Hw 3

1 (3.2)

$$y=\frac{1}{1+e^{-(w^Tx+b)}}$$

$$\frac{\partial y}{\partial w}=\frac{x\,e^{-(w^Tx+b)}}{(1+e^{-(w^Tx+b)})^2}=\vec{x}\,y(1-y).$$

$$\frac{\partial^2 y}{\partial w^T\partial w}=\frac{\partial}{\partial w^T}\frac{\partial y}{\partial w}=\frac{\partial}{\partial w^T}(\vec{x}\,y(1-y))=y'^T(1-y)\vec{x}-yy'^T\vec{x}=x^Tx\,y(1-y)(1-2y).$$

由 $x^Tx\geq 0$，故 $y\in(0.5,1)$ 时 $\frac{\partial^2 y}{\partial w^T\partial w}<0$，此时函数非凸

而 $l(\beta)=\sum_{i=1}^{m}(-y_i\beta^Tx_i+\ln(1+e^{\beta^Tx_i}))$

$$\frac{\partial l(\beta)}{\partial \beta}=\sum_{i=1}^{m}\left(-y_i\vec{x}_i+\frac{1}{1+e^{\beta^Tx_i}}\cdot\vec{x}_i\cdot e^{\beta^Tx_i}\right)$$

$$\frac{\partial^2 l(\beta)}{\partial\beta^T\partial\beta}=\sum_{i=1}^{m}\frac{\beta^Tx_i}{(1+e^{\beta^Tx_i})^2}x_i^Tx_i\geq 0$$

所以 $l(\beta)$ 为凸函数

2(3.7) 最优ECOC 要求 任意两编码之间最小距离最大，

任意两类别之间海明距离最大，且反码之间距离最大

可采取如下编码

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
C_1	1	1	1	1	1	1	1	×	×
C_2	0	0	0	0	1	1	1	×	×
C_3	0	0	1	1	0	0	1	×	×
C_4	0	1	0	1	0	1	0	×	×

类别间海明距为4，f_8, f_9 可为任意编码

3. 令 $S_b=\sum_c m_c(\mu_c-\mu)(\mu_c-\mu)^T=AMA^T=(\mu_1-\mu,\cdots,\mu_N-\mu)\begin{pmatrix}m_1&&&\\&m_2&&\\&&\ddots&\\&&&m_N\end{pmatrix}\begin{pmatrix}\mu_1-\mu\\ \vdots\\ \mu_N-\mu\end{pmatrix}$

$\Rightarrow \text{rank}(S_b)=\text{rank}(AMA^T)=\text{rank}(AM^{\frac12}\cdot(AM^{\frac12})^T)=\text{rank}(AM^{\frac12})=\text{rank}(A).$

而 $\sum m_i(\mu_i-\mu)=0$ 线性相关.

故 $\text{rank}(S_b)=\text{rank}\,A\leq N-1$

4. 证明 · $S_b W = \lambda S_w W$

解：原问题：$\max_{w} \frac{Tr(W^T S_b W)}{Tr(W^T S_w W)}$

$\Longleftrightarrow \max_{w} Tr(W^T S_b W)$, s.t. $tr(W^T S_w W) = 1$.

$\Longleftrightarrow L(W,\lambda) = -tr(W^T S_b W) + \lambda(tr(W^T S_w W) - 1)$

$\frac{\partial L}{\partial W} = 0$

$\Rightarrow \lambda\left(\frac{\partial tr(W^T S_w W)}{\partial W}\right) = \frac{\partial tr(W^T S_b W)}{\partial W}$

$\Rightarrow \lambda(S_w + S_w^T)W = (S_w + S_w^T)W$

$\Rightarrow S_b W = \lambda S_w W$

5. 令 $P = X(X^T X)^{-1} X^T$

$P^T = X(X(X^T X)^{-1})^T = X((X^T X)^T)^{-1} X^T$

$= X(X^T X)^{-1} X^T = P.$

$P^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = P.$

所以 P 是投影矩阵

解释：对任意向量 y，$\hat{y} = X(X^T X)^{-1} X^T y$ 是在 y 空间的线性投影

Hw4.

1.(4.1). 反证：假设不存在，那么决策树必然有与训练集冲突的数据，这与假设矛盾，

2.(4.9). $\text{Gini}(D,a) = \rho \times \text{Gini\_index}(\tilde{D},a)$

$$= \rho \sum_{v=1}^{|V|} \tilde{r}_v \,\text{Gini}(\tilde{D}^v)$$

$$= \rho \sum_{v=1}^{|V|} \tilde{r}_v \left(1-\sum_{i=1}^{K} \tilde{p}_k^2\right)$$

3. 设随机变量 $X \in \{1,\cdots,N\}$，取值为 k 概率：$p(X=k)=p_k$

信息熵为：$H = -\sum_{k=1}^{N} p_k \log_2 p_k \quad s.t. \ \sum_k p_k = 1.$

Lagrange：$L = -\sum_{k=1}^{N} p_k \log_2 p_k + \lambda\left(\sum_{k=1}^{N} p_k - 1\right)$

$$\Rightarrow \frac{\partial L}{\partial p_i} = -\log_2 p_i - \frac{1}{\ln 2} + \lambda = 0$$

$$\left.\begin{aligned} &\Rightarrow p_i = 2^{\lambda - \frac{1}{\ln 2}} \ (i=1,\cdots N) \\ &\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{N} p_i - 1 = 0 \end{aligned}\right\} \Rightarrow p_i = \frac{1}{N}$$

4. (1). $p^+ = \frac{1}{2}, \quad p^- = \frac{1}{2}$

$$\Rightarrow \text{Ent}(D) = -p^+\log p^+ - p^-\log p^- = 1$$

(2). $\text{Gain}(D,A) = \text{Ent}(D) - \sum_v \frac{|D^v|}{|D|}\text{Ent}(D^v)$

$$= 1 - \left[\frac{4}{10}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{6}{10}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right)\right]$$

$$= 0.125$$

$$\text{Gain}(D,B) = 1 - \left[\frac{5}{10}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{5}{10}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)\right]$$

$$= 0.029$$

(3) 选取：$c \in \{1.5 + k\}_{k=0}^{6}$ 作为二分的节点

则 $G(D,C,c_k)$ 分别为：0.108, 0.236, 0.035, 0.125, 0, 0.035, 0.108

$k=0$	1	2	3	4	5	6

(4). $Gini(D,A) = \frac{4}{10}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right) + \frac{6}{10}\left(1-\left(\frac{2}{6}\right)^2-\left(\frac{4}{6}\right)^2\right)$

$= 0.417$

$Gini(D,B) = 0.48.$

因此 A更优、

(5) 由 $\begin{cases} Gain\_r(D,A) = \frac{Gain(D,A)}{IV(D,A)} \approx 0.129. \\ Gain\_r(D,B) = \frac{Gain(D,B)}{IV(D,B)} \approx 0.029 \\ Gain\_r(D,C,2.5) = \frac{Gain(D,C,2.5)}{IV(D,C)} \approx 0.326. \end{cases}$

First. 以 (C, 2.5) 为划分.

对 $D'_{C\leq 2.5}$ 为正例,不再细分

对 $D'_{C>2.5}$ 有 继续上述操作, 得 C4.5 决策树.