ ;  $\#(x,y)+=1$ . For more information about PMI, see [PMI](#).

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console



| Tab                | Parameter                                             | Description                                                                                                                                                                               |
|--------------------|-------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Columns of Documents with Words Separated with Spaces | None                                                                                                                                                                                      |
| Parameters Setting | Minimum Frequency of Words                            | Words that appear for a number of times less than this value are filtered out. Default value: 5.                                                                                          |
|                    | Window Size                                           | The window size. For example, the value 5 refers to the five words adjacent to the right of the current word. Words that appear in the window are considered related to the current word. |
| Tuning             | Computing Cores                                       | The number of cores used for calculation. The value is automatically allocated.                                                                                                           |
|                    | Memory Size per Core (Unit: MB)                       | The size of memory required by each core. The value is automatically allocated.                                                                                                           |

- PAI command

```
PAI -name PointwiseMutualInformation
  -project algo_public
  -DinputTableName=maple_test_pmi_basic_input
  -DdocColName=doc
  -DoutputTableName=maple_test_pmi_basic_output
  -DminCount=0
  -DwindowSize=2
  -DcoreNum=1
  -DmemSizePerCore=110;
```

| Parameter       | Required | Description                                                                                     | Default value    |
|-----------------|----------|-------------------------------------------------------------------------------------------------|------------------|
| inputTableName  | Yes      | The input table.                                                                                | No default value |
| outputTableName | Yes      | The output table.                                                                               | No default value |
| docColName      | Yes      | The name of the document column after word segmentation, where words are separated with spaces. | No default value |

| Parameter            | Required | Description                                                                                                                                                                                                                       | Default value           |
|----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| windowSize           | No       | The window size. For example, the value 5 refers to the five words adjacent to the right of the current word. Words that appear in the window are considered related to the current word.                                         | All content in a row    |
| minCount             | No       | The minimum frequency of words for truncation. Words that appear for a number of times less than this value are filtered out.                                                                                                     | 5                       |
| inputTablePartitions | No       | The partitions selected from the input table for training, in the format of Partition_name=value. To specify multiple partitions, use the following format: name1=value1/name2=value2. Separate these partitions with commas (,). | Full table              |
| lifecycle            | No       | The lifecycle of the output table.                                                                                                                                                                                                | No default value        |
| coreNum              | No       | The number of cores. Valid values: [1,9999].                                                                                                                                                                                      | Automatically allocated |
| memSizePerCore       | No       | The memory size of each core. Unit: MB. Valid values: [1024,65536].                                                                                                                                                               | Automatically allocated |

## Examples

- Generate data.

```
create table maple_test_pmi_basic_input as
select * from
(
  select "w1 w2 w3 w4 w5 w6 w7 w8 w8 w9" as doc from dual
  union all
  select "w1 w3 w5 w6 w9" as doc from dual
  union all select "w0" as doc from dual
  union all
  select "w0 w0" as doc from dual
  union all
  select "w9 w1 w9 w1 w9" as doc from dual
)tmp;
```

| doc:string                    |
|-------------------------------|
| w1 w2 w3 w4 w5 w6 w7 w8 w8 w9 |
| w1 w3 w5 w6 w9                |
| w0                            |
| w0 w0                         |
| w9 w1 w9 w1 w9                |

- Run the PAI command.

```
PAI -name PointwiseMutualInformation
  -project algo_public
  -DinputTableName=maple_test_pmi_basic_input
  -DdocColName=doc
  -DoutputTableName=maple_test_pmi_basic_output
  -DminCount=0
  -DwindowSize=2
  -DcoreNum=1
  -DmemSizePerCore=110;
```

- Generate the output.

| word1 | word2 | word1_count | word2_count | co_occurrence<br>s_count | pmi                 |
|-------|-------|-------------|-------------|--------------------------|---------------------|
| w0    | w0    | 2           | 2           | 1                        | 2.0794415416798357  |
| w1    | w1    | 10          | 10          | 1                        | -1.1394342831883648 |
| w1    | w2    | 10          | 3           | 1                        | 0.06453852113757116 |

| word1 | word2 | word1_count | word2_count | co_occurrences_count | pmi                      |
|-------|-------|-------------|-------------|----------------------|--------------------------|
| w1    | w3    | 10          | 7           | 2                    | -<br>0.08961215868968704 |
| w1    | w5    | 10          | 8           | 1                    | -<br>0.916290731874155   |
| w1    | w9    | 10          | 12          | 4                    | 0.06453852113757116      |
| w2    | w3    | 3           | 7           | 1                    | 0.4212134650763035       |
| w2    | w4    | 3           | 4           | 1                    | 0.9808292530117262       |
| w3    | w4    | 7           | 4           | 1                    | 0.13353139262452257      |
| w3    | w5    | 7           | 8           | 2                    | 0.13353139262452257      |
| w3    | w6    | 7           | 7           | 1                    | -<br>0.42608439531090014 |
| w4    | w5    | 4           | 8           | 1                    | 0                        |
| w4    | w6    | 4           | 7           | 1                    | 0.13353139262452257      |
| w5    | w6    | 8           | 7           | 2                    | 0.13353139262452257      |
| w5    | w7    | 8           | 4           | 1                    | 0                        |
| w5    | w9    | 8           | 12          | 1                    | -<br>1.0986122886681098  |
| w6    | w7    | 7           | 4           | 1                    | 0.13353139262452257      |
| w6    | w8    | 7           | 7           | 1                    | -<br>0.42608439531090014 |
| w6    | w9    | 7           | 12          | 1                    | -<br>0.9650808960435872  |

| word1 | word2 | word1_count | word2_count | co_occurrences_count | pmi                  |
|-------|-------|-------------|-------------|----------------------|----------------------|
| w7    | w8    | 4           | 7           | 2                    | 0.8266785731844679   |
| w8    | w8    | 7           | 7           | 1                    | -0.42608439531090014 |
| w8    | w9    | 7           | 12          | 2                    | -0.2719337154836418  |
| w9    | w9    | 12          | 12          | 2                    | -0.8109302162163288  |

## 3.8. Network analysis

### 3.8.1. k-Core

The k-core of a graph is the subgraph that remains after all vertices with a degree less than or equal to K are removed. If a vertex belongs to the k-core but is not included in the (k+1)-core, the coreness of the vertex is k. Therefore, the coreness of a vertex whose degree is 1 must be 0. The largest coreness among the corenesses of all vertices is considered to be the coreness of the graph. This topic describes the K-Core component provided by Machine Learning Studio.

You can configure the component by using one of the following methods:

#### Machine Learning Platform for AI console

| Tab                 | Parameter                   | Description                                                                                                                                                                                                       |
|---------------------|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Settings     | Source Vertex Column        | The start vertex column in the edge table.                                                                                                                                                                        |
|                     | Target Vertex Column        | The end vertex column in the edge table.                                                                                                                                                                          |
| Parameters Settings | k Cores                     | The value of the coreness. Default value: 3. This parameter is required.                                                                                                                                          |
| Tuning              | Workers                     | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                     | Memory Size Per Worker (MB) | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |

## PAI command

```
PAI -name KCore
    -project algo_public
    -DinputEdgeTableName=KCore_func_test_edge
    -DfromVertexCol=flow_out_id
    -DtoVertexCol=flow_in_id
    -DoutputTableName=KCore_func_test_result
    -Dk=2;
```

| Parameter                | Required | Description                                                                                                                                                                                                       | Default value    |
|--------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.                                                                                                                                                                                 | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table.                                                                                                                                                                           | Full table       |
| fromVertexCol            | Yes      | The start vertex column in the input edge table.                                                                                                                                                                  | No default value |
| toVertexCol              | Yes      | The end vertex column in the input edge table.                                                                                                                                                                    | No default value |
| outputTableName          | Yes      | The name of the output table.                                                                                                                                                                                     | No default value |
| outputTablePartitions    | No       | The partitions in the output table.                                                                                                                                                                               | No default value |
| lifecycle                | No       | The lifecycle of the output table.                                                                                                                                                                                | No default value |
| workerNum                | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured   |
| workerMem                | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096             |
| splitSize                | No       | The data split size.                                                                                                                                                                                              | 64               |

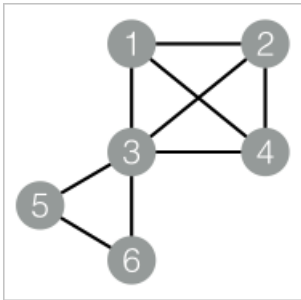
| Parameter | Required | Description             | Default value |
|-----------|----------|-------------------------|---------------|
| k         | Yes      | The number of coreness. | 3             |

## Examples

1. Generate training data.

```
drop table if exists KCore_func_test_edge;
create table KCore_func_test_edge as
select * from
(
  select '1' as flow_out_id,'2' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'3' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'4' as flow_in_id from dual
  union all
  select '2' as flow_out_id,'3' as flow_in_id from dual
  union all
  select '2' as flow_out_id,'4' as flow_in_id from dual
  union all
  select '3' as flow_out_id,'4' as flow_in_id from dual
  union all
  select '3' as flow_out_id,'5' as flow_in_id from dual
  union all
  select '3' as flow_out_id,'6' as flow_in_id from dual
  union all
  select '5' as flow_out_id,'6' as flow_in_id from dual
)tmp;
```

The following figure shows the structure of the k-core graph.



2. Set k to 2 and view training results.

```

+-----+-----+
| node1 | node2 |
+-----+-----+
| 1 | 2 |
| 1 | 3 |
| 1 | 4 |
| 2 | 1 |
| 2 | 3 |
| 2 | 4 |
| 3 | 1 |
| 3 | 2 |
| 3 | 4 |
| 4 | 1 |
| 4 | 2 |
| 4 | 3 |
+-----+-----+

```

### 3.8.2. Single-source Shortest Path

The single-source shortest path (SSSP) refers to the shortest path between a vertex and all other vertices. The shortest path is calculated by using the Dijkstra algorithm. This topic describes the Single-source Shortest Path component provided by Machine Learning Studio.

You can configure the component by using one of the following methods:

#### Machine Learning Platform for AI console

| Tab                 | Parameter                          | Description                                                                                                                                                                                                       |
|---------------------|------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Settings     | <b>Source Vertex Column</b>        | The start vertex column in the edge table.                                                                                                                                                                        |
|                     | <b>Target Vertex Column</b>        | The end vertex column in the edge table.                                                                                                                                                                          |
|                     | <b>Edge Weight Column</b>          | The edge weight column in the edge table.                                                                                                                                                                         |
| Parameters Settings | <b>Start Vertex ID</b>             | The start vertex that is used to calculate the shortest path. This parameter is required.                                                                                                                         |
| Tuning              | <b>Workers</b>                     | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                     | <b>Memory Size Per Worker (MB)</b> | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |

#### PAI command



#### PAI -name SSSP

```
-project algo_public
-DinputEdgeTableName=SSSP_func_test_edge
-DfromVertexCol=flow_out_id
-DtoVertexCol=flow_in_id
-DoutputTableName=SSSP_func_test_result
-DhasEdgeWeight=true
-DedgeWeightCol=edge_weight
-DstartVertex=a;
```

| Parameter                | Required | Description                                                                                                                                                                                                       | Default value    |
|--------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.                                                                                                                                                                                 | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table.                                                                                                                                                                           | Full table       |
| fromVertexCol            | Yes      | The start vertex column in the input edge table.                                                                                                                                                                  | No default value |
| toVertexCol              | Yes      | The end vertex column in the input edge table.                                                                                                                                                                    | No default value |
| outputTableName          | Yes      | The name of the output table.                                                                                                                                                                                     | No default value |
| outputTablePartitions    | No       | The partitions in the output table.                                                                                                                                                                               | No default value |
| lifecycle                | No       | The lifecycle of the output table.                                                                                                                                                                                | No default value |
| workerNum                | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured   |
| workerMem                | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096             |

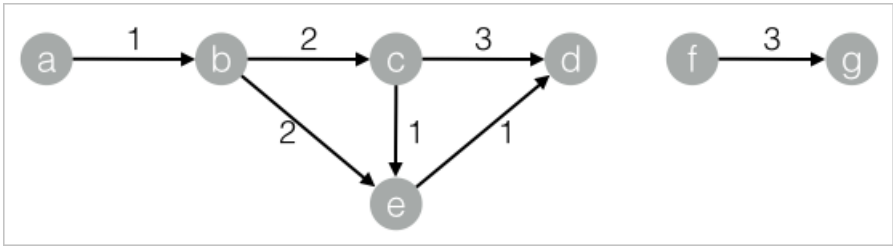
| Parameter     | Required | Description                                                       | Default value    |
|---------------|----------|-------------------------------------------------------------------|------------------|
| splitSize     | No       | The data split size.                                              | 64               |
| startVertex   | Yes      | The ID of the start vertex.                                       | No default value |
| hasEdgeWeight | No       | Specifies whether the edges in the input edge table have weights. | false            |
| edgeWeightCol | No       | The edge weight column in the input edge table.                   | No default value |

Examples

1. Generate training data.

```
drop table if exists SSSP_func_test_edge;
create table SSSP_func_test_edge as
select
  flow_out_id,flow_in_id,edge_weight
from
(
  select "a" as flow_out_id,"b" as flow_in_id,1.0 as edge_weight from dual
union all
  select "b" as flow_out_id,"c" as flow_in_id,2.0 as edge_weight from dual
union all
  select "c" as flow_out_id,"d" as flow_in_id,1.0 as edge_weight from dual
union all
  select "b" as flow_out_id,"e" as flow_in_id,2.0 as edge_weight from dual
union all
  select "e" as flow_out_id,"d" as flow_in_id,1.0 as edge_weight from dual
union all
  select "c" as flow_out_id,"e" as flow_in_id,1.0 as edge_weight from dual
union all
  select "f" as flow_out_id,"g" as flow_in_id,3.0 as edge_weight from dual
union all
  select "a" as flow_out_id,"d" as flow_in_id,4.0 as edge_weight from dual
) tmp;
```

The following figure shows the structure of the SSSP graph.



2. View training results.

```

+-----+-----+-----+-----+
|start_node|dest_node|distance|distance_cnt|
+-----+-----+-----+-----+
|a|b|1.0|1|
|a|c|3.0|1|
|a|d|4.0|3|
|a|a|0.0|0|
|a|e|3.0|1|
+-----+-----+-----+-----+

```

### 3.8.3. Page Rank

PageRank is an algorithm that is used to sort and calculate the rankings of web pages based on their link sources. This topic describes the Page Rank component provided by Machine Learning Studio.

#### Background information

The basic idea behind PageRank is that:

- A larger number of web pages that direct to a web page indicates higher importance or quality of the web page.
- In addition to the number of links that direct to a web page, the weight of the web page and the number of external links are considered during page ranking.
- For a social network of users, the edge weight is an important factor in addition to the influence of the users.

For example, a Sina Weibo user is more likely to have influence on their family, friends, classmates, and colleagues than they have on followers with a weaker relationship. In a social network, the edge weight is equivalent to the user-to-user relationship strength index.

The PageRank formula with the link weight:

$$W(A) = (1 - d) + d * (\sum_i W(i) * C(Ai))$$

- $W(i)$ : indicates the weight of Node  $i$ .
- $C(Ai)$ : indicates the link weight.
- $d$ : indicates the damping coefficient.
- $W(A)$ : indicates the influence index of each user and represents the node weight after the algorithm iteration becomes stable.

You can configure the component by using one of the following methods:

#### Machine Learning Platform for AI console

| Tab             | Parameter            | Description                                |
|-----------------|----------------------|--------------------------------------------|
| Fields Settings | Source Vertex Column | The start vertex column in the edge table. |
|                 | Target Vertex Column | The end vertex column in the edge table.   |

| Tab                        | Parameter                          | Description                                                                                                                                                                                                       |
|----------------------------|------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                            | <b>Edge Weight Column</b>          | The edge weight column in the edge table.                                                                                                                                                                         |
| <b>Parameters Settings</b> | <b>Maximum Iterations</b>          | The number of iterations before the algorithm automatically converges. Default value: 30. This parameter is optional.                                                                                             |
|                            | <b>Damping Coefficient</b>         | The probability that a user continues browsing.                                                                                                                                                                   |
| <b>Tuning</b>              | <b>Workers</b>                     | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                            | <b>Memory Size Per Worker (MB)</b> | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |

## PAI command

```
PAI -name PageRankWithWeight
-project algo_public
-DinputEdgeTableName=PageRankWithWeight_func_test_edge
-DfromVertexCol=flow_out_id
-DtoVertexCol=flow_in_id
-DoutputTableName=PageRankWithWeight_func_test_result
-DhasEdgeWeight=true
-DedgeWeightCol=weight
-DmaxIter 100;
```

| Parameter                | Required | Description                                      | Default value    |
|--------------------------|----------|--------------------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.                | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table.          | Full table       |
| fromVertexCol            | Yes      | The start vertex column in the input edge table. | No default value |
| toVertexCol              | Yes      | The end vertex column in the input edge table.   | No default value |
| outputTableName          | Yes      | The name of the output table.                    | No default value |
| outputTablePartitions    | No       | The partitions in the output table.              | No default value |

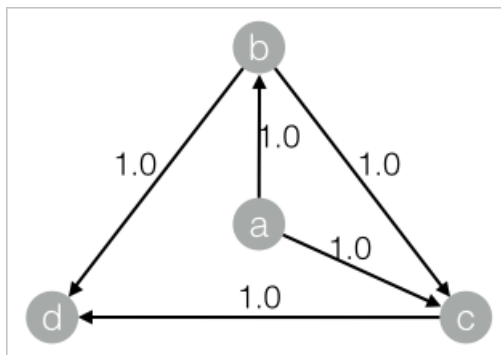
| Parameter     | Required | Description                                                                                                                                                                                                       | Default value    |
|---------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| lifecycle     | No       | The lifecycle of the output table.                                                                                                                                                                                | No default value |
| workerNum     | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured   |
| workerMem     | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096             |
| splitSize     | No       | The data split size.                                                                                                                                                                                              | 64               |
| hasEdgeWeight | No       | Specifies whether the edges in the input edge table have weights.                                                                                                                                                 | false            |
| edgeWeightCol | No       | The edge weight column in the input edge table.                                                                                                                                                                   | No default value |
| maxIter       | No       | The maximum number of iterations.                                                                                                                                                                                 | 30               |

## Examples

1. Generate training data.

```
drop table if exists PageRankWithWeight_func_test_edge;
create table PageRankWithWeight_func_test_edge as
select * from
(
  select 'a' as flow_out_id,'b' as flow_in_id,1.0 as weight from dual
  union all
  select 'a' as flow_out_id,'c' as flow_in_id,1.0 as weight from dual
  union all
  select 'b' as flow_out_id,'c' as flow_in_id,1.0 as weight from dual
  union all
  select 'b' as flow_out_id,'d' as flow_in_id,1.0 as weight from dual
  union all
  select 'c' as flow_out_id,'d' as flow_in_id,1.0 as weight from dual
)tmp;
```

The following figure shows the structure of the PageRank graph.



## 2. View training results.

```
+-----+-----+
| node | weight |
+-----+-----+
| a | 0.0375 |
| b | 0.06938 |
| c | 0.12834 |
| d | 0.20556 |
+-----+-----+
```

## 3.8.4. Label Propagation Clustering

A label propagation algorithm (LPA) is a semi-supervised machine learning algorithm. The labels of a node (community) depend on those of the neighboring nodes. The degree of dependence is determined by the similarity between nodes. Data becomes stable by iterative propagation updates. This topic describes the Label Propagation Clustering component provided by Machine Learning Studio.

### Background information

Graph clustering is used to divide a graph into subgraphs based on the topology of the graph. Therefore, the links between the nodes in a subgraph are more than the links between the subgraphs.

You can configure the component by using one of the following methods:

### Machine Learning Platform for AI console

| Tab                 | Parameter                               | Description                                                                                                                                                                                                       |
|---------------------|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Settings     | <b>Vertex Table: Vertex Column</b>      | The vertex column in the vertex table.                                                                                                                                                                            |
|                     | <b>Vertex Table: Weight Column</b>      | The vertex weight column in the vertex table.                                                                                                                                                                     |
|                     | <b>Edge Table: Source Vertex Column</b> | The start vertex column in the edge table.                                                                                                                                                                        |
|                     | <b>Edge Table: Target Vertex Column</b> | The end vertex column in the edge table.                                                                                                                                                                          |
|                     | <b>Edge Table: Weight Column</b>        | The edge weight column in the edge table.                                                                                                                                                                         |
| Parameters Settings | <b>Maximum Iterations</b>               | The default value is 30. This parameter is optional.                                                                                                                                                              |
| Tuning              | <b>Workers</b>                          | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                     | <b>Memory Size Per Worker (MB)</b>      | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |

## PAI command

```
PAI -name LabelPropagationClustering
-project algo_public
-DinputEdgeTableName=LabelPropagationClustering_func_test_edge
-DfromVertexCol=flow_out_id
-DtoVertexCol=flow_in_id
-DinputVertexTableName=LabelPropagationClustering_func_test_node
-DvertexCol=node
-DoutputTableName=LabelPropagationClustering_func_test_result
-DhasEdgeWeight=true
-DedgeWeightCol=edge_weight
-DhasVertexWeight=true
-DvertexWeightCol=node_weight
-DrandSelect=true
-DmaxIter=100;
```

| Parameter          | Required | Description                       | Default value    |
|--------------------|----------|-----------------------------------|------------------|
| inputEdgeTableName | Yes      | The name of the input edge table. | No default value |

| Parameter                  | Required | Description                                                                                                                                                                                                       | Default value    |
|----------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputEdgeTablePartitions   | No       | The partitions in the input edge table.                                                                                                                                                                           | Full table       |
| fromVertexCol              | Yes      | The start vertex column in the input edge table.                                                                                                                                                                  | No default value |
| toVertexCol                | Yes      | The end vertex column in the input edge table.                                                                                                                                                                    | No default value |
| inputVertexTableName       | Yes      | The name of the input vertex table.                                                                                                                                                                               | No default value |
| inputVertexTablePartitions | No       | The partitions in the input vertex table.                                                                                                                                                                         | Full table       |
| vertexCol                  | Yes      | The vertex column in the input vertex table.                                                                                                                                                                      | No default value |
| outputTableName            | Yes      | The name of the output table.                                                                                                                                                                                     | No default value |
| outputTablePartitions      | No       | The partitions in the output table.                                                                                                                                                                               | No default value |
| lifecycle                  | No       | The lifecycle of the output table.                                                                                                                                                                                | No default value |
| workerNum                  | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured   |
| workerMem                  | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096             |
| splitSize                  | No       | The data split size.                                                                                                                                                                                              | 64               |
| hasEdgeWeight              | No       | Specifies whether the edges in the input edge table have weights.                                                                                                                                                 | false            |



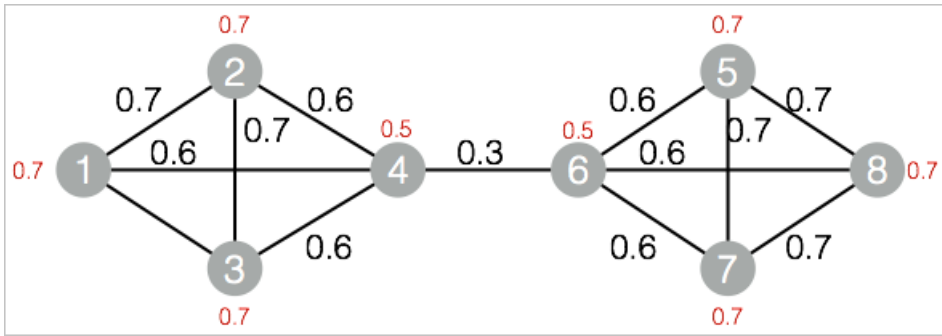
| Parameter       | Required | Description                                                            | Default value    |
|-----------------|----------|------------------------------------------------------------------------|------------------|
| edgeWeightCol   | No       | The edge weight column in the input edge table.                        | No default value |
| hasVertexWeight | No       | Specifies whether the vertices in the input vertex table have weights. | false            |
| vertexWeightCol | No       | The vertex weight column in the input vertex table.                    | No default value |
| randSelect      | No       | Specifies whether the maximum label value is to be randomly selected.  | false            |
| maxIter         | No       | The maximum number of iterations.                                      | 30               |

## Examples

1. Generate training data.

```
drop table if exists LabelPropagationClustering_func_test_edge;
create table LabelPropagationClustering_func_test_edge as
select * from
(
  select '1' as flow_out_id,'2' as flow_in_id,0.7 as edge_weight from dual
  union all
  select '1' as flow_out_id,'3' as flow_in_id,0.7 as edge_weight from dual
  union all
  select '1' as flow_out_id,'4' as flow_in_id,0.6 as edge_weight from dual
  union all
  select '2' as flow_out_id,'3' as flow_in_id,0.7 as edge_weight from dual
  union all
  select '2' as flow_out_id,'4' as flow_in_id,0.6 as edge_weight from dual
  union all
  select '3' as flow_out_id,'4' as flow_in_id,0.6 as edge_weight from dual
  union all
  select '4' as flow_out_id,'6' as flow_in_id,0.3 as edge_weight from dual
  union all
  select '5' as flow_out_id,'6' as flow_in_id,0.6 as edge_weight from dual
  union all
  select '5' as flow_out_id,'7' as flow_in_id,0.7 as edge_weight from dual
  union all
  select '5' as flow_out_id,'8' as flow_in_id,0.7 as edge_weight from dual
  union all
  select '6' as flow_out_id,'7' as flow_in_id,0.6 as edge_weight from dual
  union all
  select '6' as flow_out_id,'8' as flow_in_id,0.6 as edge_weight from dual
  union all
  select '7' as flow_out_id,'8' as flow_in_id,0.7 as edge_weight from dual
)tmp
;
drop table if exists LabelPropagationClustering_func_test_node;
create table LabelPropagationClustering_func_test_node as
select * from
(
  select '1' as node,0.7 as node_weight from dual
  union all
  select '2' as node,0.7 as node_weight from dual
  union all
  select '3' as node,0.7 as node_weight from dual
  union all
  select '4' as node,0.5 as node_weight from dual
  union all
  select '5' as node,0.7 as node_weight from dual
  union all
  select '6' as node,0.5 as node_weight from dual
  union all
  select '7' as node,0.7 as node_weight from dual
  union all
  select '8' as node,0.7 as node_weight from dual
)tmp;
```

The following figure shows the structure of the label propagation clustering graph.



2. View training results.

| +-----+-----+   |   |  |
|-----------------|---|--|
| node   group_id |   |  |
| +-----+-----+   |   |  |
| 1               | 1 |  |
| 2               | 1 |  |
| 3               | 1 |  |
| 4               | 1 |  |
| 5               | 5 |  |
| 6               | 5 |  |
| 7               | 5 |  |
| 8               | 5 |  |
| +-----+-----+   |   |  |

### 3.8.5. Label Propagation Classification

Label propagation classification is a semi-supervised classification algorithm. It uses the label information of labeled nodes to predict the label information for unlabeled nodes. This topic describes the Label Propagation Classification component provided by Machine Learning Studio.

#### Background information

The basic idea behind the label propagation classification algorithm is that: During algorithm execution, the labels of each node are propagated to the neighboring nodes based on the similarity between the nodes. In each step of propagation, a node updates its labels based on the labels of the neighboring nodes. A higher similarity indicates a higher labeling influence that the neighboring nodes have on the node. In this case, the labels are easy to be propagated. During label propagation, the labels of the labeled data remain unchanged. These labels serve as sources for propagation to the unlabeled data. After the iterations end, the probability distributions of similar nodes tend to be similar. These nodes can be classified into the same category. This completes the label propagation.

You can configure the component by using one of the following methods:

#### Machine Learning Platform for AI console

| Tab | Parameter                          | Description                            |
|-----|------------------------------------|----------------------------------------|
|     | <b>Vertex Table: Vertex Column</b> | The vertex column in the vertex table. |
|     |                                    |                                        |

| Tab                 | Parameter                               | Description                                                                                                                                                                                                       |
|---------------------|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Settings     | <b>Vertex Table: Label Column</b>       | The vertex label column in the vertex table.                                                                                                                                                                      |
|                     | <b>Vertex Table: Weight Column</b>      | The vertex weight column in the vertex table.                                                                                                                                                                     |
|                     | <b>Edge Table: Source Vertex Column</b> | The start vertex column in the edge table.                                                                                                                                                                        |
|                     | <b>Edge Table: Target Vertex Column</b> | The end vertex column in the edge table.                                                                                                                                                                          |
|                     | <b>Edge Table: Weight Column</b>        | The edge weight column in the edge table.                                                                                                                                                                         |
| Parameters Settings | <b>Maximum Iterations</b>               | The maximum number of iterations. Default value: 30. This parameter is optional.                                                                                                                                  |
|                     | <b>Damping Coefficient</b>              | The damping coefficient. Default value: 0.8.                                                                                                                                                                      |
|                     | <b>Convergence Coefficient</b>          | The convergence coefficient. Default value: 0.000001.                                                                                                                                                             |
| Tuning              | <b>Workers</b>                          | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                     | <b>Memory Size Per Worker (MB)</b>      | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |

## PAI command

```
PAI -name LabelPropagationClassification
-project algo_public
-DinputEdgeTableName=LabelPropagationClassification_func_test_edge
-DfromVertexCol=flow_out_id
-DtoVertexCol=flow_in_id
-DinputVertexTableName=LabelPropagationClassification_func_test_node
-DvertexCol=node
-DvertexLabelCol=label
-DvertexWeightCol=label_weight
-DoutputTableName=LabelPropagationClassification_func_test_result
-DhasEdgeWeight=true
-DedgeWeightCol=edge_weight
-DhasVertexWeight=true
-DvertexWeightCol=label_weight
-Dalpha=0.8
-Depsilon=0.000001;
```

| Parameter                  | Required | Description                                                                                                                                                                                                       | Default value    |
|----------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputEdgeTableName         | Yes      | The name of the input edge table.                                                                                                                                                                                 | No default value |
| inputEdgeTablePartitions   | No       | The partitions in the input edge table.                                                                                                                                                                           | Full table       |
| fromVertexCol              | Yes      | The start vertex column in the input edge table.                                                                                                                                                                  | No default value |
| toVertexCol                | Yes      | The end vertex column in the input edge table.                                                                                                                                                                    | No default value |
| inputVertexTableName       | Yes      | The name of the input vertex table.                                                                                                                                                                               | No default value |
| inputVertexTablePartitions | No       | The partitions in the input vertex table.                                                                                                                                                                         | Full table       |
| vertexCol                  | Yes      | The vertex column in the input vertex table.                                                                                                                                                                      | No default value |
| outputTableName            | Yes      | The name of the output table.                                                                                                                                                                                     | No default value |
| outputTablePartitions      | No       | The partitions in the output table.                                                                                                                                                                               | No default value |
| lifecycle                  | No       | The lifecycle of the output table.                                                                                                                                                                                | No default value |
| workerNum                  | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured   |
| workerMem                  | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096             |
| splitSize                  | No       | The data split size.                                                                                                                                                                                              | 64               |

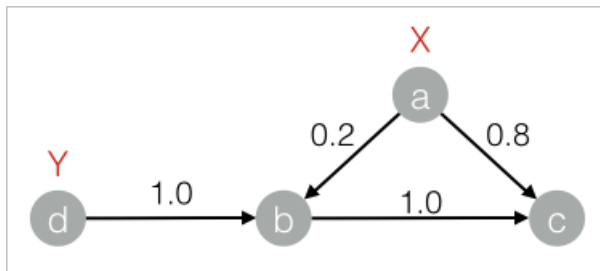
| Parameter       | Required | Description                                                            | Default value    |
|-----------------|----------|------------------------------------------------------------------------|------------------|
| hasEdgeWeight   | No       | Specifies whether the edges in the input edge table have weights.      | false            |
| edgeWeightCol   | No       | The edge weight column in the input edge table.                        | No default value |
| hasVertexWeight | No       | Specifies whether the vertices in the input vertex table have weights. | false            |
| vertexWeightCol | No       | The vertex weight column in the input vertex table.                    | No default value |
| alpha           | No       | The damping coefficient.                                               | 0.8              |
| epsilon         | No       | The convergence coefficient.                                           | 0.000001         |
| maxIter         | No       | The maximum number of iterations.                                      | 30               |

## Examples

### 1. Generate training data.

```
drop table if exists LabelPropagationClassification_func_test_edge;
create table LabelPropagationClassification_func_test_edge as
select * from
(
  select 'a' as flow_out_id, 'b' as flow_in_id, 0.2 as edge_weight from dual
  union all
  select 'a' as flow_out_id, 'c' as flow_in_id, 0.8 as edge_weight from dual
  union all
  select 'b' as flow_out_id, 'c' as flow_in_id, 1.0 as edge_weight from dual
  union all
  select 'd' as flow_out_id, 'b' as flow_in_id, 1.0 as edge_weight from dual
)tmp
;
drop table if exists LabelPropagationClassification_func_test_node;
create table LabelPropagationClassification_func_test_node as
select * from
(
  select 'a' as node, 'X' as label, 1.0 as label_weight from dual
  union all
  select 'd' as node, 'Y' as label, 1.0 as label_weight from dual
)tmp;
```

The following figure shows the structure of the label propagation classification graph.



2. View training results.

```

+-----+-----+
| node | tag | weight |
+-----+-----+
| a | X | 1.0 |
| b | X | 0.16667 |
| b | Y | 0.83333 |
| c | X | 0.53704 |
| c | Y | 0.46296 |
| d | Y | 1.0 |
+-----+-----+

```

## 3.8.6. Modularity

Modularity is a metric that is used to evaluate the structure of communities in a network. It is designed to measure the strength of division of a network into communities. Values greater than 0.3 indicate a strong community structure. This topic describes the Modularity component provided by Machine Learning Studio.

You can configure the component by using one of the following methods:

### Machine Learning Platform for AI console

| Tab            | Parameter                   | Description                                                                                                                                           |
|----------------|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Source Vertex Column        | The start vertex column in the edge table.                                                                                                            |
|                | Initial Vertex Label Column | The group of start vertices in the edge table.                                                                                                        |
|                | Target Vertex Column        | The end vertex column in the edge table.                                                                                                              |
|                | Target Vertex Label Column  | The group of end vertices in the edge table.                                                                                                          |
|                | Workers                     | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter. |

| Tuning Tab | Parameter                     | Description                                                                                                                                                                                                       |
|------------|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|            | <b>Memory Size per Worker</b> | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |

## PAI command

```
PAI -name Modularity
-project algo_public
-DinputEdgeTableName=Modularity_func_test_edge
-DfromVertexCol=flow_out_id
-DfromGroupCol=group_out_id
-DtoVertexCol=flow_in_id
-DtoGroupCol=group_in_id
-DoutputTableName=Modularity_func_test_result;
```

| Parameter                | Required | Description                                          | Default value    |
|--------------------------|----------|------------------------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.                    | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table.              | Full table       |
| fromVertexCol            | Yes      | The start vertex column in the input edge table.     | No default value |
| fromGroupCol             | Yes      | The group of start vertices in the input edge table. | No default value |
| toVertexCol              | Yes      | The end vertex column in the input edge table.       | No default value |
| toGroupCol               | Yes      | The group of end vertices in the input edge table.   | No default value |
| outputTableName          | Yes      | The name of the output table.                        | No default value |
| outputTablePartitions    | No       | The partitions in the output table.                  | No default value |
| lifecycle                | No       | The lifecycle of the output table.                   | No default value |



| Parameter | Required | Description                                                                                                                                                                                                       | Default value  |
|-----------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| workerNum | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured |
| workerMem | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096           |
| splitSize | No       | The data split size.                                                                                                                                                                                              | 64             |

## Examples

1. Generate training data.

The data is similar to that of the label propagation algorithm. For more information, see .

2. View training results.

```
+-----+
| val   |
+-----+
| 0.4230769 |
+-----+
```

### 3.8.7. Maximum Connected Subgraph

In Undirected Graph G, Vertex A is connected to Vertex B if a path exists between the two vertices. Undirected Graph G contains several subgraphs. Each vertex is connected to other vertices in the same subgraph. Vertices in different subgraphs are not connected. In this case, the subgraphs in Undirected Graph G are called maximum connected subgraphs. This topic describes the Maximum Connected Subgraph component provided by Machine Learning Studio.

You can configure the component by using one of the following methods:

#### Machine Learning Platform for AI console

| Tab | Parameter    | Description                                |
|-----|--------------|--------------------------------------------|
|     | Start Vertex | The start vertex column in the edge table. |
|     |              |                                            |

| Fields Setting Tab | Parameter                     | Description                                                                                                                                                                                                       |
|--------------------|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                    | <b>End Node</b>               | The end vertex column in the edge table.                                                                                                                                                                          |
| <b>Tuning</b>      | <b>Workers</b>                | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                    | <b>Memory Size per Worker</b> | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |
|                    | <b>Data Split Size</b>        | The data split size. Default value: 64.                                                                                                                                                                           |

## PAI command

```
PAI -name MaximalConnectedComponent
-project algo_public
-DinputEdgeTableName=MaximalConnectedComponent_func_test_edge
-DfromVertexCol=flow_out_id
-DtoVertexCol=flow_in_id
-DoutputTableName=MaximalConnectedComponent_func_test_result;
```

| Parameter                | Required | Description                                      | Default value    |
|--------------------------|----------|--------------------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.                | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table.          | Full table       |
| fromVertexCol            | Yes      | The start vertex column in the input edge table. | No default value |
| toVertexCol              | Yes      | The end vertex column in the input edge table.   | No default value |
| outputTableName          | Yes      | The name of the output table.                    | No default value |
| outputTablePartitions    | No       | The partitions in the output table.              | No default value |
| lifecycle                | No       | The lifecycle of the output table.               | No default value |

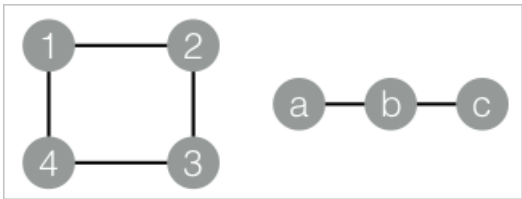
| Parameter | Required | Description                                                                                                                                                                                                       | Default value  |
|-----------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| workerNum | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured |
| workerMem | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096           |
| splitSize | No       | The data split size.                                                                                                                                                                                              | 64             |

## Examples

### 1. Generate training data.

```
drop table if exists MaximalConnectedComponent_func_test_edge;
create table MaximalConnectedComponent_func_test_edge as
select * from
(
  select '1' as flow_out_id,'2' as flow_in_id from dual
  union all
  select '2' as flow_out_id,'3' as flow_in_id from dual
  union all
  select '3' as flow_out_id,'4' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'4' as flow_in_id from dual
  union all
  select 'a' as flow_out_id,'b' as flow_in_id from dual
  union all
  select 'b' as flow_out_id,'c' as flow_in_id from dual
)tmp;
drop table if exists MaximalConnectedComponent_func_test_result;
create table MaximalConnectedComponent_func_test_result
(
  node string,
  grp_id string
);
```

The following figure shows the structure of the maximum connected subgraph.



2. View training results.

```
+-----+-----+
| node | grp_id|
+-----+-----+
| 1    | 4    |
| 2    | 4    |
| 3    | 4    |
| 4    | 4    |
| a    | c    |
| b    | c    |
| c    | c    |
+-----+-----+
```

### 3.8.8. Vertex Clustering Coefficient

A vertex clustering coefficient is used to calculate the peripheral density of a vertex in Undirected Graph G. The density of a star network is 0, and that of a fully meshed network is 1. This topic describes the Vertex Clustering Coefficient component provided by Machine Learning Studio.

You can configure the component by using one of the following methods:

#### Machine Learning Platform for AI console

| Tab                | Parameter                   | Description                                                                                                                                                                                                       |
|--------------------|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting     | Start Vertex                | The start vertex column in the edge table.                                                                                                                                                                        |
|                    | End Vertex                  | The end vertex column in the edge table.                                                                                                                                                                          |
| Parameters Setting | Largest Vertex Degree       | If the vertex degree is greater than the value of this parameter, sampling is required. This parameter is optional. Default value: 500.                                                                           |
| Tuning             | Workers                     | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                    | Memory Size per Worker (MB) | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |
|                    | Data Split Size (MB)        | The data split size. Default value: 64.                                                                                                                                                                           |

## PAI command

```
PAI -name NodeDensity
    -project algo_public
    -DinputEdgeTableName=NodeDensity_func_test_edge
    -DfromVertexCol=flow_out_id
    -DtoVertexCol=flow_in_id
    -DoutputTableName=NodeDensity_func_test_result
    -DmaxEdgeCnt=500;
```

| Parameter                | Required | Description                                                                                                                                           | Default value    |
|--------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.                                                                                                                     | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table.                                                                                                               | Full table       |
| fromVertexCol            | Yes      | The start vertex column in the input edge table.                                                                                                      | No default value |
| toVertexCol              | Yes      | The end vertex column in the input edge table.                                                                                                        | No default value |
| outputTableName          | Yes      | The name of the output table.                                                                                                                         | No default value |
| outputTablePartitions    | No       | The partitions in the output table.                                                                                                                   | No default value |
| lifecycle                | No       | The lifecycle of the output table.                                                                                                                    | No default value |
| maxEdgeCnt               | No       | If the vertex degree is greater than the value of this parameter, sampling is required.                                                               | 500              |
| workerNum                | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter. | Not configured   |

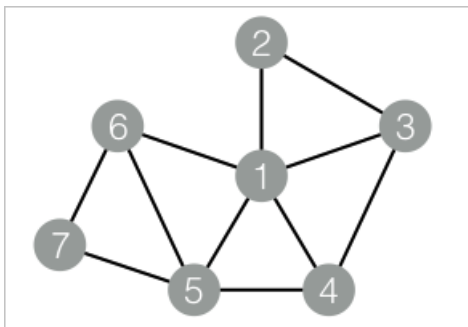
| Parameter | Required | Description                                                                                                                                                                                                       | Default value |
|-----------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| workerMem | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096          |
| splitSize | No       | The data split size.                                                                                                                                                                                              | 64            |

## Examples

1. Generate training data.

```
drop table if exists NodeDensity_func_test_edge;
create table NodeDensity_func_test_edge as
select * from
(
  select '1' as flow_out_id, '2' as flow_in_id from dual
  union all
  select '1' as flow_out_id, '3' as flow_in_id from dual
  union all
  select '1' as flow_out_id, '4' as flow_in_id from dual
  union all
  select '1' as flow_out_id, '5' as flow_in_id from dual
  union all
  select '1' as flow_out_id, '6' as flow_in_id from dual
  union all
  select '2' as flow_out_id, '3' as flow_in_id from dual
  union all
  select '3' as flow_out_id, '4' as flow_in_id from dual
  union all
  select '4' as flow_out_id, '5' as flow_in_id from dual
  union all
  select '5' as flow_out_id, '6' as flow_in_id from dual
  union all
  select '5' as flow_out_id, '7' as flow_in_id from dual
  union all
  select '6' as flow_out_id, '7' as flow_in_id from dual
)tmp;
drop table if exists NodeDensity_func_test_result;
create table NodeDensity_func_test_result
(
  node string,
  node_cnt bigint,
  edge_cnt bigint,
  density double,
  log_density double
);
```

The following figure shows the structure of the vertex clustering coefficient graph.



2. View training results.

```

1,5,4,0.4,1.45657
2,2,1,1.0,1.24696
3,3,2,0.66667,1.35204
4,3,2,0.66667,1.35204
5,4,3,0.5,1.41189
6,3,2,0.66667,1.35204
7,2,1,1.0,1.24696

```

### 3.8.9. Edge Clustering Coefficient

An edge clustering coefficient is used to calculate the edge density in Undirected Graph G. This topic describes the Edge Clustering Coefficient component provided by Machine Learning Studio.

You can configure the component by using one of the following methods:

#### Machine Learning Platform for AI console

| Tab            | Parameter                   | Description                                                                                                                                                                                                       |
|----------------|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Start Vertex                | The start vertex column in the edge table.                                                                                                                                                                        |
|                | End Vertex                  | The end vertex column in the edge table.                                                                                                                                                                          |
| Tuning         | Workers                     | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                | Memory Size per Worker (MB) | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |
|                | Data Split Size (MB)        | The data split size. Default value: 64.                                                                                                                                                                           |

#### PAI command

```

PAI -name EdgeDensity
-project algo_public
-DinputEdgeTableName=EdgeDensity_func_test_edge
-DfromVertexCol=flow_out_id
-DtoVertexCol=flow_in_id
-DoutputTableName=EdgeDensity_func_test_result;

```

| Parameter                | Required | Description                             | Default value    |
|--------------------------|----------|-----------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.       | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table. | Full table       |



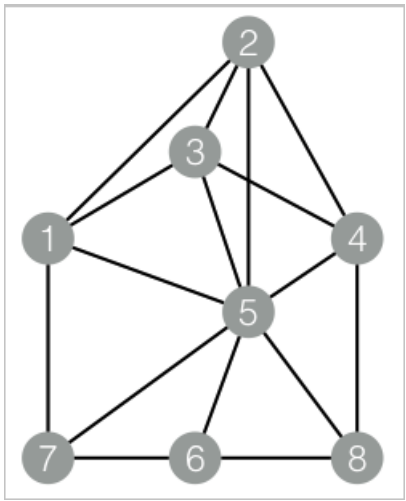
| Parameter             | Required | Description                                                                                                                                                                                                       | Default value    |
|-----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| fromVertexCol         | Yes      | The start vertex column in the input edge table.                                                                                                                                                                  | No default value |
| toVertexCol           | Yes      | The end vertex column in the input edge table.                                                                                                                                                                    | No default value |
| outputTableName       | Yes      | The name of the output table.                                                                                                                                                                                     | No default value |
| outputTablePartitions | No       | The partitions in the output table.                                                                                                                                                                               | No default value |
| lifecycle             | No       | The lifecycle of the output table.                                                                                                                                                                                | No default value |
| workerNum             | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured   |
| workerMem             | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096             |
| splitSize             | No       | The data split size.                                                                                                                                                                                              | 64               |

## Examples

1. Generate training data.

```
drop table if exists EdgeDensity_func_test_edge;
create table EdgeDensity_func_test_edge as
select * from
(
  select '1' as flow_out_id,'2' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'3' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'5' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'7' as flow_in_id from dual
  union all
  select '2' as flow_out_id,'5' as flow_in_id from dual
  union all
  select '2' as flow_out_id,'4' as flow_in_id from dual
  union all
  select '2' as flow_out_id,'3' as flow_in_id from dual
  union all
  select '3' as flow_out_id,'5' as flow_in_id from dual
  union all
  select '3' as flow_out_id,'4' as flow_in_id from dual
  union all
  select '4' as flow_out_id,'5' as flow_in_id from dual
  union all
  select '4' as flow_out_id,'8' as flow_in_id from dual
  union all
  select '5' as flow_out_id,'6' as flow_in_id from dual
  union all
  select '5' as flow_out_id,'7' as flow_in_id from dual
  union all
  select '5' as flow_out_id,'8' as flow_in_id from dual
  union all
  select '7' as flow_out_id,'6' as flow_in_id from dual
  union all
  select '6' as flow_out_id,'8' as flow_in_id from dual
)tmp;
drop table if exists EdgeDensity_func_test_result;
create table EdgeDensity_func_test_result
(
  node1 string,
  node2 string,
  node1_edge_cnt bigint,
  node2_edge_cnt bigint,
  triangle_cnt bigint,
  density double
);
```

The following figure shows the structure of the edge clustering coefficient graph.



2. View training results.

```
1,2,4,4,2,0.5
2,3,4,4,3,0.75
2,5,4,7,3,0.75
3,1,4,4,2,0.5
3,4,4,4,2,0.5
4,2,4,4,2,0.5
4,5,4,7,3,0.75
5,1,7,4,3,0.75
5,3,7,4,3,0.75
5,6,7,3,2,0.66667
5,8,7,3,2,0.66667
6,7,3,3,1,0.33333
7,1,3,4,1,0.33333
7,5,3,7,2,0.66667
8,4,3,4,1,0.33333
8,6,3,3,1,0.33333
```

### 3.8.10. Counting Triangle

A counting triangle is used to determine the number of triangles that pass through each vertex in Undirected Graph G. This topic describes the Counting Triangle component provided by Machine Learning Studio.

You can configure the component by using one of the following methods:

#### Machine Learning Platform for AI console

| Tab            | Parameter    | Description                                |
|----------------|--------------|--------------------------------------------|
| Fields Setting | Start Vertex | The start vertex column in the edge table. |
|                | End Vertex   | The end vertex column in the edge table.   |
|                |              |                                            |

| Tab                | Parameter                          | Description                                                                                                                                                                                                       |
|--------------------|------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | <b>Largest Vertex Degree</b>       | If the vertex degree is greater than the value of this parameter, sampling is required. This parameter is optional. Default value: 500.                                                                           |
| Tuning             | <b>Workers</b>                     | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                    | <b>Memory Size per Worker (MB)</b> | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |
|                    | <b>Data Split Size (MB)</b>        | The data split size. Default value: 64.                                                                                                                                                                           |

## PAI command

```
PAI -name TriangleCount
-project algo_public
-DinputEdgeTableName=TriangleCount_func_test_edge
-DfromVertexCol=flow_out_id
-DtoVertexCol=flow_in_id
-DoutputTableName=TriangleCount_func_test_result;
```

| Parameter                | Required | Description                                      | Default value    |
|--------------------------|----------|--------------------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.                | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table.          | Full table       |
| fromVertexCol            | Yes      | The start vertex column in the input edge table. | No default value |
| toVertexCol              | Yes      | The end vertex column in the input edge table.   | No default value |
| outputTableName          | Yes      | The name of the output table.                    | No default value |
| outputTablePartitions    | No       | The partitions in the output table.              | No default value |
| lifecycle                | No       | The lifecycle of the output table.               | No default value |

| Parameter  | Required | Description                                                                                                                                                                                                       | Default value  |
|------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| maxEdgeCnt | No       | If the vertex degree is greater than the value of this parameter, sampling is required.                                                                                                                           | 500            |
| workerNum  | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured |
| workerMem  | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096           |
| splitSize  | No       | The data split size.                                                                                                                                                                                              | 64             |

## Examples

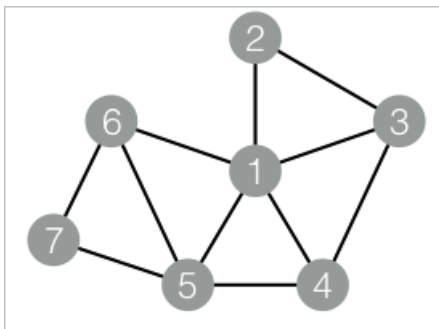
1. Generate training data.

```

drop table if exists TriangleCount_func_test_edge;
create table TriangleCount_func_test_edge as
select * from
(
  select '1' as flow_out_id,'2' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'3' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'4' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'5' as flow_in_id from dual
  union all
  select '1' as flow_out_id,'6' as flow_in_id from dual
  union all
  select '2' as flow_out_id,'3' as flow_in_id from dual
  union all
  select '3' as flow_out_id,'4' as flow_in_id from dual
  union all
  select '4' as flow_out_id,'5' as flow_in_id from dual
  union all
  select '5' as flow_out_id,'6' as flow_in_id from dual
  union all
  select '5' as flow_out_id,'7' as flow_in_id from dual
  union all
  select '6' as flow_out_id,'7' as flow_in_id from dual
)tmp;
drop table if exists TriangleCount_func_test_result;
create table TriangleCount_func_test_result
(
  node1 string,
  node2 string,
  node3 string
);

```

The following figure shows the structure of the counting triangle.



## 2. View training results.

```

1,2,3
1,3,4
1,4,5
1,5,6
5,6,7

```

## 3.8.11. Tree Depth

In a tree network, the Tree Depth component generates the depth of each vertex in a tree and the tree ID. This topic describes the Tree Depth component provided by Machine Learning Studio.

You can configure the component by using one of the following methods:

### Machine Learning Platform for AI console

| Tab            | Parameter                       | Description                                                                                                                                                                                                       |
|----------------|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Edge Table: Start Vertex Column | The start vertex column in the edge table.                                                                                                                                                                        |
|                | Edge Table: End Vertex Column   | The end vertex column in the edge table.                                                                                                                                                                          |
| Tuning         | Workers                         | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             |
|                | Memory Size per Worker          | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. |
|                | Data Split Size                 | The data split size. Default value: 64.                                                                                                                                                                           |

### PAI command

```
PAI -name TreeDepth
  -project algo_public
  -DinputEdgeTableName=TreeDepth_func_test_edge
  -DfromVertexCol=flow_out_id
  -DtoVertexCol=flow_in_id
  -DoutputTableName=TreeDepth_func_test_result;
```

| Parameter                | Required | Description                                      | Default value    |
|--------------------------|----------|--------------------------------------------------|------------------|
| inputEdgeTableName       | Yes      | The name of the input edge table.                | No default value |
| inputEdgeTablePartitions | No       | The partitions in the input edge table.          | Full table       |
| fromVertexCol            | Yes      | The start vertex column in the input edge table. | No default value |
| toVertexCol              | Yes      | The end vertex column in the input edge table.   | No default value |

| Parameter             | Required | Description                                                                                                                                                                                                       | Default value    |
|-----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| outputTableName       | Yes      | The name of the output table.                                                                                                                                                                                     | No default value |
| outputTablePartitions | No       | The partitions in the output table.                                                                                                                                                                               | No default value |
| lifecycle             | No       | The lifecycle of the output table.                                                                                                                                                                                | No default value |
| workerNum             | No       | The number of vertices for parallel job execution. The parallelism level and framework communication costs increase with the value of this parameter.                                                             | Not configured   |
| workerMem             | No       | The maximum size of memory that a single job can use. By default, the system allocates 4,096 MB for each job. If the used memory size exceeds the value of this parameter, the OutOfMemory exception is reported. | 4096             |
| splitSize             | No       | The data split size.                                                                                                                                                                                              | 64               |

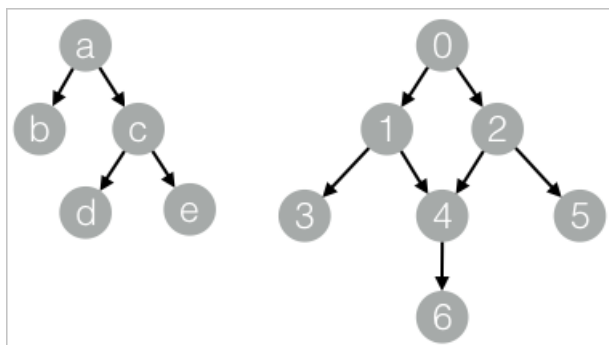
## Examples

1. Generate training data.



```
drop table if exists TreeDepth_func_test_edge;
create table TreeDepth_func_test_edge as
select * from
(
  select '0' as flow_out_id, '1' as flow_in_id from dual
  union all
  select '0' as flow_out_id, '2' as flow_in_id from dual
  union all
  select '1' as flow_out_id, '3' as flow_in_id from dual
  union all
  select '1' as flow_out_id, '4' as flow_in_id from dual
  union all
  select '2' as flow_out_id, '4' as flow_in_id from dual
  union all
  select '2' as flow_out_id, '5' as flow_in_id from dual
  union all
  select '4' as flow_out_id, '6' as flow_in_id from dual
  union all
  select 'a' as flow_out_id, 'b' as flow_in_id from dual
  union all
  select 'a' as flow_out_id, 'c' as flow_in_id from dual
  union all
  select 'c' as flow_out_id, 'd' as flow_in_id from dual
  union all
  select 'c' as flow_out_id, 'e' as flow_in_id from dual
)tmp;
drop table if exists TreeDepth_func_test_result;
create table TreeDepth_func_test_result
(
  node string,
  root string,
  depth bigint
);
```

The following figure shows the tree depth structure.



2. View training results.

```
0,0,0
1,0,1
2,0,1
3,0,2
4,0,2
5,0,2
6,0,3
a,a,0
b,a,1
c,a,1
d,a,2
e,a,2
```

## 3.9. Tools

This topic describes the tool-type components provided by Machine Learning Studio. The components include SQL Script and Semantic Vector Distance (Double Table).

### SQL Script

You can use the SQL script editor to write SQL statements. For more information, see [SQL overview](#).

You can configure the SQL Script component only in the Machine Learning Platform for AI (PAI) console. The following table describes the parameters that must be configured for the component.

| Parameter    | Description                                                    |
|--------------|----------------------------------------------------------------|
| Input Source | The name of the input table.                                   |
| SQL Script   | The SQL script that is used to implement the required feature. |

### Example on how to use SQL Script

1. In the left-side navigation tree, click **Data Source/Target**. Then, drag and drop the **Read MaxCompute Table** component onto the canvas and click the component. In the right-side pane, configure parameters.

The screenshot shows the 'Table Selection' tab of a configuration window. It has two tabs: 'Table Selection' (active) and 'Fields Information'. Under 'Table Selection', there is a 'Table Name' field with the text 'Cross-project reference: Project ...' and a text input box containing 'bank\_data'. To the right of the input box is a small circle with an 'x' icon. Below this is a 'Partition' checkbox, which is currently unchecked and has a light gray background, indicating it is disabled.

- If the input table is a partitioned table, the system automatically selects the **Partition** check box. You can select or specify only one partition. If you clear the Partition check box or do not specify a partition, the full table is used as the input.
  - If the input table is a non-partitioned table, the Partition check box is dimmed.
2. In the left-side navigation tree, click **Tools**. Then, drag and drop the **SQL Script** component onto the canvas and connect the **Read MaxCompute Table** component to the SQL Script component.
  3. Click the **SQL Script** component on the canvas. In the **SQL Script** field of the **Parameters**

**Setting** section, enter the SQL script that is used to implement the required feature.

**Parameters Setting**

**Input Source**

t1 fromName"bank\_data"

t2

t3

t4

**SQL script** The last clause in the script must b...

Save Format Maximize

```

1  --The input data has been a
2  -- You can also directly in
3  -- To read the partition da
4  -- Components do not automa
5  --This component supports a
6  select count(*) from ${t1}
    
```

Description of the SQL Script component:

- The component supports one to four inputs and one output.
  - Input tables are automatically mapped to tables t1 to t4. You can directly call \${t1} or \${t2} without specifying the table names.
  - You can execute various SQL statements in the middle of an SQL script. However, the last statement must be a SELECT statement. The output table contains the execution result of the SELECT statement.
  - The sample SQL script is used to collect the number of rows in the input tables.
4. In the left-side navigation tree, click **Data Source/Target**. Then, drag and drop the **Write MaxCompute Table** component onto the canvas and click the component. In the **Select Table** section, specify **New Table Name**. Then, the system creates a table based on the name that you specified. If you want to write data to a partitioned table, you must first create the table.

**Table Selection**

**New Table Name** To write a partitioned table, f...

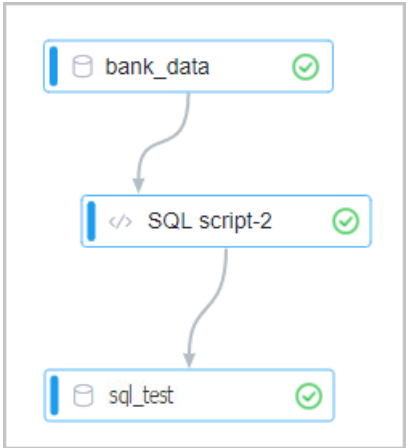
sql\_test

☐ Partition

**Lifecycle** No lifecycle is set if you do not enter ...

28

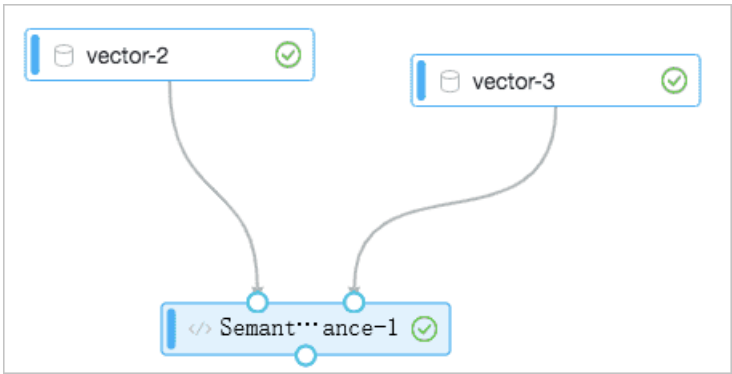
5. Connect the SQL Script component to the Write MaxCompute Table component. Then, click **Run** in the top toolbar.



6. After the experiment is executed, right-click the **Write MaxCompute Table** component on the canvas and select **View Data** to view the data that is written.

Semantic Vector Distance (Double Table)

- Input
- The Semantic Vector Distance (Double Table) component allows you to specify two input tables. The left input port of the component is for a query table and the right input port is for a dictionary table.



You can configure the component in the PAI console. The following table describes the parameters that must be configured for the component.

| Tab            | Parameter                       | Description                                                                                                                                                                                                                                                                          |   |   |   |   |   |   |   |   |   |   |   |   |
|----------------|---------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|
| Fields Setting | Vector Column                   | <div>The vector values. You must write the vector to one field. Separate all values with spaces.</div> <div><table><tr><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>2</td><td>2</td><td>2</td><td>2</td></tr><tr><td>3</td><td>3</td><td>3</td><td>3</td></tr></table></div> | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 |
|                | 1                               | 1                                                                                                                                                                                                                                                                                    | 1 | 1 |   |   |   |   |   |   |   |   |   |   |
|                | 2                               | 2                                                                                                                                                                                                                                                                                    | 2 | 2 |   |   |   |   |   |   |   |   |   |   |
| 3              | 3                               | 3                                                                                                                                                                                                                                                                                    | 3 |   |   |   |   |   |   |   |   |   |   |   |
| ID Column      | The primary key of each column. |                                                                                                                                                                                                                                                                                      |   |   |   |   |   |   |   |   |   |   |   |   |
|                |                                 |                                                                                                                                                                                                                                                                                      |   |   |   |   |   |   |   |   |   |   |   |   |

| Tab                | Parameter                           | Description                                                                                                                                                                                                          |
|--------------------|-------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Distance Calculation Method         | Valid values: <b>Euclidean</b> and <b>Cosine</b> .                                                                                                                                                                   |
|                    | Number of Highest Similarity Scores | The value of this parameter must be a positive integer.                                                                                                                                                              |
| Tuning             | Cores                               | The number of CPU cores that you want to use for computing. Default value: 3. If an out-of-memory (OOM) error occurs during computing, you can increase the values of <b>Cores</b> and <b>Memory Size per Core</b> . |
|                    | Memory Size per Core                | The memory size of each CPU core. Default value: 2046. Unit: MB. If an OOM error occurs during computing, you can increase the values of <b>Cores</b> and <b>Memory Size per Core</b> .                              |

- Output

The following figure shows the output of the Semantic Vector Distance (Double Table) component. The output table contains the top N distances between data in the query table and that in the dictionary table and the sorting of the distances.

| original_id ▲ | near_id ▲ | distance ▲  | rank ▲ |
|---------------|-----------|-------------|--------|
| 2             | 3         | 2.236067... | 2      |
| 2             | 2         | 0           | 1      |
| 3             | 2         | 2.236067... | 2      |
| 3             | 3         | 0           | 1      |
| 1             | 2         | 2.236067... | 2      |
| 1             | 1         | 0           | 1      |

- Recommendations

- The component is mainly used to calculate the Cartesian product distances between data in the two input tables and to sort the distances. Therefore, we recommend that you do not use more than tens of millions of samples.
- By default, a small number of resources are specified on the **Tuning** tab. If an OOM error occurs, you can increase the resources.
- If you use the Cosine distance calculation method for DOUBLE-type data, negative numbers may exist in the output table. This is normal.

## 3.10. Financials

This topic describes the financial components provided by Machine Learning Studio. These components include Binning, Data Conversion Module, Scorecard Training, Scorecard Prediction, and Population Stability Index (PSI).

### Binning

The Binning component is used for feature discretization. Feature discretization is a process of converting continuous data into multiple discrete intervals. The Binning component supports equal frequency binning, equal width binning, and automated binning. You can configure the component by using one of the following methods:

- Machine Learning Platform for AI (PAI) console

| Tab            | Parameter                                      | Description                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|----------------|------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | <b>Feature Columns</b>                         | Only columns of the STRING, BIGINT, or DOUBLE data type are supported.                                                                                                                                                                                                                                                                                                                                                                               |
|                | <b>Label Column</b>                            | This parameter is required only for binary classification.                                                                                                                                                                                                                                                                                                                                                                                           |
|                | <b>Positive Value</b>                          | This parameter is valid only when <b>Label Column</b> is specified.                                                                                                                                                                                                                                                                                                                                                                                  |
|                | <b>Binning Parameter Source</b>                | Valid values: <b>Parameters in Parameter Settings</b> and <b>Manual Binning or Custom JSON</b> .                                                                                                                                                                                                                                                                                                                                                     |
|                | <b>Reserve Unselected Feature Columns</b>      | If you set Binning Parameter Source to Manual Binning or Custom JSON and set Reserve Unselected Feature Columns to <b>Yes</b> , the columns that are not specified for <b>Feature Columns</b> remain unchanged in the output. If you set Binning Parameter Source to Manual Binning or Custom JSON but set Reserve Unselected Feature Columns to <b>No</b> , the columns that are not specified for Feature Columns are not contained in the output. |
|                | <b>Upload Binning and Constraint JSON Code</b> | This parameter is valid only when you set <b>Binning Parameter Source</b> to <b>Manual Binning or Custom JSON</b> .                                                                                                                                                                                                                                                                                                                                  |
|                | <b>Append Binning Files</b>                    | If the files you specify contain new features, the new features are appended to binning results. If the files do not contain new features, the features in the files take effect.                                                                                                                                                                                                                                                                    |
|                | <b>Bins</b>                                    | If you set this parameter to 10, continuous features are converted into 10 discrete intervals.                                                                                                                                                                                                                                                                                                                                                       |
|                |                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                      |

| Tab                | Parameter                                    | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|--------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | <b>Custom Bins</b>                           | <p>You can specify the numbers of bins for specific columns. The setting of this parameter takes precedence over the setting of the Bins parameter. If a specific column is not included in the selected columns, this column is also used in binning. For example, columns col0 and col1 are selected for data binning. The number of bins customized for column col0 is 3 and that customized for col2 is 5. The Bins parameter is set to 10. Binning is performed based on col0:3,col1:10,col2:5.</p> <p>Specify this parameter in the format of Column name 1: Number of bins, Column name 2: Number of bins.</p> |
|                    | <b>Custom Discrete Value Count Threshold</b> | Specify this parameter in the col0:3 format.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|                    | <b>Interval Type</b>                         | Valid values: <b>Left-open</b> , <b>Right-closed</b> and <b>Left-closed</b> , <b>Right-open</b> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|                    | <b>Binning Mode</b>                          | Valid values: <b>Equal Frequency</b> , <b>Equal Width</b> , and <b>Automatic Binning</b> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|                    | <b>Discrete Value Count Threshold</b>        | If a value is less than this threshold, the value is distributed to the else bin.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Tuning             | <b>Cores</b>                                 | The number of cores. By default, the number is allocated by the system.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|                    | <b>Memory Size per Core</b>                  | The memory size of each core. By default, the memory size is allocated by the system.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |

- PAI command

```
PAI -name binning
-project algo_public
-DinputTableName=input
-DoutputTableName=output
```

| Parameter       | Description                   | Required | Default value    |
|-----------------|-------------------------------|----------|------------------|
| inputTableName  | The name of the input table.  | Yes      | No default value |
| outputTableName | The name of the output table. | Yes      | No default value |

| Parameter            | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Required | Default value                                                                              |
|----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|--------------------------------------------------------------------------------------------|
| selectedColNames     | The columns that are selected from the input table for binning.                                                                                                                                                                                                                                                                                                                                                                                                                                               | No       | All columns except the label column (if no label column exists, all columns are selected.) |
| labelColumn          | The label column.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | No       | No default value                                                                           |
| validTableName       | The name of the validation table. This parameter is required when binningMethod is set to auto.                                                                                                                                                                                                                                                                                                                                                                                                               | No       | No default value                                                                           |
| validTablePartitions | The partitions that are selected from the validation table.                                                                                                                                                                                                                                                                                                                                                                                                                                                   | No       | Full table                                                                                 |
| inputTablePartitions | The partitions that are selected from the input table.                                                                                                                                                                                                                                                                                                                                                                                                                                                        | No       | Full table                                                                                 |
| inputBinTableName    | The input binning table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | No       | No default value                                                                           |
| selectedBinColNames  | The columns that are selected from the input binning table.                                                                                                                                                                                                                                                                                                                                                                                                                                                   | No       | No default value                                                                           |
| positiveLabel        | Specifies whether the samples are positive samples.                                                                                                                                                                                                                                                                                                                                                                                                                                                           | No       | 1                                                                                          |
| nDivide              | The number of bins. The value of this parameter must be a positive integer.                                                                                                                                                                                                                                                                                                                                                                                                                                   | No       | 10                                                                                         |
| colsNDivide          | The numbers of bins for specific columns. Specify this parameter in the format of Column name 1: Number of bins, Column name 2: Number of bins, such as col0:3,col2:5. If the columns that are specified for colsNDivide are not included in those specified for selectedColNames, the columns are also used in binning. For example, selectedColNames is set to col0,col1, colsNDivide is set to col0:3,col2:5, and nDivide is set to 10. In this case, binning is performed based on col0:3,col1:10,col2:5. | No       | No default value                                                                           |



| Parameter           | Description                                                                                                                                                                                                                                                     | Required | Default value           |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|-------------------------|
| isLeftOpen          | The interval type. Valid values: <ul style="list-style-type: none"><li>◦ {true}: indicates left-open, right-closed intervals.</li><li>◦ {false}: indicates left-closed, right-open intervals.</li></ul>                                                         | No       | true                    |
| stringThreshold     | The threshold for discrete values in the else bin.                                                                                                                                                                                                              | No       | No default value        |
| colsStringThreshold | The threshold for specific columns. Specify this parameter in the same format as the colsNDivide parameter.                                                                                                                                                     | No       | No default value        |
| binningMethod       | The binning mode. Valid values: <ul style="list-style-type: none"><li>◦ quantile: indicates equal frequency binning.</li><li>◦ bucket: indicates equal width binning.</li><li>◦ auto: indicates that the system automatically selects a binning mode.</li></ul> | No       | quantile                |
| lifecycle           | The lifecycle of the output table. The value of this parameter must be a positive integer.                                                                                                                                                                      | No       | No default value        |
| coreNum             | The number of cores. The value of this parameter must be a positive integer.                                                                                                                                                                                    | No       | Automatically allocated |
| memSizePerCore      | The memory size of each core. The value of this parameter must be a positive integer.                                                                                                                                                                           | No       | Automatically allocated |

The Binning component must be used with the Scorecard Training component. During scorecard training, the Binning component converts continuous features into multiple discrete dummy variables to achieve feature engineering. You can specify constraints for the weights of the dummy variables. The following information describes the constraints:

- Ascending order: Weights must be added to the dummy variables of a feature based on index values in ascending order. This indicates that a dummy variable with a large index value has a high weight.
- Descending order: Weights must be added to the dummy variables of a feature based on index values in descending order. This indicates that a dummy variable with a large index value has a low weight.
- Same weight: The weights of two dummy variables of a feature must be the same.
- Zero weight: The weight of a dummy variable must be 0.
- Specific weight: The weight of a dummy variable must be a specific floating-point value.
- WOE order: Weights must be added to the dummy variables of a feature based on the weight of evidence (WOE) values in ascending order. This indicates that a dummy variable with a large WOE value has a high weight.

## Data Conversion Module

The Data Conversion Module component performs normalization, discretization, indexation, or WOE conversion on data. You can configure the component by using one of the following methods:

- PAI console

| Tab            | Parameter                              | Description                                                                                                                                                                                                                                                                                                                                                                  |
|----------------|----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | <b>Feature Columns in Input Table</b>  | The feature columns that are selected from the input table. By default, all columns in the input table are selected.                                                                                                                                                                                                                                                         |
|                | <b>Columns without Data Conversion</b> | The columns on which data conversion is not required. The selected columns in the output are the same as those in the input. You can specify labels in the columns.                                                                                                                                                                                                          |
|                | <b>Data Conversion Mode</b>            | Valid values: <b>Normalization</b> , <b>Discretization</b> , <b>WOE Conversion</b> , and <b>Index</b> .                                                                                                                                                                                                                                                                      |
|                | <b>Default WOE Value</b>               | This parameter is valid only when the <b>Data Conversion Mode</b> parameter is set to <b>WOE Conversion</b> .<br><br>If this parameter is specified, when a sample value falls into a bin without WOE values, this value is used as the WOE value. If this parameter is not specified, when a sample value falls into a bin without WOE values, the system reports an error. |
| Tuning         | <b>Number of Cores</b>                 | The number of CPU cores that are required. By default, the number is allocated by the system.                                                                                                                                                                                                                                                                                |
|                | <b>Memory Size per Core</b>            | The memory size of each CPU core. By default, the memory size is allocated by the system.                                                                                                                                                                                                                                                                                    |

- PAI command

```
PAI -name data_transform
-project algo_public
-DinputFeatureTableName=feature_table
-DinputBinTableName=bin_table
-DoutputTableName=output_table
-DmetaColNames=label
-DfeatureColNames=feaname1,feaname2
```

| Parameter             | Description                           | Required | Default value    |
|-----------------------|---------------------------------------|----------|------------------|
| inputFeatureTableName | The name of the input feature table.  | Yes      | No default value |
| inputBinTableName     | The name of the binning result table. | Yes      | No default value |

| Parameter                   | Description                                                                                                                                                                       | Required | Default value           |
|-----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|-------------------------|
| inputFeatureTablePartitions | The partitions that are selected from the input feature table.                                                                                                                    | No       | Full table              |
| outputTableName             | The name of the output table.                                                                                                                                                     | Yes      | No default value        |
| featureColNames             | The feature columns that are selected from the input table.                                                                                                                       | No       | All columns             |
| metaColNames                | The columns that do not need to be converted. These columns in the output are the same as those in the input. You can specify labels and sample_id in the columns.                | No       | No default value        |
| transformType               | The type of data conversion. Valid values: <ul style="list-style-type: none"> <li>normalize: normalization</li> <li>dummy: discretization</li> <li>woe: WOE conversion</li> </ul> | No       | dummy                   |
| itemDelimiter               | The delimiter that is used to separate features. This parameter is valid only when the transformType parameter is set to dummy.                                                   | No       | ,                       |
| kvDelimiter                 | The delimiter that is used to separate keys and values. This parameter is valid only when the transformType parameter is set to dummy.                                            | No       | :                       |
| lifecycle                   | The lifecycle of the output table.                                                                                                                                                | No       | No default value        |
| coreNum                     | The number of CPU cores that are required.                                                                                                                                        | No       | Automatically allocated |
| memSizePerCore              | The memory size of each CPU core. Unit: MB.                                                                                                                                       | No       | Automatically allocated |

To implement normalization, the Data Conversion Module component converts variable values into values between 0 and 1 based on input binning information, and sets missing values to 0. The following algorithm is used:

```

if feature_raw_value == null or feature_raw_value == 0 then
    feature_norm_value = 0.0
else
    bin_index = FindBin(bin_table, feature_raw_value)
    bin_width = round(1.0 / bin_count * 1000) / 1000.0
    feature_norm_value = 1.0 - (bin_count - bin_index - 1) * bin_width

```

The Data Conversion Module component can convert different types of data into different formats.

- For normalization and WOE conversion, the component generates a common table.
- During discretization in which data is converted into dummy variables, the component generates a table in the key-value format. Each variable in the table is in the `${feaname}}\_bin\_${bin_id}` format. In the following example, the `sns` variable is used:
  - If `sns` falls into the second bin, the generated variable is `[sns]_bin_2`.
  - If `sns` does not have a value, it falls into the empty bin, and the generated variable is `[sns]_bin_null`.
  - If `sns` has a value but does not fall into a defined bin, it falls into the else bin, and the generated variable is `[sns]_bin_else`.

## Scorecard Training

Scorecard is a common modeling tool that is used in the credit risk assessment field. The scorecard performs binning to implement the discretization of variables and uses linear models (linear or logistic regression models) to train a model. The model training process includes feature selection and score transformation. The scorecard also allows you to add constraints to the variables during model training.

 **Note** If you use the scorecard without binning, scorecard training is equivalent to logistic or linear regression.

The following description provides the terms involved in scorecard training:

- Feature engineering

The main difference between the scorecard and normal linear models is that the scorecard performs feature engineering before it trains linear models. The Scorecard Training component supports the following methods for feature engineering:

- The Binning component is used to implement feature discretization. Then, one-hot encoding is performed for each variable based on binning results to generate `N` dummy variables. `N` represents the number of bins.

 **Note** When you convert original variables into dummy variables, you can specify constraints for these dummy variables.

- The Binning component is used to implement feature discretization. Then, WOE conversion is performed to replace the original value of a variable with the WOE value of the bin into which the variable falls.

- Score transformation

In scenarios such as credit scoring, you must perform a linear transformation to convert the predicted sample odds into a score. The following formula is used for linear transformation:

$$\log(odds) = \sum_i w_i x_i = a * scaled\_score + b$$

You can use the following parameters to specify the linear transformation relationship:

- `scaledValue`: specifies a scaled score.
- `odds`: specifies the odds of the scaled score.
- `pdo`: specifies the points at which the odds are doubled.

For example, `scaledValue` is 800, `odds` is 50, and `pdo` is 25. In this case, the following two dots are

determined for a line:

```
log(50)=a×800+b  
log(100)=a×825+b
```

Calculate the values of a and b, and perform a linear transformation to obtain the scores of a and b.

The scaling information is specified in the JSON format by using the `-Dscale` parameter.

```
{"scaledValue":800,"odds":50,"pdo":25}
```

If you specify the `-Dscale` parameter, you must also specify `scaledValue`, `odds`, and `pdo`.

- Constraint addition during training

During scorecard training, you can add constraints to variables. For example, you can specify the score of a specific bin to a fixed value, specify the scores of two bins to a specific proportion, or limit the scores between bins, such as sorting bin scores by WOE value. The implementation of constraints depends on the underlying optimization algorithm that contains constraints. You can specify the constraints in the Binning component in the PAI console. After the constraints are specified, the Binning component generates JSON-formatted constraints and automatically transfers them to its connected training component. The system supports the following JSON-formatted constraints:

- "<": The weights of variables must be sorted in ascending order.
- ">": The weights of variables must be sorted in descending order.
- "=": The weight of a specific variable must be a fixed value.
- "%": The weights of two variables must meet a proportional relationship.
- "UP": the upper limit for the weights of variables.
- "LO": the lower limit for the weights of variables.

Each JSON-formatted constraint is stored in a table as a string. The table contains only one row and one column.

```
{  
  "name": "feature0",  
  "<": [  
    [0,1,2,3]  
  ],  
  ">": [  
    [4,5,6]  
  ],  
  "=": [  
    "3:0","4:0.25"  
  ],  
  "%": [  
    ["6:1.0","7:1.0"]  
  ]  
}
```

- Built-in constraints

Each original variable has a built-in constraint: For each variable, the average score of a population must be 0. Based on this constraint, `scaled_weight` in the intercept options of the scorecard model equals the average score of the population in terms of all variables.

- Optimization algorithms

On the Parameters Setting tab, select Advanced Options. Then, you can configure the optimization algorithm that is used during scorecard training. The system supports the following optimization algorithms:

- L-BFGS: This algorithm is a first-order optimization algorithm that is used to process large amounts of feature data. The algorithm does not contain constraints. If you select this algorithm, the system automatically ignores the constraints that you specify.
- Newton's method: This algorithm is a classic second-order optimization algorithm. It is fast in convergence and accurate. However, the algorithm is not suitable for processing large amounts of feature data because it needs to calculate a second-order Hessian matrix. The algorithm does not contain constraints. If you select this algorithm, the system automatically ignores the constraints that you specify.
- Barrier method: This algorithm is a second-order optimization algorithm. If the algorithm does not contain constraints, it is completely equivalent to the Newton's method algorithm. The barrier method algorithm provides almost the same computing performance and accuracy as the SQP algorithm. In most cases, we recommend that you select SQP.
- SQP

This algorithm is a second-order optimization algorithm. If the algorithm does not contain constraints, it is completely equivalent to the Newton's method algorithm. The SQP algorithm provides almost the same computing performance and accuracy as the barrier method algorithm. In most cases, we recommend that you select SQP.

 **Note**

- L-BFGS and Newton's method are optimization algorithms without constraints. Barrier method and SQP are optimization algorithms with constraints.
- If you are not familiar with optimization algorithms, we recommend that you set Optimization Method to Auto Selection. In this case, the system selects the most appropriate algorithm based on the data volume and constraints.

- **Feature selection**

The Scorecard Training component supports stepwise feature selection. Stepwise feature selection is a combination of forward and backward selection. Each time the system performs a forward selection to select a new variable and adds it to the model, the system also performs a backward selection. The backward selection is used to remove the variables whose significance does not meet requirements. Stepwise feature selection supports various functions and feature transformation methods. Therefore, stepwise feature selection also supports multiple selection standards. The following standards are supported:

- **Marginal contribution:** This standard can be applied to all functions and feature engineering methods.

For this standard, two models must be trained: Model A and Model B. Model A does not contain Variable X, and Model B contains Variable X in addition to all the variables of Model A. The difference between the functions of the two models in final convergence is the marginal contribution of Variable X to all the other variables in Model B. In scenarios where variables are converted into dummy variables, the marginal contribution of Variable X is the difference between the functions of all dummy variables in Model A and the functions of all dummy variables in Model B. Therefore, the marginal contribution standard is supported by all feature engineering methods.

Marginal contribution is flexible and is not limited to a specific type of model. Only variables that contribute to functions are passed to the model. Marginal contribution has disadvantages when compared with statistical significance. Typically, 0.05 is used as the threshold for statistical significance. Marginal contribution does not provide a recommended threshold for beginners. We recommend that you set the threshold to 10E-5.

- **Score test:** This standard supports only WOE conversion and logistic regression without feature engineering.

During a forward selection, a model with only intercept options is trained first. In each subsequent iteration, the score chi-squares of the variables that are not passed to the model are measured. The variable with the largest score chi-square is passed to the model. In addition, the p-value of the variable with the largest score chi-square is calculated based on chi-square distribution. If the p-value of the variable is greater than the given SLENTY value, the variable is not passed to the model, and feature selection is terminated.

After the forward selection is completed, a backward selection is performed for the variable that is passed to the model. The Wald chi-square of the variable and the related p-value are calculated. If the p-value is greater than the given SLSTAY value, the variable is removed from the model. The system then starts a new iteration.

- **F test:** This standard supports only WOE conversion and linear regression without feature engineering.

During a forward selection, a variable with only intercept options is trained first. In each subsequent iteration, the F-values of the variables that are not passed to the model are calculated. F-value calculation is similar to marginal contribution calculation. Two models must be trained to calculate the F-value of a variable. The F-value follows F distribution. The related p-value can be calculated based on the probability density function of F distribution. If the p-value is greater than the given SLENTY value, the variable is not passed to the model, and the forward selection is terminated.

During the backward selection, the F-value is used to calculate the significance of a variable in a way similar to a score test.

- **Forced selection of the variables that you want to pass to a model**

Before a feature selection is performed, you can specify the variables that you want to forcibly pass to the model. No forward or backward selection is performed for the specified variables. These variables are directly passed to the model regardless of their significance. You can specify the number of iterations and significance thresholds by using the -Dselected parameter. Specify this parameter in the JSON format. Example:

```
{"max_step":2, "slentry": 0.0001, "slstay": 0.0001}
```

If the -Dselected parameter is left empty or the max\_step parameter is set to 0, no feature selection

is performed.

You can configure the Scorecard Training component in the PAI console or by running a PAI command. The following code provides a sample PAI command:

```
pai -name=linear_model -project=algo_public
-DinputTableName=input_data_table
-DinputBinTableName=input_bin_table
-DinputConstraintTableName=input_constraint_table
-DoutputTableName=output_model_table
-DlabelColName=label
-DfeatureColNames=feaname1,feaname2
-Doptimization=barrier_method
-Dloss=logistic_regression
-Dlifecycle=8
```

| Parameter                | Description                                                                                                                                                                                | Required | Default value                       |
|--------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|-------------------------------------|
| inputTableName           | The name of the input feature table.                                                                                                                                                       | Yes      | No default value                    |
| inputTablePartitions     | The partitions that are selected from the input feature table.                                                                                                                             | No       | Full table                          |
| inputBinTableName        | The name of the binning result table. If you specify this parameter, the system automatically performs discretization for features based on the binning rules in the binning result table. | No       | No default value                    |
| featureColNames          | The feature columns that are selected from the input table.                                                                                                                                | No       | All columns except the label column |
| labelColName             | The name of the label column.                                                                                                                                                              | Yes      | No default value                    |
| outputTableName          | The name of the output model table.                                                                                                                                                        | Yes      | No default value                    |
| inputConstraintTableName | The name of the table that stores constraints. The constraints are a JSON string that is stored in a cell of the table.                                                                    | No       | No default value                    |



| Parameter            | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Required | Default value           |
|----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|-------------------------|
| optimization         | <p>The optimization algorithm. Valid values:</p> <ul style="list-style-type: none"> <li>lbfgs</li> <li>newton</li> <li>barrier_method</li> <li>sqp</li> <li>auto</li> </ul> <p>Only sqp and barrier_method support constraints. If you set the optimization parameter to auto, the system automatically selects an appropriate optimization algorithm based on user data and related parameter settings. If you are not familiar with optimization algorithms, we recommend that you set the optimization parameter to auto.</p> | No       | auto                    |
| loss                 | The loss type. Valid values: logistic_regression and least_square.                                                                                                                                                                                                                                                                                                                                                                                                                                                               | No       | logistic_regression     |
| iterations           | The maximum number of iterations for optimization.                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | No       | 100                     |
| l1Weight             | The parameter weight of L1 regularization. Only L-BFGS supports this parameter.                                                                                                                                                                                                                                                                                                                                                                                                                                                  | No       | 0                       |
| l2Weight             | The parameter weight of L2 regularization.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | No       | 0                       |
| m                    | The historical step size for optimization that is performed by using the L-BFGS algorithm. Only the L-BFGS algorithm supports this parameter.                                                                                                                                                                                                                                                                                                                                                                                    | No       | 10                      |
| scale                | The weight scaling information of the scorecard.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | No       | No default value        |
| selected             | Specifies whether to enable feature selection during scorecard training.                                                                                                                                                                                                                                                                                                                                                                                                                                                         | No       | No default value        |
| convergenceTolerance | The convergence tolerance.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | No       | 1e-6                    |
| positiveLabel        | Specifies whether the samples are positive samples.                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | No       | 1                       |
| lifecycle            | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | No       | No default value        |
| coreNum              | The number of cores.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | No       | Automatically allocated |
| memSizePerCore       | The memory size of each core. Unit: MB.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | No       | Automatically allocated |

The Scorecard Training component generates data to a model report. The model report contains basic model evaluation statistics, such as binning information, binning constraints, WOE values, and marginal contribution information. The following table describes the columns in a model report.

| Column            | Data type | Description                                                                                                                       |
|-------------------|-----------|-----------------------------------------------------------------------------------------------------------------------------------|
| feaname           | STRING    | The feature name.                                                                                                                 |
| binid             | BIGINT    | The bin ID.                                                                                                                       |
| bin               | STRING    | The description of the bin, which indicates the interval of the bin.                                                              |
| constraint        | STRING    | The constraints that are added to the bin during training.                                                                        |
| weight            | DOUBLE    | The weight of a binning variable. For a non-scorecard model without binning, this field indicates the weight of a model variable. |
| scaled_weight     | DOUBLE    | The score that is linearly transformed from the weight of a binning variable in scorecard training.                               |
| woe               | DOUBLE    | A statistical metric. It indicates the WOE value of a bin in the training set.                                                    |
| contribution      | DOUBLE    | A statistical metric. It indicates the marginal contribution value of a bin in the training set.                                  |
| total             | BIGINT    | A statistical metric. It indicates the total number of samples in a bin in the training set.                                      |
| positive          | BIGINT    | A statistical metric. It indicates the number of positive samples in a bin in the training set.                                   |
| negative          | BIGINT    | A statistical metric. It indicates the number of negative samples in a bin in the training set.                                   |
| percentage_pos    | DOUBLE    | A statistical metric. It indicates the proportion of positive samples in a bin to total positive samples in the training set.     |
| percentage_neg    | DOUBLE    | A statistical metric. It indicates the proportion of negative samples in a bin to total negative samples in the training set.     |
| test_woe          | DOUBLE    | A statistical metric. It indicates the WOE value of a bin in the testing set.                                                     |
| test_contribution | DOUBLE    | A statistical metric. It indicates the marginal contribution value of a bin in the testing set.                                   |
| test_total        | BIGINT    | A statistical metric. It indicates the total number of samples in a bin in the testing set.                                       |

| Column              | Data type | Description                                                                                                                  |
|---------------------|-----------|------------------------------------------------------------------------------------------------------------------------------|
| test_positive       | BIGINT    | A statistical metric. It indicates the number of positive samples in a bin in the testing set.                               |
| test_negative       | BIGINT    | A statistical metric. It indicates the number of negative samples in a bin in the testing set.                               |
| test_percentage_pos | DOUBLE    | A statistical metric. It indicates the proportion of positive samples in a bin to total positive samples in the testing set. |
| test_percentage_neg | DOUBLE    | A statistical metric. It indicates the proportion of negative samples in a bin to total negative samples in the testing set. |

## Scorecard Prediction

The Scorecard Prediction component predicts credit scores based on the model report that is generated by the Scorecard Training component. You can configure the Scorecard Prediction component by using one of the following methods:

- PAI console

| Tab            | Parameter             | Description                                                                                                                                                   |
|----------------|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fields Setting | Feature Columns       | The feature columns that are used in prediction. By default, all feature columns are selected.                                                                |
|                | Reserved Columns      | The columns that are appended to the prediction result table without processing, such as the ID and Target columns.                                           |
|                | Output Variable Score | Specifies whether to generate a score for each feature variable. The total predicted score is the score of intercept options plus the score of each variable. |
| Tuning         | Cores                 | The number of CPU cores that are required. By default, the number is allocated by the system.                                                                 |
|                | Memory Size per Core  | The memory size of each CPU core. By default, the memory size is allocated by the system.                                                                     |

- PAI command

```

pai -name=lm_predict
  -project=algo_public
  -DinputFeatureTableName=input_data_table
  -DinputModelTableName=input_model_table
  -DmetaColNames=sample_key,label
  -DfeatureColNames=fea1,fea2
  -DoutputTableName=output_score_table

```

| Parameter                   | Description                                                                                                                                                                                                                                                                                                                                                         | Required | Default value           |
|-----------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|-------------------------|
| inputFeatureTableName       | The name of the input feature table.                                                                                                                                                                                                                                                                                                                                | Yes      | No default value        |
| inputFeatureTablePartitions | The partitions that are selected from the input feature table.                                                                                                                                                                                                                                                                                                      | No       | Full table              |
| inputModelTableName         | The name of the input model table.                                                                                                                                                                                                                                                                                                                                  | Yes      | No default value        |
| featureColNames             | The feature columns that are selected from the input table.                                                                                                                                                                                                                                                                                                         | No       | All columns             |
| metaColNames                | The columns that do not need to be converted. These columns in the output are the same as those in the input. You can specify labels and sample_id in the columns.                                                                                                                                                                                                  | No       | No default value        |
| outputFeatureScore          | Specifies whether to generate the scores of variables in the prediction results. Valid values:<br><ul style="list-style-type: none"> <li>◦ true: indicates that the system generates the scores of variables in the prediction results.</li> <li>◦ false: indicates that the system does not generate the scores of variables in the prediction results.</li> </ul> | No       | false                   |
| outputTableName             | The name of the output table.                                                                                                                                                                                                                                                                                                                                       | Yes      | No default value        |
| lifecycle                   | The lifecycle of the output table.                                                                                                                                                                                                                                                                                                                                  | No       | No default value        |
| coreNum                     | The number of cores.                                                                                                                                                                                                                                                                                                                                                | No       | Automatically allocated |
| memSizePerCore              | The memory size of each core. Unit: MB.                                                                                                                                                                                                                                                                                                                             | No       | Automatically allocated |

The following figure shows a score table generated by the Scorecard Prediction component.

| churn ▲ | prediction_score ▲  | prediction_prob ▲   | prediction_detail ▲                 |
|---------|---------------------|---------------------|-------------------------------------|
| 1       | -2.152581613419945  | 0.10409022740823... | {"0":0.8959097726,"1":0.1040902274} |
| 1       | 0.40321295914989297 | 0.599459362718963   | {"0":0.4005406373,"1":0.5994593627} |
| 0       | -5.9448781609701316 | 0.00261237905306... | {"0":0.9973876209,"1":0.0026123791} |
| 1       | -1.4235254136279643 | 0.19410950481015... | {"0":0.8058904952,"1":0.1941095048} |
| 0       | -0.4354127766052662 | 0.3928345539282477  | {"0":0.6071654461,"1":0.3928345539} |
| 0       | -2.322642905577369  | 0.08926496638122... | {"0":0.9107350336,"1":0.0892649664} |
| 1       | -1.8095060182152187 | 0.14069783849895... | {"0":0.8593021615,"1":0.1406978385} |
| 0       | 0.09435077042706097 | 0.5235702098997645  | {"0":0.4764297901,"1":0.5235702099} |
| 0       | -2.460605112010114  | 0.0786664686456529  | {"0":0.9213335314,"1":0.0786664686} |

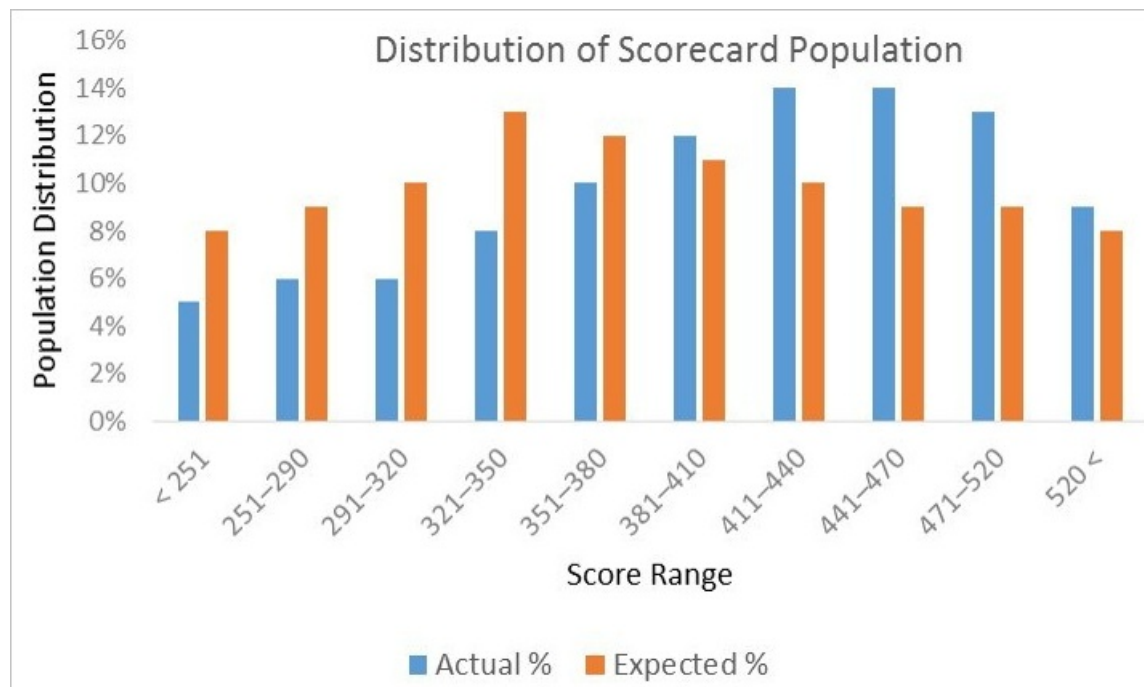
The churn column is the column appended to the result table from the input table. The data in this column does not affect the prediction results. The other three columns display the prediction results. The following table describes the three columns.

| Column            | Data type | Description                                                                                                                                                                                                                                                                            |
|-------------------|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| prediction_score  | DOUBLE    | The column that contains predicted scores. In a linear model, the feature values and model weight values are multiplied and summed up to obtain the predicted scores. In a scorecard model, if score transformation is performed, the transformed scores are generated in this column. |
| prediction_prob   | DOUBLE    | The column that contains the probability values of positive samples in binary classification. The probability values are transformed from the original scores (without score transformation) by using the sigmoid function.                                                            |
| prediction_detail | STRING    | The column that contains the probability values of positive and negative samples described in JSON strings. The value 0 represents negative and the value 1 represents positive. Example: {"0":0.1813110520,"1":0.8186889480}.                                                         |

## PSI

PSI is an important metric to identify a shift in two samples of a population. For example, you can use it to measure whether the changes in the population within two months are stable. A PSI value less than 0.1 indicates insignificant changes. A PSI value between 0.1 and 0.25 indicates minor changes. A PSI value greater than 0.25 indicates major changes.

If the changes in a population over time are unstable, you can use charts to identify the changes. You can perform binning for variables, calculate the number and proportion of the samples in each bin, and present the statistics in a column chart. The following figure shows a sample chart.



The preceding method can directly show whether a variable in two samples changes significantly. However, the shift in these changes cannot be measured by using this method. Therefore, the population stability cannot be automatically monitored. To resolve this issue, you can use the PSI component. The following figure shows the formula that is used to calculate PSI values.

$$PSI = \sum((Actual \% - Expected \%) \times (\ln(\frac{Actual \%}{Expected \%})))$$

You can configure the PSI component by using one of the following methods:

- PAI console

| Tab            | Parameter                    | Description                                                                                   |
|----------------|------------------------------|-----------------------------------------------------------------------------------------------|
| Fields Setting | Features for PSI Calculation | The feature columns that are required for PSI value calculation.                              |
| Tuning         | Cores                        | The number of CPU cores that are required. By default, the number is allocated by the system. |
|                | Memory Size                  | The memory size of each CPU core. By default, the memory size is allocated by the system.     |

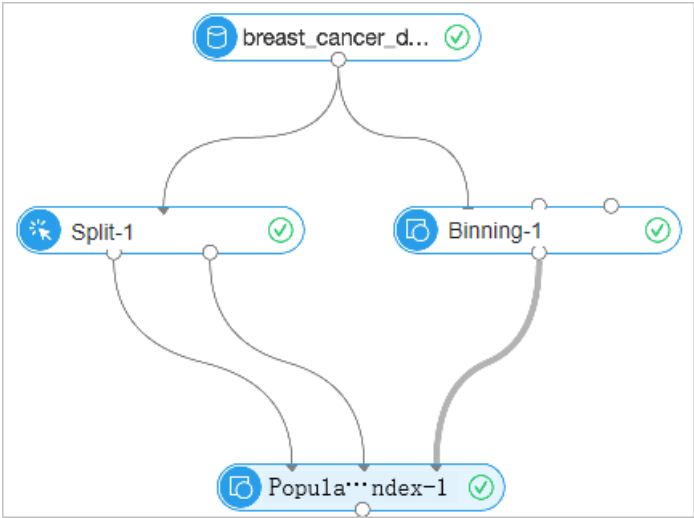
- PAI command

```
PAI -name psi
-project algo_public
-DinputBaseTableName=psi_base_table
-DinputTestTableName=psi_test_table
-DoutputTableName=psi_bin_table
-DinputBinTableName=pai_index_table
-DfeatureColNames=fea1,fea2,fea3
-Dlifecycle=7
```

| Parameter                | Description                                                                                                             | Required | Default value           |
|--------------------------|-------------------------------------------------------------------------------------------------------------------------|----------|-------------------------|
| inputBaseTableName       | The name of the base table. The shift of the population is calculated based on the samples in the base and test tables. | Yes      | No default value        |
| inputBaseTablePartitions | The partitions that are selected from the base table.                                                                   | No       | Full table              |
| inputTestTableName       | The name of the test table. The shift of the population is calculated based on the samples in the base and test tables. | Yes      | No default value        |
| inputTestTablePartitions | The partitions that are selected from the test table.                                                                   | No       | Full table              |
| inputBinTableName        | The name of the binning result table.                                                                                   | Yes      | No default value        |
| featureColNames          | The feature columns that are required for PSI value calculation.                                                        | No       | Full table              |
| outputTableName          | The name of the output table.                                                                                           | Yes      | No default value        |
| lifecycle                | The lifecycle of the output table.                                                                                      | No       | No default value        |
| coreNum                  | The number of CPU cores that are required.                                                                              | No       | Automatically allocated |
| memSizePerCore           | The memory size of each CPU core. Unit: MB.                                                                             | No       | Automatically allocated |

## Example on how to use PSI

Use the Binning component to perform binning for features. Then, connect the PSI component to the two sample datasets that you want to compare and the Binning component, as shown in the following figure. Specify the **Features for PSI Calculation** parameter.



The following figure shows the calculation results of PSI.

|     | Feature ▲  | Bin ▲    | Test % ▲ | Base % ▲ | Test - Base ▲ | ln(Test/Base) ▲ | PSI ▲  |
|-----|------------|----------|----------|----------|---------------|-----------------|--------|
| [-] | age        | -        | -        | -        | -             | -               | 0.0475 |
|     |            | (-inf,1] | 21.35    | 19.35    | 1.99          | 0.0979          | 0.0019 |
|     |            | (1,3]    | 23.1     | 21.99    | 1.11          | 0.049           | 0.0005 |
|     |            | (3,4]    | 10.23    | 12.9     | -2.67         | -0.2318         | 0.0062 |
|     |            | (4,5]    | 16.37    | 21.11    | -4.74         | -0.2542         | 0.0121 |
|     |            | (5,7]    | 7.89     | 8.5      | -0.61         | -0.0744         | 0.0005 |
|     |            | (7,9]    | 10.82    | 6.16     | 4.66          | 0.5635          | 0.0263 |
|     |            | (9,+inf) | 10.23    | 9.97     | 0.26          | 0.0261          | 0.0001 |
| [-] | menopause  | -        | -        | -        | -             | -               | 0.0588 |
|     |            | (-inf,1] | 56.14    | 53.08    | 3.06          | 0.0561          | 0.0017 |
|     |            | (1,2]    | 4.68     | 8.5      | -3.83         | -0.5976         | 0.0229 |
|     |            | (2,4]    | 15.2     | 11.14    | 4.06          | 0.3107          | 0.0126 |
|     |            | (4,6]    | 9.06     | 7.04     | 2.03          | 0.253           | 0.0051 |
|     |            | (6,9]    | 6.73     | 8.8      | -2.07         | -0.2686         | 0.0056 |
|     |            | (9,+inf) | 8.19     | 11.44    | -3.25         | -0.3343         | 0.0109 |
| [-] | tumor_size | -        | -        | -        | -             | -               | 0.0409 |



## 4.Components in deep learning frameworks

### 4.1. Overview and enabling of deep learning frameworks

Machine Learning Platform for AI (PAI) of Alibaba Cloud supports deep learning frameworks and provides GPU computing clusters with various features. You can use deep learning algorithms based on these frameworks and hardware resources.

#### Prerequisites

A project is created. For more information, see [Create a project](#).

#### Background information

Deep learning supports the following frameworks: MXNet 0.9.5, Caffe rc3, TensorFlow 1.4, and TensorFlow 1.8. TensorFlow and MXNet support custom coding in Python. Caffe supports customized net files.

Before you use a deep learning framework to train models, you must upload your data to Alibaba Cloud Object Storage Service (OSS). The algorithms read data from specified OSS directories when you run the algorithms. If you use algorithms to access OSS data in the same region, no costs are generated for the traffic. However, if you use algorithms to access OSS data in other regions, costs are generated for the traffic.

 **Note** GPU clusters are deployed only in the **China (Shanghai)** and **China (Beijing)** regions for PAI.

#### Enable deep learning

To use deep learning, you must select **By usage** in the **Open GPU** column for a specific project.

1. Log on to the [PAI console](#).
2. In the left-side navigation pane, choose **Model Training > Studio-Modeling Visualization**.
3. On the **PAI Visualization Modeling** page, find the created project and select **By usage** in the **Open GPU** column.

The projects for which you have selected **By usage** are allocated to a public resource pool. This way, the projects can use the underlying GPU computing resources.

### 4.2. Caffe

Caffe is an open source deep learning framework. This topic describes how to train models by using Caffe in Machine Learning Platform for AI (PAI).

#### Format conversion

The **Caffe** component does not support training data in a customized format. You must convert the data to the required format by using the **Format Conversion** component.

The input port of the **Format Conversion** component is connected to the **Read File Data** component.

- **Read File Data**

To use this component, you must set the OSS Data Path parameter. The following code shows the format of the `file_list` training data in the `bucket.hz.aliyun.com/train_img/train_file_list.txt` directory:

```
bucket/ilsvrc12_val/ILSVRC2012_val_00029021.JPEG 817
bucket/ilsvrc12_val/ILSVRC2012_val_00021046.JPEG 913
bucket/ilsvrc12_val/ILSVRC2012_val_00041166.JPEG 486
bucket/ilsvrc12_val/ILSVRC2012_val_00029527.JPEG 327
bucket/ilsvrc12_val/ILSVRC2012_val_00042825.JPEG 138
```

- **Format Conversion**

Configure Output OSS Directory and other parameters. `bucket_name.oss-cn-hangzhou-zmf.aliyuncs.com/ilsvrc12_val_convert` is used in this example. The Format Conversion component generates the converted `data_file_list.txt` and related data files. Format of `data_file_list`:

```
bucket/ilsvrc12_val_convert/train_data_00_01
bucket/ilsvrc12_val_convert/train_data_00_02
```

- Machine Learning Platform for AI console
  - OSS Path for Storing Image Lists: The OSS path in which images are stored.
  - File Prefix: The default value is `data`.
  - `resize_height`: The default value is 256.
  - `resize_width`: The default value is 256.
  - Encoding Type: You can select JPG, PNG, or Raw from the drop-down list.
  - Output OSS Directory: The directory to which output data is stored.
  - Shuffle: This option is selected by default.
  - Gray: specifies whether the image is grayscale. This option is cleared by default.
  - Image Mean: specifies whether to generate the image mean file. This option is cleared by default.
- PAI command

```
PAI -name convert_image_oss2oss
    -project algo_public_dev
    -Darn=acs:ram::1607128916545079:role/test-1
    -DossImageList=bucket_name.oss-cn-hangzhou-zmf.aliyuncs.com/image_list.txt
    -DossOutputDir=bucket_name.oss-cn-hangzhou-zmf.aliyuncs.com/your/dir
    -DencodeType=jpg
    -Dshuffle=true
    -DdataFilePrefix=train
    -DresizeHeight=256
    -DresizeWidth=256
    -DisGray=false
    -DimageMeanFile=false
```

| Parameter      | Description                                        | Value                                                                                                                                  | Required/Default value                                                                                                 |
|----------------|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| ossHost        | The endpoint of the OSS bucket.                    | Example: <code>oss-test.aliyun-inc.com</code> .                                                                                        | No. The default value is <code>oss-cn-hangzhou-zmf.aliyuncs.com</code> , which indicates the internal endpoint of OSS. |
| arn            | The ARN of the OSS role.                           | Example: <code>acs:ram::XXXXXXXXXXXX:role/ossaccessroleforodps</code> . XXX indicates the 16 digits generated for the ARN of the role. | Yes.                                                                                                                   |
| ossImageList   | The list of image files.                           | Example: <code>bucket_name/image_list.txt</code> .                                                                                     | Yes.                                                                                                                   |
| ossOutputDir   | The directory to which data is stored.             | Example: <code>bucket_name/your/dir</code> .                                                                                           | Yes.                                                                                                                   |
| encodeType     | The type of the encoding.                          | Example: JPG, PNG, or Raw.                                                                                                             | No. The default value is JPG.                                                                                          |
| shuffle        | Specifies whether to shuffle the data.             | A value of the BOOLEAN type.                                                                                                           | No. The default value is true.                                                                                         |
| dataFilePrefix | The prefix of the data file.                       | A value of the STRING type. Example: train or val.                                                                                     | Yes.                                                                                                                   |
| resizeHeight   | The height of the resized image.                   | A value of the int type. Users can customize the value.                                                                                | No. The default value is 256.                                                                                          |
| resizeWidth    | The width of the resized image.                    | A value of the int type. Users can customize the value.                                                                                | No. The default value is 256.                                                                                          |
| isGray         | Specifies whether the image is grayscale.          | A value of the BOOLEAN type.                                                                                                           | No. The default value is false.                                                                                        |
| imageMeanFile  | Specifies whether to generate the image mean file. | A value of the BOOLEAN type.                                                                                                           | No. The default value is false.                                                                                        |

## Caffe

Caffe is a deep learning framework that features clarity, high readability, and agility. For more information, see [The official website](#). You can configure parameters of the Caffe component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter         | Description                                                                                                     |
|--------------------|-------------------|-----------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Solver Path       | The OSS path in which the data is stored.                                                                       |
|                    | Limit Job Runtime | After you select this option, you can set the Limit Job Runtime parameter. Valid values: 1 to 168. Unit: hours. |
| Tuning             | GPUs              | The number of GPUs. Default value: 1.                                                                           |

- PAI command

```
PAI -name pluto_train_oss
-project algo_public_dev
-DossHost=oss-cn-hangzhou-zmf.aliyuncs.com
-Darn=acs:ram::1607128916545079:role/test-1
-DsolverPrototxtFile=bucket_name.oss-cn-hangzhou-zmf.aliyuncs.com/solver.prototxt
-DgpuRequired=1
```

| Parameter          | Required | Description                                                                                                                                                         | Default value                    |
|--------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|
| ossHost            | No       | The endpoint of the OSS bucket.                                                                                                                                     | oss-cn-hangzhou-zmf.aliyuncs.com |
| arn                | Yes      | The ARN of the OSS role. Example: <code>acs:ram::XXXXXXXXXXXXXXXX:role/ossaccessroleforodps</code> . XXX indicates the 16 digits generated for the ARN of the role. | No default value                 |
| solverPrototxtFile | Yes      | The OSS path of the solver file. The path must start with the bucket name.                                                                                          | No default value                 |
| gpuRequired        | No       | The number of GPUs.                                                                                                                                                 | 1                                |

Based on the optimized parallel computing, Solver is slightly different from Caffe. Differences:

- net: The net file is stored in the OSS path.
- type: This parameter is set to ParallelSGD, which is a string value.
- model\_average\_iter\_interval: the synchronous frequency in multiple GPUs. The value 1 indicates that the data is synchronized each round.
- snapshot\_prefix: The OSS output directory of the model.

```
net: "bucket/alexnet/train_val.prototxt"
test_iter: 1000
test_interval: 1000
base_lr: 0.01
lr_policy: "step"
gamma: 0.1
stepsize: 100000
display: 20
max_iter: 450000
momentum: 0.9
weight_decay: 0.0005
snapshot: 10000
snapshot_prefix: "bucket/snapshot/alexnet_train"
solver_mode: GPU
type: "ParallelSGD"
model_average_iter_interval: 1
```

Select `BinaryDataLayer` for `data_layer` in `train_val`. Example:

```

layer {
  name: "data"
  type: "BinaryData"
  top: "data"
  top: "label"
  include {
    phase: TRAIN
  }
  transform_param {
    mirror: true
    crop_size: 227
    mean_file: "bucket/imagenet_mean.binaryproto"
  }
  binary_data_param {
    source: "bucket/ilsvrc12_train_binary/data_file_list.txt"
    batch_size: 256
    num_threads: 10
  }
}
layer {
  name: "data"
  type: "BinaryData"
  top: "data"
  top: "label"
  include {
    phase: TEST
  }
  transform_param {
    mirror: false
    crop_size: 227
    mean_file: "bucket/imagenet_mean.binaryproto"
  }
  binary_data_param {
    source: "bucket/ilsvrc12_val_binary/data_file_list.txt"
    batch_size: 50
    num_threads: 10
  }
}

```

The name of the new **data Layer** is **BinaryData**. You can also use **transform param** to transform the input image data. The parameters are consistent with the native parameters of Caffe.

**binary\_data\_param** specifies the parameter settings for the data layer, including the following parameters:

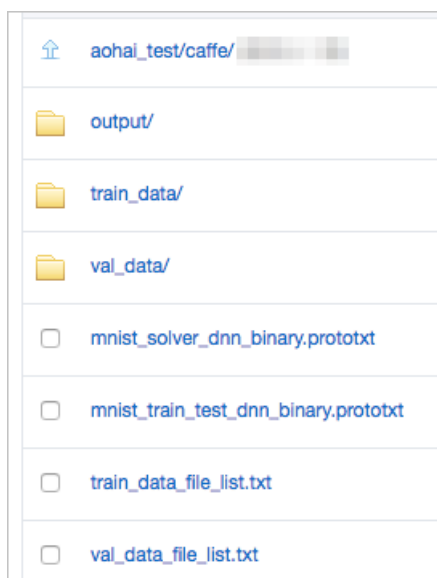
- **source**: the data source. The path of the data source is the same as the path that is specified in filelist. The value starts with a bucket name and does not contain `oss://`.
- **num\_threads**: the number of concurrent threads for reading OSS data. The default value is 10. You can modify this parameter as needed.

## Example

The following example shows how to train a model with MNIST data by using Caffe.

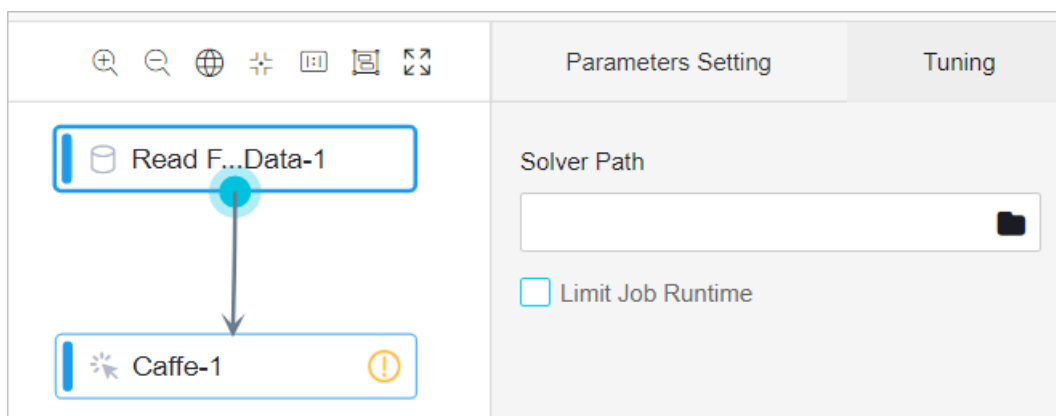
### 1. Prepare data sources

Download and extract Caffe data. For more information, see [Download Caffe resources](#). Upload the data to OSS. The path settings shown in the following figure are for reference only.



### 2. Run the experiment

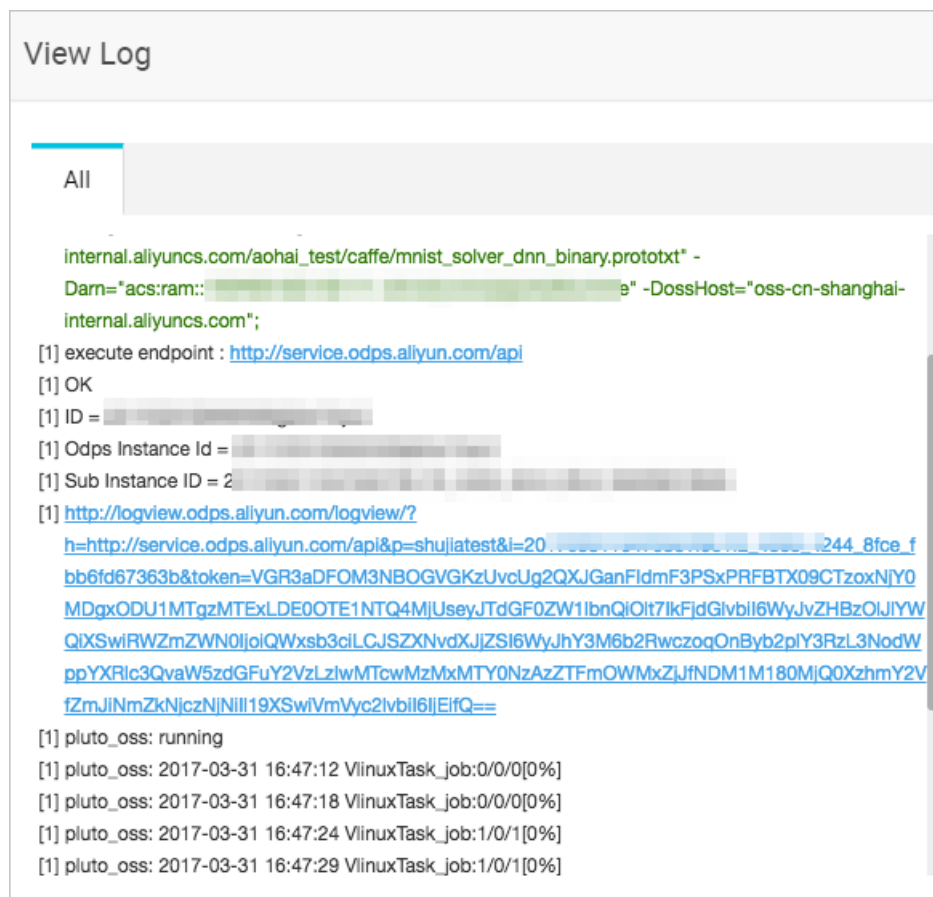
Drag the Caffe component and connect it with the Read File Data component.



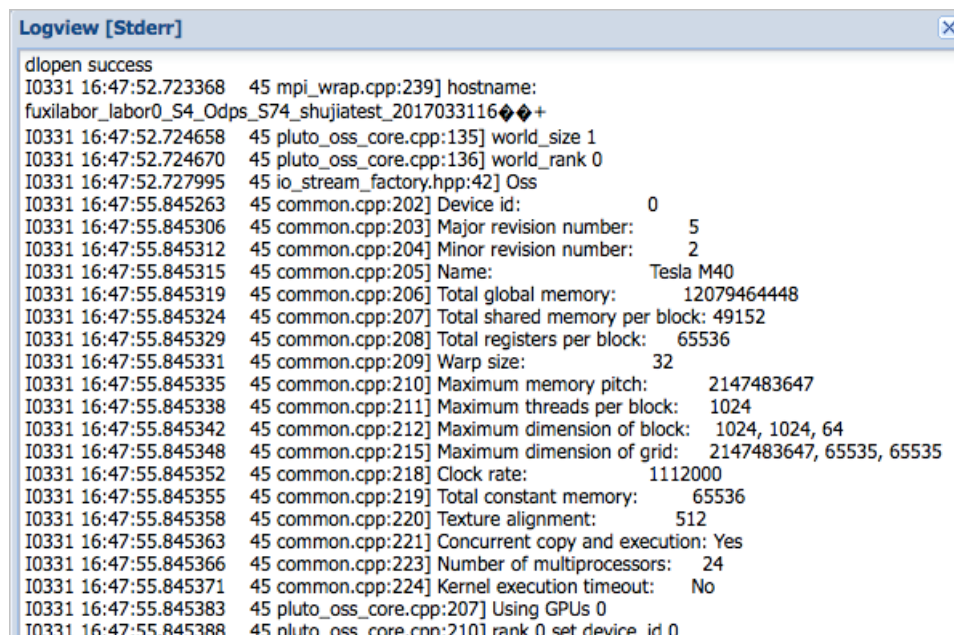
Set **Solver Path** to `mnist_solver_dnn_binary.prototxt` and click **Run** in the upper-left corner of canvas.

### 3. View Logs

Right-click the Caffe component and select **View Log**.



In the View Log message, click the **logview** link. On the page that appears, choose **ODPS Tasks > VlinuxTask > StdErr** to view the logs generated during the training.



## 4.3. MXNet



MXNet is a deep learning framework that supports imperative and symbolic programming. You can run MXNet on CPU or GPU clusters. This topic describes how to use the MXNet component in Machine Learning Platform for AI (PAI).

## Configure the component

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                                                | Description                                                                                                                                          |
|--------------------|----------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Python Code Files                                        | The program execution file. Multiple files can be packaged to a TAR.GZ file and then uploaded.                                                       |
|                    | Primary Python File                                      | The primary file in a code file package.                                                                                                             |
|                    | Data Source Directory                                    | The path to data sources in Object Storage Service (OSS).                                                                                            |
|                    | Configuration File Hyperparameters and Custom Parameters | MXNet allows you to use commands to pass in hyperparameter settings. You can test different learning rates and batch sizes during model experiments. |
|                    | Output Directory                                         | The output directory of the model.                                                                                                                   |
|                    | Limit Job Runtime                                        | After you select this option, you can enter the maximum scheduled time of job execution. Valid values: 1 to 168. Unit: hours.                        |
| Tuning             | GPUs                                                     | The number of GPUs. Default value: 1.                                                                                                                |

- PAI command

```
PAI -name mxnet_ext
-Ddescript="oss://imagenet.oss-cn-shanghai-internal.aliyuncs.com/mxnet-ext-code/mxnet_cifar10_demo.tar.gz"
-project algo_public_dev
-DentryFile="train_cifar10.py"
-Dbuckets="oss://imagenet.oss-cn-shanghai-internal.aliyuncs.com"
-DcheckpointDir="oss://imagenet.oss-cn-shanghai-internal.aliyuncs.com/mxnet-ext-model/"
-DhyperParameters="oss://imagenet.oss-cn-shanghai-internal.aliyuncs.com/mxnet-ext-code/hyperparam.txt.single"
-Darn="acs:ram::1664081855183111:role/role-for-pai";
```

You do not need to configure all parameters. We recommend that you do not directly copy the preceding command. The following table describes the parameters.

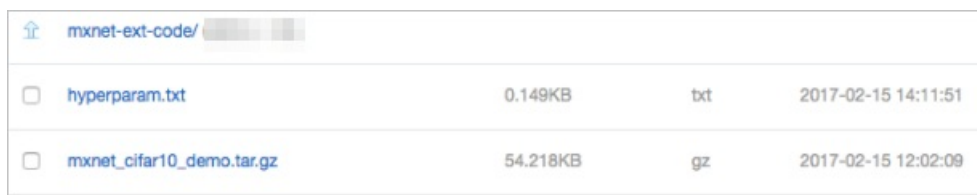
| Parameter | Required | Description | Default value |
|-----------|----------|-------------|---------------|
|-----------|----------|-------------|---------------|

| Parameter       | Required | Description                                                                                                 | Default value                                                                 |
|-----------------|----------|-------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| script          | Yes      | The MXNet algorithm file. This file can be a single file or TAR.GZ file.                                    | oss://imagenet.oss-cn-shanghai-internal.aliyuncs.com/smoke_mxnet/mnist_ext.py |
| entryFile       | No       | The entry file for the algorithm. If the script is a TAR.GZ file, this parameter is required.               | No default value                                                              |
| buckets         | No       | The input bucket. Separate multiple buckets with commas (,). Each bucket must end with a forward slash (/). | No default value                                                              |
| hyperParameters | No       | The path of the command line hyperparameters.                                                               | No default value                                                              |
| gpuRequired     | No       | The number of used GPUs.                                                                                    | 100                                                                           |
| checkpointDir   | No       | The directory of the checkpoint.                                                                            | No default value                                                              |

## Example

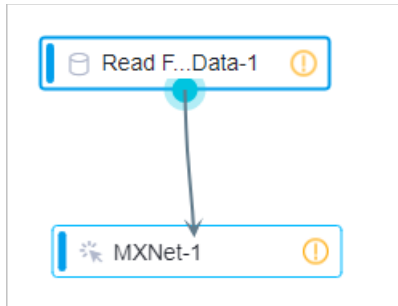
The CIFAR-10 dataset is provided by MXNet and contains 60,000  $32 \times 32$  color images in 10 different categories. This dataset is commonly used to train machine learning algorithms to recognize objects and sort them into airplanes, cars, birds, cats, deers, dogs, frogs, horses, ships, or trucks. For more information, visit [The CIFAR-10 dataset](#).

1. Upload the Python execution file and training dataset to OSS. In this example, a bucket named tfmnist is created in the China (Shanghai) region.

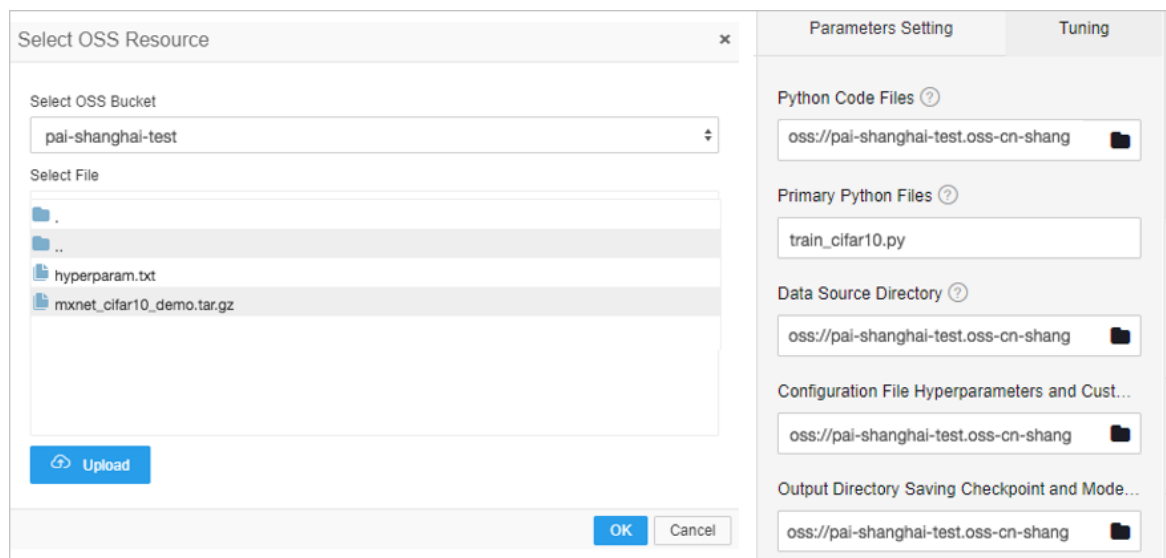


|                          |                           |          |     |                     |
|--------------------------|---------------------------|----------|-----|---------------------|
| mxnet-ext-code/          |                           |          |     |                     |
| <input type="checkbox"/> | hyperparam.txt            | 0.149KB  | txt | 2017-02-15 14:11:51 |
| <input type="checkbox"/> | mxnet_cifar10_demo.tar.gz | 54.218KB | gz  | 2017-02-15 12:02:09 |

2. Drag a MXNet component and connect it with a Read File Data component. Specify the region for the OSS bucket and complete RAM authorization.



- Configure MXNet component parameters. You can set the paths of the Python execution file and the data source file based on the following figure.



- Select a TAR.GZ file for Python Code Files.
  - Select an entry file in the tar file for Primary Python File.
  - Select a file in the .txt.single format for Configuration File Hyperparameters and Custom Parameters.
  - The checkpoint directory is the output directory of the model.
- Click **Run** in the upper-left corner of the canvas and wait for the component running task to complete.
  - Right-click the MXNet component and select **View Log** to view the operational logs.

6. The following figure shows the model generated in the checkpoint directory.

## 4.4. PyTorch

## Procedure

The screenshot displays the DataWorks console interface. On the left, a workflow diagram shows a task named 'Read F...Data-4' (with a Read icon) connected by an arrow to a task named 'Pytorch-1' (with a PyTorch icon). Both tasks have a green checkmark in the top right corner. On the right, the 'Fields Setting' panel is visible. It contains the following information:

- OSS Data Path**: A text input field containing the path `oss://aohailibo.oss-cn-beijing-internal.i...` and a folder icon on the right.
- Role Name**: A text input field containing the role name `AliyunODPSPAIDefaultRole`.
- ARN**: A text input field containing the ARN `acs:ram::...`.

You can configure the component by using one of the following methods:

- Machine Learning Platform for AI console

| Tab                | Parameter                                                | Description                                                                                                                                                                                                                              |
|--------------------|----------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Parameters Setting | Python Version                                           | Valid values: 2.7 and 3.6.                                                                                                                                                                                                               |
|                    | Python Code Files                                        | The path of the code file. If you upload a <i>tar.gz</i> file, you must specify Primary Python File. If you upload a single PY file, Primary Python File is not required.                                                                |
|                    | Primary Python File                                      | Optional. The name of the primary code file. If you upload a single file, this parameter is not required. If you upload a package, you must set this parameter to the path in which the file is stored. Example: <i>train/train.py</i> . |
|                    | Data Source Directory                                    | The OSS path in which data sources are stored.                                                                                                                                                                                           |
|                    | Configuration File Hyperparameters and Custom Parameters | The hyperparameter file, in the <b>key-value</b> pair format.                                                                                                                                                                            |
|                    | Checkpoint Output Directory/Model Input Directory        | The output directory of the model.                                                                                                                                                                                                       |
|                    | Limit Job Runtime                                        | Specifies whether to limit the time for job execution.                                                                                                                                                                                   |
|                    | Maximum Scheduled Job Runtime                            | Unit: hours. The default value is 24 hours.                                                                                                                                                                                              |
| Tuning             | GPUs per Worker                                          | The number of GPUs for each worker.                                                                                                                                                                                                      |
|                    | Workers                                                  | The number of distributed machines.                                                                                                                                                                                                      |

- PAI command

```
PAI -name pytorch_ext -DossHost="oss-cn-beijing-internal.aliyuncs.com"
-Dcluster="{\"worker\":{\"gpu\":100}}" -DworkerCount="2"
-Dpython="3.6"
-Dinputs="oss://${The name of the OSS bucket}.oss-cn-beijing-internal.aliyuncs.com/mnist/"
-Darn="acs:ram::168069136****:role/aliyunodpspaidefaultrole"
-Dscript="oss://${The name of the OSS bucket}.oss-cn-beijing-internal.aliyuncs.com/pytorch/pytorch_
dist_mnist.py"
-DcheckpointDir="oss://${The name of the OSS bucket}.oss-cn-beijing-internal.aliyuncs.com/pytorch/"
;
```

| Parameter        | Description                                                                                                |
|------------------|------------------------------------------------------------------------------------------------------------|
| DossHost         | The endpoint of the OSS bucket.                                                                            |
| Dcluster         | The number of GPUs for each worker. The value 100 indicates one GPU, and the value 200 indicates two GPUs. |
| DworkerCount     | The number of workers.                                                                                     |
| Dpython          | The Python version. Valid values: 2.7 and 3.6.                                                             |
| Dinputs          | The input path of data sources.                                                                            |
| Darn             | The ARN of the OSS role.                                                                                   |
| Dscript          | The name of the code file.                                                                                 |
| DcheckpointDir   | The path in which the model is stored.                                                                     |
| DhyperParameters | The hyperparameter file.                                                                                   |

```
import argparse
parser = argparse.ArgumentParser()
parser.add_argument("--inputs", type=str, default='inputspath')
parser.add_argument("--checkpointDir", type=str, default='Checkpoint dir')

args = parser.parse_args()
```

To obtain parameters configured on the web side, you must specify parser in the code. For example, if the path of the OSS data source is specified in the component parameters, you can specify inputs in the code to obtain the path. For more information, see [Example](#).

## Example

1. Download the [Distributed MNIST file code processed by PyTorch](#) and upload the file to OSS. You must enter your AccessKey pair.

```
import argparse
parser = argparse.ArgumentParser()
parser.add_argument("--inputs", type=str, default='inputspath')
parser.add_argument("--checkpointDir", type=str, default='Checkpoint dir')

args = parser.parse_args()

ak = "" # YourAccessKey
akSec = "" # YourAccessKeySecret
```

2. Download the MNIST training file and MNIST test file, and upload them to the OSS folder.
3. Configure the parameters as needed.

The following figure shows an example of parameter settings.

#### 4. Configure resources.

The sample code is distributed code, and the number of workers must be greater than 1. The following figure shows an example of parameter settings.

## 4.5. Deep learning code downloads

This topic provides links to download sample code of deep learning.

- [Download TensorFlow resource](#)
- [Download MXNet resources](#)
- [Download Caffe resources](#)

## 5.AutoML

### 5.1. Use AutoML for automatic parameter tuning

This topic describes the algorithm used by AutoML for automatic parameter tuning and how to use AutoML.

#### Procedure

1. Log on to the [Machine Learning Platform for AI \(PAI\) console](#) and go to the Algorithm Platform tab.
2. In the left-side navigation pane, click **Experiments** and select an experiment.  
The **Air Quality Prediction** experiment is used in this example.
3. On the tab that appears, choose **Auto ML > Auto Parameter Tuning**.
4. In the **Auto Parameter Tuning** dialog box, set the Select Algorithm parameter and click **Next**.

Auto Parameter Tuning

1 Select Algorithm

☐ Random Forest-1

☒ Logistic Regression for Binary Classification-1

2 Configure Parameter Tuning

3 Configure Model Output

Cancel Next

**Note** If multiple algorithms are used in an experiment, select only one algorithm.

5. In the **Configure Parameter Tuning** step, set the **Parameter Tuning Method** parameter and click **Next**. Alibaba Cloud PAI provides the following tuning methods:
  - o EVOLUTIONARY\_OPTIMIZER
    - a. Randomly select A parameter candidate sets. A indicates the value of **Exploration Samples**.



- b. Use the N parameter candidate sets that have high evaluation metrics as the parameter candidate sets for the next iteration.
- c. Continue the exploration within R times as the standard deviation range around these parameters to explore new parameter sets. The new parameter sets replace the last (A - N) parameter sets by evaluation metric in the previous round. R indicates the value of **Convergence Coefficient**.
- d. Iterate the exploration for M rounds based on the preceding logic until the optimal parameter set is obtained. M indicates the value of **Explorations**.

Based on the preceding principle, the final number of models is calculated by using the following formula:  $A + (A - N) \times M$ .

 **Notice** The first value of N is defined as  $A/2 - 1$ , and subsequent values of N are  $N/2 - 1$  (rounded up).

**Auto Parameter Tuning**
✕

1

Select Algorithm  
Logistic Regression for Binary Classification-1

2

Configure Parameter Tuning

Data Splitting Ratio:

0.7

Parameter Tuning Method:

EVOLUTIONARY\_OPTIMIZER

Exploration Samples:

5

Explorations:

2

Convergence Coefficient:

0.5

| Parameters                   | Default Range | Custom Range                                                                                                                                                                                                                                                |
|------------------------------|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Regularization Type          | None          | <div style="border: 1px solid #ccc; padding: 2px 5px;">None</div>                                                                                                                                                                                           |
| Regularization Coefficient   | 1             | <div style="display: flex; align-items: center;"> <div style="border: 1px solid #ccc; padding: 2px 5px; flex: 1;">0.1</div> <div style="margin: 0 5px;">-</div> <div style="border: 1px solid #ccc; padding: 2px 5px; flex: 1;">2</div> </div>              |
| Minimum Convergence Deviance | 0.000001      | <div style="display: flex; align-items: center;"> <div style="border: 1px solid #ccc; padding: 2px 5px; flex: 1;">0.00000001</div> <div style="margin: 0 5px;">-</div> <div style="border: 1px solid #ccc; padding: 2px 5px; flex: 1;">0.00001</div> </div> |
| Maximum Iterations           | 100           | <div style="display: flex; align-items: center;"> <div style="border: 1px solid #ccc; padding: 2px 5px; flex: 1;">50</div> <div style="margin: 0 5px;">-</div> <div style="border: 1px solid #ccc; padding: 2px 5px; flex: 1;">500</div> </div>             |

3

Configure Model Output

Previous

Next

- **Data Splitting Ratio:** the ratio of training sets to evaluation sets. Input data sources are divided into training and evaluation sets. The value 0.7 indicates that 70% of the data is used for model training and 30% for evaluation.

- Exploration Samples: the number of parameter sets for each iteration. A high value indicates high accuracy and large calculation. This parameter must be set to a positive integer in the range of 5 to 30.
  - Explorations: the number of iterations. A high value indicates high accuracy and large calculation. This parameter must be set to a positive integer in the range of 1 to 10.
  - Convergence Coefficient: used to tune exploration ranges. A small value indicates fast convergence. However, optimal parameters may be missed. This parameter must be set to a value in the range of 0.1 to 1.
  - Custom Range: You must enter the tuning range for each parameter. If a parameter range is not configured, the default range is used, and this parameter is not included in automatic parameter tuning.
- RANDOM\_SEARCH
- a. Randomly select a value for each parameter within the parameter range.
  - b. Enter random values into a set of parameters for model training.
  - c. Perform M rounds and sort the output models. M indicates the value of **Explorations**.

Auto Parameter Tuning
✕

✓

Select Algorithm

Logistic Regression for Binary Classification-1

2

Configure Parameter Tuning

Data Splitting Ratio: 0.7

Parameter Tuning Method: RANDOM\_SEARCH

Iterations 5

| Parameters                   | Default Range | Custom Range                                                                                                                                         |
|------------------------------|---------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| Regularization Type          | None          | <span style="border: 1px solid #ccc; padding: 2px 10px;">None</span>                                                                                 |
| Regularization Coefficient   | 1             | <span style="border: 1px solid #ccc; padding: 2px 10px;">0.1</span> - <span style="border: 1px solid #ccc; padding: 2px 10px;">2</span>              |
| Minimum Convergence Deviance | 0.000001      | <span style="border: 1px solid #ccc; padding: 2px 10px;">0.00000001</span> - <span style="border: 1px solid #ccc; padding: 2px 10px;">0.00001</span> |
| Maximum Iterations           | 100           | <span style="border: 1px solid #ccc; padding: 2px 10px;">50</span> - <span style="border: 1px solid #ccc; padding: 2px 10px;">500</span>             |

3

Configure Model Output

Previous
Next

- Iterations: the number of searches in the configured range. This parameter must be set to a positive integer in the range of 2 to 50.

- Data Splitting Ratio: the ratio of training sets to evaluation sets. Input data sources are divided into training and evaluation sets. The value 0.7 indicates that 70% of data is used for model training and 30% for evaluation.
  - Custom Range: You must enter the tuning range for each parameter. If a parameter range is not configured, the default range is used, and this parameter is not included in automatic parameter tuning.
- GRID\_SEARCH
    - a. Split the value range of each parameter into N segments. N indicates the value of **Grids**.
    - b. Randomly select a value from the N segments. Assume that M parameters exist.  $N^M$  parameter sets can be combined.
    - c. Generate a total of  $N^M$  models based on the  $N^M$  parameter sets. Then, sort the models.

**Auto Parameter Tuning**
✕

✓

Select Algorithm

Logistic Regression for Binary Classification-1

2

Configure Parameter Tuning

Data Splitting Ratio:

Parameter Tuning Method:

Grids

| Parameters                   | Default Range | Custom Range                                                                                                            |
|------------------------------|---------------|-------------------------------------------------------------------------------------------------------------------------|
| Regularization Type          | None          | <input style="width: 100px;" type="text" value="None"/>                                                                 |
| Regularization Coefficient   | 1             | <input style="width: 50px;" type="text" value="0.1"/> - <input style="width: 50px;" type="text" value="2"/>             |
| Minimum Convergence Deviance | 0.000001      | <input style="width: 50px;" type="text" value="0.0000001"/> - <input style="width: 50px;" type="text" value="0.00001"/> |
| Maximum Iterations           | 100           | <input style="width: 50px;" type="text" value="50"/> - <input style="width: 50px;" type="text" value="500"/>            |

3

Configure Model Output

- Grids: the number of split grids. This parameter must be set to a positive integer in the range of 2 to 10.
- Data Splitting Ratio: the ratio of training sets to evaluation sets. Input data sources are divided into training and evaluation sets. The value 0.7 indicates that 70% of data is used for model training and 30% for evaluation.

- Custom Range: You must enter the tuning range for each parameter. If a parameter range is not configured, the default range is used, and this parameter is not included in automatic parameter tuning.
- UserDefine

Auto Parameter Tuning
✕

✓

Select Algorithm

Logistic Regression for Binary Classification-1

2

Configure Parameter Tuning

Data Splitting Ratio: 0.7 ▼

Parameter Tuning Method: UserDefine ▼

| Parameters                   | Default Range | Custom Range                                                                         |
|------------------------------|---------------|--------------------------------------------------------------------------------------|
| Regularization Type          | None          | <span style="border: 1px solid #ccc; padding: 2px 10px;">None</span>                 |
| Regularization Coefficient   | 1             | <span style="border: 1px solid #ccc; padding: 2px 10px;">[0.1,2]</span>              |
| Minimum Convergence Deviance | 0.000001      | <span style="border: 1px solid #ccc; padding: 2px 10px;">[0.00000001,0.00001]</span> |
| Maximum Iterations           | 100           | <span style="border: 1px solid #ccc; padding: 2px 10px;">[50,500]</span>             |

3

Configure Model Output

Previous
Next

**Custom Range:** The system calculates the score based on all combinations of your parameter values. If the ranges are not configured, the default ranges of parameters are used.

? **Note** In PAI V2.0, the number of parameter tuning algorithms increases from 4 to 7. The following table describes the algorithms.

| Algorithm | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| GAUSE     | This algorithm indicates Gaussian processes. Gaussian processes are examples of Bayesian nonparametric models and are widely used for hyperparameter tuning. Gaussian processes allow the system to observe hyperparameter configuration performance to fit the proxy model, and strengthen decision-making by using model prediction. This way, the system can purposefully select the appropriate hyperparameter results from a limited number of attempts. |

| Algorithm              | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SAMPLE                 | This algorithm is developed by the PAI team and DAMO Academy. If large amounts of data are processed, this algorithm allows the system to estimate the final result that a set of hyperparameters can obtain based on part of the data. The SAMPLE algorithm, in combination with the population-based training (PBT) algorithm, allows the system to gradually improve the sampling ratio with the increase in the number of hyperparameters. This way, broader and faster explorations can be performed.                                                                                                                                                                                                                                                                                                                                                        |
| EVOLUTIONARY_OPTIMIZER | This algorithm is developed by the PAI team based on the techniques of the PBT algorithm. In the EVOLUTIONARY_OPTIMIZER algorithm logic, parameter tuning is considered a multi-round iteration to progressively explore an optimal solution. The Exploration Samples parameter indicates the number of samples in each iteration. The Explorations parameter indicates the number of iterations. The Convergence Coefficient parameter indicates the step size of each iteration. The EVOLUTIONARY_OPTIMIZER algorithm allows the system to discard the exploration samples that have unsatisfactory results after each iteration. Then, the system expands exploration samples in the set of exploration samples that have satisfactory results to form the sample set for the next iteration. This process is repeated until all the iterations are completed. |
| PBT                    | PBT is an evolutionary algorithm based on the concept of population. In the PBT algorithm logic, hyperparameter configurations are considered a population and the search process is considered a dynamic environment. In iterations, this algorithm allows the system to filter hyperparameter configurations based on the principle of survival of the fittest. This way, the hyperparameter configurations with the best performance can be obtained. This algorithm is simple and adapts to different data structures to help you obtain optimal results in deep learning model training.                                                                                                                                                                                                                                                                     |
| GRID_SEARCH            | This algorithm allows the system to equally split the parameters that are used in parameter tuning into several sets based on a specific ratio. Then, the system randomly combines the split parameter sets to generate parameter candidate sets for calculation and comparison.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |

| Algorithm     | Description                                                                                                                                                                     |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| RANDOM_SEARCH | The RANDOM_SEARCH algorithm allows the system to randomly obtain and combine samples from each parameter space to form parameter candidate sets for calculation and comparison. |
| UserDefine    | Users define parameter sets.                                                                                                                                                    |

6. In the **Configure Model Output** step, configure model output parameters and click **Next**.

Auto Parameter Tuning

✓

Select Algorithm

Logistic Regression for Binary Classification-1

✓

Configure Parameter Tuning

Data Splitting Ratio: 0.7  
 Parameter Tuning Method: Evolutionary Optimizer  
 Exploration Samples: 5  
 Explorations: 2  
 Convergence Coefficient: 0.5

| Parameters                   | Value    |
|------------------------------|----------|
| Regularization Type          | None     |
| Regularization Coefficient   | [0.1,2]  |
| Minimum Convergence Deviance | 0.000001 |
| Maximum Iterations           | 100      |

3

Configure Model Output

Generated Models: 11 Formula

Algorithm Type

Binary Classification Evaluation

Evaluation Criteria

AUC

Saved Models

5

AUC:

The most common indicator used in the two classification algorithm, which indicates the area under the ROC curve. The larger the value, the better the model.

Pass Down Model:

☒

Note: The model with the optimal evaluation among the saved models will be passed down.

Previous

Next

- Evaluation Criteria: Valid values are **AUC**, **F1-score**, **Precision**, and **Recall**.
- Saved Models: Valid values are 1, 2, 3, 4, and 5. The system ranks models based on the value of the Evaluation Criteria parameter that you set and saves the top-ranked models based on the value of the **Saved Models** parameter.
- Pass Down Model: This switch is turned on by default. If the switch is turned off, the model generated by the default parameters of the current component are passed down to the node of the subsequent component. If the switch is turned on, the optimal model generated by automatic parameter tuning is passed down to the node of the subsequent component.

- After the configuration, click **Run** in the upper-left corner of the canvas. After the preceding configuration is completed, the **Auto ML** switch of the related algorithm is turned on. You can turn the switch on or off as needed.
- (Optional) Right-click the model and select **Edit AutoML parameters** to modify AutoML parameters.

## Result

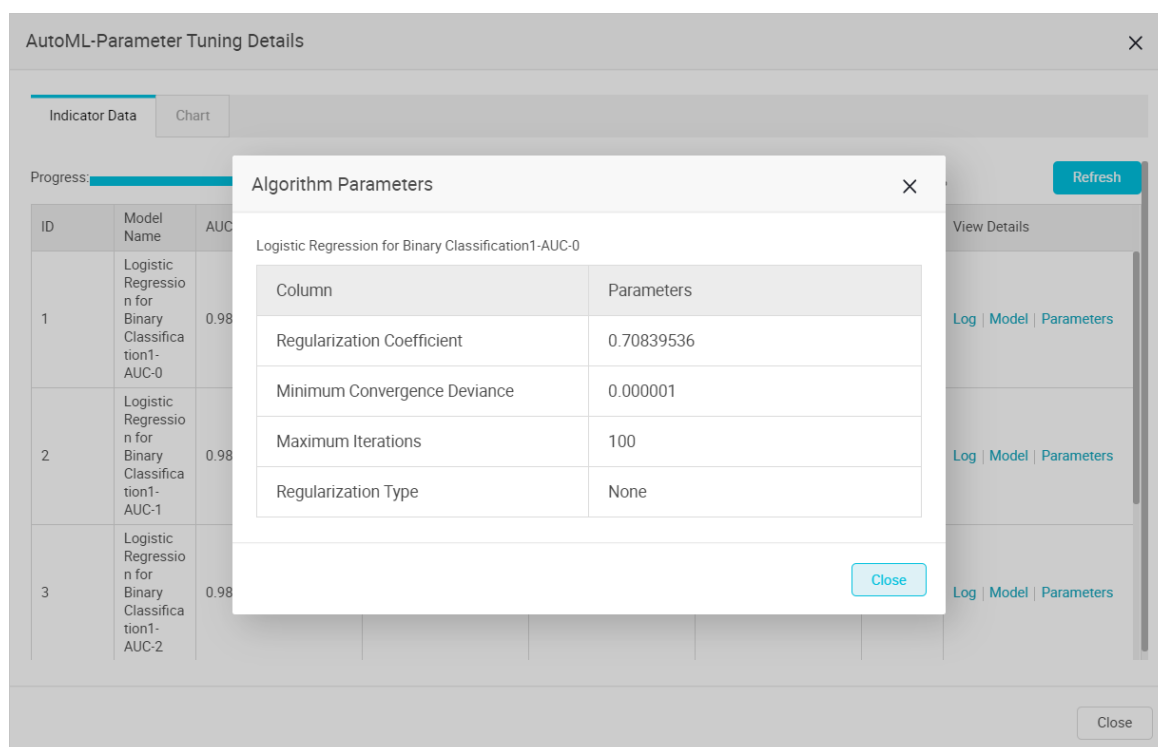
Model output:

- During parameter tuning, right-click the model component and select **Parameter Tuning Details**.
- In the **AutoML-Parameter Tuning Details** dialog box, click the **Metrics** tab to view the tuning progress and the running status of each model.

| AutoML-Parameter Tuning Details |                                                      |                             |                     |                      |                   |            |                                                                          |
|---------------------------------|------------------------------------------------------|-----------------------------|---------------------|----------------------|-------------------|------------|--------------------------------------------------------------------------|
| Indicator Data                  |                                                      | Chart                       |                     |                      |                   |            |                                                                          |
| Progress:                       |                                                      | <div><div></div></div> 100% |                     |                      |                   |            | <button>Refresh</button>                                                 |
| ID                              | Model Name                                           | AUC $\uparrow$              | F1 Score $\uparrow$ | PRECISION $\uparrow$ | RECALL $\uparrow$ | Status     | View Details                                                             |
| 1                               | Logistic Regression for Binary Classification1-AUC-0 | 0.984970773381295           | 0.823529411764706   | 0.8727272727272727   | 0.75              | Successful | <a href="#">Log</a>   <a href="#">Model</a>   <a href="#">Parameters</a> |
| 2                               | Logistic Regression for Binary Classification1-AUC-1 | 0.984970773381295           | 0.823529411764706   | 0.8727272727272727   | 0.75              | Successful | <a href="#">Log</a>   <a href="#">Model</a>   <a href="#">Parameters</a> |
| 3                               | Logistic Regression for Binary Classification1-AUC-2 | 0.984970773381295           | 0.823529411764706   | 0.8727272727272727   | 0.75              | Successful | <a href="#">Log</a>   <a href="#">Model</a>   <a href="#">Parameters</a> |

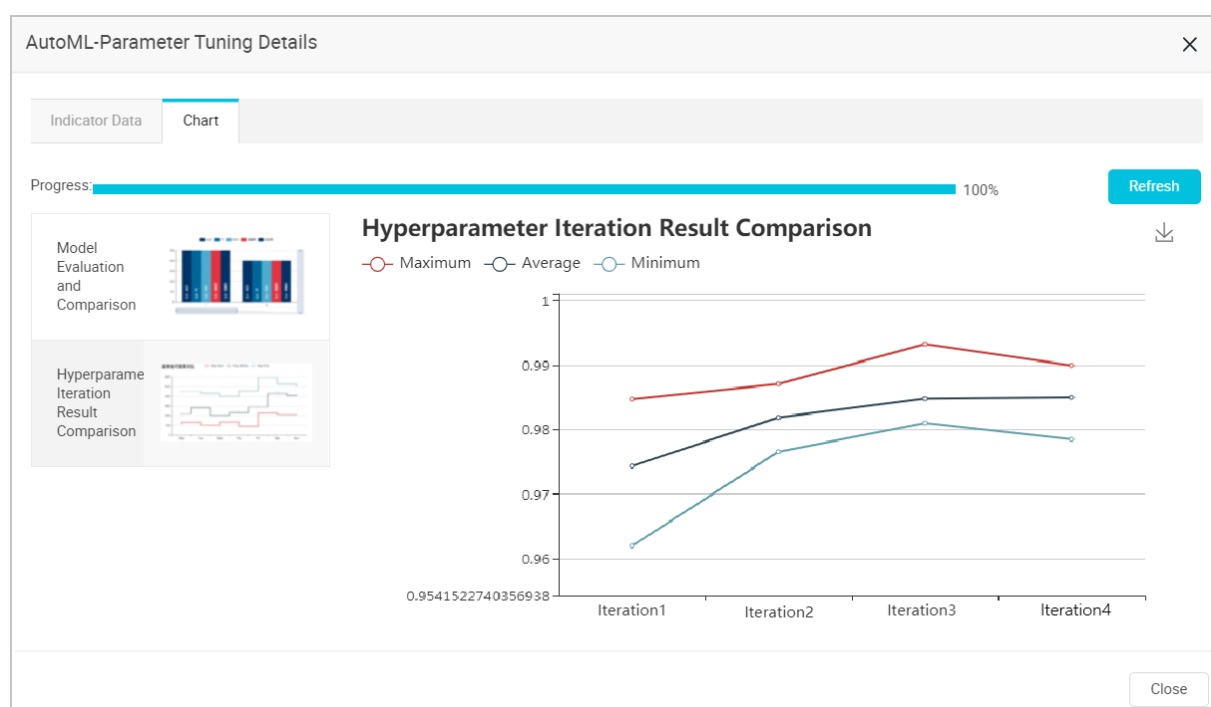
- Sort candidate models based on metrics. The metrics include **AUC**, **F1-score**, **Accuracy**, and **Recall Rate**.
- In the **View Details** column, click **Log** or **Parameter** to view the logs and parameters of each candidate model.





Display of parameter tuning effect

You can view the growth trend of the evaluation metrics of updated parameters in the **Hyperparameter Iteration Result Comparison** chart.



Model storage:

1. In the left-side navigation pane of the Algorithm Platform tab, click **Models**.
2. Click **Experiment Models**.
3. Click the required experiment folder to view the model saved by using AutoML.

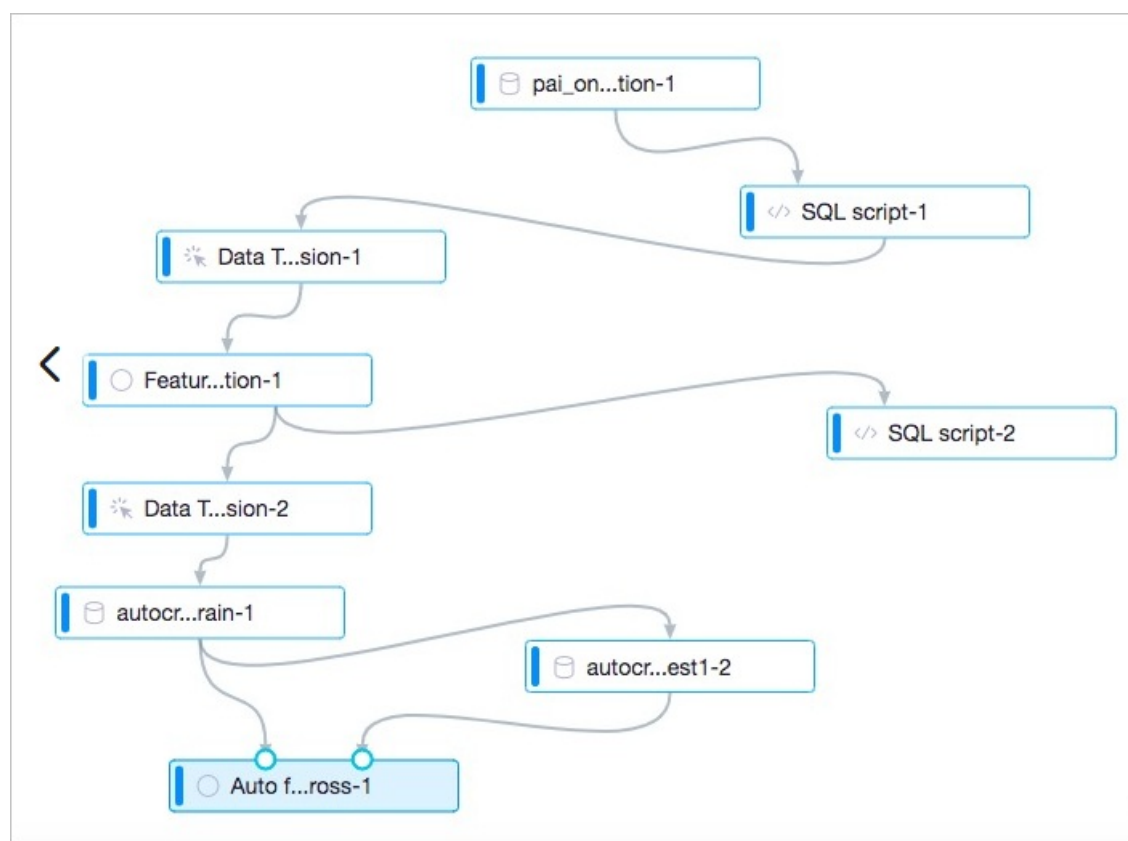
## 5.2. Use AutoML for automatic feature engineering

Feature engineering is essential to model training in machine learning. Feature engineering helps find feature crosses that are beneficial to models. Generally, algorithm engineers must spend a lot of efforts in feature engineering. Machine Learning Platform for AI (PAI) provides the Auto Feature Cross component to help you find effective feature crosses. This topic describes how to use the Auto Feature Cross component.

### Flowchart

The Auto Feature Cross component is developed based on the deep learning framework TensorFlow. This component involves intensive parallel computing at the underlying layer and requires GPU resources. The Auto Feature Cross component is supported only in the China (Beijing) and China (Shanghai) regions.

The following flowchart shows the process of automatic feature engineering.



**Note** You can create an experiment that contains the Auto Feature Cross component by using a template in the **Templates** section on the **Home** page. In this case, you must set the **Output path** parameter of the Auto Feature Cross component to the URL of the Object Storage Service (OSS) bucket under your account.

### 1. Authorize PAI to access your GPU resources and OSS bucket

1. Log on to the [PAI console](#). In the left-side navigation pane, choose **Model Training > Studio-**

**Modeling Visualization.** On the page that appears, find the project in which you want to perform operations and click **Machine Learning** in the **Operation** column.

- On the page that appears, click **Settings** in the left-side navigation pane. On the **Settings** page, select the **Authorize Machine Learning Platform for AI to access my OSS resources and Pay by used** check boxes in the **General** pane.

## 2. Bin data

The Auto Feature Cross component supports only the BIGINT data type. However, raw data in most business scenarios are of the DOUBLE data type, as shown in the following figure.

| age ▲ | sex ▲  | cp ▲   | trestbps ▲ | chol ▲ | fbs ▲ | restecg ▲ | thalach ▲ | exang ▲ | oldpeak ▲ | slop ▲ | ca ▲ | thal ▲ | status ▲ | style ▲ |
|-------|--------|--------|------------|--------|-------|-----------|-----------|---------|-----------|--------|------|--------|----------|---------|
| 63.0  | male   | an...  | 145.0      | 233.0  | true  | hyp       | 150.0     | false   | 2.3       | down   | 0.0  | fix    | buff     | H       |
| 67.0  | male   | as...  | 160.0      | 286.0  | false | hyp       | 108.0     | true    | 1.5       | flat   | 3.0  | norm   | sick     | S2      |
| 67.0  | male   | as...  | 120.0      | 229.0  | false | hyp       | 129.0     | true    | 2.6       | flat   | 2.0  | rev    | sick     | S1      |
| 37.0  | male   | not... | 130.0      | 250.0  | false | norm      | 187.0     | false   | 3.5       | down   | 0.0  | norm   | buff     | H       |
| 41.0  | female | ab...  | 130.0      | 204.0  | false | hyp       | 172.0     | false   | 1.4       | up     | 0.0  | norm   | buff     | H       |
| 56.0  | male   | ab...  | 120.0      | 236.0  | false | norm      | 178.0     | false   | 0.8       | up     | 0.0  | norm   | buff     | H       |
| 62.0  | female | as...  | 140.0      | 268.0  | false | hyp       | 160.0     | false   | 3.6       | down   | 2.0  | norm   | sick     | S3      |
| 57.0  | female | as...  | 120.0      | 354.0  | false | norm      | 163.0     | true    | 0.6       | up     | 0.0  | norm   | buff     | H       |
| 63.0  | male   | as...  | 130.0      | 254.0  | false | hyp       | 147.0     | false   | 1.4       | flat   | 1.0  | rev    | sick     | S2      |
| 53.0  | male   | as...  | 140.0      | 203.0  | true  | hyp       | 155.0     | true    | 3.1       | down   | 0.0  | rev    | sick     | S1      |
| 57.0  | male   | as...  | 140.0      | 192.0  | false | norm      | 148.0     | false   | 0.4       | flat   | 0.0  | fix    | buff     | H       |

In this case, you must use the SQL Script or One Hot Encoding component to convert the raw data from the DOUBLE type to the BIGINT type. In addition, you must use the Feature Discretization component to decompose feature data in different intervals into different bins. The following figure shows the data after binning.

| sex ▲ | cp ▲ | fbs ▲ | restecg ▲ | exang ▲ | slop ▲ | thal ▲ | age ▲ | trestbps ▲ | chol ▲ | thalach ▲ | oldpeak ▲ | ca ▲ | ifhealth ▲ |
|-------|------|-------|-----------|---------|--------|--------|-------|------------|--------|-----------|-----------|------|------------|
| 4     | 1    | 4     | 4         | 1       | 4      | 3      | 4     | 3          | 2      | 4         | 2         | 1    | 0          |
| 4     | 4    | 1     | 4         | 4       | 3      | 1      | 4     | 4          | 2      | 2         | 2         | 4    | 1          |
| 4     | 4    | 1     | 4         | 4       | 3      | 4      | 4     | 2          | 2      | 3         | 3         | 4    | 1          |
| 4     | 3    | 1     | 1         | 1       | 4      | 1      | 1     | 2          | 2      | 4         | 3         | 1    | 0          |
| 1     | 4    | 1     | 4         | 1       | 1      | 1      | 2     | 2          | 1      | 4         | 2         | 1    | 0          |
| 4     | 4    | 1     | 1         | 1       | 1      | 1      | 3     | 2          | 2      | 4         | 1         | 1    | 0          |
| 1     | 4    | 1     | 4         | 1       | 4      | 1      | 4     | 3          | 2      | 4         | 3         | 4    | 1          |
| 1     | 4    | 1     | 1         | 4       | 1      | 1      | 3     | 2          | 3      | 4         | 1         | 1    | 0          |
| 4     | 4    | 1     | 4         | 1       | 3      | 4      | 4     | 2          | 2      | 3         | 2         | 2    | 1          |
| 4     | 4    | 4     | 4         | 4       | 4      | 4      | 3     | 3          | 1      | 4         | 3         | 1    | 1          |
| 4     | 4    | 1     | 1         | 1       | 3      | 3      | 3     | 3          | 1      | 3         | 1         | 1    | 0          |
| 1     | 4    | 1     | 4         | 1       | 3      | 1      | 3     | 3          | 2      | 4         | 2         | 1    | 0          |

## 3. Determine the range of feature values

The basic process of feature crossing includes the following steps: generate feature vectors for features, create feature crosses, verify the feature crosses, and then select effective feature crosses. Before you generate feature vectors, you must know the maximum feature value in each feature space. Example:

- The maximum value of the thalach feature is 4.
- The maximum value of the oldpeak feature is 3.
- The maximum value of the ca feature is 4.

| thalach ▲ | oldpeak ▲ | ca ▲ |
|-----------|-----------|------|
| 4         | 2         | 1    |
| 2         | 2         | 4    |
| 3         | 3         | 4    |

You can execute the following SQL statement to obtain the maximum value of each feature:

```
select max(feature) from table;
```

In the sample data of this topic, the maximum value of all features after binning is 4.

[illegible]

You must specify the **Feature length** parameter of the Auto Feature Cross component in the format shown in the following figure. In the format, 5 indicates a left-closed right-open interval  $[0,5)$  that includes 4.

[illegible]

#### 4. Prepare training data and test data

In this topic, the training data is the same as the test data. In actual use, the test data can differ from the training data, provided that the fields in the test data are the same as the fields in the training data.

## 5. Configure the Auto Feature Cross component

- Set the parameters on the **Fields Setting** tab

In the Auto Feature Cross component, the input port on the left is used to import training data and the input port on the right is used to import test data. You must set the following parameters on the Fields Setting tab based on your needs:

- Feature selection: the feature fields that will be involved in feature crossing.

- if sparse data: specifies whether the data is sparse data. This check box is cleared by default, which means that the data is dense data.
- Label: the label column that is used to determine whether a feature cross is effective.
- Output path: the URL of the OSS bucket that stores the generated model.
- Set the parameters on the **Parameters Setting** tab

The screenshot shows the 'Parameters Setting' tab with the following values:

- Ergodic number: 1500
- Ergodic number: 3
- Ergodic number: 16
- Threshold: 0.5
- Steps of attenuation: 30000

- Ergodic number: the number of iterations.
- Feature order: the maximum number of features in each feature cross. For example, 3 indicates that each feature cross involves a maximum of three features.

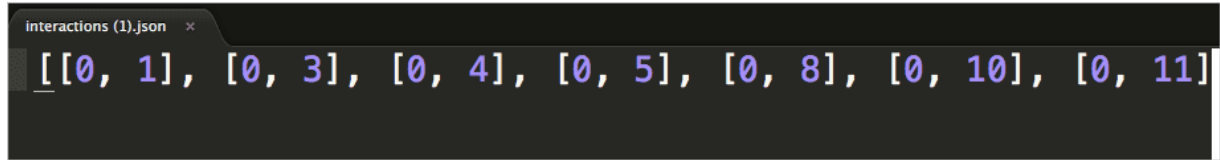
In this example, the Auto Feature Cross component runs the following PAI command to create feature crosses:

```
PAI -name fives_ext -project algo_public
-DlabelColName="ifhealth" // The label column that is used to determine whether a feature cross is effective.
-Dmetric_file="metric_log.log" // The name of the system log file.
-Dfeature_meta="[5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5]"
-DtrainTable="odps://Project name/tables/Table name"
-Dbuckets="oss://{oss_bucket}/"
-Dthreshold="0.5"
-Dk="3"
-DossHost="oss-cn-beijing-internal.aliyuncs.com" // The region in which OSS is activated.
-Demb_dims="16"
-DenableSparse="0"
-Dtemp_anneal_steps="30000"
-DfeatureColName="sex,cp,fbs,restecg,exang,slop,thal,age,trestbps,chol,thalach,oldpeak,ca" // The feature fields that will be involved in feature crossing.
-DtestTable="odps://Project name/tables/Table name"
-Darn="acs:ram::*:role/aliyunodpspaidefaultrole" //rolearn
-Depochs="1500"
-DcheckpointDir="oss://{oss_bucket}/{path}";
```

## View the feature crosses

In the root directory of your OSS bucket, find the interactions.json file. The root directory of your OSS bucket is specified by buckets.

The file shows the effective feature crosses that are created by the Auto Feature Cross component.



```
interactions (1).json x
[[0, 1], [0, 3], [0, 4], [0, 5], [0, 8], [0, 10], [0, 11]]
```

You can create other feature crosses based on the feature crosses in the file. Example:

- [0,1] indicates that the cross of the first and second features is effective. The feature order in each feature cross is the same as the feature order in the input table.
- [8, 6, 5] indicates that the cross of the ninth, seventh, and fifth features is effective.

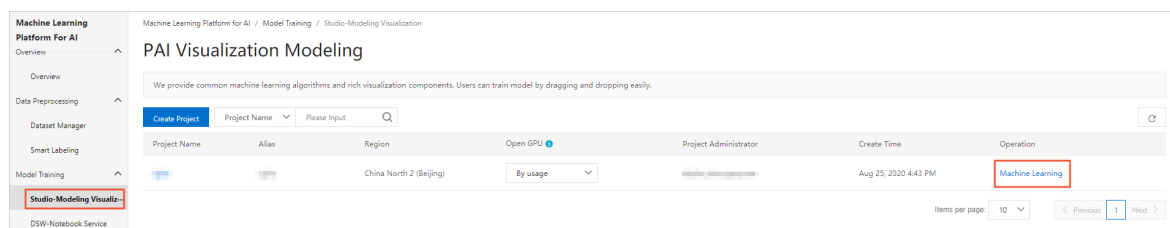
## 6.Periodic scheduling

### 6.1. Import PAI models to OSS for a scheduling task

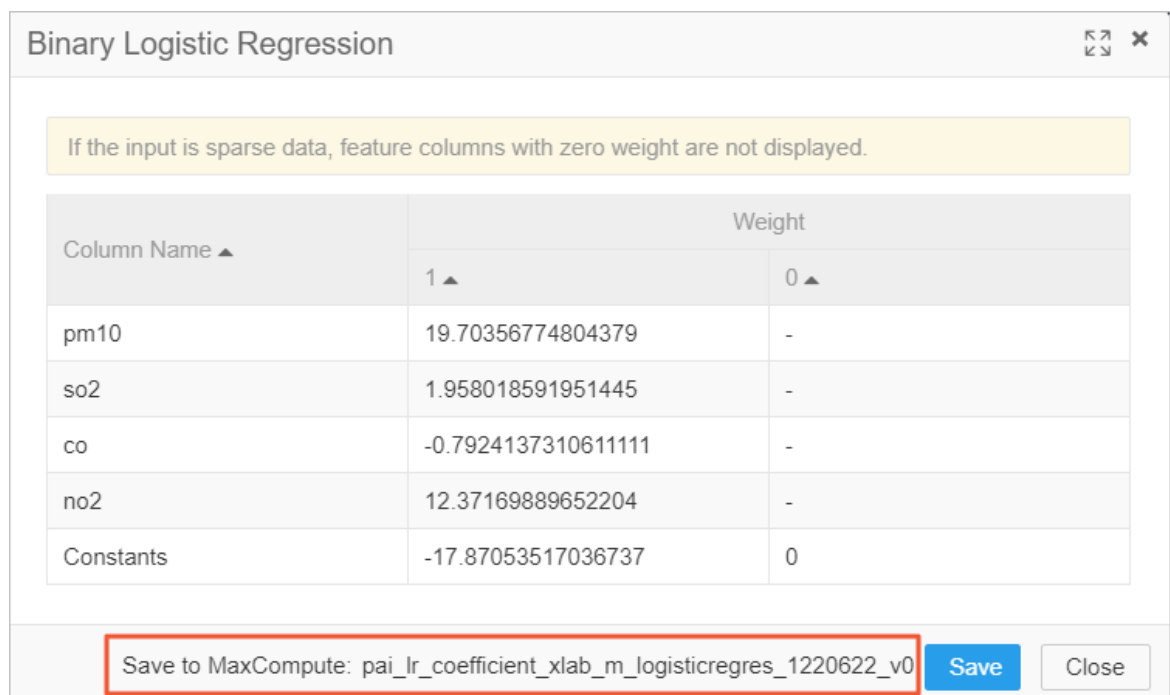
This topic describes how to automatically import Machine Learning Platform for AI (PAI) models to Object Storage Service (OSS) for a scheduling task.

#### Procedure

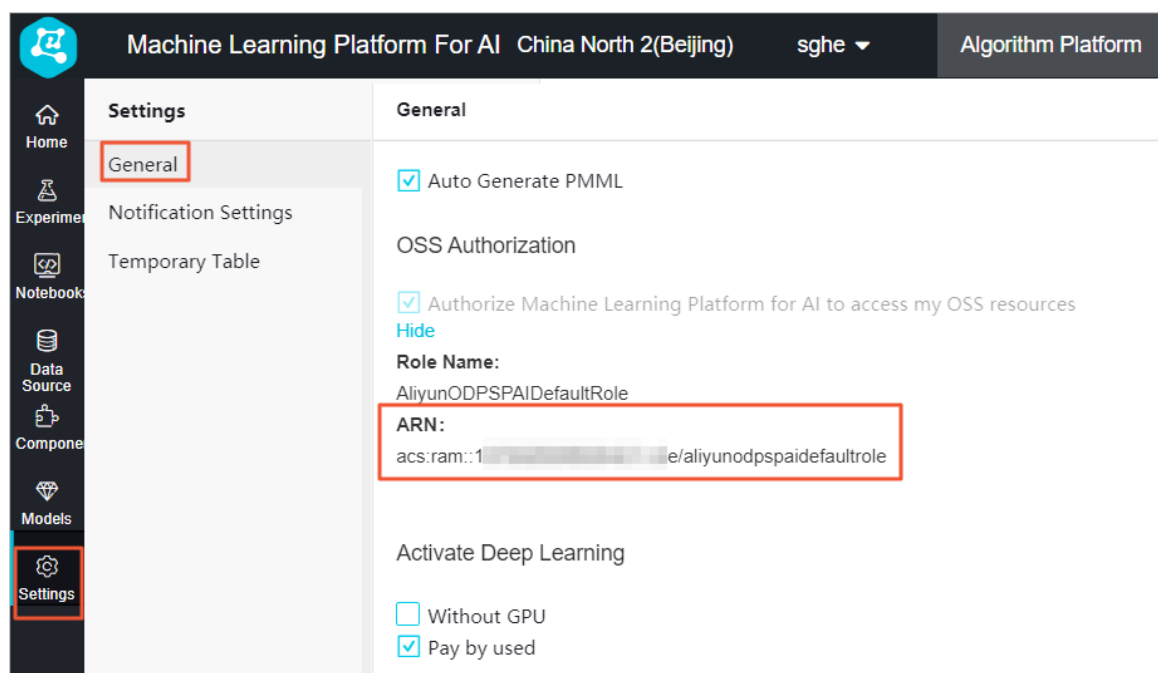
1. Log on to the **PAI console**.
2. In the left-side navigation pane, choose **Model Training > Studio-Modeling Visualization** to go to the **PAI Visualization Modeling** page. When you create a project, we recommend that you select **By usage** for the project. PAI-TensorFlow tasks can run only by using GPU resources.



3. Find the required project and click **Machine Learning** in the **Operation** column.
4. In the left-side navigation pane of the tab that appears, click **Models**. On the page that appears, drag and drop the model that you want to manage to the canvas, right-click the model, and choose **Model Option > Show Model** to obtain the model name.
5. In the dialog box that appears, find the model name and remove the prefix `pair_coefficient` from the name.



- Determine the Alibaba Cloud resource name (ARN) that is the unique credential used by PAI to access OSS.



- Run the command to save the model. To periodically import models to OSS, you must first create an SQL node in DataWorks and run the following command:

```
pai -name modeltransfer2oss
-DmodelName=xlab_m_logisticregres_1006580_v0(Model name)
-DossPath="oss://images-beijing.oss-cn-beijing-internal.aliyuncs.com/aohai/"(OSS path)
-Darn="acs:ram::1664081855183111:role/aliyunodpspaidefaultrole"(ARN)
-Dformat=pmml
-Doverwrite=true/false;
```

## 6.2. Perform offline scheduling

DataWorks task scheduling is widely used in machine learning scenarios. It allows you to periodically run DataWorks tasks to update your model, which helps you create a model training pipeline. Machine Learning Platform for AI (PAI) can work with DataWorks to periodically schedule machine learning experiments.

### Context

When all nodes in an experiment are up and running, you can deploy the experiment to DataWorks and schedule DataWorks to periodically run the experiment.

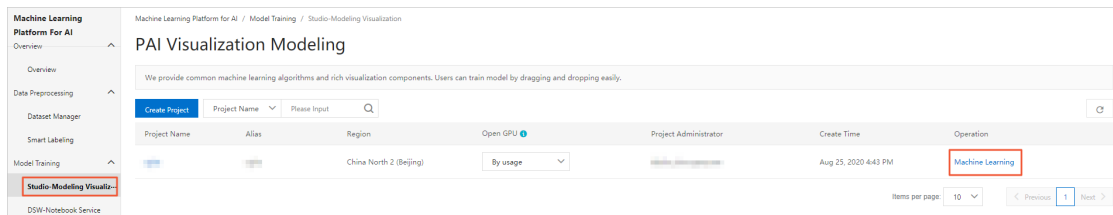
**Note** Before you schedule nodes offline, make sure that all nodes in an experiment are up and running and DataWorks is activated. For more information, see [Create a workspace](#).

### Procedure

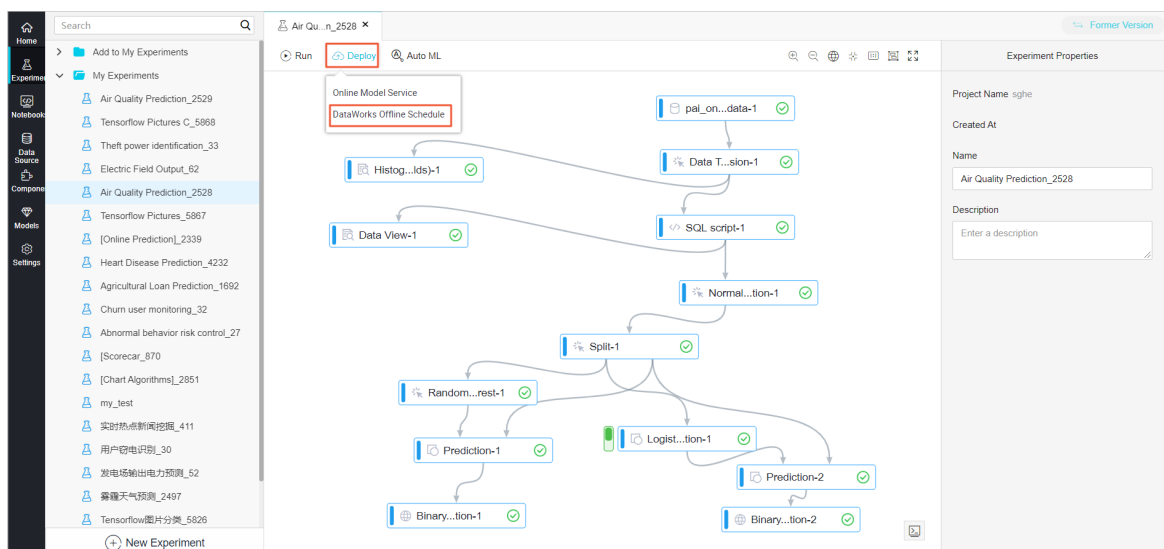
- Go to the Experiments page of the PAI console.
  - Log on to the [PAI console](#).



- ii. In the left-side navigation pane, choose **Model Training > Studio-Modeling Visualization** to go to the **PAI Visualization Modeling** page. When you create a project, we recommend that you select **By usage** for the project. PAI-TensorFlow tasks can run only by using GPU resources.



- iii. Find the required project and click **Machine Learning** in the **Operation** column.
  - iv. In the left-side navigation pane, click **Experiments**. On the page that appears, find the required experiment and double-click it. The Heart Disease Prediction experiment is used in this example.
- On the tab that appears, choose **Deploy > DataWorks Offline Schedule** to go to the Data Analytics page of DataWorks.



- Create a PAI node.
  - i. In the **Create Node** dialog box, configure **Node Name** and **Location**. If you do not create a node in the dialog box that appears, move the point over the  icon on the **Data Analytics** page and choose **Machine Learning > PAI Experiment**.

**Note** The node name can be a maximum of 128 characters in length and can contain letters, digits, underscores (\_), and periods (.).

- ii. Click **Commit**.
- On the tab that appears, select the PAI experiment that you created from the **Experiment** drop-down list. To edit the loaded PAI experiment, click **Edit** in **PAI Console** to edit the experiment.
  - In the right-side navigation pane, click the **Properties** tab. In the panel that appears, configure the properties for the node. For more information, see [Basic properties](#). Configure task scheduling parameters, including the recurrence, input parameters, and output parameters.
  - Save and commit the node.

 **Notice** You must configure **Rerun** and **Parent Nodes** before you commit the node.

- i. Click the  icon in the toolbar to save the node.
- ii. Click the  icon in the toolbar.
- iii. In the **Commit Node** dialog box, set **Change description**.
- iv. Click **OK**.

In a workspace in standard mode, you must click **Publish** in the upper-right corner after you commit the node. For more information, see [Deploy nodes](#).

7. In the upper-right corner, click **Operation** to view the running status and system logs of the PAI task. You can also perform other operations, such as generating retroactive data and testing the experiment. For more information, see [View auto triggered nodes](#).

## 7. Use the online notebook feature of Machine Learning Studio

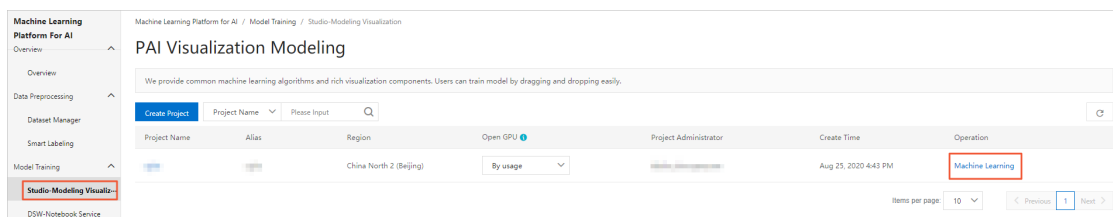
Alibaba Cloud Machine Learning Platform for AI (PAI) provides the online notebook feature to help you edit and debug TensorFlow code. The online notebook feature can work with Object Storage Service (OSS) and the underlying compute clusters of PAI to debug code on the cloud.

### Context

- Before you start the online notebook feature, make sure that the size of the files stored in the selected OSS path and all its subdirectories is no greater than 80 MB. If a connection is established for PAI to read data from OSS, the connection times out after two minutes, and PAI attempts to establish a connection again.
- Before you start the online notebook feature, make sure that the number of the .py, .tar.gz, or .zip files that are stored in the selected OSS path and all its subdirectories is no greater than 500.
- If you do not perform any operations on the notebook editing page for more than 20 minutes after you enable the online notebook feature, the resources are released. To use these resources again, you must restart the online notebook feature. We recommend that you save your data in a timely manner.

### Procedure

1. Go to the homepage of the PAI console.
  - i. Log on to the [PAI console](#).
  - ii. In the left-side navigation pane, choose **Model Training > Studio-Modeling Visualization**. The **PAI Visualization Modeling** page appears. When you create a project, we recommend that you select the pay-as-you-go billing method and enable GPU-accelerated computing. PAI-TensorFlow tasks can run only by using GPU resources.

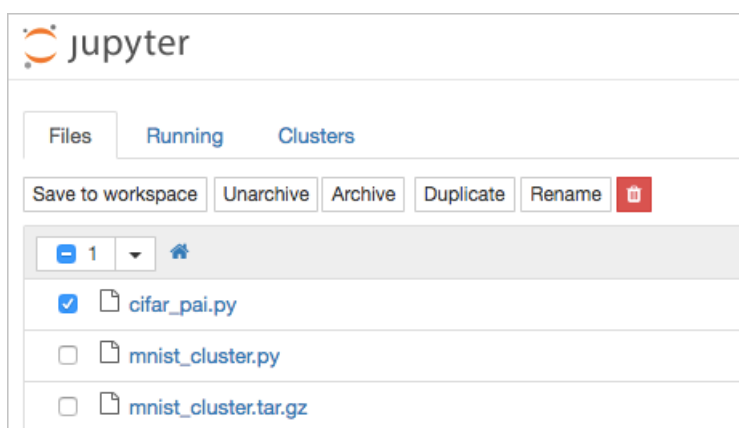


- iii. Find the required project and click **Machine Learning** in the **Operation** column.
  - iv. In the left-side navigation pane of the tab that appears, click **Home**.
2. Create a notebook.
    - i. In the upper-right corner of the **Templates** page, click **New** and select **New Notebook**.
    - ii. In the **New Notebook** dialog box, specify the **Name**, **Description**, and **Save To** parameters.
    - iii. Click **Next**.

- iv. In the dialog box that appears, specify the Select OSS Bucket and Select Computing Resources parameters.

Take note of the following points:

- The online notebook feature of Machine Learning Studio reads only the tar.gz, .py, and .zip files in the specified directory.
  - If the OSS bucket and the experiment project are in different regions, additional costs are generated.
  - The valid values of Select Computing Resources are CPU and GPU. If you select GPU, you can specify the number of GPUs.
3. In the left-side navigation pane, click **Notebooks** to view the newly created notebook.
  4. Right-click the required notebook and select **Open Notebook**. Wait for 2 to 5 minutes to start the notebook on the cloud.
  5. Click **Click to view**.
  6. Edit the notebook. Use the notebook in a way similar to that of an open source notebook. Take note of the following points:
    - The notebook provides the file compression and decompression features. A tar.gz file is a TAR archive that contains compressed cluster files for machine learning.
    - Files are saved to the required directory in OSS only if you click **Save to workspace**.



## 8. Configure and use global variables

If you use Machine Learning Platform for AI (PAI) to configure components, global variables can be used to reduce the workload of repeat parameter configurations. You can create and delete global variables. You can also use the configuration copy feature to obtain the reference methods of global variables. This topic describes how to configure and use global variables.

### Scenarios

- Multiple components in an experiment use the same parameters.

For more information, see [Example 1: Configure the parameters that are used by all the components in an experiment](#).

- You want to replace scheduled scheduling parameters.

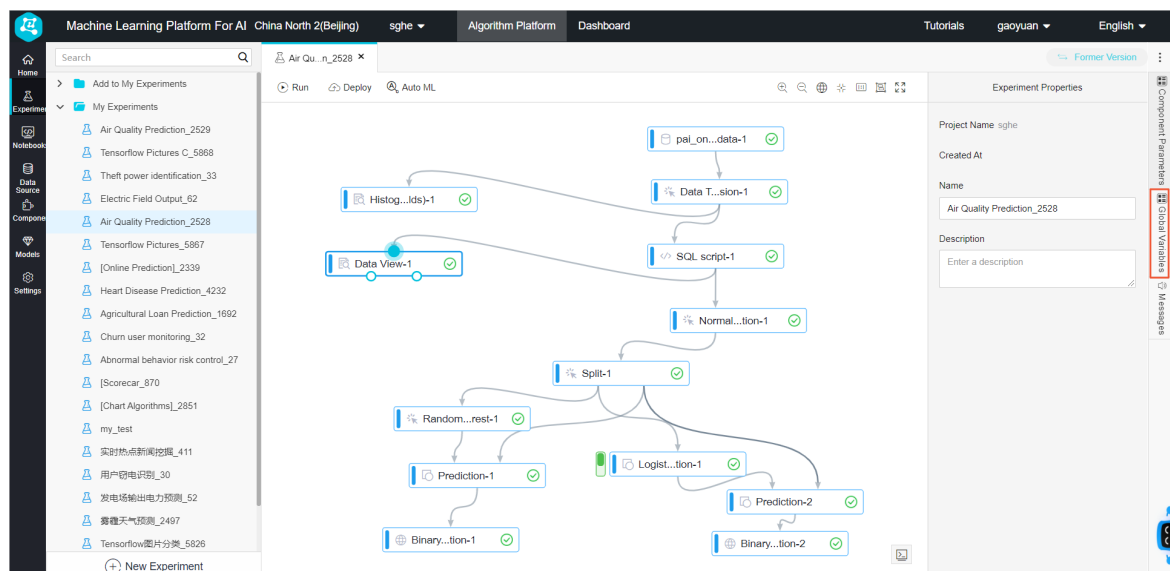
For more information, see [Example 2: Replace scheduled scheduling parameters](#).

### Prerequisites

A project and an experiment are created. For more information, see [Quick Start](#).

### Go to the global variable configuration page

- Log on to the [PAI console](#).
- In the left-side navigation pane, choose **Model Training** > **Studio-Modeling Visualization**.
- On the **PAI Visualization Modeling** page, find the required project and click **Machine Learning** in the Operation column.
- In the left-side navigation pane, click **Experiments**. On the page that appears, click the required experiment.
- Click a component in the canvas.
- On the right side of the page, click the **Global Variables** tab. The Global Variables pane appears.



## Example 1: Configure the parameters that are used by all the components in an experiment

The Heart Disease Prediction experiment is used in this example. If you configure the parameters, you can use global variables. The variables can also be used for the components that use the same parameters.

### 1. Configure global variables.

- i. [Go to the global variable configuration page.](#)
- ii. In the lower part of the Global Variable pane, click **New Global Variable**.
- iii. Configure the **Name** and **Value** parameters.

In this example, set **Name** to param. Then, the system automatically generates `${param}` for **Reference Method**. If you replace a parameter with `${param}`, you can reference this variable. The valid values of **Value** include cp and fbs.

| Global Variable |                                        |
|-----------------|----------------------------------------|
| Name            | param                                  |
| Value           | sex, cp, fbs, restecg, exang, slop, th |
| Reference       | <code>\${param}</code>                 |

### 2. Use global variables.

Replace the parameter that uses a variable with the value of **Reference Method** for this variable.

| Selected                                                                            |                        | List | Edit |
|-------------------------------------------------------------------------------------|------------------------|------|------|
|  | Column                 | Type |      |
|  | <code>\${param}</code> | -    |      |

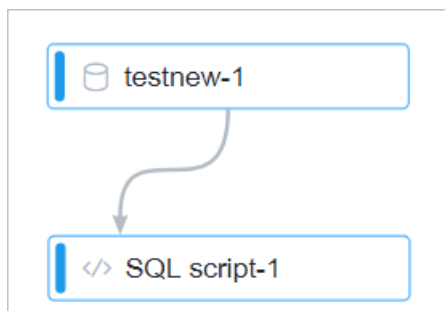
### 3. Check whether the parameter is automatically replaced with the value you entered.

```

[3] OK
[3] run subjob: #3[sql]([dw_scheduler_pre_dev] drop table if exists pai_temp_67621_976589_1)
    successfully!
[4] start subjob: #4[odpscmd]
[4] Start OdpsCmdHandler:jobId=976589-1601308
[4] local log file =
    /home/admin/alisaTasknode/taskinfo//20190515/phoenixdev/17/08/03/zm00ddktgu7ef73xe3f
    s5cwb//T3_0805992405_jcs.log
[4] user accessId :LTAlsaKqcavxJLis
[4] execute command : set biz_id=13[redacted]; PAI -name
    type_transform_v1 -project algo_public -Dlifecycle="3" -DreserveOldFeat="false" -
    Ddefault_int_value="0" -DoutputTable="pai_temp_67621_976589_1" -
    DinputTable "pai_temp_67621_976589_1" Ddefault_double_value "0.0"
    Dcols_to_double="sex,cp,fbs,restecg,exang,slop,thal,ifhealth,age,trestbps,chol,thalach,oldpeak,ca
    ",
[4] execute endpoint : http://service.odps.aliyun-inc.com/api
[4] Data Health Manager:Your health synthesize score is 86, so, your job priority is 3
[4] OK
[4] ID = [redacted]
[4] Odps Instance Id = [redacted]
[4] OdpsInstanceId: 20[redacted]
[4] Sub Instance ID = [redacted]
[4] SubOdpsInstance[redacted]
    callback succ
[4] http://logview.odps.aliyun-inc.com:8080/logview/?h=http://service.odps.aliyun-inc.com/api&p=dw\_scheduler\_pre\_dev&i=\[redacted\]
  
```

## Example 2: Replace scheduled scheduling parameters

Global variables can also be used for scheduling experiments associated with dates. In a PAI experiment, the configured values for global variables can only be used to run the experiment online. If scheduling is performed offline, the parameter values that you configured for the scheduling are used. In this example, the source data contains the data of two partitions.



1. Prepare source data.

|    | A   | B   | C     | D        |
|----|-----|-----|-------|----------|
| 1  | t1  | t2  | label | ds       |
| 2  | 1.0 | 1.1 | 'A'   | 20190519 |
| 3  | 1.0 | 1.0 | 'A'   | 20190519 |
| 4  | 0.0 | 0.0 | 'B'   | 20190519 |
| 5  | 0.0 | 0.1 | 'B'   | 20190519 |
| 6  | 1.2 | 1.1 | 'A'   | 20190519 |
| 7  | 0.1 | 0.1 | 'C'   | 20190519 |
| 8  | 0.1 | 0.1 | 'C'   | 20190519 |
| 9  | 0.1 | 0.1 | 'D'   | 20190519 |
| 10 | 0.1 | 0.1 | 'E'   | 20190519 |
| 11 | 1.0 | 1.1 | F     | 20190520 |

## 2. Configure global variables.

For more information, see [Example 1: Configure the parameters that are used by all the components in an experiment](#). In this example, **Name** is set to `data`, and **Value** is set to `20190520`.

Global Variable

×

Name

Value

Reference

📄

## 3. Use global variables in SQL scripts.

Save Format Maximize

```

1 --The input data has been automatically mapped to t1 through t4. Use statement select * from ${t1} to query the data.
2 -- You can also directly input MaxCompute tables.
3 -- To read the partition data, you must specify the partition conditions in the SQL script.
4 -- Components do not automatically retrieve partition parameters from the parent data source components.
5 --This component supports all MaxCompute SQL syntax.
6 select * from ${t1} where ds=${data}

```

## 4. Run the experiment online and view the result.

| t1 ▲ | t2 ▲ | label ▲ | ds ▲     |
|------|------|---------|----------|
| 1    | 1.1  | F       | 20190520 |

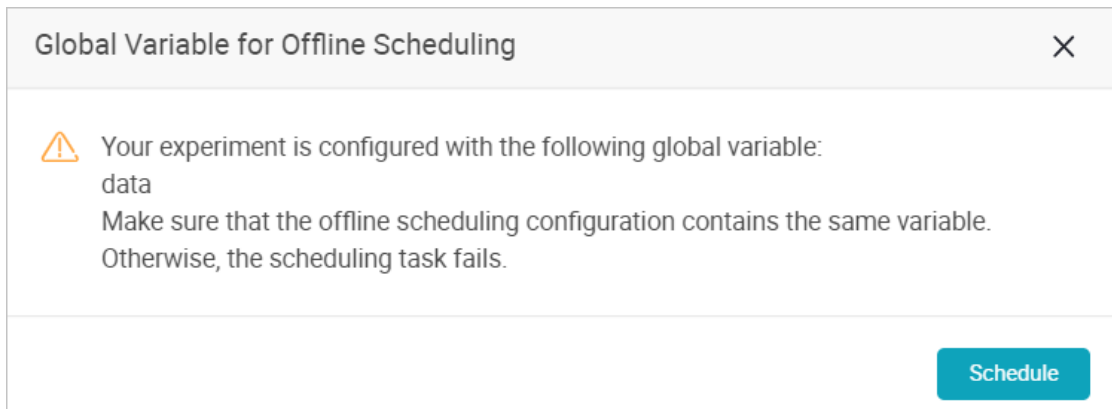
## 5. Use global variables to configure offline scheduling.

 **Note** For more information, see [Perform offline scheduling](#).

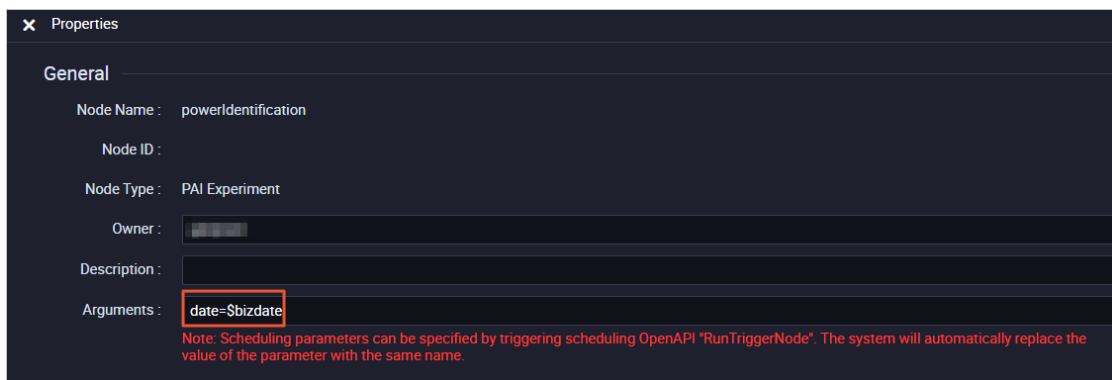


- i. Go to the offline scheduling page.

When you go to the offline scheduling page, the system asks you to configure a parameter that has the same name as the global variable that you want to use.

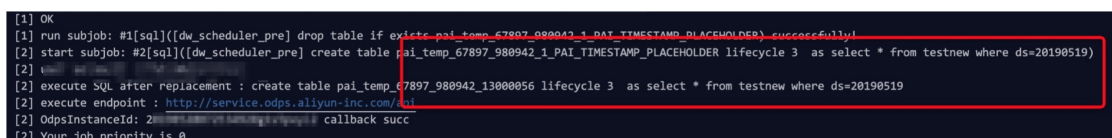


- ii. Use global variables to configure an offline scheduling task.



- iii. Run the scheduling task and view the result.

The scheduling result shows that the variables in the generated instances are replaced with 20190519.



## 9. FastNN

Fast Neural Networks (FastNN) in Machine Learning Platform for AI (PAI) is a distributed neural network warehouse based on the PAISoar framework. FastNN supports classic algorithms, such as Inception, ResNet, and VGG. It will support more advanced algorithms in the future. FastNN is embedded in Machine Learning Studio. You can use it on this platform.

### Prepare data sources

To facilitate the usage of FastNN in the **PAI console**, CIFAR-10, MNIST, and Flowers datasets are downloaded, converted to TFRecord datasets, and saved in a public Object Storage Service (OSS) bucket. You can use the **Read MaxCompute Table** or **OSS Data Synchronization** component provided by PAI to access the datasets. The following table lists the storage paths in OSS.

| Dataset | Number of classes | Training dataset | Test dataset | Storage path                                                                                                                                                                                                                                                 |
|---------|-------------------|------------------|--------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| mnist   | 10                | 3320             | 350          | <ul style="list-style-type: none"> <li>China (Beijing): oss://pai-online-beijing.oss-cn-beijing-internal.aliyuncs.com/fastnn-data/mnist/</li> <li>China (Shanghai): oss://pai-online.oss-cn-shanghai-internal.aliyuncs.com/fastnn-data/mnist/</li> </ul>     |
| cifar10 | 10                | 50000            | 10000        | <ul style="list-style-type: none"> <li>China (Beijing): oss://pai-online-beijing.oss-cn-beijing-internal.aliyuncs.com/fastnn-data/cifar10/</li> <li>China (Shanghai): oss://pai-online.oss-cn-shanghai-internal.aliyuncs.com/fastnn-data/cifar10/</li> </ul> |
| flowers | 5                 | 60000            | 10000        | <ul style="list-style-type: none"> <li>China (Beijing): oss://pai-online-beijing.oss-cn-beijing-internal.aliyuncs.com/fastnn-data/flowers/</li> <li>China (Shanghai): oss://pai-online.oss-cn-shanghai-internal.aliyuncs.com/fastnn-data/flowers/</li> </ul> |

FastNN allows you to read data in the TFRecord format. It uses TFRecordDataset to build dataset pipelines, which can be used for model training. This saves time for data preprocessing. FastNN does not provide fine-grained data partitioning. We recommend that you evenly distribute data to each server.

- Make sure that the number of samples for each TFRecord file is the same.

- Make sure that the number of TFRecord files that each worker processes is the same.

If the data is in the TFRecord format, you can build dataset pipelines by referring to the files, such as `cifar10`, `mnist`, and `flowers` in the `datasets` directory. The `cifar10` file is used in this example.

The following code shows the `key_to_features` format in the `cifar10` file:

```
features={
    'image/encoded': tf.FixedLenFeature([], tf.string, default_value=''),
    'image/format': tf.FixedLenFeature([], tf.string, default_value='png'),
    'image/class/label': tf.FixedLenFeature(
        [], tf.int64, default_value=tf.zeros([], dtype=tf.int64)),
}
```

1. In the `datasets` directory, create a file named `cifar10.py` for data parsing and edit the file.

```
"""Provides data for the Cifar10 dataset.
The dataset scripts used to create the dataset can be found at:
datasets/download_and_convert_data/download_and_convert_cifar10.py
"""

from __future__ import division
from __future__ import print_function
import tensorflow as tf

"""Expect func_name is 'parse_fn'
"""

def parse_fn(example):
    with tf.device("/cpu:0"):
        features = tf.parse_single_example(
            example,
            features={
                'image/encoded': tf.FixedLenFeature([], tf.string, default_value=''),
                'image/format': tf.FixedLenFeature([], tf.string, default_value='png'),
                'image/class/label': tf.FixedLenFeature(
                    [], tf.int64, default_value=tf.zeros([], dtype=tf.int64)),
            }
        )
        image = tf.image.decode_jpeg(features['image/encoded'], channels=3)
        label = features['image/class/label']
        return image, label
```

2. Configure `dataset_map` in the `datasets/dataset_factory.py` file.

```
from datasets import cifar10
dataset_map = {
    'cifar10': cifar10,
}
```

3. When you run the task script, specify the `dataset_name=cifar10` and `train_files=cifar10_train.tfrecord` parameters. After the configuration, you can use the `cifar10` file for model training.

 **Note** If you want to read data in other formats, you must build a dataset pipeline based on the required logic. For more information, see the `utils/dataset_utils.py` file.

## Hyperparameters

FastNN supports the following types of hyperparameters:

- Dataset hyperparameters: the parameters used to determine the basic attributes of training datasets, such as `dataset_dir` in which a training dataset is stored
- Data preprocessing hyperparameters: the data preprocessing functions and dataset pipeline parameters
- Modeling hyperparameters: the basic parameters for model training, including `model_name` and `batch_size`
- Learning rate hyperparameters: the learning rate parameters and tuning parameters
- Optimizer hyperparameters: the optimizer parameters
- Log hyperparameters: the output log parameters
- Performance tuning hyperparameters: the tuning parameters, such as mixed precision

The following example shows the format of a hyperparameter file:

```
enable_paisora=True
batch_size=128
use_fp16=True
dataset_name=flowers
dataset_dir=oss://pai-online-beijing.oss-cn-beijing-internal.aliyuncs.com/astnn-data/flowers/
model_name=inception_resnet_v2
optimizer=sgd
num_classes=5
job_name=worker
```

- Dataset hyperparameters

| Parameter                         | Type    | Description                                                                                                                                                                                              |
|-----------------------------------|---------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>dataset_name</code>         | string  | The name of the input data parsing file. Valid values: mock, cifar10, mnist, and flowers. For more information, see all data parsing files in the <i>images/datasets</i> directory. Default value: mock. |
| <code>dataset_dir</code>          | string  | The absolute path of the input dataset. Default value: None.                                                                                                                                             |
| <code>num_sample_per_epoch</code> | integer | The total number of dataset samples. This parameter is used with learning rate decay.                                                                                                                    |
| <code>num_classes</code>          | integer | The number of sample classes. Default value: 100.                                                                                                                                                        |
| <code>train_files</code>          | string  | The file name of all training data. Separate multiple names with commas (,). Example: 0.tfrecord,1.tfrecord.                                                                                             |

- Data preprocessing hyperparameters

| Parameter                 | Type    | Description                                                                                                                                                                                                                                                                                                                 |
|---------------------------|---------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| preprocessing_name        | string  | This parameter is used together with <code>model_name</code> to specify the name of the data preprocessing method. For more information about valid values, see the <i>preprocessing_factory</i> file in the <i>images/preprocessing</i> directory. Default value: None, which indicates that the data is not preprocessed. |
| shuffle_buffer_size       | integer | The size of the buffer pool for sample-based shuffles when a data pipeline is created. Default value: 1024.                                                                                                                                                                                                                 |
| num_parallel_batches      | integer | The number of parallel threads multiplied by <code>batch_size</code> to equal <code>map_and_batch</code> . This parameter helps specify the parallel granularity of parsing samples. Default value: 8.                                                                                                                      |
| prefetch_buffer_size      | integer | The number of batches of data prefetched by the data pipeline. Default value: 32.                                                                                                                                                                                                                                           |
| num_preprocessing_threads | integer | The number of threads used by the data pipeline to prefetch data in parallel. Default value: 16.                                                                                                                                                                                                                            |
| datasets_use_caching      | bool    | Specifies whether to enable caching for compressed input data with memory overheads. Default value: False, which indicates that caching is disabled.                                                                                                                                                                        |

- Modeling hyperparameters

| Parameter    | Type    | Description                                                                                                                                                                                                                                 |
|--------------|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| task_type    | string  | The type of the task. Valid values: <ul style="list-style-type: none"> <li>◦ <code>pretrain</code>: model pre-training. This is the default value.</li> <li>◦ <code>finetune</code>: model optimization.</li> </ul>                         |
| model_name   | string  | The model to be trained. Valid values include all models in <i>images/models</i> . You can set this parameter based on all models defined in the <i>images/models/model_factory</i> file. Default value: <code>inception_resnet_v2</code> . |
| num_epochs   | integer | The number of training rounds for the training dataset. Default value: 100.                                                                                                                                                                 |
| weight_decay | float   | The weight decay factor during model training. Default value: 0.00004.                                                                                                                                                                      |

| Parameter         | Type    | Description                                                                                                                                                           |
|-------------------|---------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| max_gradient_norm | float   | Specifies whether to perform gradient clipping based on the global normalization value. Default value: None, which indicates that gradient clipping is not performed. |
| batch_size        | integer | The amount of data that one card processes in an iteration. Default value: 32.                                                                                        |
| model_dir         | string  | The path to reload the checkpoint. Default value: None, which indicates that model tuning is not performed.                                                           |
| ckpt_file_name    | string  | The name of the file that reloads the checkpoint. Default value: None.                                                                                                |

- Learning rate hyperparameters

| Parameter                  | Type    | Description                                                                                                                                                                                                                                                                                                                                                                                                                    |
|----------------------------|---------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| warmup_steps               | integer | The number of iterations for inverse learning rate decay. Default value: 0.                                                                                                                                                                                                                                                                                                                                                    |
| warmup_scheme              | string  | The scheme of inverse learning rate decay. Set the value to t2t (Tensor2Tensor). This value indicates that the learning rate is initialized to be 1/100 of the specified learning rate and then is exponentiated to inverse-decay to the specified learning rate.                                                                                                                                                              |
| decay_scheme               | string  | The scheme of learning rate decay. Valid values: <ul style="list-style-type: none"> <li>luong234: Start 4 rounds of decay with a factor of 1/2 after 2/3 of total iterations are completed.</li> <li>luong5: Start 5 rounds of decay with a factor of 1/2 after 1/2 of total iterations are completed.</li> <li>luong10: Start 10 rounds of decay with a factor of 1/2 after 1/2 of total iterations are completed.</li> </ul> |
| learning_rate_decay_factor | float   | The factor of learning rate decay. Default value: 0.94.                                                                                                                                                                                                                                                                                                                                                                        |
| learning_rate_decay_type   | string  | The type of learning rate decay. Valid values: fixed, exponential, and polynomial. Default value: exponential.                                                                                                                                                                                                                                                                                                                 |
| learning_rate              | float   | The initial learning rate. Default value: 0.01.                                                                                                                                                                                                                                                                                                                                                                                |
| end_learning_rate          | float   | The minimum learning rate during decay. Default value: 0.0001.                                                                                                                                                                                                                                                                                                                                                                 |

- Optimizer hyperparameters

| Parameter                         | Type   | Description                                                                                                                              |
|-----------------------------------|--------|------------------------------------------------------------------------------------------------------------------------------------------|
| optimizer                         | string | The name of the optimizer. Valid values: adadelta, adagrad, adam, ftrl, momentum, sgd, rmsprop, adamweightdecay. Default value: rmsprop. |
| adadelta_rho                      | float  | The decay factor of Adadelta. Default value: 0.95.                                                                                       |
| adagrad_initial_accumulator_value | float  | The initial value of the Adagrad accumulator. Default value: 0.1. This parameter is specific to the Adagrad optimizer.                   |
| adam_beta1                        | float  | The exponential decay rate in primary momentum prediction. Default value: 0.9. This parameter is specific to the Adam optimizer.         |
| adam_beta2                        | float  | The exponential decay rate in secondary momentum prediction. Default value: 0.999. This parameter is specific to the Adam optimizer.     |
| opt_epsilon                       | float  | The offset of the optimizer. Default value: 1.0. This parameter is specific to the Adam optimizer.                                       |
| ftrl_learning_rate_power          | float  | The idempotent parameter of the learning rate. Default value: -0.5. This parameter is specific to the FTRL optimizer.                    |
| ftrl_initial_accumulator_value    | float  | The starting point of the FTRL accumulator. Default value: 0.1. This parameter is specific to the FTRL optimizer.                        |
| ftrl_l1                           | float  | The regularization term of FTRL l1. Default value: 0.0. This parameter is specific to the FTRL optimizer.                                |
| ftrl_l2                           | float  | The regularization term of FTRL l2. Default value: 0.0. This parameter is specific to the FTRL optimizer.                                |
| momentum                          | float  | The momentum parameter of MomentumOptimizer. Default value: 0.9. This parameter is specific to the Momentum optimizer.                   |
| rmsprop_momentum                  | float  | The momentum parameter of RMSPropOptimizer. Default value: 0.9.                                                                          |
| rmsprop_decay                     | float  | The decay factor of RMSProp. Default value: 0.9.                                                                                         |

- Log hyperparameters

| Parameter | Type | Description |
|-----------|------|-------------|
|-----------|------|-------------|

| Parameter              | Type    | Description                                                                                                                 |
|------------------------|---------|-----------------------------------------------------------------------------------------------------------------------------|
| stop_at_step           | integer | The total number of training iterations. Default value: 100.                                                                |
| log_loss_every_n_iters | integer | The iterative frequency at which the loss information is printed. Default value: 10.                                        |
| profile_every_n_iters  | integer | The iterative frequency at which the timeline is printed. Default value: 0.                                                 |
| profile_at_task        | integer | The index of the machine that generates the timeline. Default value: 0, which corresponds to the index of the chief worker. |
| log_device_placement   | bool    | Specifies whether to print the device placement information. Default value: False.                                          |
| print_model_statistics | bool    | Specifies whether to print the trainable variable information. Default value: false.                                        |
| hooks                  | string  | The training hooks. Default value: StopAtStepHook, ProfilerHook, LoggingTensorHook, CheckpointSaverHook.                    |

- Performance tuning hyperparameters

| Parameter      | Type   | Description                                                                             |
|----------------|--------|-----------------------------------------------------------------------------------------|
| use_fp16       | bool   | Specifies whether to perform semi-precision training. Default value: True.              |
| loss_scale     | float  | The factor of the loss scale during training. Default value: 1.0.                       |
| enable_paisoar | bool   | Specifies whether to use the PAISoar framework. Default value: True.                    |
| protocol       | string | By default, the grpc.rdma cluster can use grpc+verbs to improve data access efficiency. |

## Develop a main file

If the existing models cannot meet your requirements, you can use the dataset, models, and preprocessing APIs for further development. Before that, you must understand the basic process of FastNN. In this example, images are used and the code entry file is *train\_image\_classifiers.py*. The following code shows the code architecture:



```
# Initialize the required model based on model_name to obtain network_fn. The input parameter train_image_size may be returned.
network_fn = nets_factory.get_network_fn(
    FLAGS.model_name,
    num_classes=FLAGS.num_classes,
    weight_decay=FLAGS.weight_decay,
    is_training=(FLAGS.task_type in ['pretrain', 'finetune']))
# Initialize the required data preprocessing function based on model_name or preprocessing_name to obtain preprocess_fn.
preprocessing_fn = preprocessing_factory.get_preprocessing(
    FLAGS.model_name or FLAGS.preprocessing_name,
    is_training=(FLAGS.task_type in ['pretrain', 'finetune']))
# Select the valid TFRecord format based on dataset_name and synchronously call preprocess_fn to parse the dataset and obtain dataset_iterator.
dataset_iterator = dataset_factory.get_dataset_iterator(FLAGS.dataset_name,
    train_image_size,
    preprocessing_fn,
    data_sources,
)
# Call network_fn and dataset_iterator to define the function loss_fn that is used to calculate the loss.
def loss_fn():
    with tf.device('/cpu:0'):
        images, labels = dataset_iterator.get_next()
        logits, end_points = network_fn(images)
        loss = tf.losses.sparse_softmax_cross_entropy(labels=labels, logits=tf.cast(logits, tf.float32), weights=1.0)
        if 'AuxLogits' in end_points:
            loss += tf.losses.sparse_softmax_cross_entropy(labels=labels, logits=tf.cast(end_points['AuxLogits'], tf.float32), weights=0.4)
        return loss
# Call the PAISoar API to encapsulate the native optimizer of loss_fn and tf.
opt = paioar.ReplicatedVarsOptimizer(optimizer, clip_norm=FLAGS.max_gradient_norm)
loss = optimizer.compute_loss(loss_fn, loss_scale=FLAGS.loss_scale)
# Define training tensors based on opt and loss.
train_op = opt.minimize(loss, global_step=global_step)
```