机器学习PAI

产品简介

产品简介

什么是机器学习PAI

什么是机器学习

机器学习指机器通过统计学算法，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。目前机器学习主要在以下方面发挥作用：

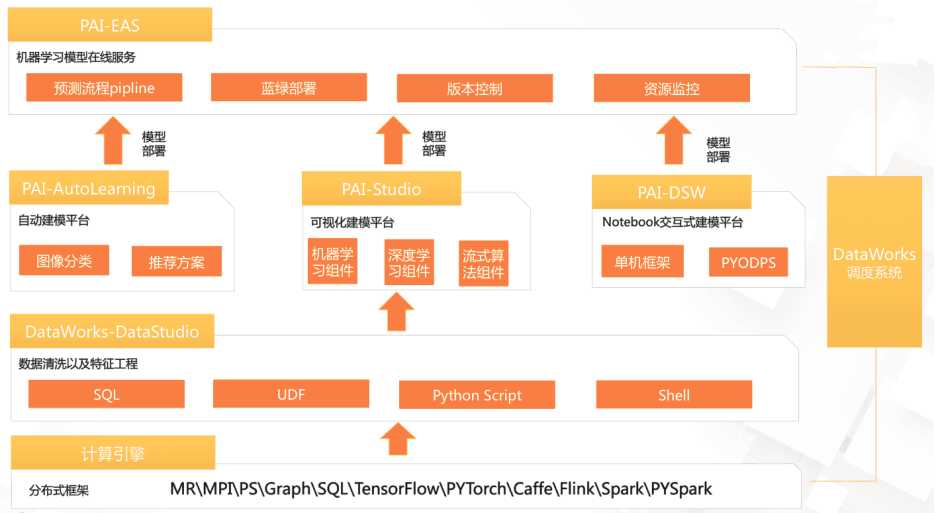
- 营销类场景：商品推荐、用户群体画像、广告精准投放
- 金融类场景：贷款发放预测、金融风险控制、股票走势预测、黄金价格预测
- SNS关系挖掘：微博粉丝领袖分析、社交关系链分析
- 文本类场景：新闻分类、关键词提取、文章摘要、文本内容分析
- 非结构化数据处理场景：图片分类、图片文本内容提取OCR
- 其它各类预测场景：降雨预测、足球比赛结果预测

笼统地讲，机器学习可以分为三类：

- 有监督学习（Supervised Learning）：指每个样本都有对应的期望值，通过模型搭建，完成从输入的特征向量到目标值的映射。典型的案例就是回归和分类问题。
- 无监督学习（Unsupervised Learning）：指在所有的样本中没有任何目标值，期望从数据本身发现一些潜在的规律，例如一些简单的聚类。
- 增强学习（Reinforcement Learning）：相对来说比较复杂，是指一个系统和外界环境不断地交互，获得外界反馈，然后决定自身的行为，达到长期目标的最优化。其中典型的案例就是阿法狗下围棋，或者无人驾驶。

什么是机器学习平台PAI

PAI起初是一个定位于服务阿里集团的机器学习平台，致力于让AI技术更加高效、简洁、标准的被公司内部开发者使用。对集团内，PAI服务了淘宝、支付宝、高德等部门的业务。随着PAI的算法的不断积累，2015年底PAI作为天池大赛的官方比赛平台在阿里云正式上线，也成为了国内最早的云端机器学习平台之一。随着PAI在阿里云的业务的不断发展，2018年PAI平台正式商业化，目前已经在公有云积累了数万的企业客户以及个人开发者，是目前国内领先的云端机器学习平台之一。



PAI底层支持多种计算框架：有流式算法框架Flink，基于开源版本深度优化的深度学习框架TensorFlow，支持千亿特征千亿样本的大规模并行化计算框架Parameter Server，同时也兼容Spark、PYSpark、MapReduce等业内主流开源框架。

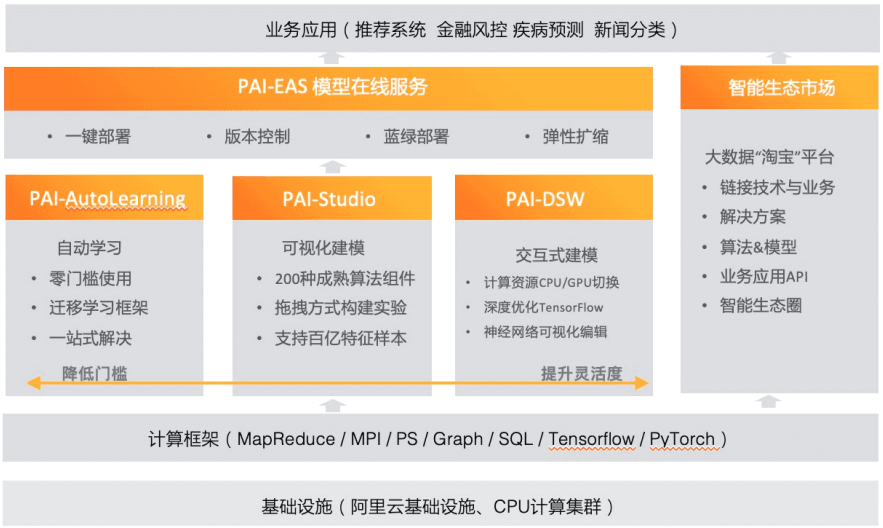
PAI平台提供：PAI-STUDIO（可视化建模和分布式训练）、PAI-DSW（notebook交互式AI研发）、PAI-AutoLearning（自动化建模）、PAI-EAS（在线预测服务）四套服务，每个服务既可单独使用，也可相互打通。用户可以从数据上传、数据预处理、特征工程、模型训练、模型评估，到最终的模型发布到离线或者在线环境，一站式完成建模，有效的提升开发效率。在数据预处理方面，PAI跟阿里云DataWorks（一站式大数据智能云研发平台）也是无缝打通的，支持SQL、UDF、UDAF、MR等多种数据处理开发方式，灵活性较高。在PAI平台上训练模型，生成的模型可以通过EAS部署到线上环境，整个实验流程支持周期性调度，可以发布到DataWorks与其它上下游任务节点打通依赖关系，另外调度任务区分生产环境以及开发环境，可以做到数据安全隔离。

一站式的机器学习平台意味着只要训练数据准备好（存放到OSS或MaxCompute中），用户就不需要额外的迁移工作，所有的建模工作都可以通过PAI来实现。

产品架构

概述

阿里云机器学习PAI平台的产品架构及上下游关系如下图所示。



上述架构图包括了整个AI业务的四个流程层：

- 基础设施层：CPU计算集群。
- 计算框架层：包括MapReduce、SQL、MPI等计算方式，分布式计算架构主要执行并行化计算分发任务。
- 核心产品功能层：即PAI提供的产品的核心能力。
- 业务应用层：阿里巴巴内部的搜索系统、推荐系统、蚂蚁金服等项目在进行数据挖掘工作时，都是依赖机器学习平台产品。机器学习平台的业务场景包含了金融、医疗、教育、交通、安全等各个领域。

这里我们重点介绍PAI的核心功能：

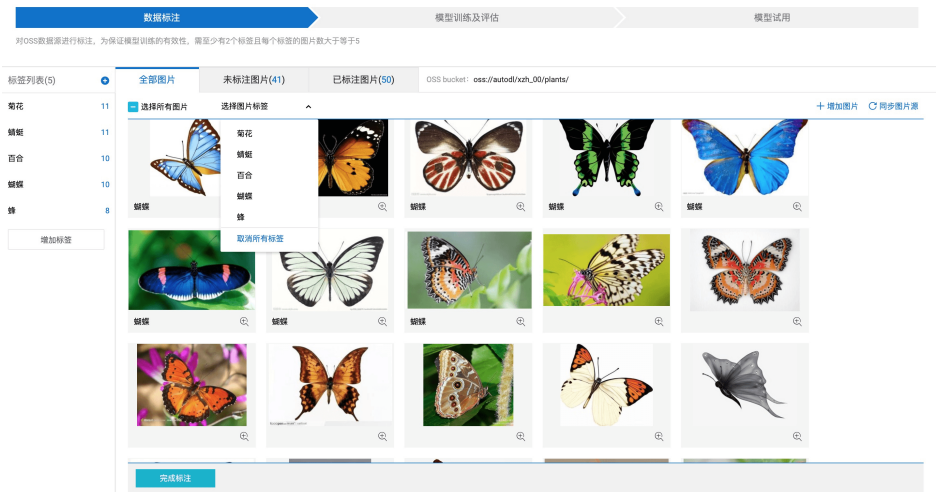
如上图所示，PAI提供PAI-AutoLearning、PAI-Studio、PAI-DSW三种建模方式，从左到右，建模的灵活度更高。从右到左，建模的技术要求降低。其中Studio中包括了数据预处理、特征工程、机器学习算法、深度学习等基本组件。所有算法组件全部脱胎于阿里巴巴集团内部成熟的算法体系，经受过PB级别业务数据的锤炼。

此外，PAI在模型建模基础上，提供模型在线服务一键部署功能，解决了用户模型部署使用的最后一公里问题。

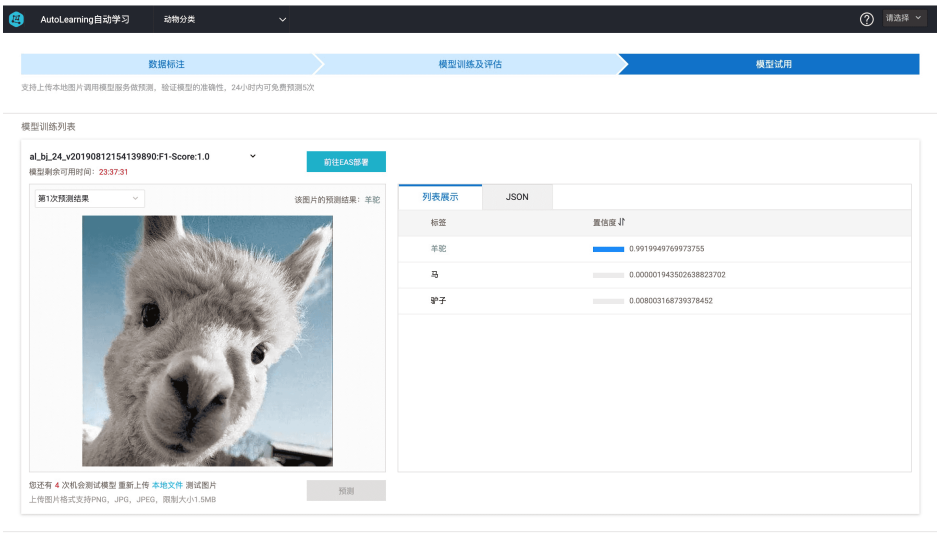
最后，PAI还给用户提供了智能生态市场功能，用户可以通过在智能生态市场快速获取业务解决方案或模型算法，进行相关业务与技术的高效对接。

PAI-AutoLearning:自动学习

PAI-AutoLearning自动化建模平台拟在为用户提供低门槛的偏场景化的机器学习建模服务，目前该平台已经内置了图像分类、推荐召回（即将上线）两款经典的机器学习业务场景，用户只需要在产品中做些基础的配置，无需对机器学习建模理论有深入的了解即可完成模型训练。

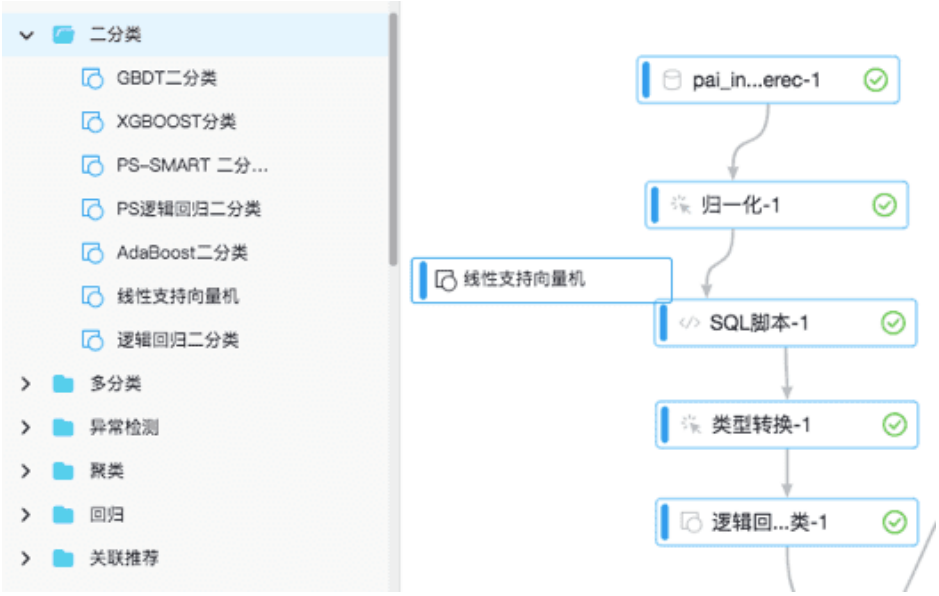


以图像分类为例，用户只需要在平台上标注不同的图片的类别，AutoLearning服务会基于迁移学习框架自动生成图像分类模型，该模型可以一键式部署到PAI-EAS上形成可调用的Restful服务，实现零门槛的机器学习使用体验。



PAI-Studio:可视化建模

PAI-Studio拖拽式建模平台，机器学习的真正门槛来自于对底层算法原理的理解，以及复杂的计算机实现。为了解决这种问题，PAI平台将200余种经典算法进行封装，让用户可以通过拖拽的方式搭建机器学习实验。

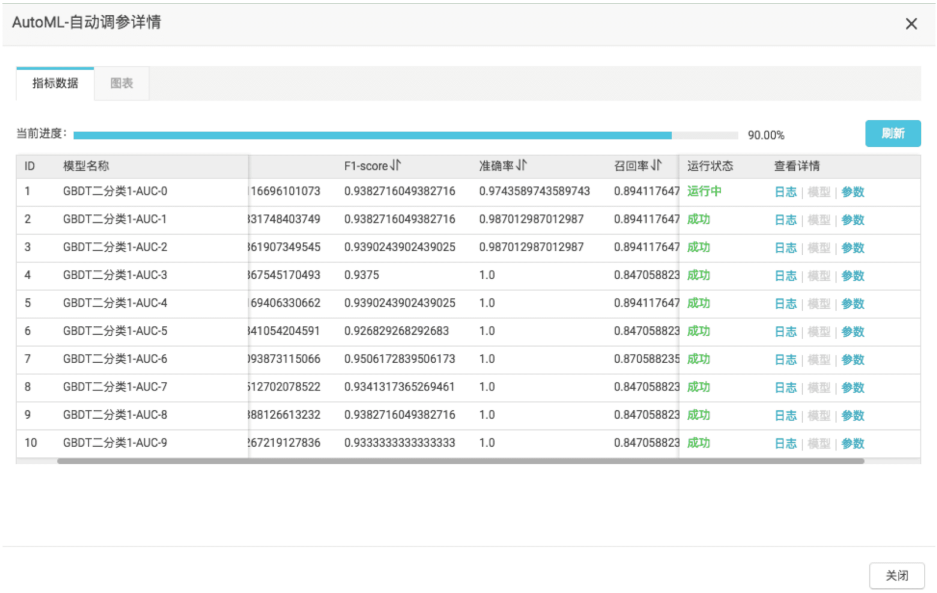


PAI-Studio中的所有算法都经历过阿里巴巴集团许多业务、EP级数据的锤炼。根据算法的不同特点选用MapReduce、MPI、ParameterSever、Flink等不同框架进行实现，真正做到成熟、稳定、简单、易用。

同时，在调参方面，如何探寻算法最优的超参数组合是一直以来困扰算法工程师的难题，调参工作不仅考验算法工程师对于算法推导认知的功底，还会带来大量手动尝试的工作量，工作效率很低。PAI-Studio内置的AutoML技术通过智能化的方式降低机器学习实验搭建的复杂度，通过自研的进化式调参等方式彻底解放用户的调参工作，实现模型参数自动探索、效果自动评估、模型自动向下传导，实现模型优化全链路零干预，大大降低机器学习门槛，节约计算成本。

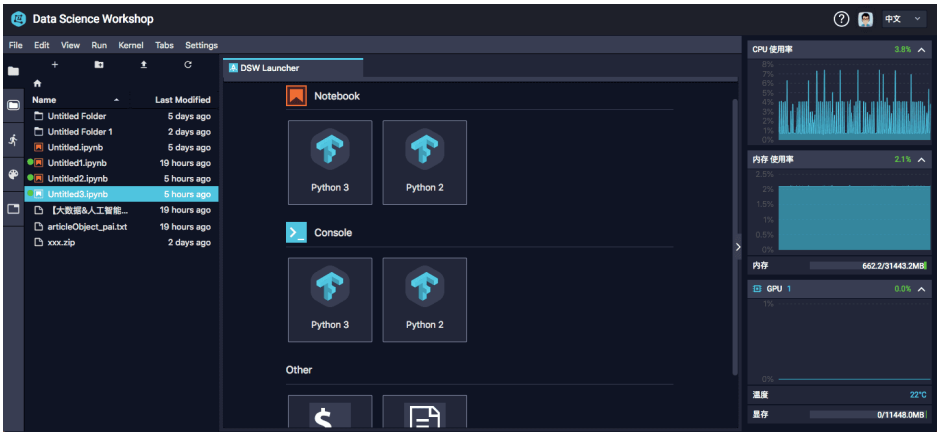
PAI-AutoML自动调参引擎自上线以来，已经收到国内外客户不错的反响。AutoML不仅包含基于Parallel Search思想的Grid search、Random search两种传统调参模式，还包含PAI团队基于Population Based Training理论原创的Evolutionary Optimizer调参模式，这种调参方式可以渐进式的帮助用户以最小代价探寻最优参数组合。

与此同时，Evolutionary Optimizer在调参过程中保留所有参数的表现以备追溯，并且调参模式与训练流程打通，做到自动选参、自动训练、自动评估、自动部署的整个链路自动化。



PAI-DSW:交互式代码建模

PAI-DSW交互式建模平台基于原生JupyterLab做了大量定制化工作，可以实现交互式的建模工作。支持用户绑定自己的云端存储资源，同时底层可以灵活动态的选用不同类型的GPU机器。

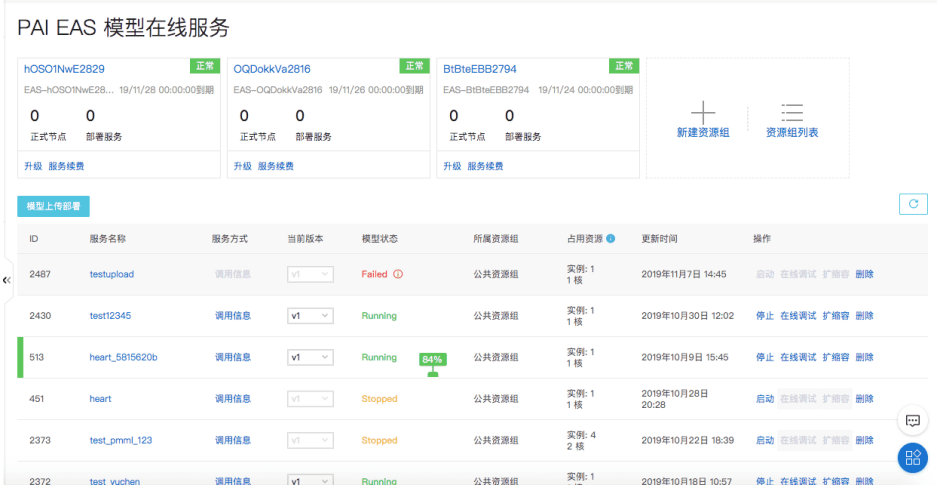


另外DSW还提供了可视化深度学习神经网络开发功能以及GPU资源可视化功能，DSW内置了大量的开源数据集以及模型文件供开发者使用，支持用户自己安装Python依赖文件。如果想使用弹性的GPU资源，快速构建深度学习代码，DSW是最合适的选择。

PAI-EAS:在线服务部署

PAI-EAS模型在线服务引擎提供了机器学习模型在线服务功能，支持基于异构硬件(CPU/GPU)的模型加载和数据请求的实时响应。您可以通过多种部署方式将您的模型发布成为在线的Restful API接口，同时我们提供的资源监控、弹性扩缩、蓝绿部署、版本控制等特性可以支撑您以最低的资源成本获取高并发、稳定的在线算法模型服务。

用户可以将Studio、DSW、Autolearning服务生成的模型一键式的发布到PAI-EAS形成Restful服务，通过EAS服务与用户自己业务系统打通，解决模型和客户业务最后一公里的问题。



目前EAS公共云支持区域：华北2(北京)、华东2(上海)、华东1(杭州)、华南1(深圳)、新加坡。

EAS公共云上整体通过资源占用量收费。提供公共资源组及专属资源组两种资源占用模式。在公共资源组中按照每个模型服务占用的资源计量计费，在专属资源组中根据资源组管理的机器资源包年包月或按量付费。

AI市场:数据智能技术商城

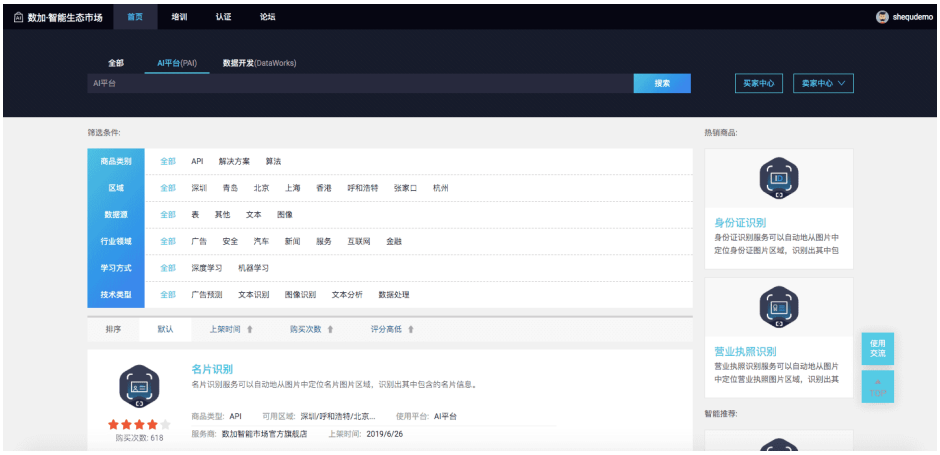
PAI平台内置了数加生态市场，用户可以基于PAI-Studio的自定义算法功能开发算法并在市场开店和上架，实现产品和生态的融合。

用户可以将数加智能市场看作大数据与AI领域的“淘宝”交易平台。市场旨在促进大数据与AI技术产品的开发创新与应用：一方面，帮助更多的开发者基于Dataworks和PAI去开发应用，并且将应用售卖给更多用户；另一方面，帮助更多有业务需求的客户，在市场中找到解决自己问题的答案。

数加智能市场的商品类目包括大数据领域的解决方案、人工智能领域的图像识别及文本识别等应用API、机器学习封装算法等，并在不断扩展中。

此外，市场不光承载着数据产品的交易功能，更多地还承载着培养整个生态的使命。所以数加智能市场还提供了认证、培训和论坛三大板块。通过培训板块和阿里云认证体系，让大数据与AI爱好者可以在此获得更多、更全面的学习机会，为社区培养更多生态开发者。论坛版块又给开发者们提供了一个交流切磋，相互提高的平台

场所。

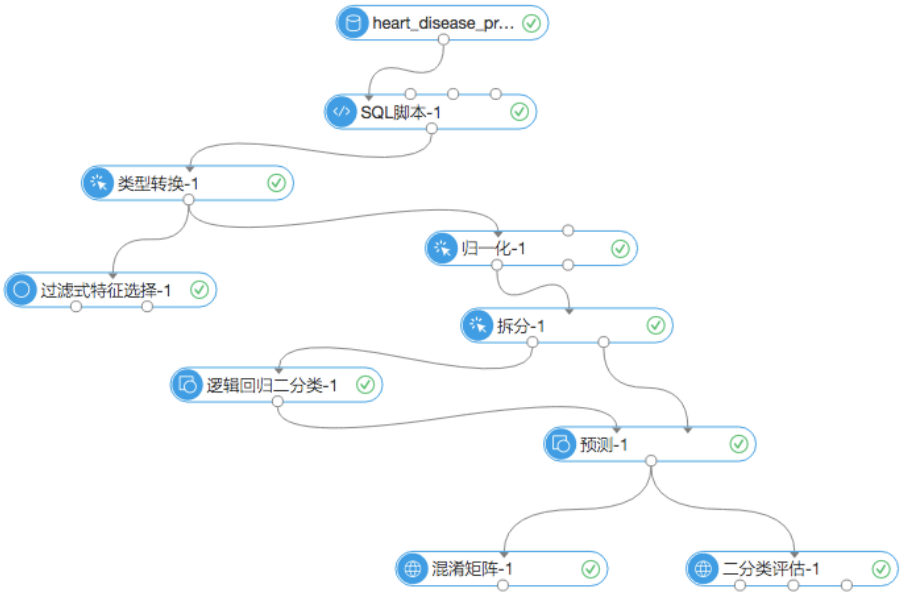


产品优势

阿里云机器学习平台的产品主要优势如下。

良好的交互设计

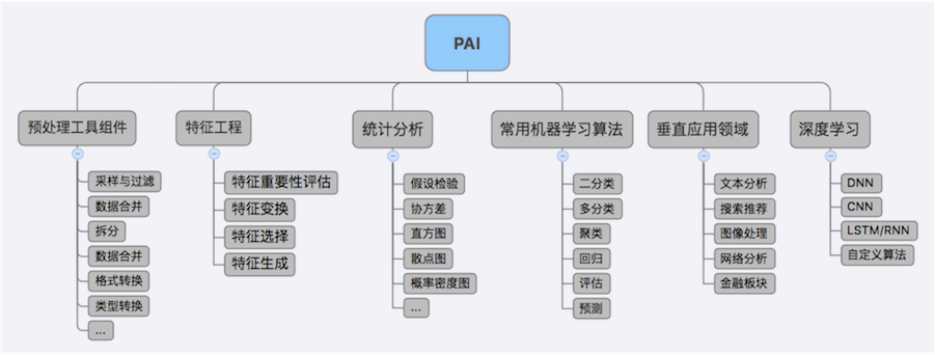
通过对底层的分布式算法封装，提供拖拉拽的可视化操作环境，让数据挖掘的创建过程像搭积木一样简单。缩短了您与数据的距离，真正实现了数据的触手可及。同时也提供了命令行工具，方便您将算法嵌入到自己的工



程中。

优质、丰富的机器学习算法

机器学习平台上的算法都是经过阿里大规模业务锤炼而成的。从算法的丰富性角度来看，阿里云机器学习平台不仅提供了基础的聚类、回归类等机器学习算法，也提供了文本分析、特征处理等比较复杂的算法。



与阿里产品完美配合

使用阿里云机器学习平台计算的模型直接存储在MaxCompute（原ODPS）上，可以配合其它阿里云的产品组件加以利用。

一站式的机器学习体验

阿里云机器学习平台除了提供模型训练功能，还提供在线预测以及离线调度功能，让机器学习训练结果和业务可以无缝衔接。

支持主流深度学习框架

阿里云机器学习平台已经包含了Tensorflow、Caffe、MXNet这三款主流的机器学习框架，底层提供M40型号的GPU卡进行训练。

优质的技术保障

阿里云机器学习算法平台的背后是阿里巴巴的算法科学家和阿里云的技术保障团队，在使用过程中遇到任何问题都可以到工单系统提交工单或者直接与相关接口人联系，具体请参见用户交流。

基本概念

本文将为您介绍机器学习PAI中的基本概念，各概念标题的括号中注明了该概念属于哪一个子产品版块。

PAI-Studio

项目

PAI Studio依赖于项目，在PAI-Studio的一个Region中，用户可以创建多个项目，用于实现资源、权限、实验的隔离与管理（主账号可对子账号进行项目授权）。如何创建项目可参考创建PAI-STUDIO项目 。

实验

在一个机器学习项目中，用户可以创建多个实验，用以构建算法模型。离线调度以实验为单位。

表

机器学习中的表是存储在MaxCompute中的，即MaxCompute的数据存储单元。它在逻辑上也是由行和列组成的二维结构，每行代表一条记录，每列表示相同数据类型的一个字段，一条记录可以包含一个或多个列，各个列的名称和类型构成这张表的Schema。
您可以在机器学习平台上创建、收藏表并导入数据，该表会自动存储在MaxCompute平台上。需要进入MaxCompute平台删除所创建的表。

PAI-DSW

实例

实例是DSW用户进行开发（如数据读取、算法开发、模型训练）的基本操作空间单元，也是资源以及存储的关联基本单元。用户在开始代码编辑之前，需要先创建DSW实例。

PAI-EAS

资源组

在EAS中，集群资源会被分为不同的资源组进行隔离，在新建模型服务的时候，用户可以选择部署在默认的公共资源组还是部署在用户自行购买创建的专属资源组中。

模型服务

模型文件+在线预测逻辑部署成的常驻服务，用户可以对模型服务进行创建、更新、停止、启动、扩缩容等操作。

模型文件

指您通过离线训练得到的离线模型，基于不同的框架会得到不同的模型格式，通常和Processor一起部署得到模型服务。

Processor

Processor是包含在线预测逻辑的程序包，通常和模型文件一起部署得到模型服务，EAS已经针对常用的PMML、TensorFlow(SavedModel)、Caffe模型提供了内置的官方Processor。

自定义processor

EAS内置Processor无法覆盖到用户所有的服务部署需求，可通过自定义processor进行更灵活的拓展，EAS支持通过C++/Java/Python开发自定义processor进行部署。

服务实例

每个服务可以部署多个服务实例以提升服务可支持的并发请求，如果资源组中有多台机器资源，EAS会自动将不同实例分布到不同机器上以更好的保障服务高可用性。

VPC高速直连

在用户的专属资源组和用户自有VPC打通之后，用户可以在自有VPC中通过客户端高速直连访问服务的每个单独的实例。

地域和可用区

在本文中将为您介绍机器学习PAI的各个产品的可用区域。

PAI-Studio

PAI-Studio目前可使用的区域包括：

	可用区域（Region）
亚太	华东1（杭州）、华东2（上海）、华北2（北京）、华南1（深圳）、中国（香港）、新加坡、澳大利亚（悉尼）、马来西亚（吉隆坡）、印度尼西亚（雅加达）、日本（东京）、
欧洲与美洲	德国（法兰克福）、美国（硅谷）、美国（弗吉尼亚）
中东与印度	印度（孟买）、阿联酋（迪拜）

PAI-DSW

PAI-DSW目前可使用的区域包括：

	可用区域（Region）
后付费	华东1（杭州）、华东2（上海）、华北2（北京）、华南1（深圳）、新加坡
预付费	华东2（上海）、华北2（北京）

PAI-EAS

PAI-EAS目前可使用的区域包括：

	可用区域（Region）
亚太	华东1（杭州）、华东2（上海）、华北2（北京）、华南1（深圳）、新加坡、印度尼西亚（雅加达）
欧洲与美洲	德国（法兰克福）
中东与印度	印度（孟买）

PAI-AutoLearning

PAI-AutoLearning目前可使用的区域包括：

可用区域（Region）
华东1（杭州）、华北2（北京）

PAI-Blade（公测中）

PAI-Blade目前可使用的区域包括：

可用区域（Region）
华东2（上海）

PAI权限管理文档

PAI成员权限管理文档

PAI分为PAI-AutoLearning、PAI-Studio、PAI-DSW、PAI-EAS 4款子产品。其中PAI-AutoLearning是个自动化建模平台，PAI-Studio为拖拽式的模型训练平台，PAI-DSW为Notebook交互式模型训练平台，PAI-EAS为模型在线服务平台。4款产品各有各自的权限点。具体权限点请分别参考下方目录：

- AutoLearning权限管理说明
- Studio权限管理说明
- DSW权限管理说明
- EAS权限管理说明