

（深圳）

实验报告

开课学期： 2021秋季

课程名称： 大数据导论

实验名称：数据理解、数据预处理及决策树的应用

实验性质： 设计型

实验学时： 2 地点：

学生班级：

学生学号：

学生姓名：

评阅教师：

报告成绩：

实验与创新实践教育中心制

2021年11月

# 实验目的

1. 学会理解数据并对数据进行预处理；

2. 理解决策树的原理并掌握其构建方法。

# 实验内容

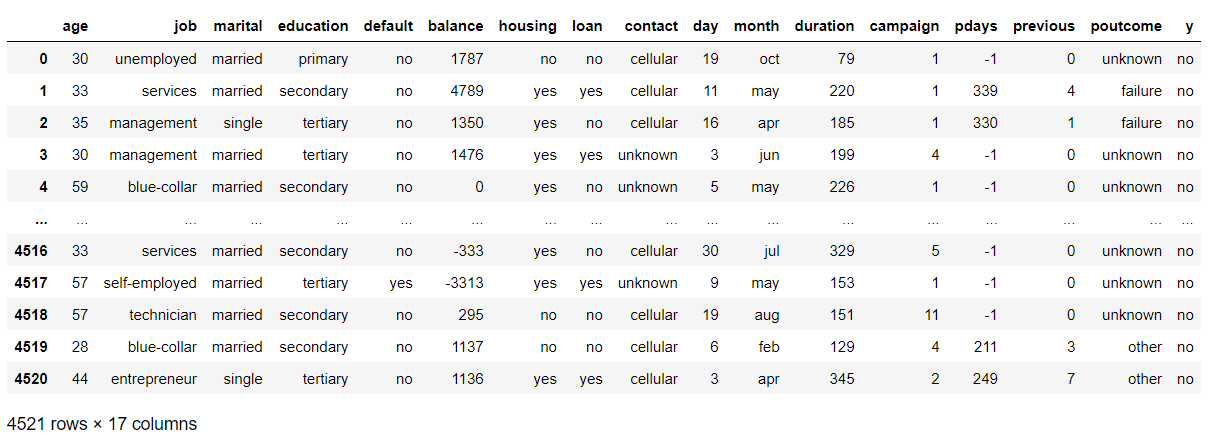
1．熟悉 Pandas 的安装和使用，并对数据进行预处理和相关分析；

2. 编写代码实现一种决策树算法，解决银行精准营销的分类问题。

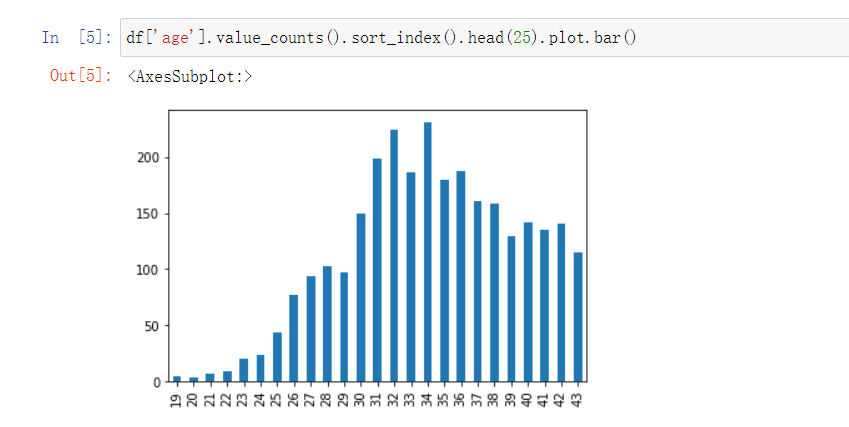
# 实验过程

**1.数据预处理**

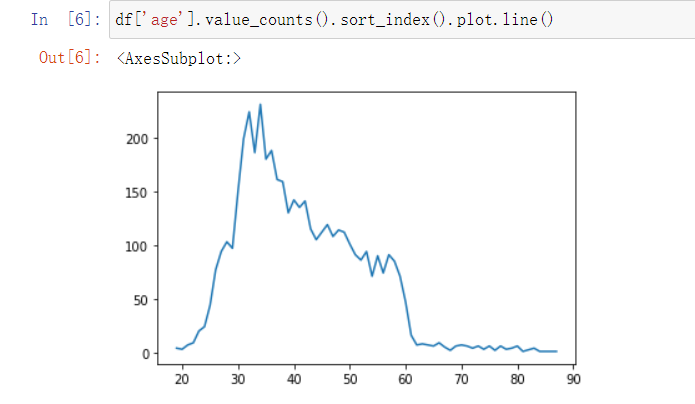
（1）首先读取数据集bank.csv，具体结构如下：



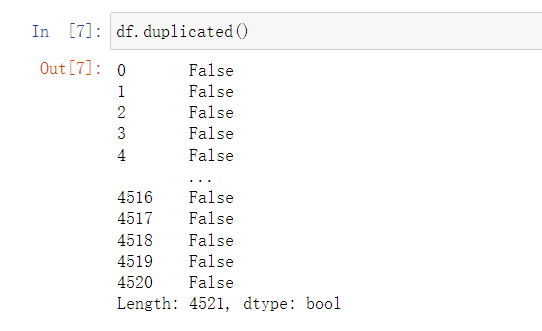
（2）对其中的’age’属性进行可视化，柱状图：



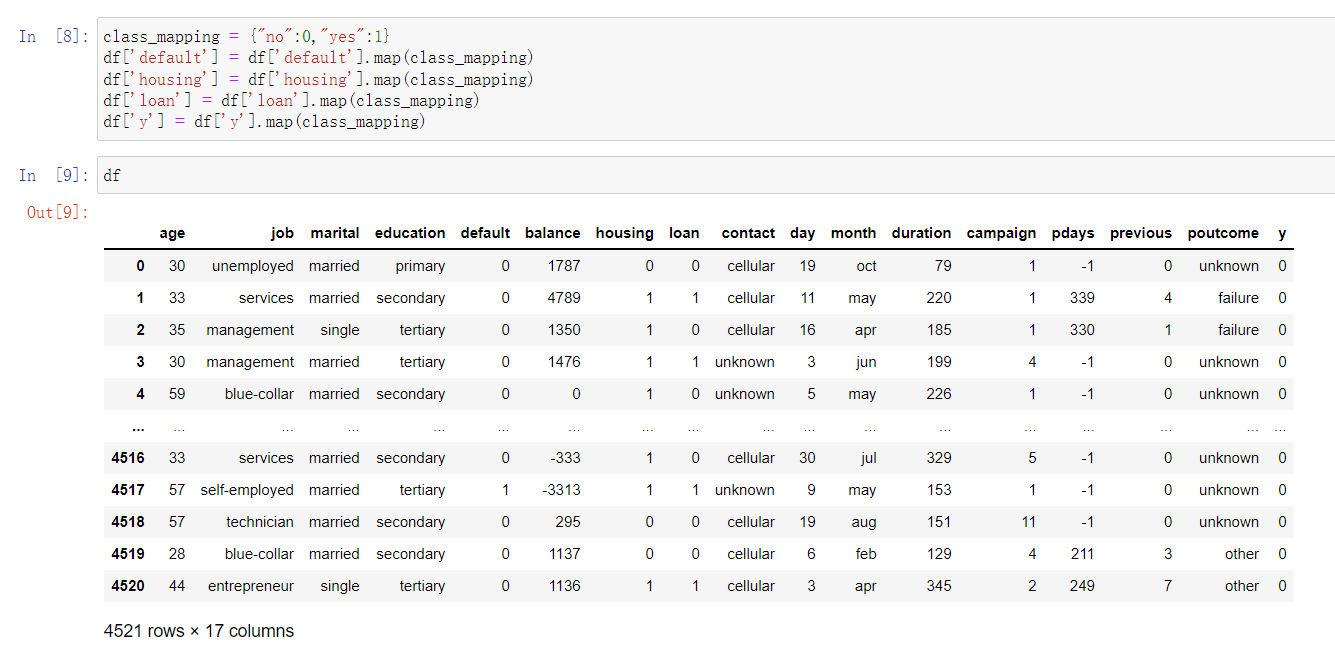
折线图：



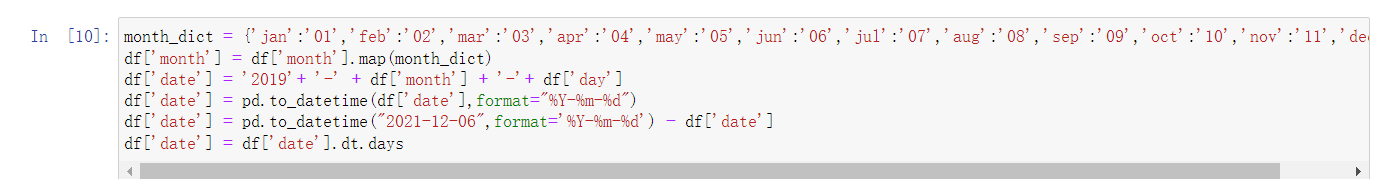
（3）筛选重复值并删除

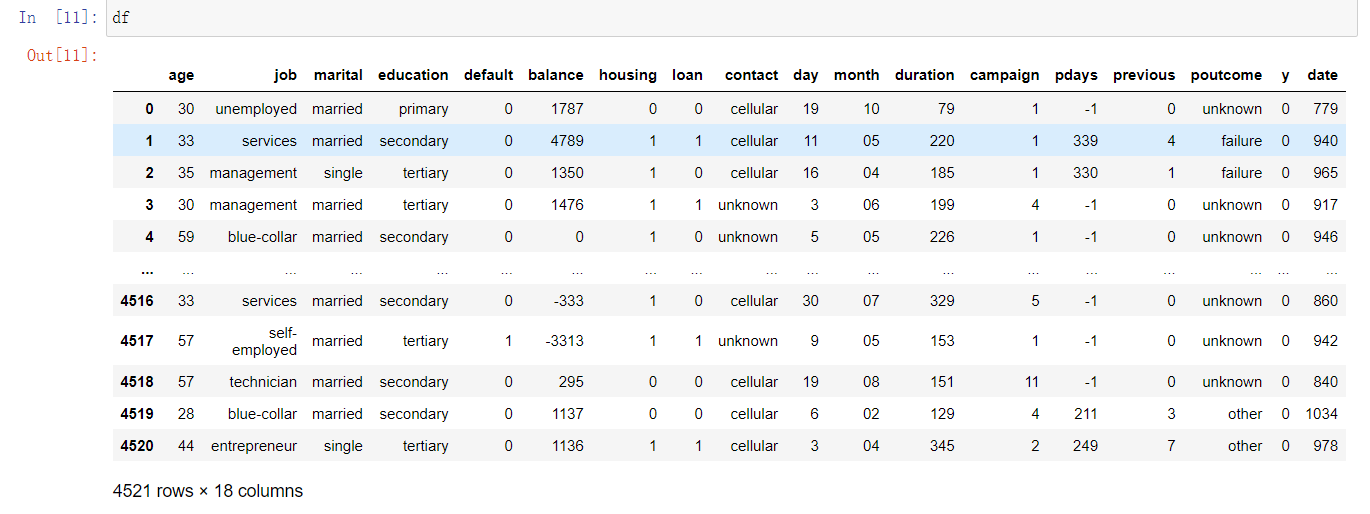


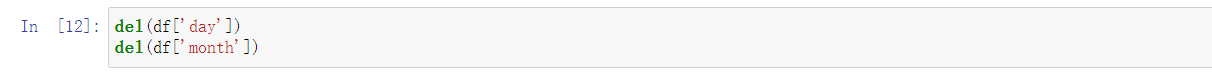
（4）将yes/no映射为0/1



（5）将月份和日期转化为距离现在的天数，并删除原有的月份和日期

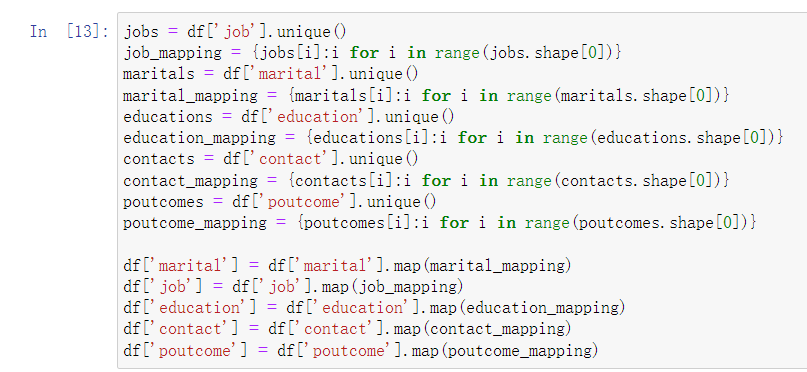


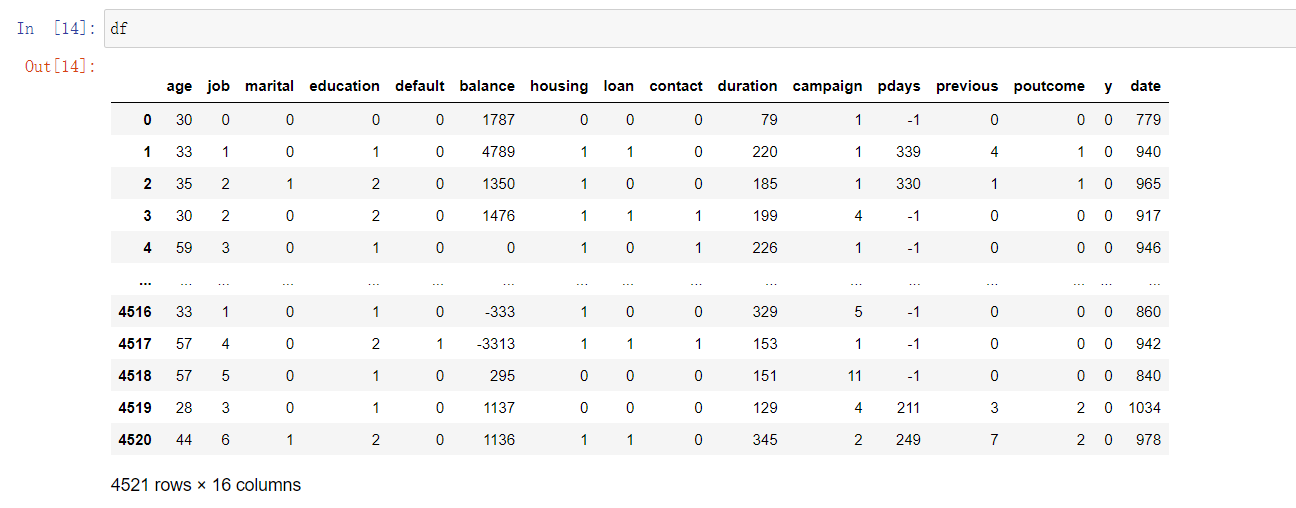




（6）将字符标签类变量转换为数值型，同时处理缺失值

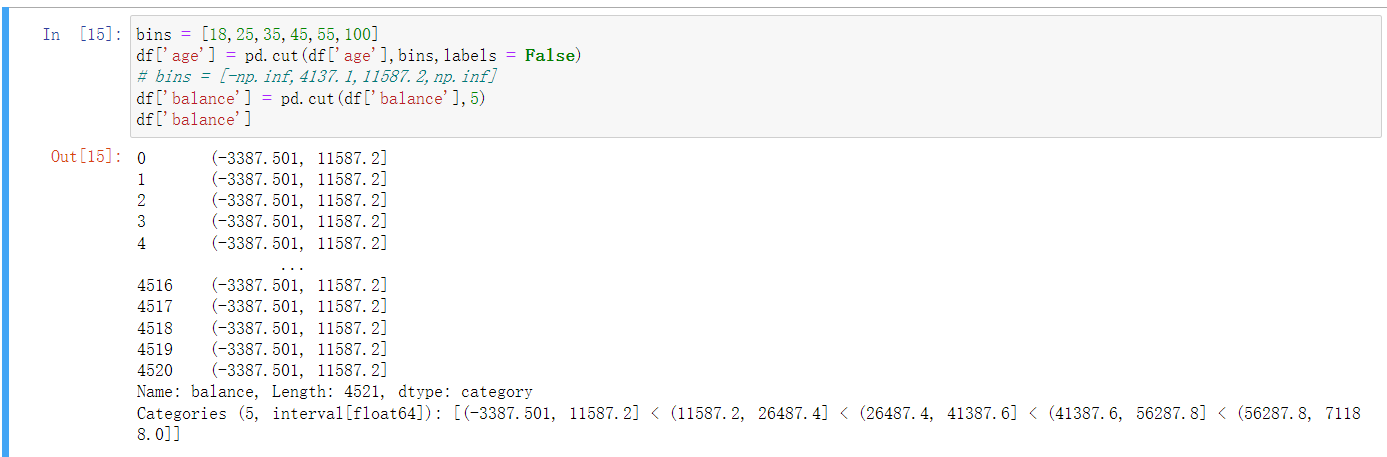
处理job,marital,education,contact,poutcome：

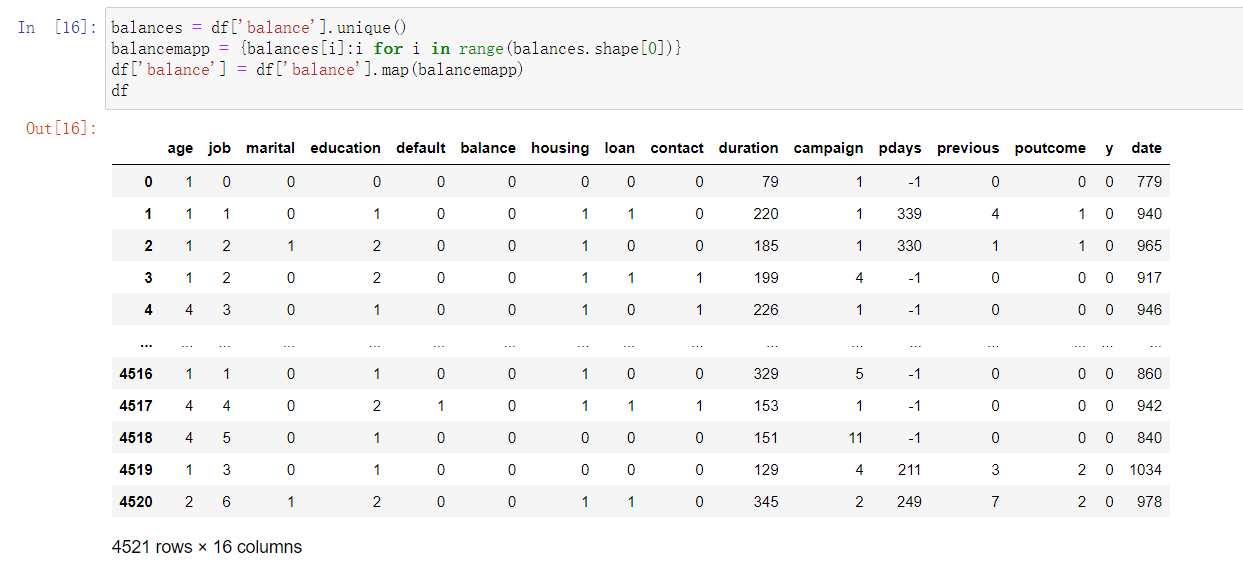




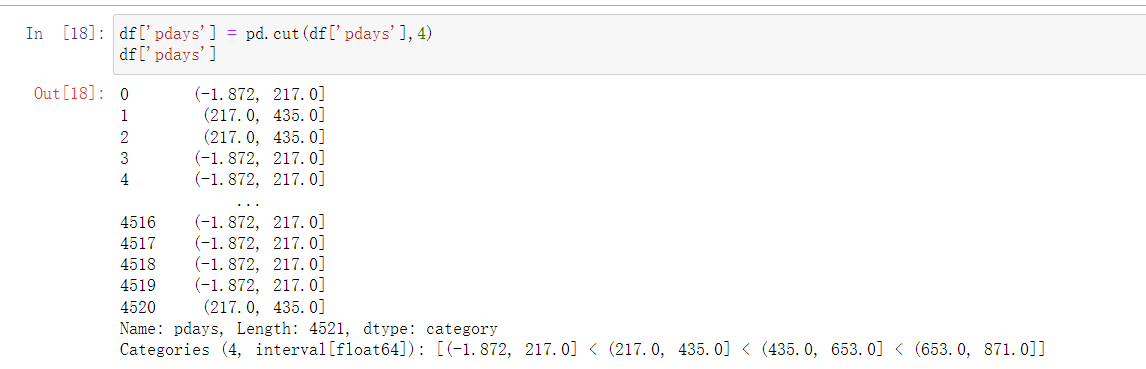
（7）采用分箱法将连续属性离散化

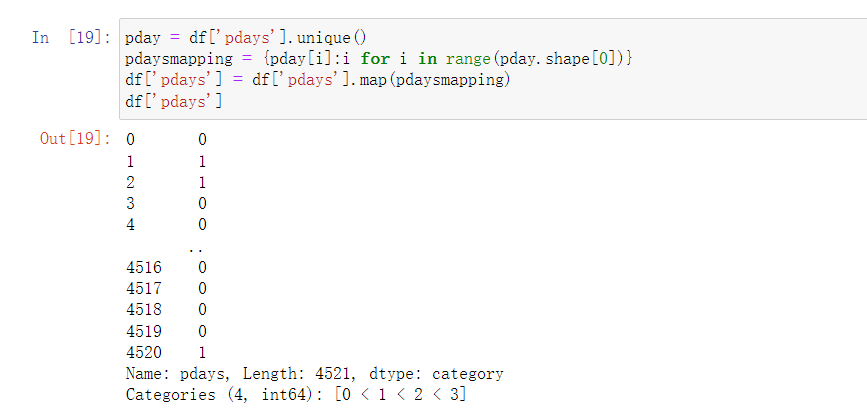
处理age和balance：



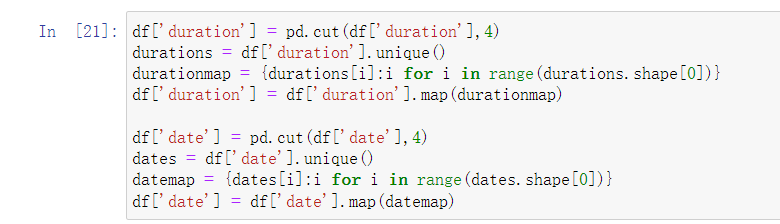


处理pdays：



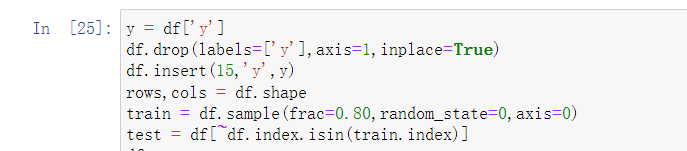


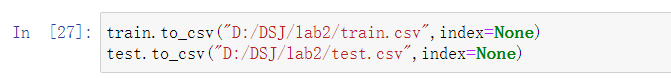
处理duration和date：



（8）划分训练集和测试集，保存文件

将原数据集按照4：1划分为训练集和测试集，保存到train.csv和test.csv中





**2.决策树构建**

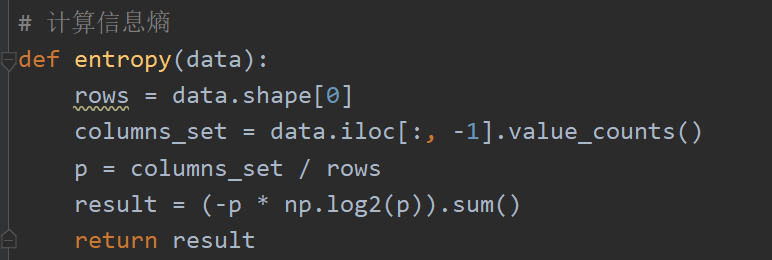
实现ID3算法的决策树，并在构建过程中完成预剪枝处理。

主要思路：递归构造决策树时，每次依次从当前剩余的所有属性中选择一个属性作为划分的属性，计算按照该属性划分后的信息熵，选择其中使得熵增最大的属性作为本次划分的属性，将数据集划分后即可递归的创建决策树。

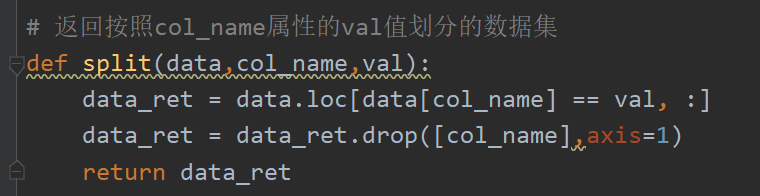
预剪枝：由于按照熵增最大的策略选择属性，若不进行剪枝处理，生成的决策树可能会又“高”又“胖”。故在构建时进行预剪枝，每次选择一个最好的属性时，若按照该属性进行划分后预测的精确度比没进行划分还要低，则进行剪枝处理。

实现：使用python实现预剪枝处理的ID3决策树，使用字典递归的构造决策树。主要函数如下：

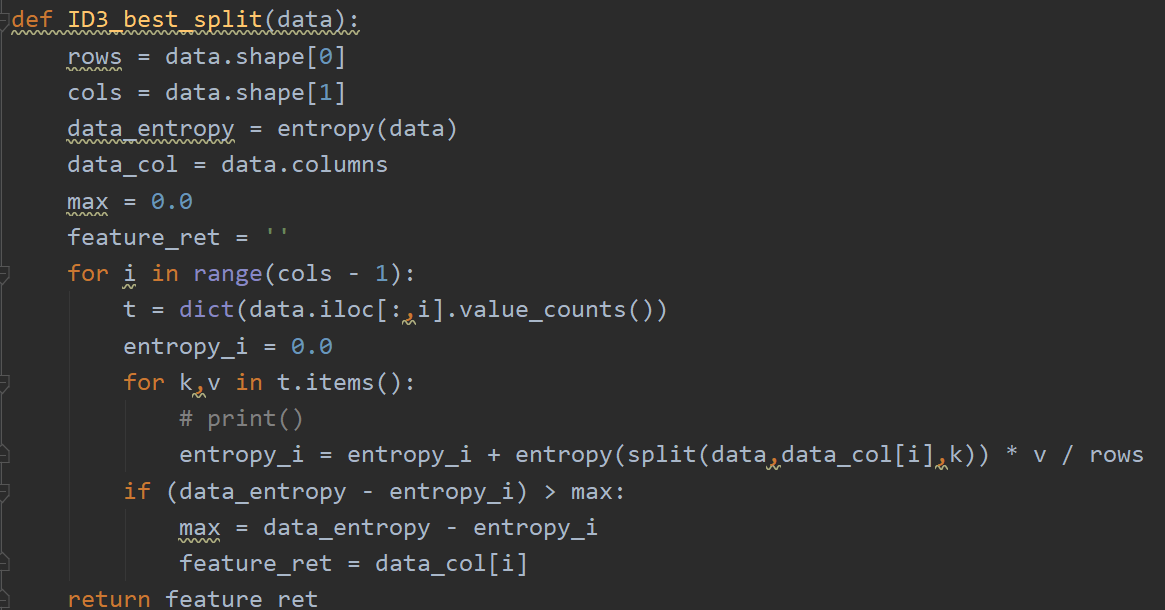
计算信息熵的函数entropy:



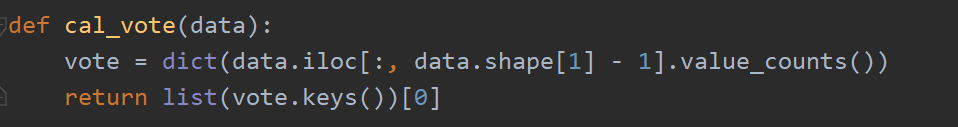
按照某一属性的某一属性值分裂并返回该数据集的函数split：



按照ID3算法，返回使熵增最大的划分的属性

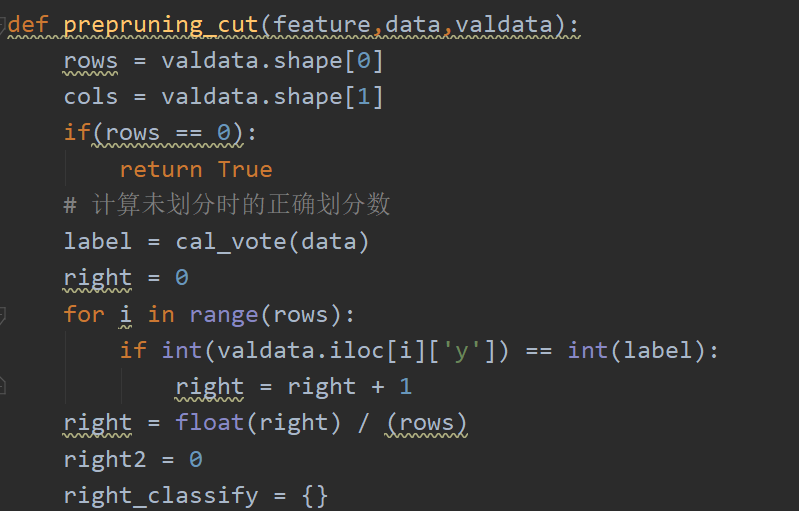


计算当前属性中，类别最多的函数

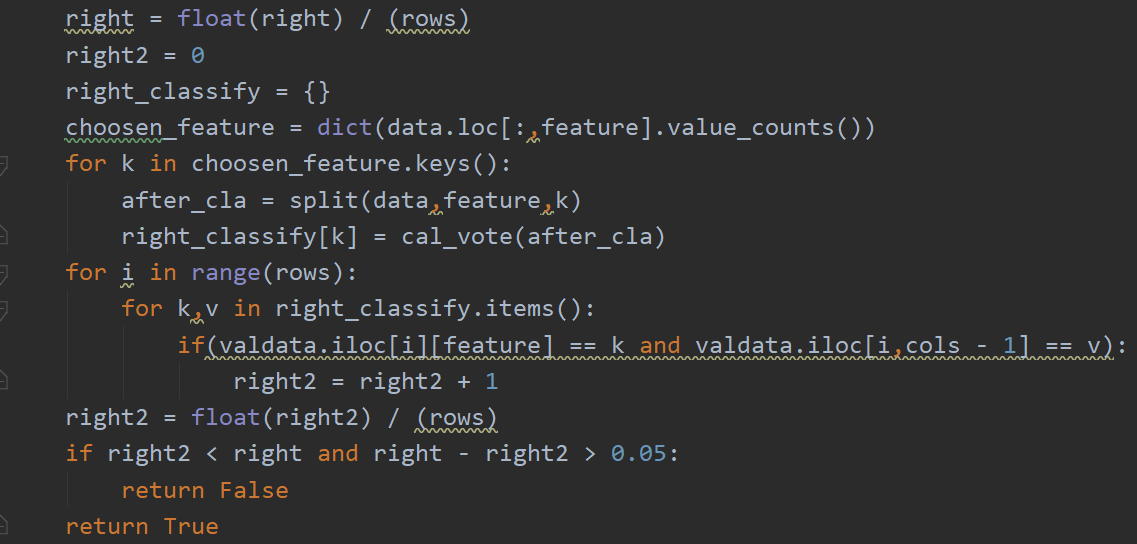


预剪枝处理函数：

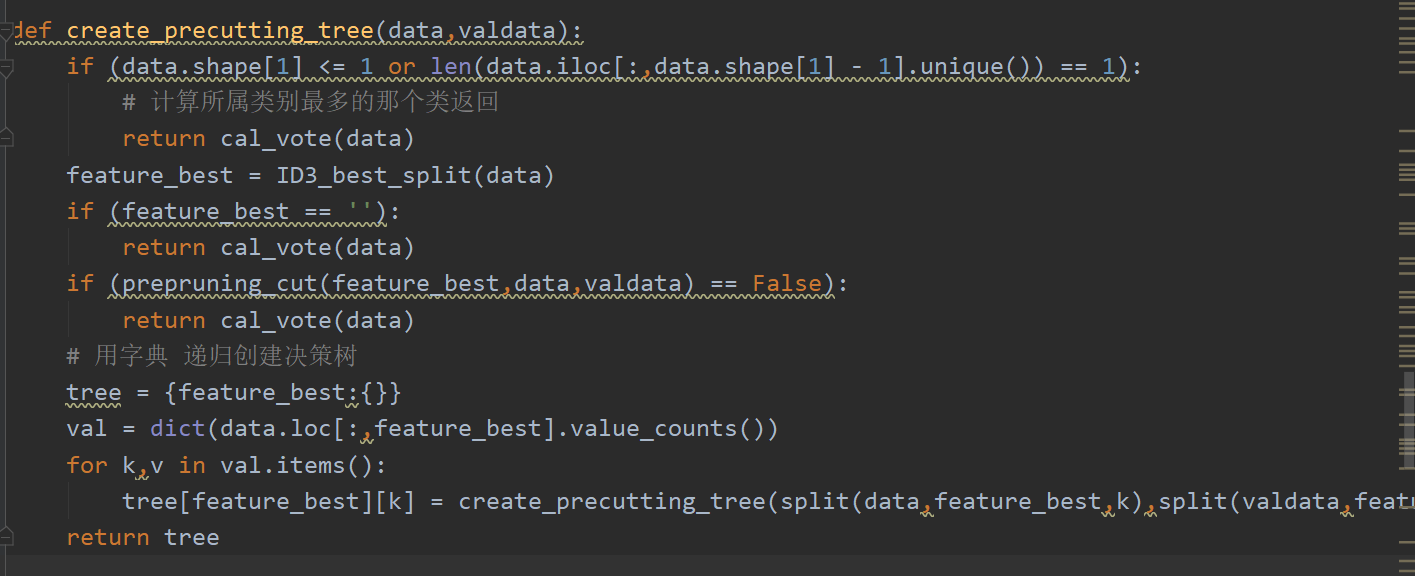
先计算当前预测的正确率



随后计算按照当前最好属性划分后预测的正确率，判断其大小关系从而决定是否剪枝。

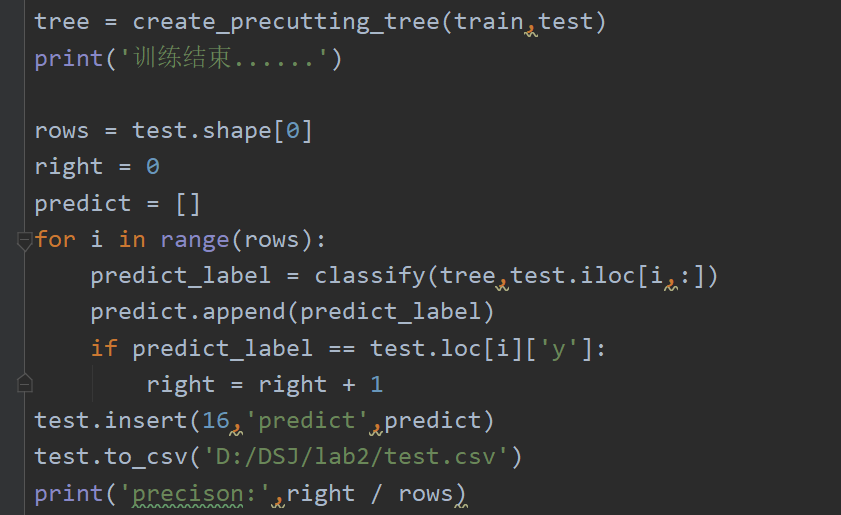


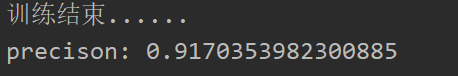
根据ID3和预剪枝策略，使用字典递归地构造决策树：



# 实验结果与分析

用测试集测试构造的决策树，得到其准确率:





精确度约为0.917

个人签名 ： 王铭

2021年 12月 9 日