**《时空数据处理与组织课程实习》**

**实验 5实习报告**

**学 院: 遥感信息工程学院**

**班 级: 20F10**

**学 号: 2020302131249**

**姓 名: 马文卓**

**实习地点: 教学实验大楼101**

**指导教师: 李晓雷**

**2022 年 5 月 28 日**

1. **实验目的**
2. 通过实验熟悉Spark ML的基本编程方法；
3. 熟悉用ML解决一些常见的数据分析问题，包括数据导入、分类和预测等。
4. **实验环境**

基于Windows的Spark环境，采用Python编程。

数据集：下载Adult数据集(http://archive.ics.uci.edu/ml/datasets/

Adult)。数据从美国1994年人口普查数据库抽取而来，可用来预测居民收入是否超过50K$/year。该数据集类变量为年收入是否超过50k$，属性变量包含年龄、工种、学历、职业、人种等重要信息。

1. **实验内容**
2. 从文件中导入数据，并转化为DataFrame。
3. 训练决策树模型，用于预测居民收入是否超过50K；
4. 对Test数据集进行验证，输出模型的准确率。
5. **步骤**

在开始创建决策树训练模型之前，首先需要做的任然是导入实验所需要的第三方库，以构建好训练环境。这里除开spark基本的一些第三方库如pyspark之外我们还导入了其中的pyspark.sql以及pyspark.ml以使用sql语句和机器学习的相关内容。（在最开始不要忘记findspark.init()初始化spark环境）



1. **数据读入和预处理**

本步骤的目的是将训练集（training\_data.txt）和测试集（test\_data.txt）中的数据读入，转化为我们需要的DataFrame格式。由于数据的内部组织方式不一定和我们的需求一致，因此需要进行数据的预处理。具体操作如下：

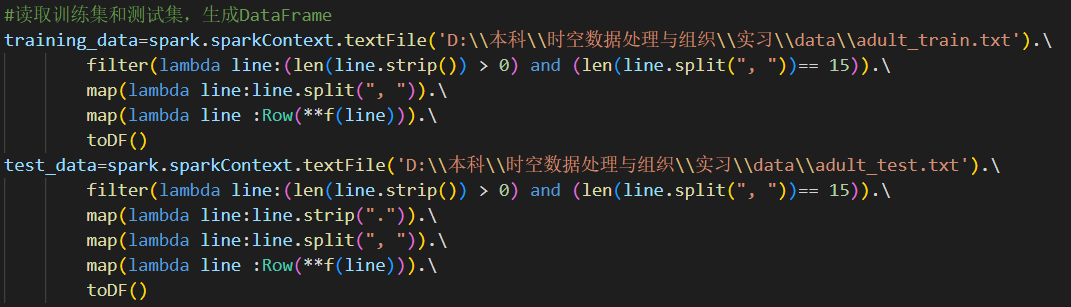
首先是使用【spark.sparkContext.textFile】函数将txt数据读入形成rdd对象（与之前的实验一致，在此就不在赘述）。然后使用filter函数去除异常数据，限定条件就是【去除空格之后长度不为0】&&【按照‘逗号空格’为分隔符进行分隔后的元素个数为15个】。然后使用map函数进行映射，以‘逗号空格’为分隔符进行分割每一行，然后使用map对每一行数据使用自定义的数据预处理函数f，最后使用toDF函数转为DataFrame格式。

测试集读取和处理方式和训练集类似，但是需要在按分割符分割之前去除行末的‘.’（**可以仔细观察测试集数据，其每行末都有一个点，如果不去除会导致精度评定报错**）

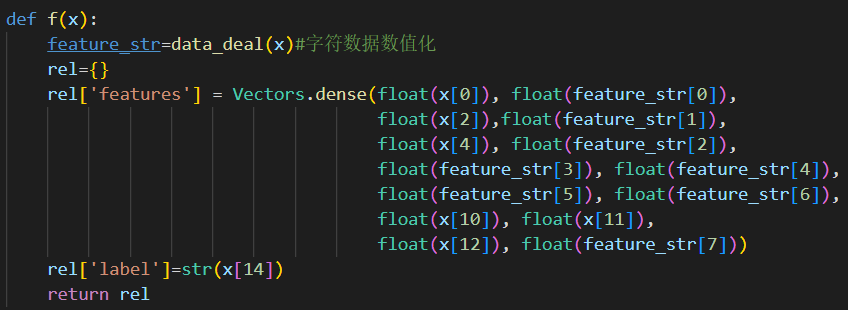
\*需要注意的是：

①异常数据必须去除，否者后面运行会报错，因为异常数据格式不对

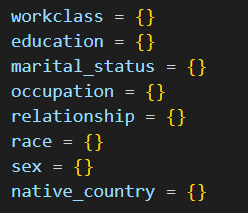
②仔细观察数据可以发现分割符是逗号加空格，仅按照逗号为分割符会导致 后面的字符特征数值化报错



数据预处理函数f个构造如下图所示，首先调用自定义的data\_deal函数将每一行中的字符数据转为数值特征，以方便我们构建特征向量。然后定义字典rel，其包含了特征关键字和标签关键字。其中特征关键字对应的值即为前14列特征，将其全部转为float类型，其中的字符串特征都用转换后的对应数值特征代替。标签即为最后一列的数据。

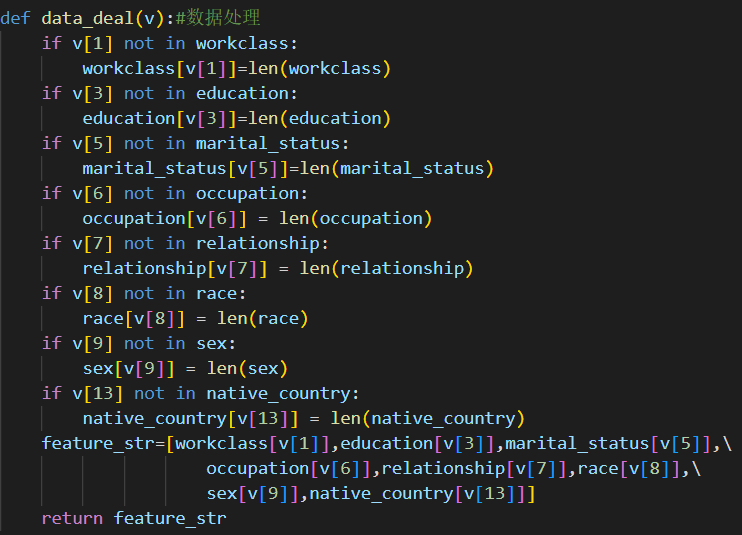


其中data\_deal函数构造如下图所示，主要功能就是给每一个字符类型的特征提供一个可以唯一标识其特征的数值型特征，并返回该数值特征列表。首先预先定义7个空字典，用来存储对应的字符特征。（由于14个特征中仅有7个字符特征）



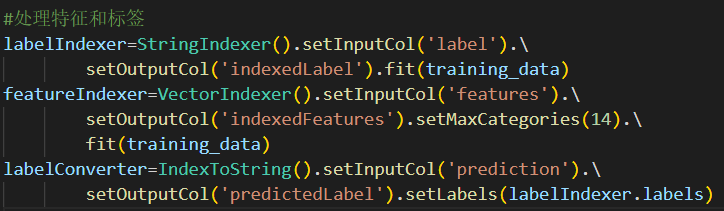
对于一行数据v而言，分别判断其7个字符特征是否已经在对应的字典中，如果不在，则在相应的字典中添加相应的关键字，并且其对应的值为当前字典的长度。这样转换之后，每个字典中的不同关键字就对应了可以唯一标识该关键字的值了。

最后将转换后的数字特征放到一个列表中，返回即可。



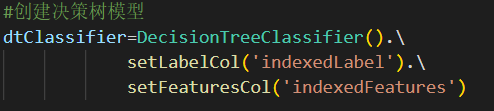
1. **处理特征与标签**

设置好训练特征和标签（这个步骤调用ml中的相关函数即可，就不过多赘述）



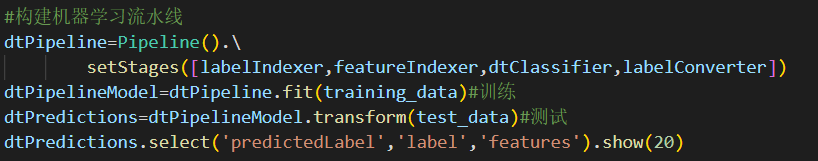
1. **创建决策树模型**

使用DecisionTreeClassifier创建决策树模型，设置好标签列和特征列。



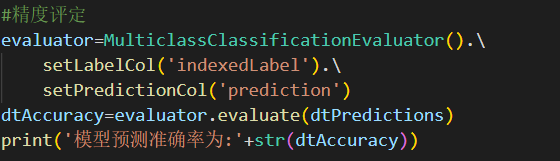
1. **构建机器学习流水线**

使用Pipeline构建流水线，把之前创建的标签、特征和预测都输入其中。然后使用fit方法进行训练，得到流水线模型，然后使用transform函数进行预测，最后使用select选择输出前20行。

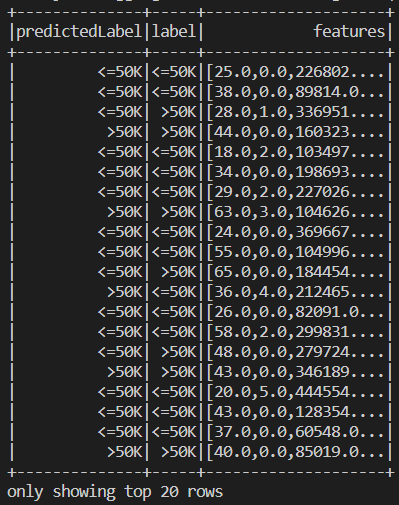


1. **精度评定**

使用MulticlassClassificationEvaluator函数创建进度评定对象，使用该对象的evaluate评定预测精度，并输出。



1. **结果**
2. **预测结果**



1. **预测精度**



1. **感悟**

本次实验熟悉了spark环境下的Ml编程实现机器学习的算法流程。最为直观的感受就是机器学习虽然有着比较复杂的运算逻辑，但是经过Ml库包装之后变成了非常结构化、简洁化、公式化的流程。事实上，只要我们将数据处理成标准格式，后面我们只需要按部就班地创建决策树模型，进行训练，进行预测，精度评定即可。

**遇到的问题：**

①数据导入的时候测试集行末有一个‘.’没有发现，导致后面 精度评定的时候出错，通过strip函数导入测试集去除即可。

②数据导入时分割符最开始认为是‘，’，但实际上是‘， ’，这会导致后面字符特征数值化的时候出错。