机器学习PAI

快速入门

为了无法计算的价值 | [] 阿里云

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。



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。

ି ***	≡ 公共表
目贝	🖯 mnist_train
A	mnist_test
头盔	🖯 sparse_mnist_train
	sparse_mnist_test
蚁恬源	breast_cancer_data
င်္ဂာ	🖯 pai_dense
组件	farm_claim_process
\bigotimes	refund_allergy_sample
模型	🖯 wumai_data
ලා	telco_template_data
设置	adult_statistics_demo
	heart_disease_prediction
\sim	people_info
nequde	C know_point
_	≡ 全部表
し エ作空间	Q 请输入关键词
くしょう (All and a construction of the second	
~	
(?) 帮助	+ 创建表

建表		к л 12 У	×
表结构 数据			
	注意:上传的数据将追加搭到原表, 请上传 .txt 或 .csv 文件		
	选择文件		

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。

确定

关闭

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。

4	机器学习PAI	算法平台	前往运维				
€ 1							
丛城		基础		慈祥	基础	慈祥	新建 >
S Notebook		Tensorflow	图片分类	【推荐算法】商品推荐 Product	【文本分析】新闻分类	【图算法】 金融风控实验	家 新建 Tensorflow 1.2
≝ ###				Recommendation		¥	$\dot{\omega}$
97 8# 673		使用点击'查 式	君文档"了解使用方	通过协同过端算法实现商品推荐。 荐。	通过主题模型实现了整个文本分 类的流程。	利用图算法,针对个人信用,解决金融行业的风控问题。	机器学习算法计算出二氧化氮对于雾霾影响最大。
*		5597位用户		4330位用户	3533位用户	2274位用户	1834位用户
8 2 102		^{基础} 心脏病预测	则案例	^{田田} 农业贷款预测的回归算法…	^{基础} 【在线预测】中学生成绩…		

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。

ស			Q 搜索	mytest 💌	表选择	字段信息
百页	*	*	常用组件	Project Name DataWorksForPAI		1 m m m
A	•	-	保存的分组	G.		16.农石
实验	•	-	源/目标		<u> </u>	
\$			☐ 写数据表			
Notebook		- [├ 读数据表 2	E	•	3
			CSS数据同步	▲ 目 读数据表-2 ①		
数据源			Mysql数据同步	Ű		
ீ	6		数据预处理			
组件	•		特征工程			
\heartsuit	Þ		统计分析			
模型	•		机器学习			

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

G

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	
ср	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	
са	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。

择字段				2 N N N N N N N N N N N N N N N N N N N
输入关键字搜索列,包含关键字即可				Q
/ 全选		已选		列表编辑
BIGINT	~	Û	字段	类型
V sex		Û	sex	BIGINT
🗸 cp		Ē	cp	BIGINT
V fbs		Ô	fbs	BIGINT
✓ restecg		1	restecg	BIGINT
exang		Û	exang	BIGINT
✓ slop		Û	slop	BIGINT
✓ thal		Ē	thal	BIGINT
✓ ifhealth		Û	ifhealth	BIGINT
V STRING	~	11	age	STRING
v age		Û	trestbps	STRING
✓ trestbps		Ē	chol	STRING
✓ chol		ů	thalach	STRING

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。

mytest *		字段设置	参数设置	执行调优
	52 回 日 * 70 分	 副務務証列 必迭 支持OoubleInf 目标列 必迭 (fheath 正純値 必迭 eg. 0/1分供中1星I 1 屋咨稱軌款編 kv,kv 供型符 	供型字段 日选择 13 个字段 日选择 13 个字段 日述 日述 日述 日述 日述 日述 日本 日述 日本 日述 日本 日述 日本 日本 日本 日本 日本 日本 日本 日本 日本 日本 日本 日本 日本	

数据可视化

操作步骤



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。

	- pai_tomp	_110040_10	00001_1 - (IX:								23
序号▲	colname 🔺	datatype 🔺	totalcount 🔺	count 🛋	missingcount 🔺	nancount 🔺	positiveinfinitycount 🔺	negativeinfinitycount 🔺	min 🔺	max 🔺	ſ
1	age	double	303	303	0	0	0	0	29	77	
2	ca	double	303	303	0	0	0	0	0	3	
3	chol	double	303	303	0	0	0	0	126	564	
4	cp	double	303	303	0	0	0	0	0	2	
5	exang	double	303	303	0	0	0	0	0	1	
6	fbs	double	303	303	0	0	0	0	0	1	1
7	ifhealth	double	303	303	0	0	0	0	0	1	T
8	oldpeak	double	303	303	0	0	0	0	0	6.2	
9	restecg	double	303	303	0	0	0	0	0	2	T
10	sex	double	303	303	0	0	0	0	0	1	
11	slop	double	303	303	0	0	0	0	0	2	T
12	thal	double	303	303	0	0	0	0	0	2	
13	thalach	double	303	303	0	0	0	0	71	202	
14	trestbps	double	303	303	0	0	0	0	94	200	



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
------	---------	-------------	-----------	------------	---------------

(-) 阿里	□ 中国時マ Q 云服务器 ECS 云服务器 ECS 云数据岸 RD:	援索 S MySQL 版 企业邮箱 商标服务 对象存f 介 元市场 支持与服务 合作	₩ 約約年 3	= 文档 备案 控制台 pal_****t_008	
#新速 / Cione and Scientific 文字 Data Table 人工智能 Data Table 人工智能 Data Table 方式智能 Data Table 方式智能 Data Table 方式智能		公式市场 文持与服务 合件/ 副司法+定支 二月二十二支 汽車相关识明 一 「力工業原以明 二 通用法+定支 二 汽車相关识明 一 方定大原以明 三 通用文字识明 三 西京文等识明 三 西京文等识明 三 夏之义慶気识明 New 人給(明) 人給(明) 人給(明) 机器計畫(学习干台 New 「開新時 New 「「「「」」」」 三 「「「」」 三 「「」」 三 「「」」 二 「」 二 「「」」 二 「」」 二 「」」 二 「」」 二 「」」 二	火井与主素 了解阿堡去 内容安全 即片溜泉 即片溜泉 即片波泉局修満 田片な泉局修満 田片な泉局修満 田片な泉局であり 田片な泉局であり 田片な泉局であり 田片なり着いろい 大本反な奥切り 福倉学3平台 川雪家万字台70H07 大客屋衣名(公第中) 城市大路开放平台 腎細出行引澤 NEW	 	• 联系我们



开通PAI及项目准备

	114 0344361334	PAI-EAS预付表						
	_						当前配置	
区域	华北2(北京)	华东2(上海)	华南1(深圳)	新加坡	马来西亚(吉隆坡)	印度尼西亚(雅加达)	区18:	华北2(北京)
	1070 德国(法兰克福)	印度(<u>血</u> 关) 阿联茜 (迪拜)	4054 (初新) 日本(东京)	ALCHIE (KOLE)	與国(姓日)	米国(外白)6堂)	版本:	V2.0
							PAI-Studio: PAI-DSW:	可視化維機服务 Notebook建模
版本	V2.0						PAI-EAS:	在线预测服务
							配置费用:	_
PAI-Studio	可初化建模服务						¥ 0.000) /小时
	账单种类	包含算法模块				价格	立即购买	
	default	Notebook、Tensore 元/计算时收费)	oard费用及其它未同	动配置价格的算法费用	(配置失败按照最低价)	⁸¹ 1元/计算时		
	data_manipulation	数据预处理、特征工程	l.			1元/计算时		
	data_analysis	统计分析、机器学习、	时间序列、网络分标	f、金融板块		1.3元/计算时		
	text_analysis	文本分析				1.7元/计算时		阿里云应用中
	deep_learning	深度学习(目前只支持]华北2、华东2)			M40:8.4元/卡/ 小时 P100:12元/卡/		財務软件等多點产品便。 小律序編就低發9.9
		解决方案	定价 云市;	汤 支持与服务	合作伙伴与生态	了解阿里云		
最新活动 产品分	类 企业应用中心 🕯	moora m						
最新活动 产品分: 支	类 企业应用中心 4	A CANTRA A					20 phy all	
最新活动 产品分 【支	类 企业应用中心 ◀ 付	新 八 5米				开道	西完成	
最新活动 产品分	类 企业应用中心 ◀ 付	₩₩₩ ₩₩ ₩ ₩ ₩ ₩ ₩ ₩ ₩ ₩ ₩ ₩	喜,开通	文功!		ŦĬ	围完成	
最新活动 产品分	秦 金业应用中心 4 付	muute muute 资 恭	喜,开通/	, 成功!		Ŧi	善无成	
最新活动 产品分	秦 金业应用中心 ◀ 付	₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩₩	喜,开通)	文功!		7 1	8751A	
联新活动 产品分:	条 企业应用中心 4	₩XXJ# AU\17# 承	喜,开通	成功!		71	8元成	
義新活动 产品分 【文 型子	条 企业应用中心 4	a∪.j≄	喜,开通	成功!	9.Cs	Ŧ4 b	8元成 二十二、 我送利	
活动 产品分: 支	类 企业应用中心 《 付	新认订单)		开道	通完成	

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。

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



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。

	■ (-)阿里云	华北2(北京) 👻	Q 接來	创建项目		×
	①器学习PAI	机器学习PAI / Studio-可视化建模		注: 援下来会引	导您创建DataWorks项目空间,完成后您可以在此项目内继续创建PAI实验,也可	1
AUG Billingio 9.42(3) C AUG Bill		PAI 可视化建模		以在DataWorks	产品中管理使用此项目。	
 (本) (*) (*)	□-可視化建模 ∨			当前Region	华北 2 (北京)	G
terebool:2013	去发布			* 依赖检测	✓ 当前账户已完成实名认证	
Katisti C	Notebook建模			* MaxCommute	HIM CONTACTORY	
	>模型在线服务 🔤		当前Region项目列表为空,您需要先创	materia	PAI Studio底层计算依赖MaxCompute,请先开道当前Region的MaxCompute, 建议采用 按量付费 的后付费方式购买。	
	515		建模服务,请点击任	* 项目名称		
注:所有的外以菜油都原菜加菜子菜小類型菜用、類型菜和 不能超过20个学校、需要学母、中交片头、只怕20中交、学母、下型加制和数字 項目描述 - 不能超过20个学校 - 不能超过20个学校 - 不能超过20个学校 			創建项目		长度 3-27 个字符,必须以字母开头,只能包含字母下划线和数字	
本 初日日2 日本			注:所有的PAI实验都需要归属于某个项目空间,项目空间	显示名	不能超过27个字符,需要字母、中文开头,只包含中文、字母、下划线和数字	
不能超过10个学校				项目描述		
不能超过40个学行		«				
17 15/5					不能超过80个字符	
11 15/8						
117 15/8						
97 828						
9.7 858						
9.7 858						
BIT B 239						
				确定	取消	



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

创建实验

完成以上流程后,在控制台单击进入机器学习。



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。





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。