决策树：根结点 x1，内部结点 xn，叶结点 xn
- 内部结点 xn → 属性测试
- 叶结点 xn → 决策结果

## 生成决策树：递归过程

base case：
① 当前结点所包含样本属于同一类型 → 无需划分
② 属性集为空 → majority vote 利用当前结点的后验分布
③ 样本集为空 → 父节点样本集 majority vote 当前结点的先验分布.

## 划分选择

### 1. 信息增益

信息熵 Entropy：$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$　$p_k$：第 k 类样本所占比例

信息增益 $Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$

Gain ↑ 使用 a 来划分所获得纯度提升个

### 2. 增益率

$$Gain\_ratio = \frac{Gain(D,a)}{IV(a)}$$

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

对可取值数目较少的属性有偏好
候选划分属性中找出信息增益高于均值的属性
再从中选增益率最高的

### 3. 基尼指数

$$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2$$

$$Gini\_index(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$$

Gini ↓ 选

## 剪枝处理

- 预剪枝 → 欠拟合风险（贪心策略）
- 后剪枝 → 开销大

## 连续值 & 缺失值

连续值　将该属性的值排序，选取不同 t 点计算信息增益，选择最优 t 点对应划分点

缺失值　为那些无缺失值样本赋权重.

## 多变量决策树

单变量：轴平行　分类边界由若干个与坐标轴平行的分段组成

多变量：斜划分　$\sum_{i=1}^{d} w_i a_i = t$ 的线性分类器