目 录

1. 信息论基础知识
2. 决策树基础算法
3. 小批量数据手推决策树
4. 上机实验
5. 算法总结
6. 信息论基本知识
7. 信息熵（Entropy）：

定义：如果X是一个离散型随机变量，取值空间为R，概率分布为

p(x)=P(X=x),(xR),那么，X的熵H（X）为：



实际意义为：信息的混乱程度，值越大则信息越混乱，不确定性越高，熵最大的时候，随机变量的值最不确定，最难预测。

例：计算一个骰子出现面的熵。



1. 条件熵（Conditional Entropy）

定义：给定随机变量C的情况下，随机变量X的熵



实际意义：大多时候我们希望系统的不确定性越低越好，这样我们大概率就可以预测结果，所以收集到的相关信息越多，随机事件的不确定性就越小。

例：在1中给定条件，已知输出结果为偶数，计算出现面的熵。

则此时的结果只有三种情况2,4,6



相比于1中当我们知道了，结果为偶数这个条件后，系统的熵降低。也即不确定性降低了。

1. 信息增益（Gain）

综合1、2我们把知道某个条件后，熵降低的程度定义为信息增益。



例：计算得知结果为偶数后的信息增益

Gain(结果为偶数) = =1

1. 信息增益率（Gain ratio）

定义：



之所以定义信息增益率是因为使用了该衡量标准会导致分类属性，优先选择那些具有比较多的分支，而这会导致过拟合。

1. Gini系数：

同样是一种做特征选择的方式，用来衡量信息的不纯度

计算方式如下：



其中D表示数据集全体样本，pi表示没种样本出现的概率。

例：取一种极端情况，6个面全为1的骰子。那么此时p1=1，则Gini=0

此刻信息最纯，混乱度最低。



1. 决策树的基础算法

决策树的构建是一个递归的过程，所以首先需要定义边界条件。

1. 当数据集中包含的记录都属于同一个类时，停止，返回类标签，作为一个叶子节点。
2. 当决策树的深度超过我们给定的最大深度时停止。
3. 当叶子节点上的记录数少于我们给定的最小记录数时，停止，若标签类别不唯一，进行多数投票表决。
4. 所有特征都处理完成了，但类别仍不唯一，停止，叶子节点标签，使用多数表决产生

算法流程：

获取数据getData（Path）

建树creatTree（data，label）

If 满足停止条件:

创建叶子节点Node

return Node（类别）

else:

Attr=getfeature（）#获取最佳的分类特征：熵、Gini、Gain

Attr\_class = getOutput()#获取Attr的所有可能输出值

for value in Attr：

dataSet = get（）取出data中输出为value的数据集

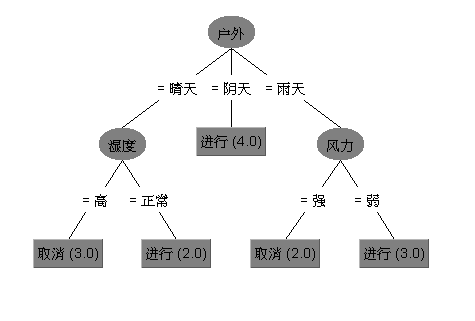
Child = creatTree（dataSet,updateLabel）

把child加入root中

1. 小批量数据集构建决策树

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | 户外 | 温度 | 湿度 | 风速 | 活动决定 |
| 1 | 晴天 | 炎热 | 高 | 弱 | 取消 |
| 2 | 晴天 | 炎热 | 高 | 强 | 取消 |
| 3 | 阴天 | 炎热 | 高 | 弱 | 进行 |
| 4 | 雨天 | 温柔 | 高 | 弱 | 进行 |
| 5 | 雨天 | 凉爽 | 正常 | 弱 | 进行 |
| 6 | 雨天 | 凉爽 | 正常 | 强 | 取消 |
| 7 | 阴天 | 凉爽 | 正常 | 强 | 进行 |
| 8 | 晴天 | 温柔 | 高 | 弱 | 取消 |
| 9 | 晴天 | 凉爽 | 正常 | 弱 | 进行 |
| 10 | 雨天 | 温柔 | 正常 | 弱 | 进行 |
| 11 | 晴天 | 温柔 | 正常 | 强 | 进行 |
| 12 | 阴天 | 温柔 | 高 | 强 | 进行 |
| 13 | 阴天 | 炎热 | 正常 | 弱 | 进行 |
| 14 | 雨天 | 温柔 | 高 | 强 | 取消 |

根据上述数据集最终建立的决策树为：



1. 上机实验。

代码各方法

Class Decision（）#决策树类

getData（） #方法，获取数据

majority（）#方法，进行多数表决

calcuEntropy（）#方法，进行信息熵计算

splitData（）#方法，切分数据集

chooseBestFeature（）#方法，选择最佳属性

createTree（）#方法，构建决策树

predict（）#方法，对测试集进行预测

\_\_init\_\_（）#构造方法

main（）#主类

1. 算法总结

根据划分结点时所使用的标准不一样，常用的决策树有三类：

ID3:使用信息增益来划分

C4.5使用信息增益率来划分

CART使用Gini系数来划分

决策树最大的优点是他可以自学习，在学习过程中不需要使用者了解过多的背景知识，只需要对训练实例进行较好的标注，就能够进行学习。速度快，计算量相对较小，可解释性强，在数据量较小的数据集上，决策树准确率比较高。

但相对于其他分类算法而言，决策树泛化能力较弱，容易针对训练数据造成过拟合。类别太多时错误的可能就会增加的比较快。对于连续性数据，有时要进行大量数据处理工作将其转换为离散值。