

（深圳）

实验报告

开课学期： 2020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： T2608

学生班级： 2018级计算机5班

学生学号： 180110505

学生姓名： 胡聪

评阅教师：

报告成绩：

实验与创新实践教育中心制

2020年7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可视化分析；
2. 熟悉 sklearn 包，调用决策树模型对数据进行训练。

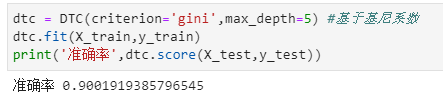
# 实验过程

1. 启动 jupyter
2. 安装 Pandas 库并熟悉其基本操作
3. 生成数据
4. 计算数据的基本信息
5. 选取特定列
6. 选取特定列行
7. 选取多行多列
8. 选取 B 列大于等于 5 的数据
9. 修改指定列
10. 数据读取及预处理
11. 安装 sklearn 并构建决策树
12. 参数调整
13. 可视化决策树模型

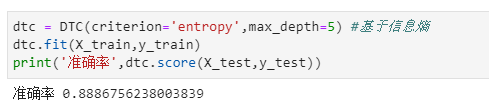
参数调整过程：

1. criterion：用以设置用信息熵还是基尼系数计算

基于基尼系数进行计算



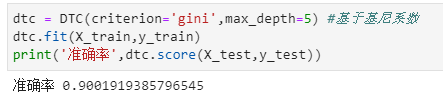
基于信息熵进行计算



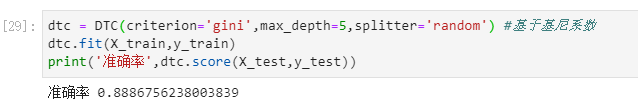
可以看到采用基尼系数进行计算，准确率更高

1. splitter：指定分支模式

默认模式为best，表示选择最优的分裂策略



改为random，表示选择最好的随机切分策略，发现准确率发生了下降



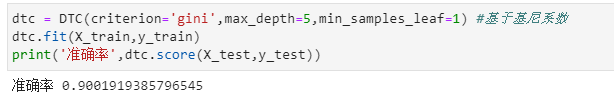
1. max\_depth：最大深度，防止过拟合



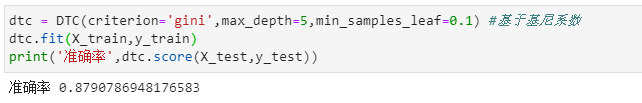
可以看到在depth为5的时候准确率最高

1. min\_samples\_leaf：限定每个节点分枝后子节点至少有多少个数据，否则就不分枝

整数：

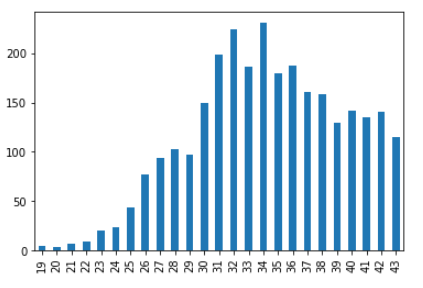


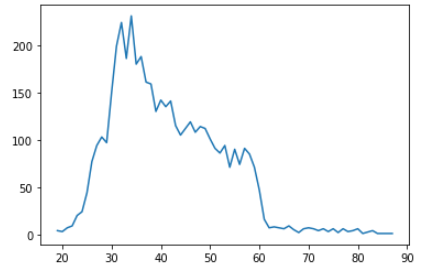
浮点数：



# 实验结果与分析

数据可视化





可视化决策树模型

图片包含 表格

描述已自动生成

个人签名 ：

2020年 12月 10日