数据挖掘

实验三：决策树——ID3算法

|  |  |
| --- | --- |
| 教 学 班： | 计算机科学与技术一班 |
| 学 号： | 19011402 |
| 姓 名： | 南佳霖 |
| 实验地点： | 理工楼913 |
| 指导教师： | 姜珊 |

2021

# 实验目的

掌握ID3算法的原理

# 实验要求

（1）掌握决策树相关的基本概念，理解信息增益

（2）熟练使用Python或其他工具实现ID3算法

# 实验器材

（1）计算机一台

（2）Python或其他编程工具

# 实验内容

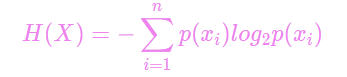
（1）选取下表为数据集：

下表显示了各种天气、温度、湿度和风速的场合下，是否进行打垒球的情况。

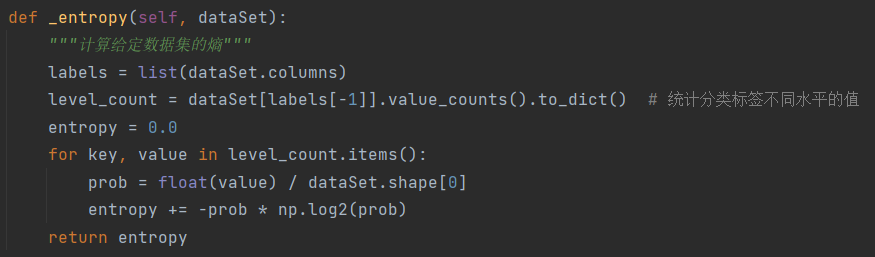
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **天气** | **温度** | **湿度** | **风速** | **活动** |
| 晴 | 炎热 | 高 | 弱 | 取消 |
| 晴 | 炎热 | 高 | 强 | 取消 |
| 阴 | 炎热 | 高 | 弱 | 进行 |
| 雨 | 适中 | 高 | 弱 | 进行 |
| 雨 | 寒冷 | 正常 | 弱 | 进行 |
| 雨 | 寒冷 | 正常 | 强 | 取消 |
| 阴 | 寒冷 | 正常 | 强 | 进行 |
| 晴 | 适中 | 高 | 弱 | 取消 |
| 晴 | 寒冷 | 正常 | 弱 | 进行 |
| 雨 | 适中 | 正常 | 弱 | 进行 |
| 晴 | 适中 | 正常 | 强 | 进行 |
| 阴 | 适中 | 高 | 强 | 进行 |
| 阴 | 炎热 | 正常 | 弱 | 进行 |
| 雨 | 适中 | 高 | 强 | 取消 |

**（2）简述熵（体现系统中的不确定性）的概念：**

信息熵是用来衡量样本纯度的指标，我们最终希望的也就是决策树中各分支结点的纯度尽可能的高。信息熵实际上就是信息量的期望



**计算熵的代码：**



**（3）简述信息增益的概念：**

**信息增益就是表示我们已知条件 X 后能得到信息 Y 的不确定性的减少程度。**

**信息增益越大，说明使用该种属性进行划分得到的纯度越大，效果越好。**

（4）以天气属性为例，计算信息增益：

未划分时，打垒球的熵为：

H(垒球|天气=晴)=－(9/14)\*log2(9/14)－(5/14)\*log2(5/14)=0.94

当天气为晴时，活动取消计数为3，活动进行计数为2，此时的条件熵为：

H(垒球|天气=晴)=－0.4\*log20.4－0.6\*log20.6=0.971

同理天气为阴时，条件熵为：

H(垒球|天气=阴)=－1\*log21=0

同理天气为雨时，条件熵为：

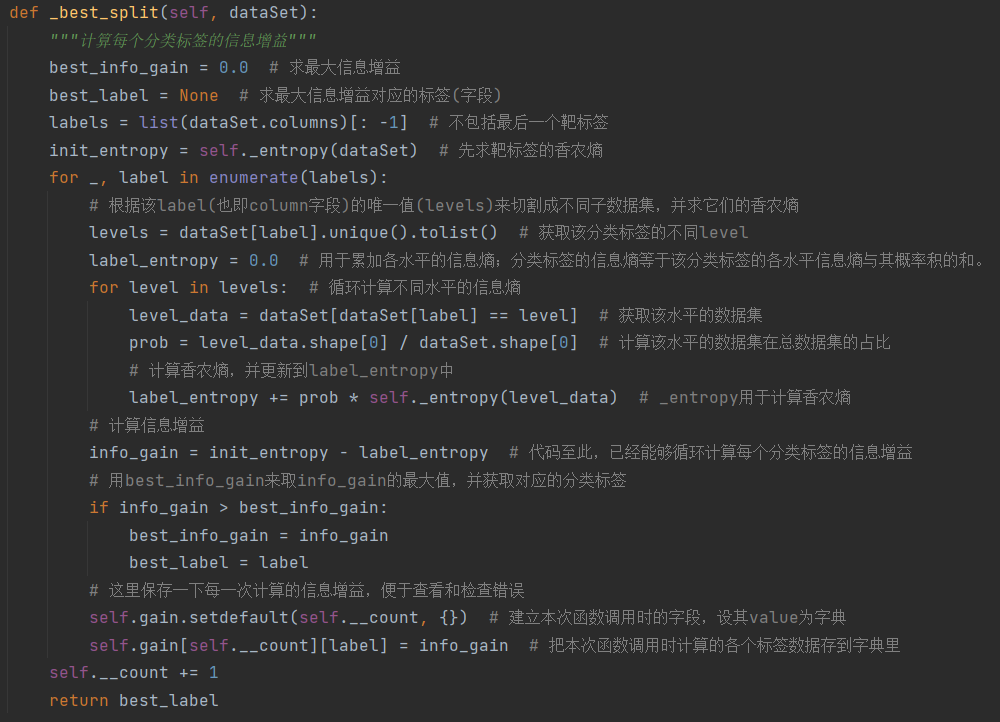
H(垒球|天气=阴)= －0.4\*log20.4－0.6\*log20.6=0.971

则按照天气属性不同取值划分时的带权平均值熵为：

H(垒球|天气)=(5/14)×H(垒球|天气=晴)＋(4/14)×H(垒球|天气=阴)＋(5/14)×H(垒球|天气=雨)=0.246

信息增益Gain=0.94-0.693=0.246

（5）计算各个属性的信息增益，并返回信息增益最大的属性，**代码：**



（6）递归处理：在天气为晴的前提下，继续在温度，湿度，风速这三个属性中找到信息增益最大的属性，以此类推。以湿度为例，计算信息增益：

未划分时，打垒球的熵为：

H(天气为晴时垒球)=－(2/5)\*log2(2/5)－(3/5)\*log2(3/5)=0.971

当天气为晴，湿度为高时，活动取消计数为3，活动进行计数为0，此时的条件熵为：

H(垒球|天气=晴，湿度=高)=－1\*log21=0

同理天气为晴，湿度为正常时，条件熵为：

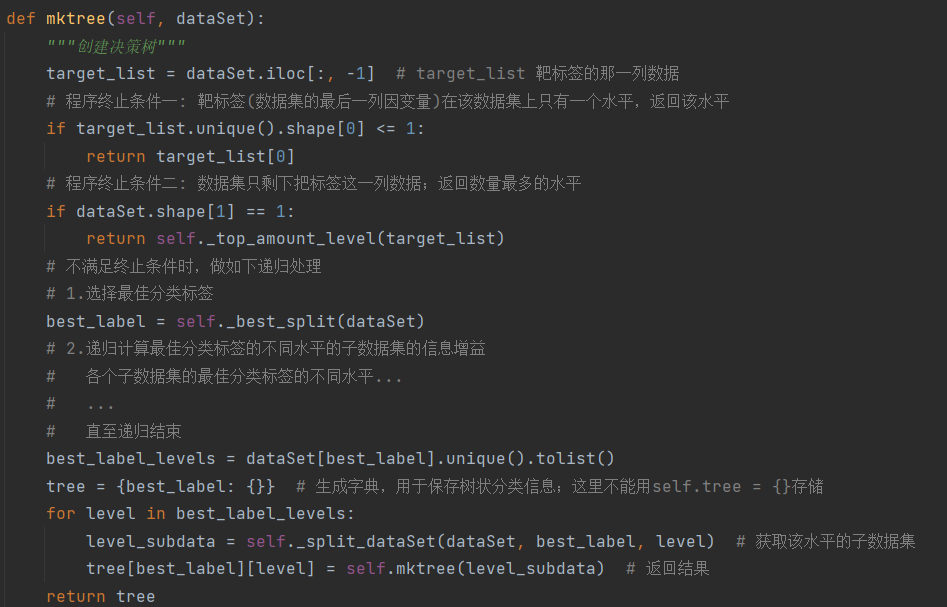
H(垒球|天气=阴，湿度=正常)=－1\*log21=0

则按照湿度属性不同取值划分时的带权平均值熵为：

H(垒球|天气，湿度)=(3/5)×H(垒球|天气=晴，湿度=高)＋(2/5)×H(垒球|天气=晴，湿度=正常) =0

信息增益Gain=0.971-0=0.971

（7）递归处理，构造决策树，**代码：**



（8）**最后得到的决策树：**



# 实验心得

ID3算法适用离散型数据，主要是根据信息增益来选择进行划分的特征，然后递归地构建决策树。它的主要优缺点如下：

优点：

计算复杂度不高，输出结果易于理解，对中间值的缺失不敏感，可以处理不相关特征数据

缺点：

没有剪枝，可能会产生过度匹配问题，需要进行剪枝

采用信息增益作为选择最优划分特征的标准，然而信息增益会偏向那些取值较多的特征