

机器学习平台操作手册

更新说明

版本	日期	更新内容	说明	撰稿人
V1.0	2017-12-20			宋广磊

目录

机器学习平台操作手册..... 1

第 1 章 机器学习平台综述..... 3

 1.1 平台服务简介..... 3

 1.2 平台服务内容..... 3

 1.3 数据限定..... 4

第 2 章 机器学习平台操作..... 5

 2.1 登录..... 5

 2.2 页面总体框架..... 5

 2.3 上传数据..... 5

 2.4 构建模型..... 6

 2.5 单例预测..... 7

 2.6 批量预测..... 7

第1章 机器学习平台综述

1.1 平台服务简介

基于 Spark2.1 ML/MLlib 库构建本机器学习平台，本着“上不碰应用，下不碰数据”的原则为用户提供构建机器学习应用的框架平台，用户上传数据、选择算法构建模型，并利用自己构建的模型进行数据挖掘。

本平台集成了 Spark ML/MLlib 库中的各种算法，包括回归算法处理数值型数据，分类算法同时处理数值型数据和文本型数据，聚类算法也同时支持处理数值型和文本型数据。本平台将数据特征处理、机器学习模型构建和预测的流程进行了良好的封装，用户仅需几步简单的页面操作即可完成一项复杂、艰巨的机器学习任务。

Spark 是大数据生态的一个分布式计算框架，它利用数据集内存缓存以及启动任务时的低延迟和低系统开销实现高性能，再者其容错性、灵活的分布式数据结构和强大的函数式编程接口，这让 Spark 在机器学习上有广泛的应用。Spark 机器学习库提供了常用机器学习算法的实现，包括聚类，分类，回归，协同过滤，维度缩减等。

1.2 平台服务内容

本平台现在集成起来、可以使用的算法包括如下：

算法类别	数据类别	算法	算法注释
回归	数值	LinearRegression	线性回归
		DecisionTreeRegression	决策树回归
		RandomForestRegression	随机森林回归
		GBTRegression	梯度迭代树回归
分类	数值	LogisticRegression	逻辑回归
		DecisionTreeClassification	决策树分类
		RandomForestClassification	随机森林分类
		GBTClassification	梯度迭代树分类
		NaiveBayes	原生贝叶斯分类
	文本	MultilayerPerceptronClassifier	多层感知器分类
聚类	数值	Kmeans	Kmeans
	文本	LDA	主题模型

为了支持多用户和数据隔离，分别创建属于每个用户单独的数据目录，并确保数据安全。同时将每个用户生成的模型存储在用户相应的目录下，这些生成的模型可以不断重复使用，避免每次都要构建。为了确保算法能够获得正确的调用，本平台提供界面配置各种参数，包括开放式优化参数，使得构建生成的模型具有可用性。

1.3 数据限定

本平台为了通用性为每种算法的输入数据进行了规范化，需要用户根据指定的格式上传数据，不过不需要用户进行特征处理，只是对数据项之间的排列和文件格式进行了限定，具体如下：

算法	数据类别	数据用途	数据格式	数据形式
回归	数值	训练	第一列为目标值，其他列为特征值，以','分割	csv/txt 文件
		批量预测	第一列为实例编号，其他列为特征值，以','分割	csv/txt 文件
		单例预测	各列均为特征值，以','分割	页面文本框输入
分类	数值	训练	第一列为类别标签，其他列为特征值，以','分割	csv/txt 文件
		批量预测	第一列为实例编号，其他列为特征值，以','分割	csv/txt 文件
		单例预测	各列均为特征值，以','分割	页面文本框输入
	文本	训练	第一列为类别标签，第二列为文本，以','分割	csv/txt 文件
		批量预测	第一列为实例编号，第二列为文本，以','分割	csv/txt 文件
		单例预测	单独的一列文本	页面文本框输入
聚类	数值	训练	第一列为实例编号，其他列为特征值，以','分割	csv/txt 文件
		批量预测	第一列为实例编号，其他列为特征值，以','分割	csv/txt 文件
		单例预测	各列均为特征值，以','分割	页面文本框输入
	文本	训练	第一列为编号，第二列为文本，以','分割	csv/txt 文件
		批量预测	第一列为编号，第二列为文本，以','分割	csv/txt 文件
		单例预测	一行文本，编号^文本，以'^'连接	页面文本框输入

第2章 机器学习平台操作

2.1 登录

打开本平台，首先展现登录页面，如下图所示：

用户：

密码：

☐ 保持登录状态

登录

在此页面上输入用户名和密码，点击登录，如果存在此用户名且密码正确，即可进入本平台系统。因为本平台未来定向为商业化，用户名和密码由管理员创建，不设置用户注册页面。

2.2 页面总体框架

本平台页面布局清晰，以构建机器学习应用为主线，其整体页面框架如下：

SaaS机器学习平台 >

左侧为 菜单列表

- 上传数据
- 构建模型
- 单例预测
- 批量预测

右侧为具体页面

数据属性设置：

数据集名：

算法类别：☒ 回归/分类 ☐ 聚类

数据类别：☒ 数值 ☐ 文本

数据用途：☒ 训练 ☐ 预测

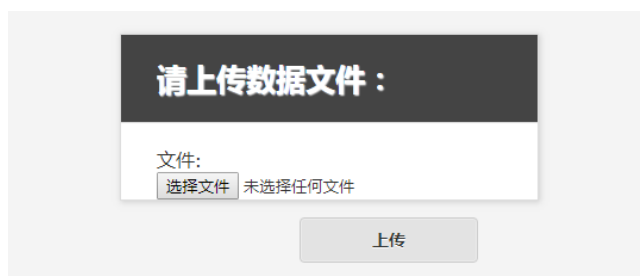
提交

2.3 上传数据

如上图所示，上传数据首先需要设置各项参数，为用户构建模型匹配训练数据创造条件。各项参数均为必填项。

数据集名：为上传的数据确定一个独一无二的名称，以后用作构建模型的一个参数

选择其他各项之后，点击提交，展现上传文件的页面：



如果文件上传成功，将在页面给出反馈结果：

```
{
  "datasetName": "dataset01",
  "status": 0,
  "time": 3.8995485305786133
}
```

status 项，0：表示上传成功，-1：表示失败。

2.4 构建模型

点击构建模型菜单项，根据各数据项提示设置各项内容，如下图所示：



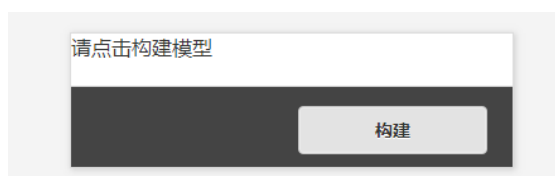
设定模型名称：为构建的模型确立一个独一无二的名称，为以后使用它进行预测创建条件。

数据集名：为刚才通过上传数据功能上传至此平台的数据集文件。

然后选择算法数据类别、具体算法，并以开放项允许用户进行各种参数的设置。

各项设置完成后，点击提交，进入模型构建页面：

首先提示由用户点击来进行模型构建：



点击上图的构建按钮，则正式进行模型构建，用时根据数据集大小和算法复杂度而有所不同，在 Spark 分布式集群环境下，一般情况下，处理速度会比较快，其构建过程的页面如下：



2.5 单例预测

点击单例预测菜单项，根据各数据项提示设置各项内容，如下图所示：

选择模型：填写刚才构建的模型名称。

选择算法类别和数据类型，其中分类包含回归和分类。然后在数据实例文本框填写具体数据内容。

点击提交进行预测，将在页面返回预测结果，如下图：

```
{
  "Prediction": "1.54",
  "prediction time": 41.029993772506714,
  "status": 0
}
```

Prediction: 预测结果

Prediction time: 预测所用时间

Status: 预测状态，0：表示成功，-1：表示失败。

2.6 批量预测

进行批量预测首先需要再次点击上传数据菜单项，如下图所示：

此时同样需要设置数据集名，并设置数据用途为预测，点击提交上传预测数据集文件。

文件上传成功后，点击批量预测菜单项，进行各项条件设置，如下图所示：

上传数据

构建模型

单例预测

批量预测

菜单列表

设置预测条件：

模型名称：model01

数据集名：dataset02

预测结果名：predict01

提交

设置具体条件

模型名称：为刚才构建的机器学习模型

数据集名：为上面步骤上传的预测数据集名

预测结果名：为临时保存预测结果的文件名，可任意设置且独一无二，点击提交，进入预测页面，如下图所示，点击预测按钮进行预测：

预测正在进行中，请等待。。。

预测

如果预测成功，则在页面展示预测结果列表，如下图所示：

预测结果列表：

实例	结果
1	1.5373
2	1.5936
3	2.9665