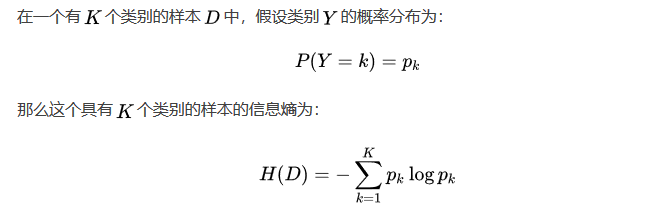
1. 基本概念
2. **信息熵**

Information Entropy，信息量大小的度量，即表示随机变量不确定性的度量；



1. **信息增益**

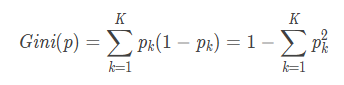
Information Gain，表示得知特征X的信息而使得类Y的信息的不确定性减少的程度。

—般地，熵H(Y)与条件熵H(Y|X)之差称为互信息(mutual information) ，决策树学习中的信息增益等价于训练数据集中**类与特征的互信息**。

1. **信息增益率**

Information Gain Ratio

**4）Gini系数**



1. **构造树的基本思想和三要素**

随着树深度的增加，节点的熵迅速地降低。熵降低的速度越快越好，这样就有望得到一颗高度最矮的决策树。

决策树是一种典型的分类方法：首先对数据进行处理，利用归纳算法生成可读的规则和 决策树；然后使用决策对新数据进行分析。本质上决策树是**通过一系列规则对数据进行分类的过程**。

决策树的三要素：**特征选择**，**决策树生成**，**决策树剪枝**

1. 决策树的组成

决策树的基本组成部分：决策结点、分支和叶子。

决策树中最上面的结点称为根结点，是整个决策树的开始。每个分支是一个新的决策结点，或者是树的叶子。每个决策结点代表一个问题或者决策，通常对应待分类对象的属性。 每个叶结点代表一种可能的分类结果。

1. 决策树与条件概率分布

决策树表示给定特征条件下类的条件概率分布，属于判别模型。

决策树学习是由训练数据集估计条件概率模型。**基于特征空间划分**的类的**条件概率模型**有无穷多个。我们选择的条件概率模型应该不仅对训练数据有很好的拟合，而且对未知数据有很好的预测.

1. 如何生成决策树？

决策树通过层次递进的属性判断，对样本进行划分。生成决策树的过程，就是确定树上的每一层（节点）应该选择什么属性来进行判断，即如何选择（每个节点的）最优划分属性。一般而言，遵循这样一个思想来选择属性进行划分：

随着划分过程不断进行，我们希望决策树的分支结点所包含的样本尽可能属于同一类别，即**结点的“纯度” (purity) 越来越高**。

1. **决策树如何进行特征选择（决策树样本集合的纯度如何度量）？**

决策树特征选择的准则：信息增益，信息增益率，基尼指。构建决策树依据：

ID3 : 信息增益

C4.5: 信息增益比率

CART: Gini系数

refer: <https://www.cnblogs.com/muzixi/p/6566803.html>

1. 介绍一下ID3、C4.5和CART决策树算法

**--> ID3算法：**

1. 概念：

ID3是最早的决策树算法，其根据信息增益(Information Gain)来寻找最佳决策特征。

当按特征A来划分数据集时的信息增益定义为：



1. 基本思想：

以信息熵为度量，用于决策树节点的属性选择，每次**优先选取信息量最多的属性**，亦即能**使熵值变为最小的属性**，以构造一颗**熵值下降最快的决策树**，到叶子节点处的熵值为0。此时，每个叶子节点对应的实例集中的实例属于同一类。

1. ID3算法特点：
2. 贪心算法
3. 每次做决策时只用一个特征
4. **没有办法处理连续特征**跟**缺失值**
5. 容易过拟合
6. 缺陷：

**信息增益倾向于选择类别数多的特征来做划分**，特别是特征属性很多，并且该特征的每个属性对应的样本个数又很少的情况。

假设有一列特征(如样本ID)类别数与样本数相等，如果以该特征来进行划分数据集，则数据集被划分成了单样本节点，每个节点的熵均为0，总熵也为0，这样一来得到了最大的信息增益，但是这种划分显然是不合理的。

引入 **信息增益率** 解决上述问题 (ID3算法 --> C4.5算法)。

**--> C4.5算法：**

1. 基本思想

C4.5在ID3的基础上改进了生成树算法，不再使用信息增益，而是使用信息增益率(gain ratio)来决定使用哪个特征来划分数据集。

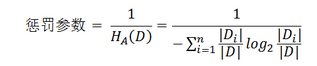
1. 信息增益率定义

特征A对训练数据集D的信息增益率GainRatio(D,A)定义为其信息增益Gain(D,A)与训练数据集D的经验熵H(D)之比：



也可以看做在信息增益的基础上乘上了一个对特征A属性个数的惩罚参数（熵的倒数），惩罚参数：

等价于数据集D以特征A作为随机变量的熵的倒数，即：将特征A取值相同的样本划分到同一个子集中。



1. C4.5算法缺陷：

信息增益率**倾向于选择类别数少的特征来做划分**。

当特征取值较少时HA(D)的值较小，因此其倒数较大，因而信息增益率较大。因而偏向取值较少的特征。

1. 优化：

考虑到C4.5算法缺陷，使用时并不是直接选择增益率最大的候选划分属性，而是使用了一个启发式：

**先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的**。

1. 特点（相对于ID3的改进）
2. 在Gain基础上引入了惩罚参数，如上所述
3. 增加了对缺失值的处理
4. 采用自底向上的剪枝策略来表面过拟合（**后剪枝？？？？**）
5. **C4.5是如何处理缺失值的？ ---> 需要再了解下**

决策树算法在处理缺失值熵需要解决三个问题：熵的计算，数据集的划分，带缺失值的预测。

1. 在做test时，当前特征为缺失值的样本不参与熵的计算；
2. 在做test时，给当前特征缺失的样本赋一个初始值为1的权重，做划分时，这些特征缺失的样本会被复制多份分配到所有子节点中，并按照叶子节点中非缺失样本的比例更新权重。设某个子结点中的非缺失样本占父节点非缺失样本比例为r，则更新该结点中每个缺失值样本的权重为 w:=w\*r。
3. 在预测带缺失值的样本时，样本在决策树中的行走路线同上，最后该样本会同时出现在多个叶子节点中，按在各叶子节点中的权重和来给出预测类别。

**--> CART算法：**

Refer1: <https://www.jianshu.com/p/4d897c0d9a05>