Decision Tree

ID3 ⇒信息增益

信息熵：$H(D)=-\sum_{k=1}^{k} P_k \log_2 P_k$

↓

样本的混乱程度。

信息熵↑，样本混乱多个，无法划分↑

信息增益：$Gain(D,A)=H(D)-H(D|A)$ → 样本在属性A上的信息增益

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

↙

属性A共有V种不同的取值

信息增益↑，使用属性A划分后样本混乱多↓，

$Gain(D,A)\uparrow = H(D)-H(D|A)\downarrow$

⇒ID3选择信息增益最大的属性划分。

不足：偏向选择取值较多的属性

↓

$H(D|A)=\sum_{v=1}^{V\uparrow}\frac{|D^v|}{|D|\downarrow}\left[-\sum_{k=1}^{k}P_k\uparrow \log_2 P_k\right]\downarrow$

C4.5 ⇒信息增益率

$Gain\text{-}ratio(D,A)=\frac{Gain(D,A)}{IV(A)}$

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

属性种类↓，D^v↑，$IV(A)\downarrow$，$Gain\text{-}Ratio(D,A)\uparrow$

CART 分类和回归树

CART是一个二叉树，对于离散和连续值都是在多种取值划分点划分的最优的。

1) 分类树

→ 表示数据的纯净，与信息熵相反。

使用gini指数作为特征：

~~Gini Index~~: $Gini(D) = \sum_{k=1}^{K} P_k(1-P_k) = 1 - \sum_{k=1}^{K} P_k^2$

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

选择gini指数最大的属性作为特征点。

1° 离散值处理

假设属性A取值有3种 {young, middle, old}，则以如下划分组合中选择最佳划分点

{((young), (middle, old)), ((middle), (young, old)), ((old), (young, middle))}

2° 连续值处理

(1) 假设属性A有V种取值 $a_1, a_2, \dots a_V$，对它们排序。

(2) 取2个特征值的中点作为分割点：

分割点1 分割点2

(3) 计算gini-index，选择最佳分割点。

2) 回归树

$$RSS = \min_{m,v} \left[\sum_{x \in R_1(m,a_v)} (y_i - \bar{y}_{R_1(m,a_v)})^2 + \sum_{x \in R_2(m,a_v)} (y_i - \tilde{y}_{R_2(m,a_v)})^2 \right]$$

第m个属性　v个割点

a_v 为样本

a_v 为样本

剪枝 ⇒ 缓解过拟合，正则化

1) 限制树的深度，叶子节点的数量，各样本子集中样本的数量。

2) 判断剪后，性能是否有所提高。

剪枝 { REP(reduced error pruning)：从下往上剪，如果性能未下降就替换节点。

Cost Complexity pruning (CCP)：生成 $T_0 \sim T_n$ 棵树，使用以下标准的方法选择树。

↙ 若干子树的大小。

剪枝后的错误率 ← $\frac{err(T_i, S) - err(T_0, S)}{\#leaves(T_0) - \#leaves(T_i)}$ → 原始的错误率

↓ 叶子节点的数量。

其它剪枝方法