## 决策树介绍

决策树是一种类似于流程图的树结构，由**内部结点、叶节点、有向边**组成。一个内部结点表示一个特征上的测试，一个叶节点表示类别标号。决策树分类就是从根节点开始，对实例的某一特征进行测试，根据测试结果将实例划分子节点，再对子结点上的剩余特征进行递归测试并分配，直到产生叶子节点。

1. 决策树优点

1）创建决策规则的过程简单。

2）决策树可以可视化，输出结果易于理解。

3）可以处理连续性和离散性数据。

可用于分类和回归

1. 缺点

1）容易产生过拟合，可以通过剪枝来缓解。

2）不能保证建立全局最优的决策树，可以通过随机森林来缓解

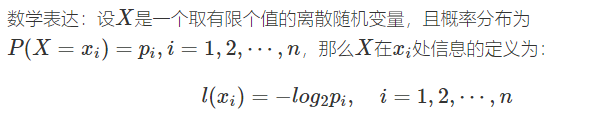
## **ID3算法**

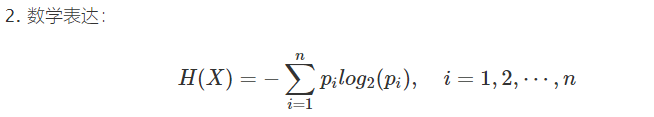
*重要概念***：**

为什么要先讲信息？决策树是使用不熟悉的数据集合，需要从中提取出一系列的规则。首要问题是，当前哪个特征对于划分数据集有较大的贡献，怎样按照一定的顺序选择特征来才能构建一颗高效的分类决策树。所以首要任务是需要对每个特征进行分析。

**2.信息(Info)：**

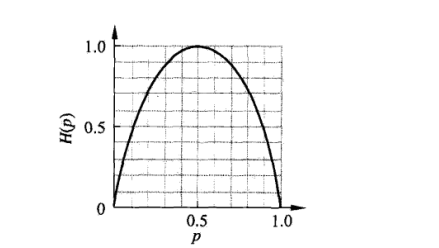
意义：信息是用于消除随机不确定性的东西



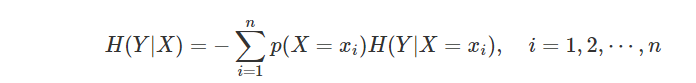
**3.熵(entropy)**意义：熵是随机变量不确定性的度量，是所有类别中所有可能值包含的信息期望。

特点：熵值越大，随机变量不确定性越大，熵值越小，随机变量不确定性越小。

如果有其中pi=1,其他等于0,那么熵为0；若所有pi都相等，所有事情概率都相等，那么熵值最大。



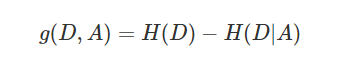
4.条件熵：表示在已知随机变量X的条件下随机变量Y的不确定性。



1. 补充：当熵和条件熵的概率由数据估计得到时，所得到的熵和条件熵分别称为经验熵和经验条件熵

**信息增益(Information Gain)**

1. 意义：信息增益是已知特征A的信息而使得类D的不确定性减少的程度(互信息)。
2. 数学表达：



1. 解释：

1) D为训练集，A为训练集的某一特征，H(D)表示对数据集D分类的不确定性，H(D|A)表示给定特征A时，数据集D分类的不确定性。

2) H(D)−H(D|A)就是信息增益，表示由于特征A而使得数据集D分类不确定性的减少程度。

3) 不同特征有不同信息增益，一般选择信息增益较大的特征来作为决策树结点。

**输入**：训练数据集、特征集、阈值

**输出**：决策树

**终止条件**：

1）子数据集中的所有类标签完全相同，实例属于同一类，返回该类。

2）使用完了所有特征仍然不能将数据集划分为唯一类别的分组，用多数表决

**否则**：

1）使用分裂准则（信息增益最大）指定分裂特征和分裂点，划分数据集。若划分的数据集信息增益小于某个阈值，停止划分，用多数表决。

2）删除已用特征，并递归产生子树。