# 决策树

## 决策树的构建流程

1. 从特征列表中选取最优特征
2. 根据特征的取值划分子树
3. 将最优特征从特征列表中删除
4. 对每个子树递归执行1~4

## 最优特征的选取方式

### 熵与信息增益

熵：表示体系/信息的混乱程度，也就是说：信息越有序，信息熵越低。

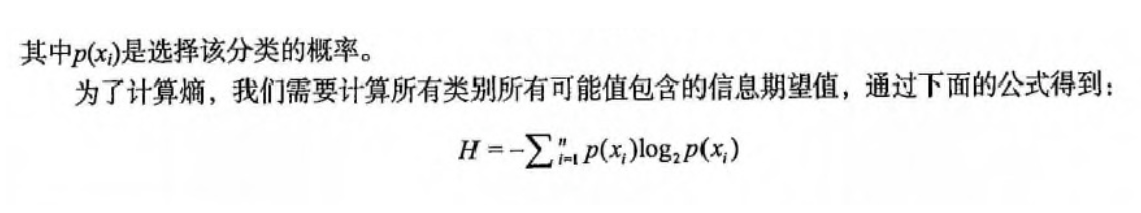
信息增益：在划分数据集前后信息发生的变化称为信息增益。

### 特征选择与熵的关系

一个好的特征，根据该特征对数据集进行划分之后，子数据集的熵要比原数据集低。因此，最优的特征就是划分成子数据集之后，信息增益最大的。

### 计算方式

数据集熵的计算方式，一个数据集包含很多类别，就是这一类出现的概率，也就是所占比例 类样本数/样本总数：



最优特征的计算方式：

1. 计算原有数据集的信息熵baseEntropy
2. 遍历所有特征
   1. 根据该特征的不同取值，划分成多个子数据集
   2. 计算每个子数据集的信息熵childEntropy和子数据集的概率childP(子数据集元素个数/总个数)
   3. 新的信息熵
   4. 该特征的信息增益
3. 信息增益最大的即为最优特征

## 构建决策树的停止条件

1. 子树里所有数据都是同一类
2. 已经用完所有特征
3. 达到人为设置的深度上限

从2、3两点可以看到，决策树的子节点可能会有多种类型的数据