

Decision Tree

ID3 \Rightarrow 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 \uparrow , 样本混乱度 \uparrow , 无法划分

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

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

信息增益 \downarrow , 使用属性A划分后样本混乱度 \downarrow , 属性A共有V种不同取值

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

\Rightarrow ID3 选择信息增益最大的属性划分

不足: 偏向选择取值较少的属性

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

C4.5 \Rightarrow 信息增益率

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

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

属性种类 \downarrow , DV \uparrow , $IV(A) \downarrow$, $Gain-ratio(D, A) \uparrow$

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中找到最优划分。

1) 分类树

使用 gini 指标作为划分依据:
 \rightarrow 表示数据的纯度, 与信息熵相反。

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

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

选择 gini 指标最大的属性作为划分点。

1° 离散值处理

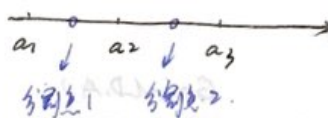
假设属性 A 取值有三种 { young, middle, old }, 则从如下划分组合中选择最佳划分点。

$$\{ ((\text{young}), (\text{middle}, \text{old})), ((\text{middle}), (\text{young}, \text{old})), ((\text{old}), (\text{young}, \text{middle})) \}$$

2° 连续值处理

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

(2) 取 2 个特征值的中间点作为分割点:



(3) 计算 gini-index, 选择最佳分割点。

2) 回归树

$$RSS = \min_{m, \text{分割点}} \left[\sum_{x \in L(m, \text{分割点})} (y_i - \bar{y}_{L(m, \text{分割点})})^2 + \sum_{x \in R(m, \text{分割点})} (y_i - \bar{y}_{R(m, \text{分割点})})^2 \right]$$

\swarrow 第 m 属性 \swarrow 左样本 \swarrow 右样本
 \swarrow V 个值点 \swarrow L 左 \swarrow R 右

剪枝 \Rightarrow 避免过拟合, 正则化

1) 限制树的深度, 叶子节点的数量, 节点与集中节点的权重.

2) 划分前后, 性能是否有所提高.

3) 后剪枝

REP (reduced error pruning): 从下往上, 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 W.T. 树评估方法选择树.

\downarrow
 若减小了树的大小,

$$\frac{\text{err}(T_i, S) - \text{err}(T_0, S)}{\#leaves(T_0) - \#leaves(T_i)}$$

剪枝后的
 错误率 \leftarrow 原树的误差率
 \downarrow
 叶子节点的数量

其它剪枝方法