机器学习PAI

快速入门

快速入门

开始使用阿里云机器学习

阿里云机器学习平台是一套基于MaxCompute（原ODPS）的数据挖掘、建模、预测的工具，为您提供算法开发、分享、模型训练、部署、监控等一站式算法服务。通过统计学算法，对大量的历史数据进行学习从而生成经验模型，利用经验模型指导业务。您可以通过可视化的操作界面来操作整个实验流程，同时也支持PAI命令，您可通过命令行来操作实验。

本文档为您介绍如何在机器学习平台上完成以下基本任务。

1. 开通机器学习服务
2. 数据准备
3. 数据预处理
4. 数据可视化
5. 算法建模
6. 模型评估

相关地址：

产品入口页

公测收费说明

算法组件文档

深度学习文档

在线预测

离线调度

案例说明

论坛交流（产品建议、心得、商务合作意向）

产品BUG反馈，工单系统

开通机器学习服务

具体请参考购买开通流程。

数据准备

机器学习平台上传数据说明

机器学习平台底层支持两种数据源，一种是MaxCompute存储数据，另一种是OSS存储数据。

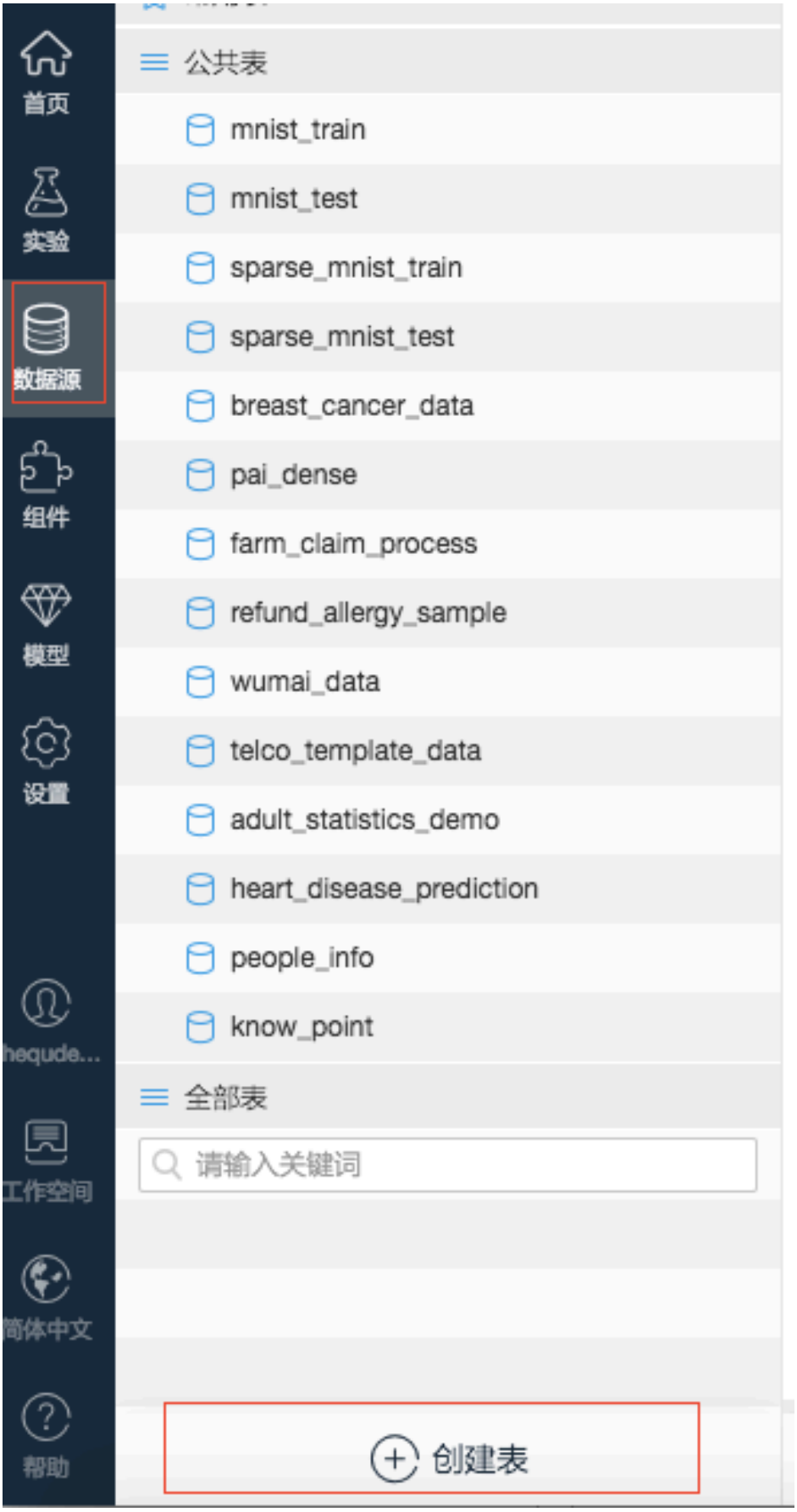
注意：使用MaxCompute作为存储，建议当数据小于20MB时使用机器学习IDE环境上传，当数据大于20MB时使用命令行工具上传。

MaxCompute上传数据：MaxCompute主要用来存储表结构数据，支持稀疏与稠密两种格式的数据，支持机器学习IDE上传和命令行工具上传两种方式。这部分的数据主要针对平台中除了深度学习以外的算法组件。

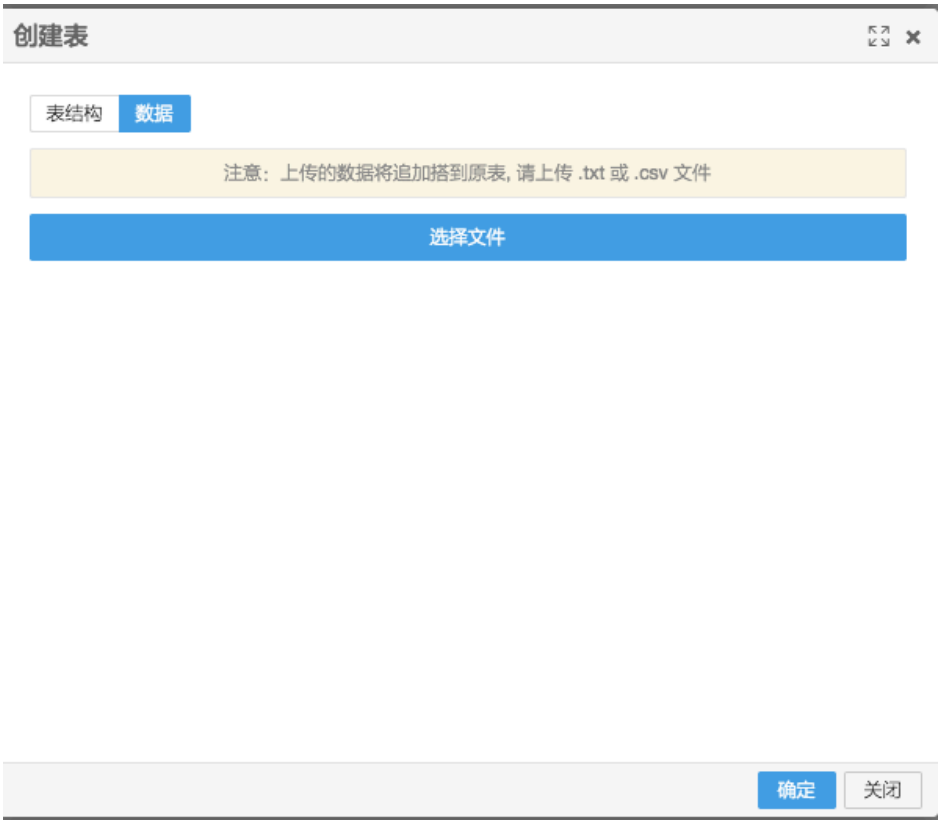
OSS上传数据：OSS数据源主要针对深度学习相关算法组件，可用来存储结构化或非结构化数据。

IDE端上传数据到MaxCompute

进入机器学习平台，单击数据源，创建表。


首页
实验
数据源
组件
模型
设置
工作空间
简体中文
帮助
公共表
mnist_train
mnist_test
sparse_mnist_train
sparse_mnist_test
breast_cancer_data
pai_dense
farm_claim_process
refund_allergy_sample
wumai_data
telco_template_data
adult_statistics_demo
heart_disease_prediction
people_info
know_point
全部表
请输入关键词
创建表

选择相应的数据源，并创建与之匹配的字段。建议使用txt格式上传，csv格式易出现特殊字符。



对于稀疏格式数据，请参考libsvm数据使用文件上传数据。

命令行工具上传数据到MaxCompute

MaxCompute提供多种数据上传方式。请参考数据迁移到MaxCompute的N种方式，选择最合适的方式上传数据。

OSS上传数据

请参见OSS上传数据。

操作步骤

开通并进入机器学习界面后，单击左边菜单栏的**首页**，选择**新建->新建空白试验**，如下图所示。

单击左边菜单栏的**组件**，打开**源/目标**文件夹，向画布中拖入**读数据表**组件，在右侧表选择栏填入对应的MaxCompute表名，如下图所示。

切换到字段信息栏，如下图所示，可以查看输入表的字段名、数据类型和前100行数据的数值分布。

表选择 字段信息

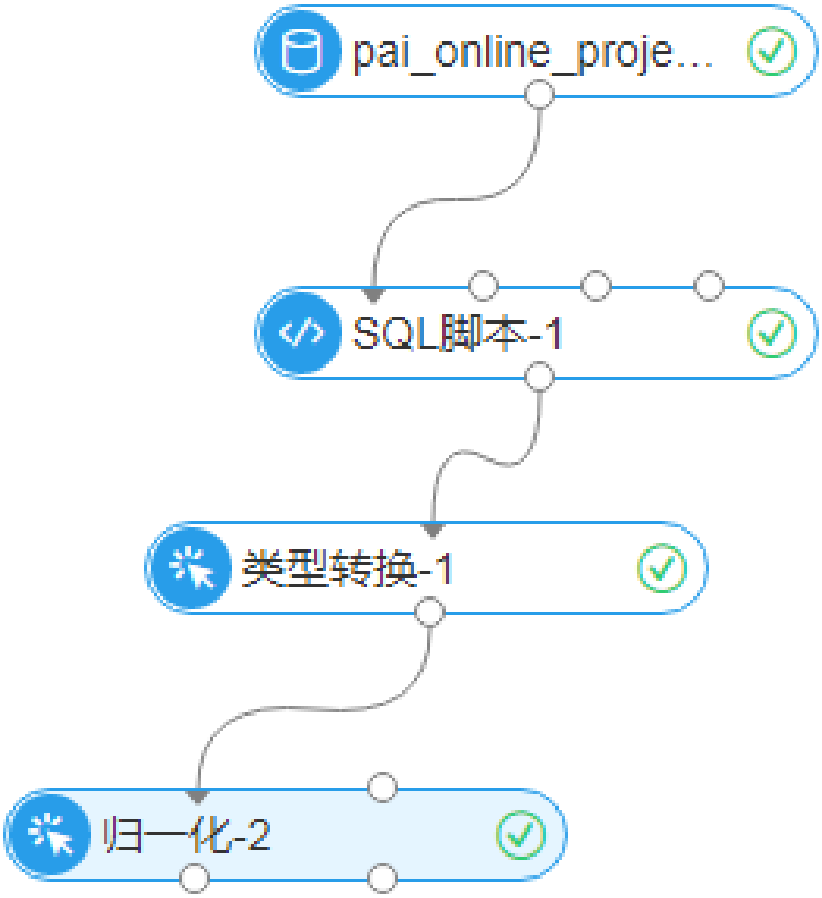
源表字段信息

字段	类型	前 100 条记录范围	
age	STRING	37.0,41.0,63.0,67.0	
sex	STRING	fem,male	
cp	STRING	abnang,angina,asympt,notang	
trestbps	STRING	120.0,130.0,145.0,160.0	
chol	STRING	204.0,229.0,233.0,250.0,286.0	
fbs	STRING	fal,true	
restecg	STRING	hyp,norm	
thalach	STRING	108.0,129.0,150.0,172.0,187.0	
exang	STRING	fal,true	
oldpeak	STRING	1.4,1.5,2.3,2.6,3.5	
slop	STRING	down,flat,up	
ca	STRING	0.0,2.0,3.0	
thal	STRING	fix,norm,rev	
status	STRING	buff,sick	
style	STRING	H,S1,S2	

数据预处理

操作步骤

数据准备完成后，单击**组件**，在**工具**和**数据预处理**文件夹下将**SQL脚本**、**类型转换**、**归一化**组件拖到画布中，并拼接成如下实验。



单击**SQL脚本**组件，在画布右侧的**SQL脚本**输入栏中输入sql语句，根据每个字段的含义将字符型转为数值。

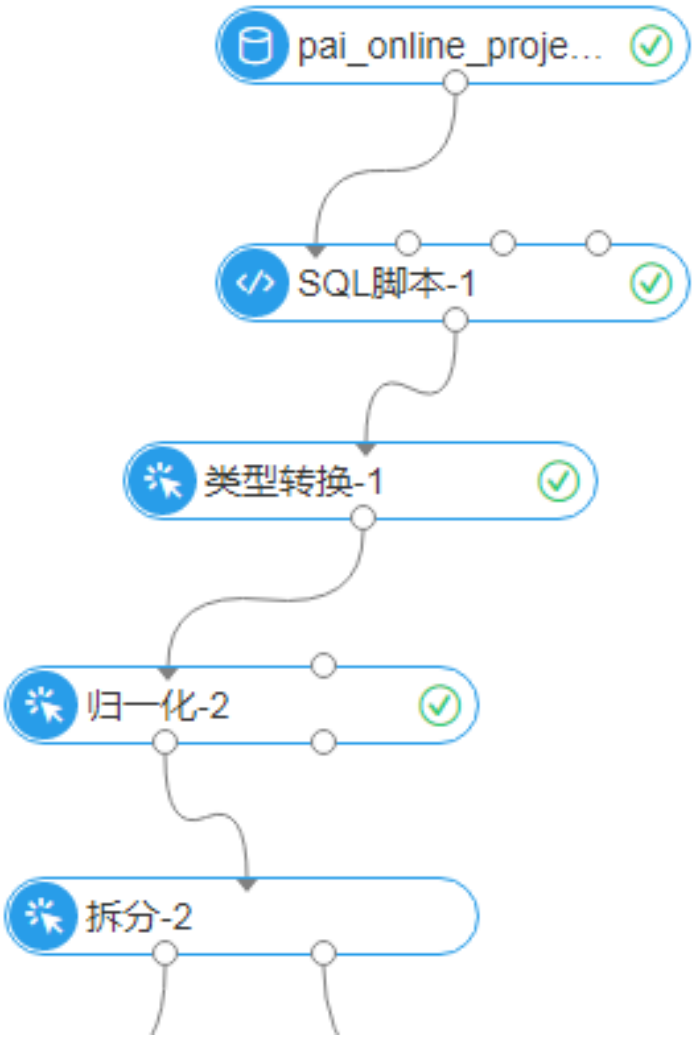
```
select age,
(case sex when 'male' then 1 else 0 end) as sex,
(case cp when 'angina' then 0 when 'notang' then 1 else 2 end) as cp,
trestbps,
chol,
(case fbs when 'true' then 1 else 0 end) as fbs,
(case restecg when 'norm' then 0 when 'abn' then 1 else 2 end) as restecg,
thalach,
(case exang when 'true' then 1 else 0 end) as exang,
oldpeak,
```

```
(case slop when 'up' then 0 when 'flat' then 1 else 2 end) as slop,
ca,
(case thal when 'norm' then 0 when 'fix' then 1 else 2 end) as thal,
(case status when 'sick' then 1 else 0 end) as ifHealth
from ${t1};
```

单击**数据转换**组件，在画布右侧的**字段设置**页签，单击**转换为double类型的列**下方的**选择字段**，选择全部字段，将字段类型转换为double。

单击**归一化**组件，在画布右侧的**字段设置**页签，选择全部字段。完成后单击画布下方的**运行**，系统将自动开始运行实验，在运行过程中可右键查看各组件的输出。

在**数据预处理**文件夹下，将**拆分**组件拖到画布中，并拼接运行，如下图所示。



说明：此步骤的目的是将数据拆分成两份，80%作为模型训练集，20%作为模型预测集。

算法建模

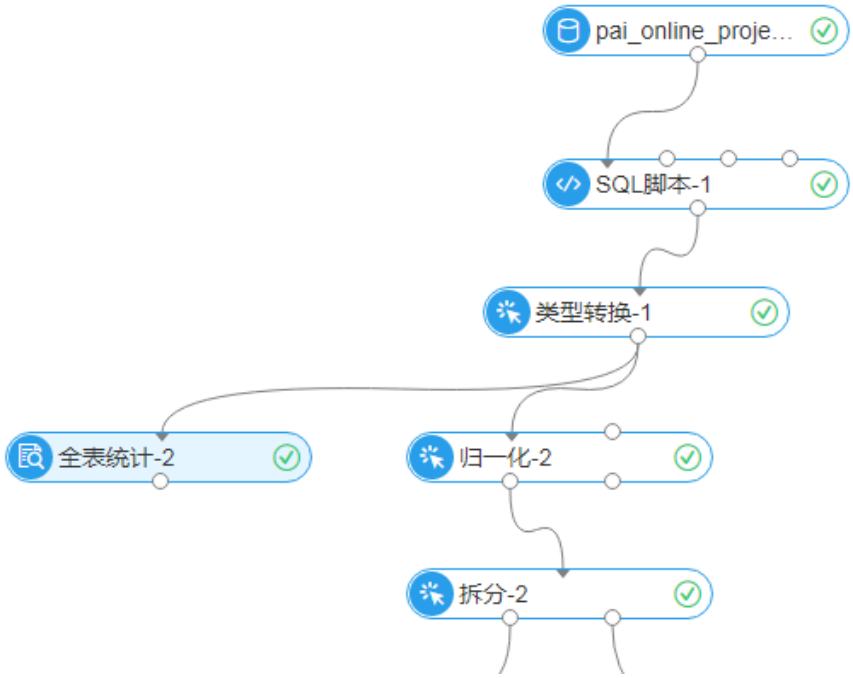
操作步骤

1. 在**机器学习->二分类**文件夹下，将**逻辑回归二分类**组件拖入画布。
2. 在右侧的**字段设置**页签，将目标列设置为**ifhealth**，**训练特征列**选择除目标列以外的全部列，并拼接运行，如下图所示。

数据可视化

操作步骤

在**统计分析**文件夹下，将**全表统计**组件拖入画布中，连接并运行，如下图所示。



待实验运行结束后，右键单击**全表统计**组件，选择**查看数据**，可看到数据的全表统计信息，如下图所示。

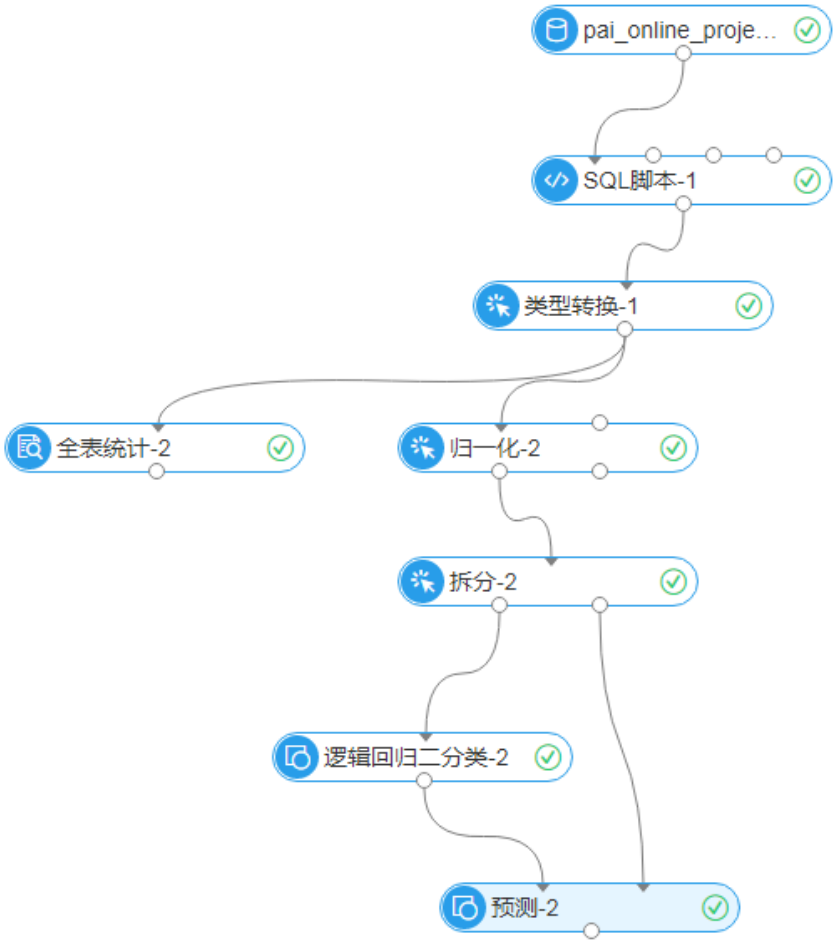

数据探查 - pai_temp_119048_1308661_1 - (仅显示前一百条)

序号	colname	datatype	totalcount	count	missingcount	nancount	positiveinfinitycount	negativeinfinitycount	min	max	m
1	age	double	303	303	0	0	0	0	29	77	5
2	ca	double	303	303	0	0	0	0	0	3	0.
3	chol	double	303	303	0	0	0	0	126	564	2
4	cp	double	303	303	0	0	0	0	0	2	1.
5	exang	double	303	303	0	0	0	0	0	1	0.
6	fbs	double	303	303	0	0	0	0	0	1	0.
7	ifhealth	double	303	303	0	0	0	0	0	1	0.
8	oldpeak	double	303	303	0	0	0	0	0	6.2	1.
9	restecg	double	303	303	0	0	0	0	0	2	0.
10	sex	double	303	303	0	0	0	0	0	1	0.
11	slop	double	303	303	0	0	0	0	0	2	0.
12	thal	double	303	303	0	0	0	0	0	2	0.
13	thalach	double	303	303	0	0	0	0	71	202	1
14	trestbps	double	303	303	0	0	0	0	94	200	1

模型评估

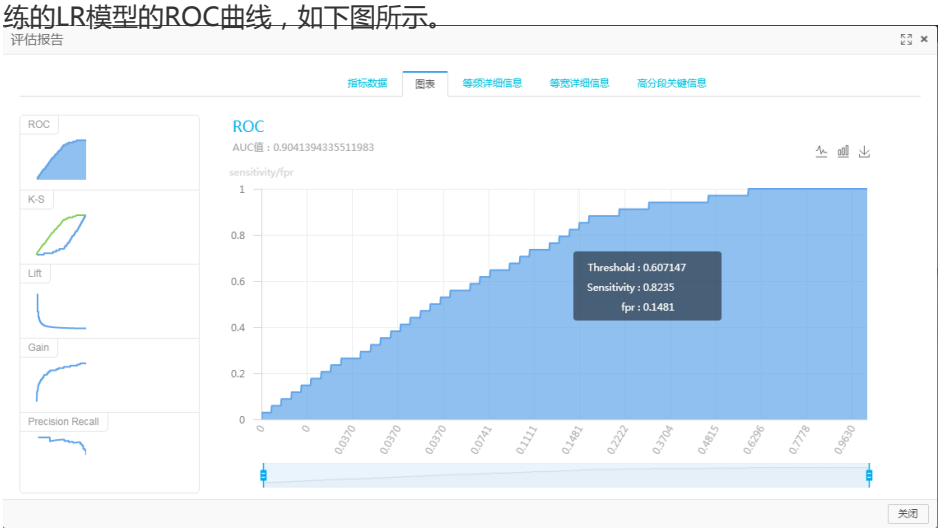
操作步骤

在**机器学习**文件夹下，将**预测**组件拖入画布，并连接对应的组件流和数据流，如下图所示。



在**机器学习->评估**文件夹下，将**二分类评估**组件拖入画布。在画布右侧的**字段设置**页签，将**原始标签列列名**设置为**ifhealth**，并连接对应的组件流和数据流。

3. 单击**运行**。完成后右键单击**二分类评估**组件，选择**查看评估报告**，单击**图表**页签，得到不同参数下训练的LR模型的ROC曲线，如下图所示。

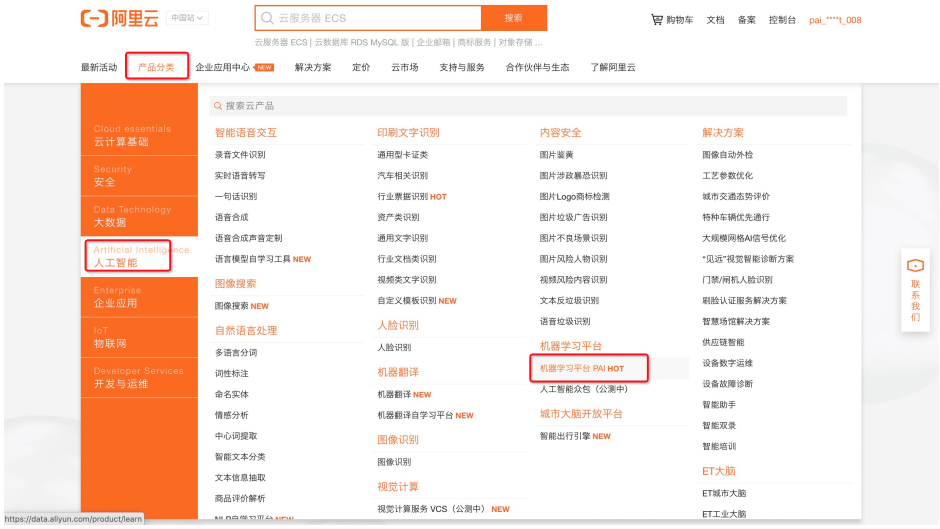


十分钟快速上手

阿里云机器学习PAI（Platform of Artificial Intelligence）平台目前有三个子产品，包括用于模型训练的可视化建模平台PAI Studio、自定义在线编程平台PAI DSW（Data Science Workshop）以及用户模型在线部署的PAI EAS（Elastic Algorithm Service）。本文将从产品入口到产品开通、PAI Studio模板实验运行生成模型全流程介绍产品功能，帮助用户快速上手。

机器学习PAI平台入口

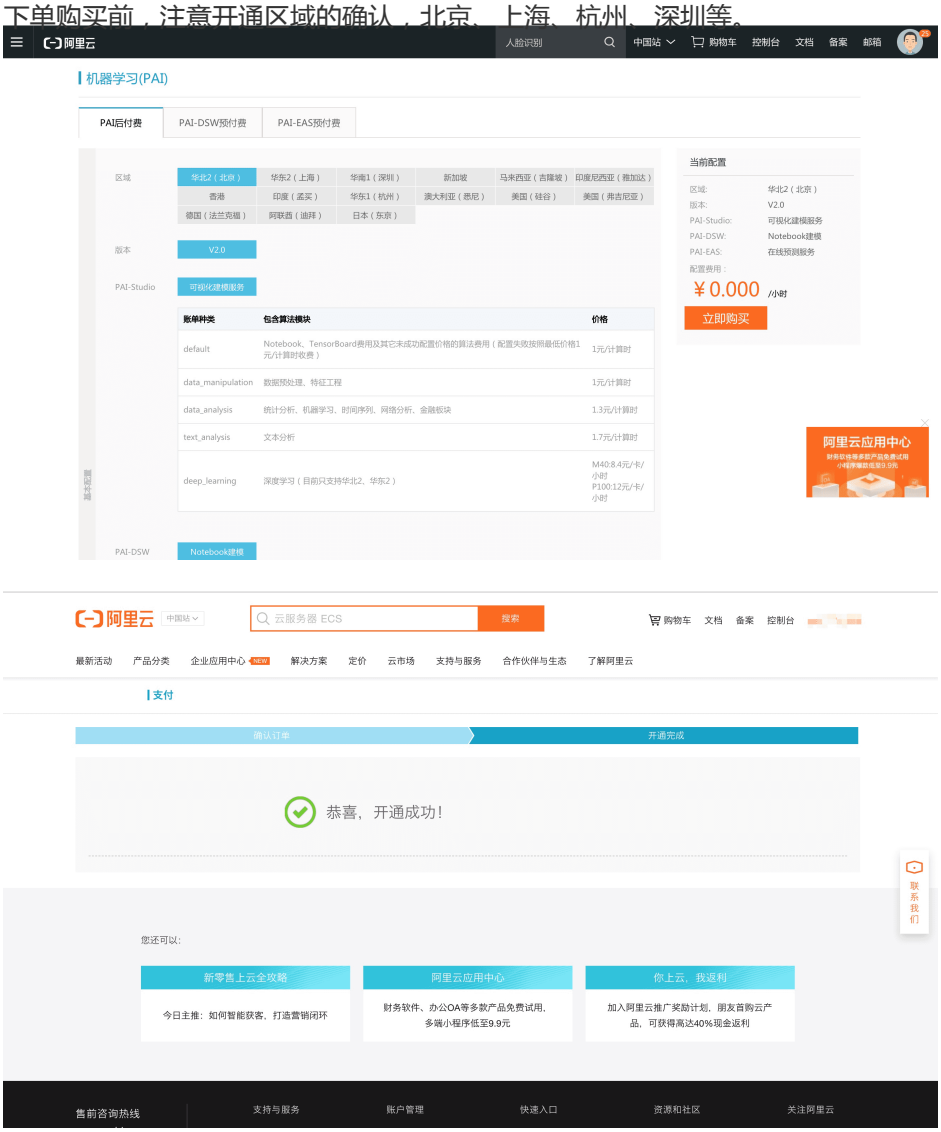
入口一： 进入阿里云官网，从上方导航栏选择产品分类->人工智能->机器学习PAI进入产品详情页，如下图所示：



入口二： 直接访问PAI Consolehttps://pai.data.aliyun.com/console

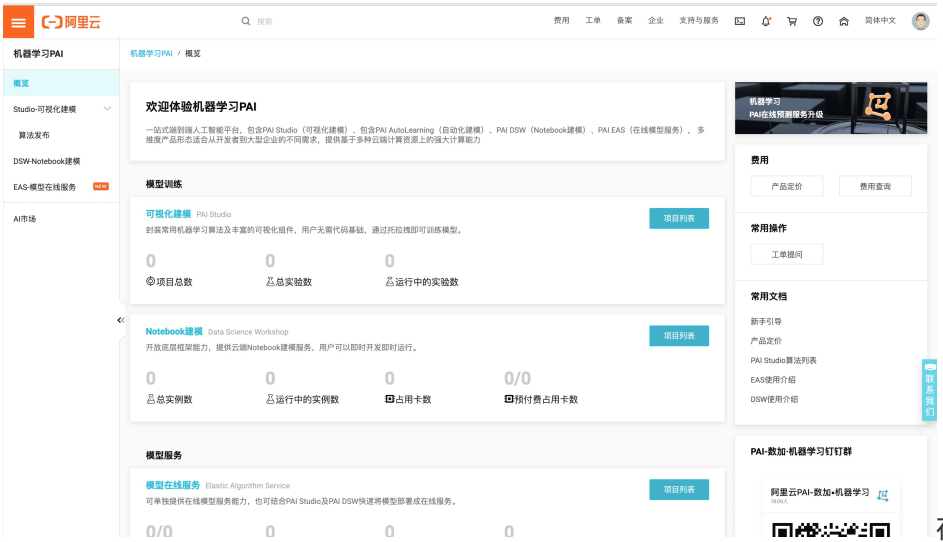
开通PAI及项目准备

访问PAI Console：https://pai.data.aliyun.com/console，点击立开通,按照下一步引导完成开通。下单购买前，注意开通区域的确认，北京、上海、杭州、深圳等。



说明：目前PAI产品提供了三款售卖产品：PAI后付费（包含PAI Studio、PAI EAS、PAI DSW）、PAI EAS预付费、PAI DSW预付费。按量付费：零元购买，按照实际使用量收费；预付费：包年包月购买。用户可根据实际业务需要和使用频度来选择合适的售卖方式。

开通成功后，回到PAI Consolehttps://pai.data.aliyun.com/console。在左侧导航切换到Studio可

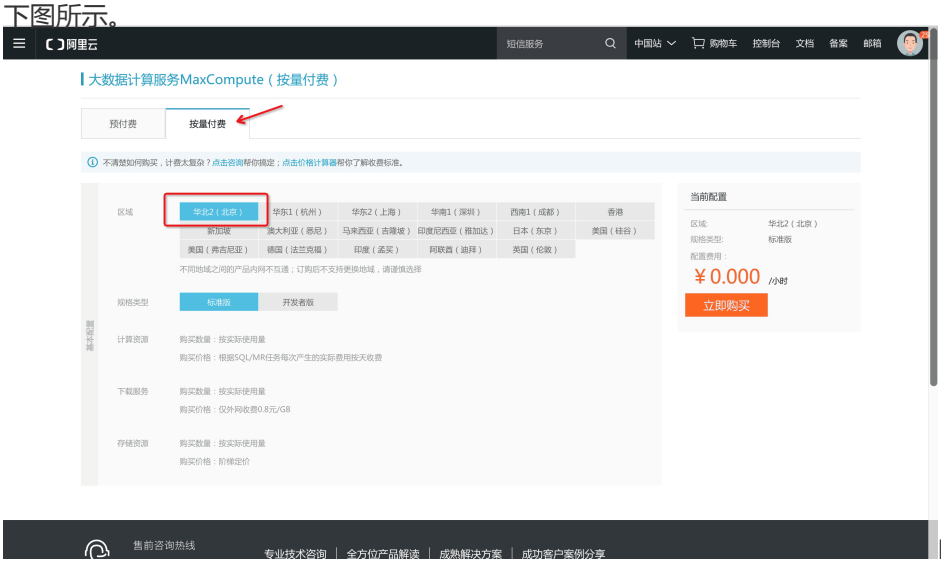


视化建模平台。在PAI Studio，切换至已购买服务的区域：

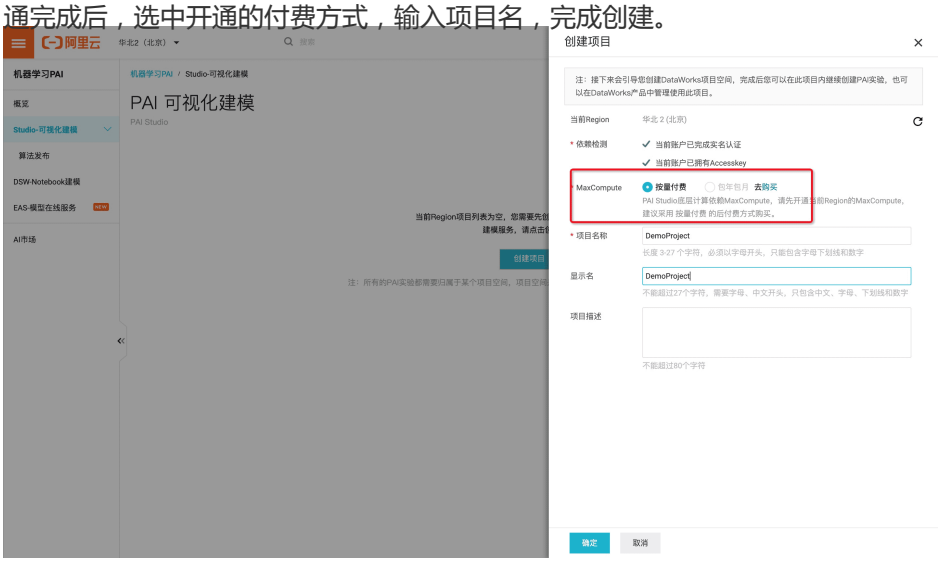
创建MaxCompute项目。需要完成以下三步：实名认证、创建AK、开通项目。 点击创建项目，按照提示开通MaxCompute服务。

点击购买

MaxCompute，前往MaxCompute购买页，注意选择**按量付费**，开通服务区域与PAI保持一致，如下图所示。



MaxCompute开通完成后，选中开通的付费方式，输入项目名，完成创建。

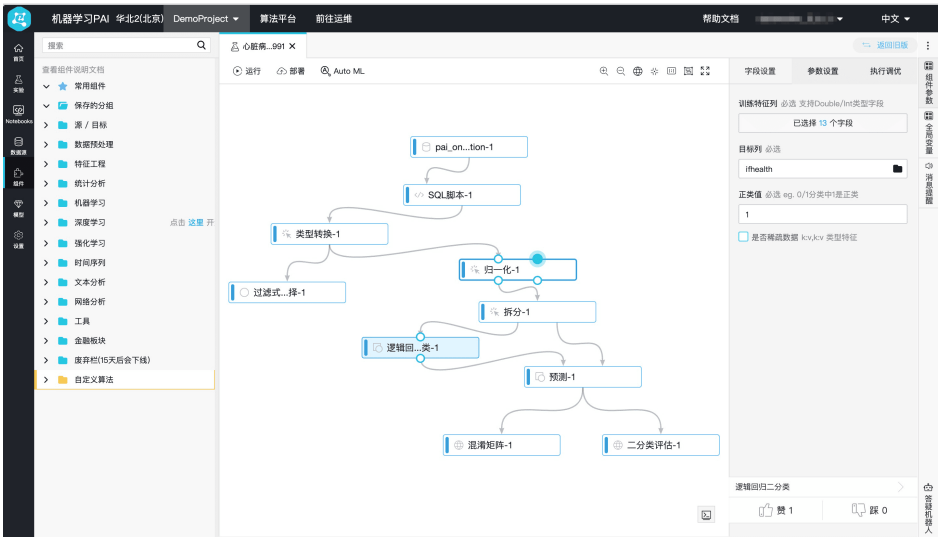


创建实验

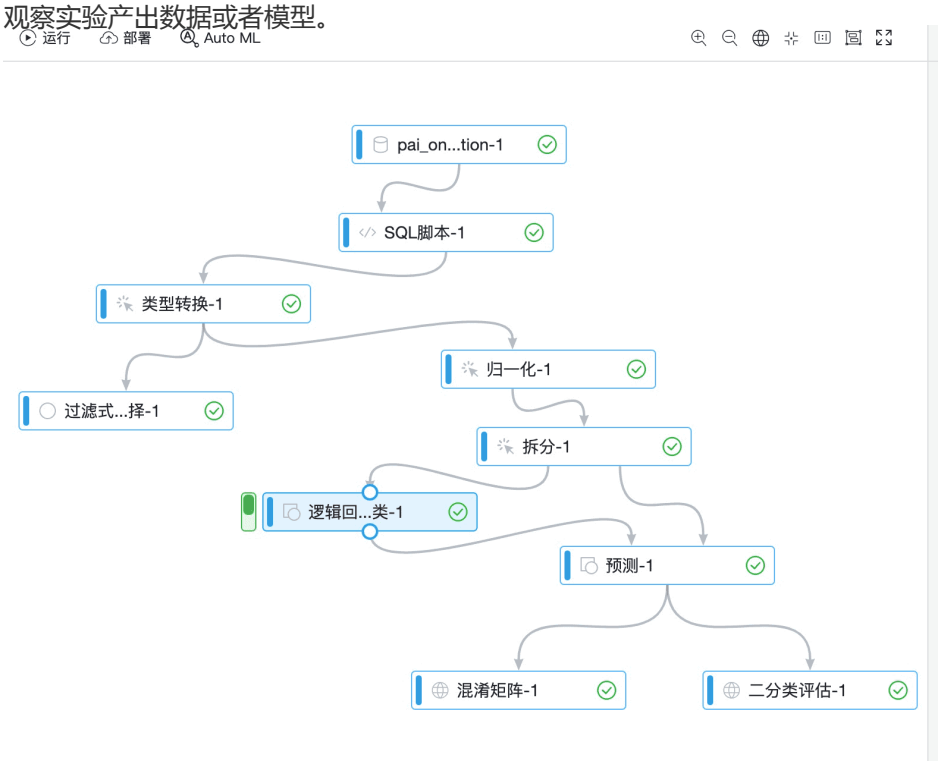
完成以上流程后，在控制台单击**进入机器学习**。

单击左上角的**首页**，进入产品首页。

选择一个模板创建。单击**从模板创建**开始创建模板，单击**查看文档**可以看到详细的案例说明。模板包含完整的实验流程以及数据，可以帮助您快速上手使用，新手建议使用**心脏病预测案例**模板，可以参照文档进行学习。



模板创建需要十秒钟左右时间，创建成功后如下图所示。单击**运行**开始实验，可以右键单击每个组件观察实验产出数据或者模型。



生成模型可通过PAI EAS部署成在线服务。

删除工作空间

如何删除已创建的工作空间可以参考文档：删除工作空间

其它文档和学习材料

阿里云机器学习的使用操作说明文档都可以通过官网学习路径获得。我们还提供了丰富的实战教学视频以及实战文章合集，帮助您快速使用常规机器学习算法和深度学习算法去解决问题。