

4.2 机器学习中的分类问题（下）

-机器学习服务操作指导

1 任务介绍

本次任务将介绍如何使用逻辑回归和SVM进行分类

2 任务执行

2.1 数据上传

如果已经学习过第三天的课程《3 机器学习中的分类问题(上)》，则无需进行数据上传，否则请按照《3.2 机器学习中的分类问题(上)_操作指导》中的数据上传指导进行操作。

2.2 创建项目

如果项目已经在第3天的课程当中创建过了，则无需创建。否则在MLS实例主页上单击“创建项目”，并写入项目名称，导入案例无需选择，完成后单击“确定”。



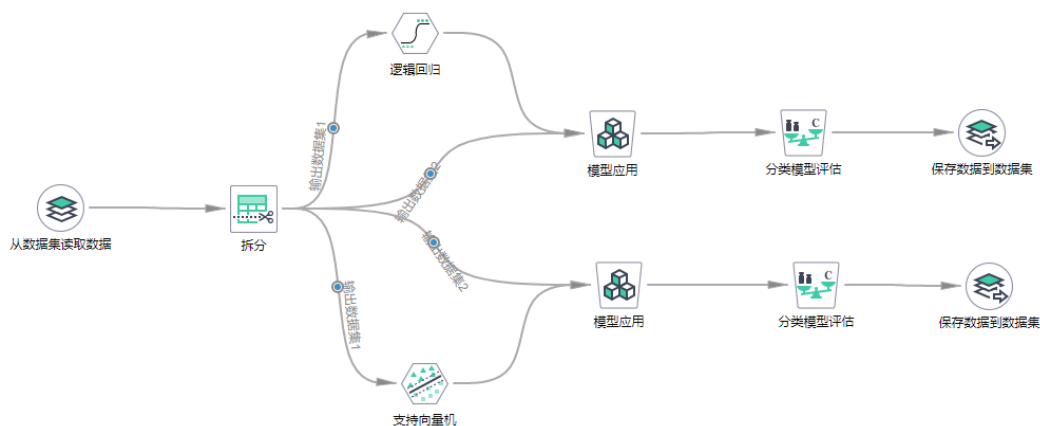
2.3 创建工作流

MLS实例主页单击“项目”—单击2.2中创建的项目名称----单击工作流-----单击“创建工作流”，工作流名称为“LRandSVM”



2.4 编辑工作流

单击“工作流”—单击2.3中创建的工作流名称----打开一个空的工作流，然后按照下图的方式进行编辑，所有的算子在工作流页面的左侧“节点库”中都可以找到。



其中：“拆分”与“逻辑回归”相连时选择“输出数据集1”；“拆分”与“支持向量机”相连时选择“输出数据集1”；“拆分”与两个“模型应用”相连时选择“输出数据集2”；

每个节点的配置如下：

1) “从数据集读取数据”：数据文件地址：/classification/new_bank.csv

元数据文件地址：/classification/bank_formal.desc

从数据集读取数据

* 数据格式:

CSV

* 数据文件:

/classification/new_ba

...

导入元数据:



* 元数据文件:

/classification/bank_fc

...

- 2) “拆分”：默认设置
- 3) “支持向量机”：默认设置
- 4) 两个“模型应用”：预测类型：分类
- 5) “逻辑回归”:

逻辑回归

* 正则化函数:

L2

* 正则化参数:

0.001

* 最大迭代次数:

100

- 6) 与“逻辑回归”分支相连的“保存数据到数据集”:

保存数据到数据集

* 文件路径:

* 文件名:

* 文件格式:

* 字段分隔符:

允许覆盖:



7) 与“支持向量机”分支相连的“保存数据到数据集”：

保存数据到数据集

* 文件路径:

* 文件名:

* 文件格式:

* 字段分隔符:

允许覆盖:



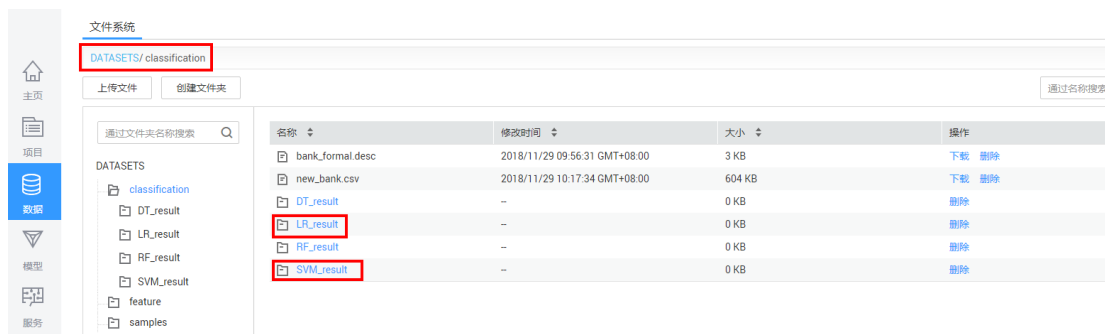
2.5 运行 workflow

1) 单击  运行 workflow。

在下方的运行日志查看运行结果。

运行日志

2) workflows运行完毕后，可以在“主页”—“数据”当中找到两个结果文件，进行查看



文件系统

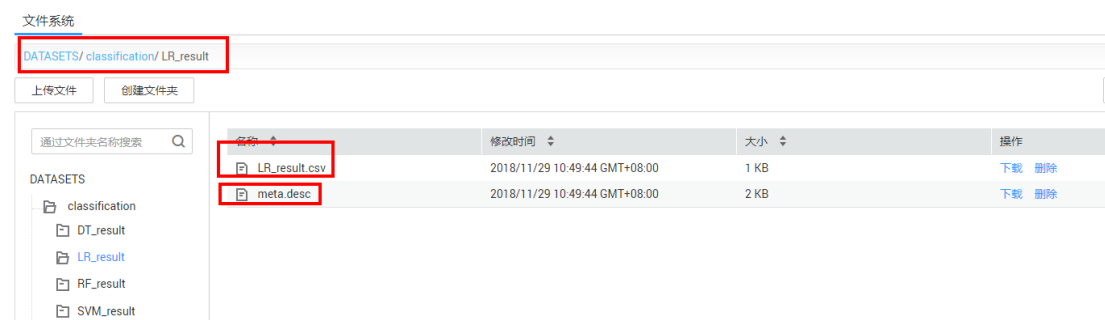
DATASETS/classification

上传文件 创建文件夹

通过名称搜索

| 名称 | 修改时间 | 大小 | 操作 |
|------------------|-------------------------------|--------|-------|
| bank_formal.desc | 2018/11/29 09:56:31 GMT+08:00 | 3 KB | 下载 删除 |
| new_bank.csv | 2018/11/29 10:17:34 GMT+08:00 | 604 KB | 下载 删除 |
| DT_result | - | 0 KB | 删除 |
| LR_result | - | 0 KB | 删除 |
| RF_result | - | 0 KB | 删除 |
| SVM_result | - | 0 KB | 删除 |

3) 比如查看“逻辑回归”的模型评估结果



文件系统

DATASETS/classification/LR_result

上传文件 创建文件夹

通过名称搜索

| 名称 | 修改时间 | 大小 | 操作 |
|---------------|-------------------------------|------|-------|
| LR_result.csv | 2018/11/29 10:49:44 GMT+08:00 | 1 KB | 下载 删除 |
| meta.desc | 2018/11/29 10:49:44 GMT+08:00 | 2 KB | 下载 删除 |

先单击meta.desc查看每一列的意义，再单击LR_result.csv查看结果。随机森林分类的结果同理。分类问题一般查看的是第一列“阈值”=0.5的那一行的结果。

数据预览

```
0.9, (0.9-1.0], 0, 0, 0.5993, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
0.8, (0.8-0.9], 0, 0, 0.5993, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
0.7, (0.7-0.8], 48, 18, 0.606, 0.0514, 0.7273, 0.0267, 0.0067, 0.0267, 1.0E-4
0.6, (0.6-0.7], 214, 110, 0.6291, 0.2392, 0.6718, 0.1455, 0.0475, 0.1455, 0.0036
0.5, (0.5-0.6], 381, 306, 0.6458, 0.4468, 0.597, 0.357, 0.1611, 0.357, 0.0321
0.4, (0.4-0.5], 446, 572, 0.6178, 0.559, 0.5198, 0.6047, 0.3734, 0.6047, 0.1342
0.3, (0.3-0.4], 391, 730, 0.5424, 0.59, 0.4602, 0.8218, 0.6444, 0.8218, 0.3275
0.2, (0.2-0.3], 282, 799, 0.4274, 0.5779, 0.4101, 0.9783, 0.941, 0.9783, 0.5945
0.1, (0.1-0.2], 39, 159, 0.4007, 0.5721, 0.4007, 1.0, 1.0, 1.0, 0.6529
0.0, [0.0-0.1], 0, 0, 0.4007, 0.5721, 0.4007, 1.0, 1.0, 1.0, 0.6529
```

在分类概率阈值为0.5时，LR的准确率为0.6458。

3 打卡任务

3.1 完成单元测试

3.2 任务截图

1、在2.4 workflow 界面进行截图：

1) 右上角为用户名、下方为“工作流运行成功”

2) 工作流与图示相同

