Chapter 4

4.1 看图4.2 P74 页的流程图，

程序终止的条件，有3个，

① $D_v$ 为空，则训练误差在这个子~~集~~树上为0

② D样本全为一个类别，训练误差也为0，~~因为类别一样~~

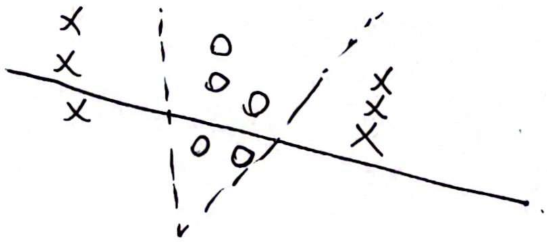
裹

③ ~~A~~ 属性集 $A=\phi$ or ~~D的类取值一样~~，A的取值相同，即属性相同。

~~假设~~ 因为不存在冲突数据，属性值相同 ⇒ 类相同，⇒ 训练误差为0。

所以一定要用剪枝防止过拟合，否则有极高概率过拟合。

4.2、 会出现以下可能，用最小训练误差做为贪心策略，



会导致本来应该用两条线分类~~数据~~的数据，用一条线分割后，再分割，不符合样本规律，泛化能力下降。

~~4.3~~、4.5、 为了避免出现4.2问题的情况，

需要让损失函数不为"最小误差"，

而是"~~最小~~ minimal entropy"，不知道怎么用logistic regression实现

loss function of:

① logistic regression : $\ell(\beta)=\sum_{i=1}^{m} y_i \cdot p(y=1|\hat{x}_i,\beta)+(1-y_i)\,p(y=0|\hat{x}_i,\beta)$

② logistic decision : $\cancel{\ell(\beta|D,A)=\sum_{x_i\in D} p(y=1|\hat{x}_i,\beta)\cdot w_i}$

$$\ell(\beta \mid D, A) = h\left(\frac{\sum_{x_i \in \tilde{D}} w_i \cdot p(y=1 \mid \hat{x}_i, \beta)}{\sum_{x_i \in D} w_i}\right) + h\left(\frac{\sum_{x_i \in D} w_i - \sum_{x_i \in \tilde{D}} w_i \cdot p_{1}}{\sum_{x_i \in D} w_i}\right)$$

$$h(x) = x \cdot \log x.$$

用熵值做为 loss function，去避免过拟合.

(这是我的方法，但是导数复杂度很高，效果可能不好)

4.9. ① $\mathrm{Gini}(D) = 1 - \sum_{k=1}^{|Y|} p_k^2$

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

②

$$\rho = \frac{\sum_{x \in \tilde{D}} w_x}{\sum_{x \in D} w_x}$$

$$\tilde{p}_k = \frac{\sum_{x \in \tilde{D}_k} w_x}{\sum_{x \in \tilde{D}} w_x} \qquad (1 \le k \le |Y|) \quad \text{class}$$

$$\tilde{r}_v = \frac{\sum_{x \in \tilde{D}_v} w_x}{\sum_{x \in \tilde{D}} w_x} \qquad (1 \le v \le V) \quad \text{attribute.}$$

$$\mathrm{Gini}(\tilde{D}) = 1 - \sum_{k=1}^{|Y|} \tilde{p}_k^2,$$

$$\mathrm{Gini\_index}(D, a) = \sum_{v=1}^{V} \tilde{r}_v \cdot \mathrm{Gini}(\tilde{D}_v)$$

for $x \in D \setminus \tilde{D}$, $w_x \to \tilde{r}_v \cdot w_x$, $\forall v \le V$. □