Decision Tree

①	X_1	X_2	
Y_1	1/4	1/4	1/2
Y_2	1/4	1/4	1/2
	1/2	1/2	

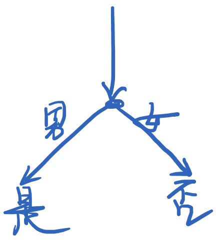
②	X_1	X_2	
Y_1	1/8	3/8	1/2
Y_2	1/8	3/8	1/2
	1/4	3/4	

③	X_1	X_2	
Y_1	1/8	1/8	1/4
Y_2	3/8	3/8	3/4
	1/2	1/2	

① $H(Y|X) = -(\frac{1}{2}\cdot\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\log_2\frac{1}{2}) = 0.5$

② $H(Y|X) = -(\frac{1}{4}\cdot\frac{1}{2}\log_2\frac{1}{2} + \frac{3}{4}\cdot\frac{1}{2}\log_2\frac{1}{2}) = 0.5$

③ $H(Y|X) = -(\frac{1}{2}\cdot\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\cdot\frac{3}{4}\log_2\frac{3}{4}) = 0.4056$



训练集D

分类 $\{C_1, \ldots, C_k\}$ $D \xrightarrow{特征A} \{D_1, D_2, \ldots, D_n\}$

$D_{i1} \leftarrow C_1$, $D_{i2} \leftarrow C_2$, …, $D_{ik} \leftarrow C_k$

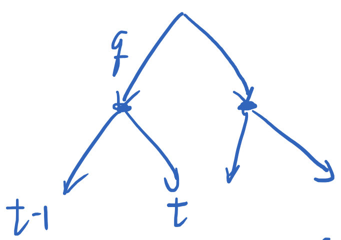
① $H(D) - H(D|A) = g(D|A)$ 信息增益 ID3 $\{$阈值ε, $g(D,A)\}$

② $\frac{g(D,A)}{H_A(D)} = g_R(D,A)$ 信息增益比 C4.5 $\{$阈值ε, $g_R(D,A)\}$

经验熵 $H(D) = -\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}$

特征A下条件熵 $H(D|A) = -\sum_{i=1}^{n}\left(\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}\right)$

$H_A(D) = -\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$



T个新叶节点 N_t $\{N_{t1} \leftarrow C_1, \ldots, N_{tk} \leftarrow C_k\}$

$C_\alpha(T_A) - C_\alpha(T_B) =$ 前 − 后

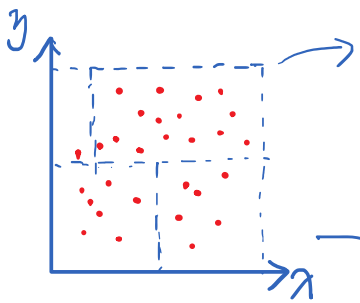
$=$ 节点数 × 分支操作减少的经验熵 $-\ \alpha$(分支后多的叶节点数 − 1)

$= -|N_q|\sum_{k=1}^{K}\frac{|N_{qk}|}{|N_q|}\log_2\frac{|N_{qk}|}{|N_q|} - \left(-\sum_{t=1}^{T}|N_t|\sum_{k=1}^{K}\frac{|N_{tk}|}{|N_t|}\cdot\log_2\frac{|N_{tk}|}{|N_t|}\right)$

$-\ \alpha$(分支后多的叶节点数 − 1)

$\frac{C_\alpha(T_A) - C_\alpha(T_B)}{|N_q|} =$ (父节点的经验熵 − 子节点的经验熵期望) $-\ \alpha\,\frac{子节点数 - 1}{|N_q|}$

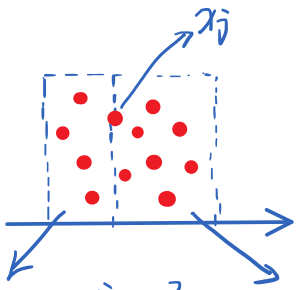
CART 回归树



分成M个单元 $R_1, \dots, R_M$，
对应一个值 $c_1, \dots, c_M$

则 $f(x)=\sum_{m=1}^{M} c_m I(x \in R_m)$

$D=\{(x_1,y_1),\dots,(x_N,y_N)\}$



① 找一点 $x^{(j)}$：

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2} (y_i - c_2)^2\right]$$

② 算对应值 s

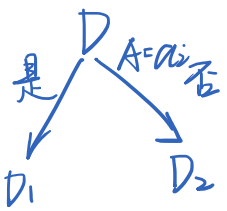
$R_1(j,s)=\{x \mid x^{(j)} \le s\}$ → $\hat{c}_1 = ave(y_i \mid x_i \in R_1(j,s))$

$R_2(j,s)=\{x \mid x^{(j)} > s\}$ → $\hat{c}_2 = ave(y_i \mid x_i \in R_2(j,s))$

$$Gini(D)=1-\sum_{k=1}^{K} p_k^2 = 1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2$$

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2)$$

CART 决策树 → 二叉树



计算 $Gini(D, A=a_i)$ 取最小的特征及对应值

$B=b_j \cdots$

停止条件：结点样本数 < 阈值；Gini指数 < 阈值

① 如何理解熵，基尼指数？

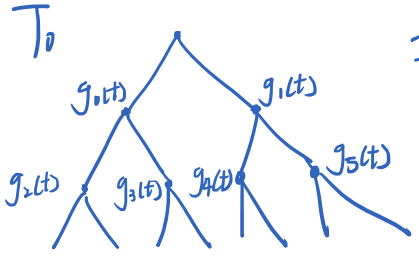
② 回归和概念区别？

③ 剪枝时，损失函数 α 如何定？（$\alpha \ge 0$，但 α 取什么最优？）

CART 决策树剪枝：

$$\alpha = g(t) = \frac{C(t) - C(T_t)}{|T_t| - 1}$$

（当父节点和子节点损失函数相等时，$\alpha = +\infty$）



重复：减枝 $\min(g_i(t) \mid i=1,\dots,5)$ → 得到 $T_1, \dots, T_n$

只有根节点和两个子节点

再从 $\{T_0, \dots, T_n\}$ 中找最优子树 T_α

设 $m \ge 2$，m 层二叉树有 $2^{(m-1)}-2$ 个内部节点

则 $n=[2, 2^{(m-1)}-2]$，n 为整数，进行 $[3, 2^{(m-1)}-1]$ 次交叉验证