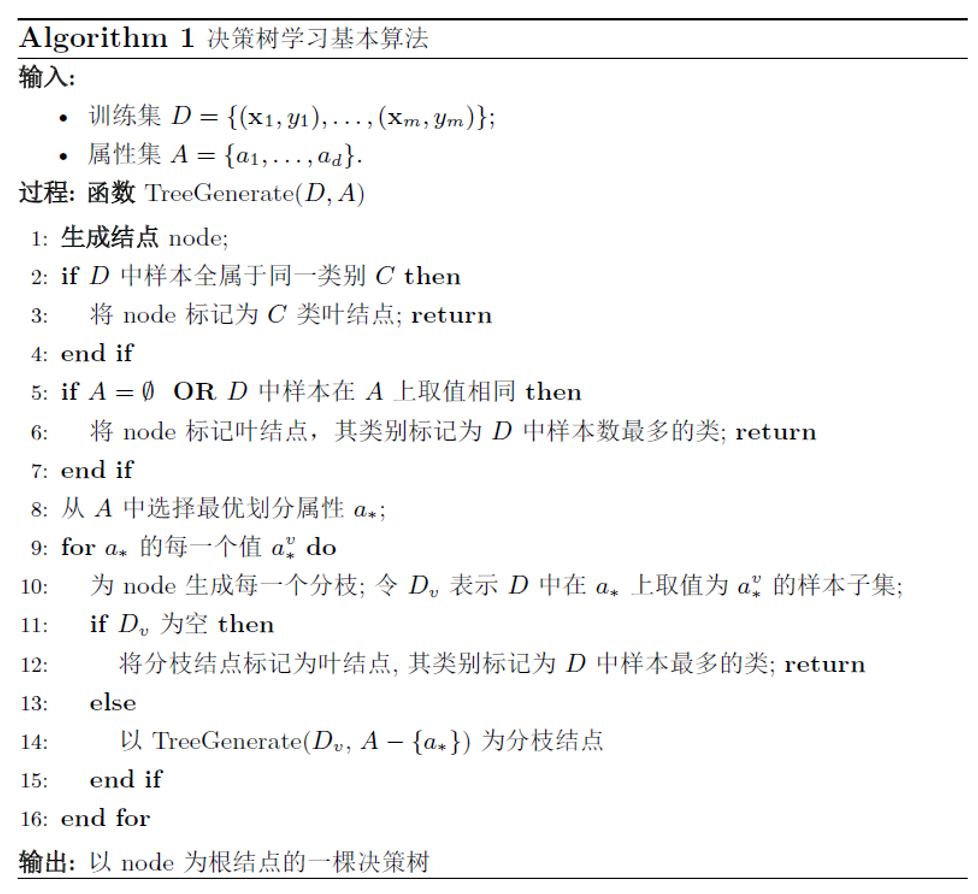
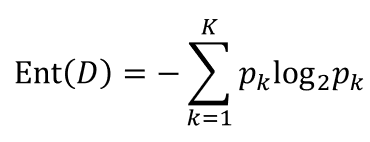
决策树

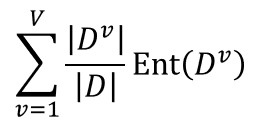
* 根结点（一个）：包含样本全集
* 内部结点（若干）：对应于一个属性的测试
* 叶结点（若干）：对应于分类结果

启发式算法：通常是一个递归地选择最优特征，并根据该特征对训练数据进行分割，使得对各个子数据集有一个最好的分类的过程。

**选择最优划分属性量化算法**

1. **信息增益 (ID3)**

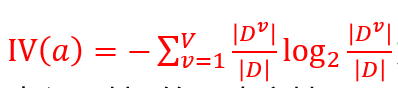
信息熵：量化信息的不确定性。（假定当前样本集合中第类样本所占的比例为）

考虑某一属性信息后的信息熵：以该属性划分的分支样本集的信息熵的加权求和。

某一属性的信息增益Grain(D,a)=总的不考虑属性信息熵Ent(D)-考虑该属性信息后的信息熵Ent(D,a)

弊端：对取值较多的属性有所偏好

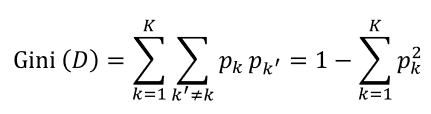
1. **信息增益率 (C4.5)**

表征属性𝑎的不确定性的固有值：

某一属性的信息增益率Gain\_ratio(D,a)= 该属性信息增益Grain(D,a)/该属性不稳定性IV(a)

（属于ID3算法的进化，一般从信息增益高于平均水平的划分属性中再选择信息增益率最高的）

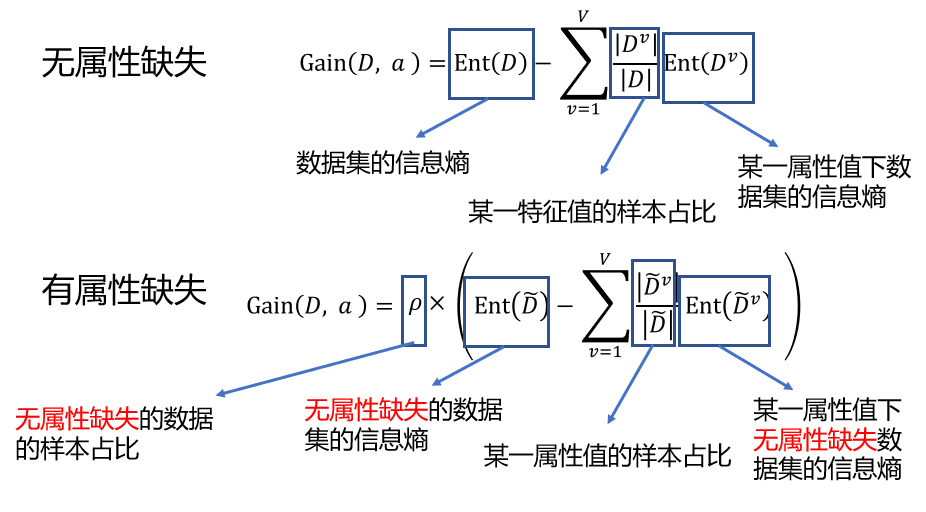
1. **基尼系数 (CART)**

样本集的基尼系数：表示从随机抽取两个样本其标记不一致的概率。

某一属性的基尼系数Gini\_index(D,a)：以该属性划分的分支样本集的信息熵的加权求和。

**连续属性离散化(二分法):**考察该属性在样本集上的不同取值，取各相邻取值的中点为候选二分划分点，取信息增益最大的划分点。

**缺失值处理**

1. 划分属性选择时面对部分样本属性缺失的情况：可使用信息增益算法的推广版本
2. 使用已知划分属性划分样本时面对样本的该属性缺失的情况：分到所有子结点，并设置权重（与该子节点样本数量成正比）。

**剪枝**：“剪”掉一些叶节点。减缓“过拟合”，避免决策分支过多。

1. 预剪枝：在决策树生成过程中，停止往下划分结点，当：

* 树达到预定的高度
* 节点包含的样本个数小于预定的阈值
* 划分当前结点不能提升泛化能力（需要输入测试集来评估）

1. 后剪枝：在决策树生成之后，自底向上地考察非叶结点，比较划分和不划分的泛化能力，来决定是否直接替换为叶结点。