Spark Mllib

刘磊

2020 年 12 月

在 StumbleUpon Evergreen 数据集上使用 Spark ML Pipeline 建立二元分类模型，用决策树预测网页是暂时的还是长青的，并使用训练验证与交叉验证找出最佳模型，提高预测准确率，最后还将介绍如何使用随机森林分类算法进一步提高准确率。

StumbleUpon 是一个个性化推荐引擎，根据用户的兴趣行为给用户推荐网页，而有些网页内容是即时性（ephemeral）的，比如新闻股票网页（用户短暂感兴趣），有些网页是长久性的（evergreen）如体育，理财等（用户持续感兴趣）。现要分辨网页是 ephemeral 的还是 evergreen 的，以便向用户推荐更加准确的网页。

下载地址：https://www.kaggle.com/c/stumbleupon/data

准备工作

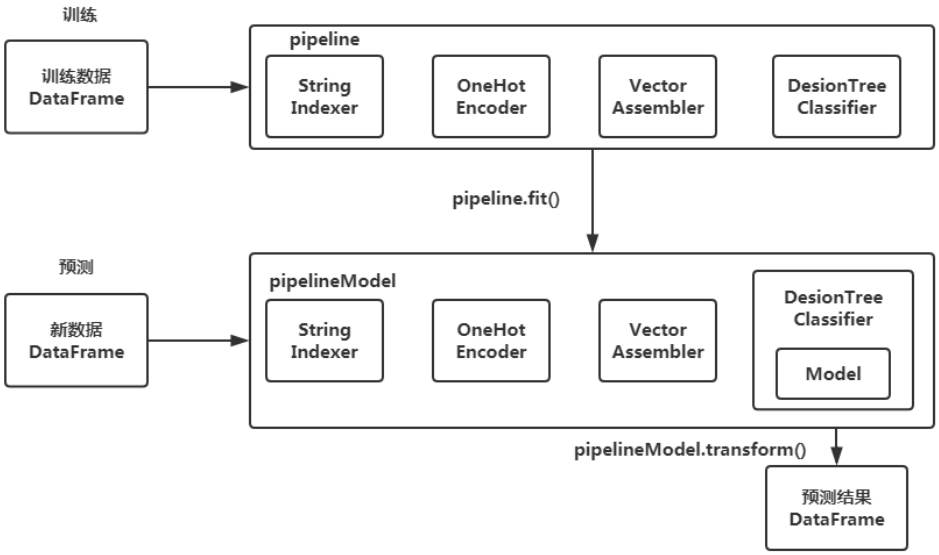
这次实验在 Jupyter 中完成，如下图所示将 PYSPARK_DRIVER_PYTHON 改为 jupyter。

```
# pyspark
export PYSPARK_DRIVER_PYTHON=jupyter
#export PYSPARK_DRIVER_PYTHON=python3
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
export PYSPARK_PYTHON=python3
export PYTHONPATH=$SPARK_HOME/python/:$SPARK_HOME/python/lib/py4j-0.10.7-src.zip
:$PYTHONPATH
```

保存退出，并使设置生效。

流程原理

Spark 机器学习工作流程（ML Pipeline）的原理就是将机器学习的每一个阶段（例如数据处理、进行训练与测试、建立 Pipeline 流程）形成机器学习工作流程，流程图如下：



（1）建立机器学习流程 pipeline：包含 4 个阶段，前 3 个是数据处理，第 4 阶段是 DesionTreeClassifier 机器学习分类算法

- StringIndexer：将文字的分类特征转化为数字
- OneHotEncoder：将一个数字的分类特征字段转为多个字段
- VectorAssembler：将所有的特征字段整合成一个 Vector 字段
- DesionTreeClassifier：进行训练并且产生模型

（2）训练：“训练数据 DataFrame”使用 pipeline.fit()进行训练。系统会按照顺序执行每一个阶段，最后产生 pipelineModel 模型。

（3）预测："新数据 DataFrame"使用 pipelineModel.trainsform()进行预测。系统会按照顺序执行每一个阶段，并使用 DecisionTree Classifier Model 进行预测。预测完成后，会产生“预测结果 DataFrame”。

启动 Hadoop

```
/apps/hadoop/sbin/start-all.sh
```

启动 Spark

```
/apps/spark/sbin/start-all.sh
```

上传数据

在 HDFS 上创建文件夹/input/mllib，并将/data 目录下的 train.tsv 文件和 test.tsv 文件

上传

```
hadoop fs -mkdir /input/mllib
hadoop fs -put /data/train.tsv /input/mllib/
hadoop fs -put /data/test.tsv /input/mllib/
```

确认是否上次成功

```
hadoop fs -ls /input/mllib
```

启动 Pyspark

创建目录~/pyspark-workspace/mllib/，在该目录中启动 Pyspark，创建一个

Notebook。