HW3.

3.2 证明对于参数 w.

(1) $y=\frac{1}{1+e^{-(w^Tx+b)}}$ 是非凸函数.

(2) $l(\beta)=\sum_{i=1}^{m}(-y_i\beta^T\hat{x}_i+\ln(1+e^{\beta^T\hat{x}_i}))$ 是凸函数.

首先联于凸函数的定理：

若 $f(x)$ 二阶连续可微，那么 $f(x)$ 是 D 上的凸函数的充要条件是，$f(x)$ 的 Hesse 矩阵在 D 上是半正定的。

(1) 对 y，求其 Hessian 矩阵。要证 y 非凸，证明该矩阵非半正定即可.

$$\frac{\partial y}{\partial w}=-[1+e^{-(w^Tx+b)}]^{-2}\cdot e^{-(w^Tx+b)}\cdot(-x)$$

$$=\frac{(1-\frac{1}{y})x}{\frac{1}{y^2}}=y(1-y)x=x(y-y^2)$$

$$\frac{\partial^2 y}{\partial w^2}=\frac{\partial}{\partial w^T}\left(\frac{\partial y}{\partial w}\right)=\frac{\partial}{\partial w^T}(x(y-y^2))=\left(\frac{\partial y}{\partial w}\right)^T\left[\frac{\partial(xy-xy^2)}{\partial y}\right]$$

$$=x^T(y-y^2)(x-2xy)$$

$$=xx^T(1-2y)(y-y^2)$$

∵ y 的值域为 $(0,1)$，xx^T 相当于 x 倍单位阵

且当 $y\in(0.5,1)$ 时，$y(y-1)(1-2y)<0$，$\frac{d}{dw^T}\left(\frac{\partial y}{\partial w}\right)$ 半负定

综上 y 为非凸函数.

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

$$\frac{\partial^2 l}{\partial \beta^2} = \frac{\partial \left(\frac{\partial l}{\partial \beta}\right)}{\partial \beta^T} = \sum_{i=1}^{m} \hat{x}_i \hat{x}_i^T p_1(\hat{x}_i; \beta)(1 - p_1(\hat{x}_i; \beta))$$

$$= X P X^T$$

其中 X 为 (n, m) 矩阵，每一列对应一个样本，P 为对角矩阵

$p_{ii} = p_1(\hat{x}_i; \beta)(1 - p_1(\hat{x}_i))$

$X P X^T$ 对任意向量 z 有：

$z^T X P X^T z = (X^T z)^T P (X^T z) = V^T P V$

$= \sum_i p_{ii} v_i^2 \geq 0$

因此此 Hessian 矩阵半正定

$l(\beta)$ 对于参数 w 是凸函数

3.7 码长为 9，类别数为 4. 求最佳 ECOC 二元码.

好的纠错码应该达到行分离、列分离，且两个分类器的编码不互为反码.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
$C_1 \rightarrow$	+1	+1	+1	+1	+1	+1	+1	-1	-1
$C_2 \rightarrow$	-1	-1	-1	-1	+1	+1	+1	+1	+1
$C_3 \rightarrow$	-1	-1	+1	+1	-1	-1	+1	-1	+1
$C_4 \rightarrow$	-1	+1	-1	+1	-1	+1	-1	+1	-1

如上的 ECOC 二元码，不存在编码互为反码的分类器

任意两组分类器之间的海明距离均大于等于 5

∴4类中不存在 f_i 和 f_j，使得 f_i 和 f_j 总是相等（这种情况，即为冗余的码长，对分类效果没有贡献）

下面证明这种情况为最优，分为2步：

①证明不存在任意两组分类器间海明距离均大于等于7的纠错码。

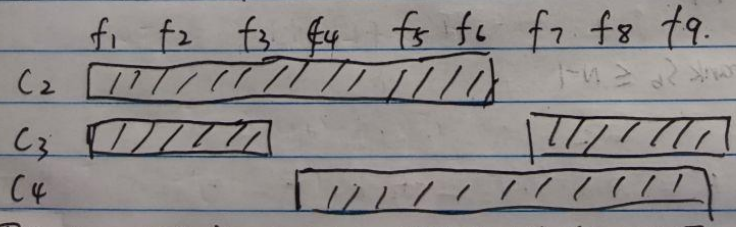
②证明存在任意两组分类器间汉明距离均为6的纠错码，但其不满足列分离。

①：以 C_1 为标准，若 C_2 与 C_1 海明距离为7，则有7位编码相异，不失一般性，不妨设为 $f_1 \sim f_7$ 7位编码相异。

由 C_3 和 C_1 海明距离为7，码长为9可知，C_3 和 C_2 必有至少 $(7+7-9)=5$ 位编码相同，但此时 C_3 和 C_2 的海明距离至多为4，无法达到7。

故命题得证。

②：同①的分析，以 C_1 为标准，若海明距离为6，则 C_1 与 C_2、C_3、C_4 分别为有6位编码相异。



上图为1例，阴影部分表示与 C_1 相异的6位编码位置。

可以看出，其满足任意两组分类器编码的海明距离为6.
但其中{f_1、f_2、f_3}、{f_4、f_5、f_6}、{f_7、f_8、f_9}三组中3个列不互相分离，不是一个好的ECOC纠错码。
综上，从开始的任意两组分类器编码的海明距离大于等于5的纠错码为最优。

补充题1. 在LDA多分类情形下，计算类间散度矩阵S_b的秩并证明.

$$S_b = \sum_{i=1}^{N} m_i (u_i - u)(u_i - u)^T$$

$$= [u_1 - u, u_2 - u, \cdots, u_n - u] \begin{pmatrix} m_1 & & \\ & \ddots & \\ & & m_N \end{pmatrix} \begin{pmatrix} (u_1 - u)^T \\ (u_2 - u)^T \\ \vdots \\ (u_n - u)^T \end{pmatrix}$$

记$M = diag(m_1, m_2, \cdots, m_N)$

$A = (u_1 - u, u_2 - u, \cdots, u_N - u)^T$

则 $rank\, S_b = rank\, A^T M A = rank\, A^T M^{\frac{1}{2}} M^{\frac{1}{2}} A = rank(A^T M^{\frac{1}{2}})(A^T M^{\frac{1}{2}})$

$= rank\,(A^T M^{\frac{1}{2}}) = rank\, A^T = rank\,(u_1 - u, u_2 - u, \cdots, u_N - u)$

其中，因为 $\sum_{i=1}^{N} m_i u_i = (\sum_{i=1}^{N} m_i) u$.

即 $\sum_{i=1}^{N} m_i (u_i - u) = 0$，故 $rank\,(u_1 - u, u_2 - u, \cdots, u_N - u) \le N - 1$

$rank\, S_b \le N - 1$

补充题2. 给出公式 3.45 的推导过程.

已知式 3.44，优化目标为 $\max_{W} \frac{tr(W^T S_b W)}{tr(W^T S_w W)}$

固定此目标的分母，即 $tr(W^T S_w W)$ 为 1，则等价于

优化问题：
$$\min_{W} \; -tr(W^T S_b W)$$
$$s.t. \quad tr(W^T S_w W) = 1$$

根据拉格朗日乘子法，可定义上述优化问题的拉格朗日函数

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

根据矩阵微分式

$$\frac{\partial tr(X^T B X)}{\partial X_{ij}} = \sum_{q=1}^{m} b_{iq} x_{qj} + \sum_{p=1}^{m} b_{pi} x_{pj} = [BX + B^T X]_{ij}$$

$$\frac{\partial tr(X^T B X)}{\partial X} = [BX + B^T X]$$

$L(W, \lambda)$ 对 W 求偏导则有：

$$\frac{\partial L(W, \lambda)}{\partial W} = -\frac{\partial (tr(W^T S_b W))}{\partial W} + \lambda \frac{\partial (tr(W^T S_w W) - 1)}{\partial W}$$

$$= -(S_b + S_b^T) W + \lambda (S_w + S_w^T) W$$

$$= -2 S_b W + 2\lambda S_w W.$$

令上式等于 0，即有：

$$-2 S_b W + 2\lambda S_w W = 0$$

$$S_b W = \lambda S_w W.$$

补充题3. 证明 $X(X^TX)^{-1}X^T$ 是投影矩阵，对线性回归模型从投影角度解释.

① 设向量 b 在矩阵 X 空间上的投影为 P，则

$$P = Xk = [x_1\ x_2\ \cdots\ x_n]\begin{bmatrix}k_1\\ \vdots\\ k_n\end{bmatrix} = x_1k_1 + \cdots + x_nk_n$$

其中 $x_1 \cdots x_n$ 为 X 的列向量.

令误差向量 $e = b - P$.

则由于 e 垂直于列空间的平面，则有：

$$\begin{cases} x_1^T(b-P)=0 \\ x_2^T(b-P)=0 \\ x_n^T(b-P)=0 \end{cases}$$

即有 $X^T(b-P)=0$

又∵ $P = Xk$

∴ $X^T(b-Xk)=0$

$X^Tb = X^TXk$

$k = (X^TX)^{-1}X^Tb$.

$P = Xk = X(X^TX)^{-1}X^Tb$

令 $W = X(X^TX)^{-1}X^T$

$P = Wb$

得出 W 为投影矩阵.

② 解释线性回归模型

可以把特征矩阵 X 看作一个向量组，每一列（特征）都是一个 n 维向量，我们有 d 个这样的向量，我们假设 $d<n$，且所有特征的线性无关。则 X 张成的空间是个 d 维度空间。真实值 y 是一个 $n\times 1$ 的向量，处于 n 维空间中。多元线性回归就是在 X 张成的 d 维空间中，寻找 n 维空间中 y 的投影。

HW 4

4.1 决策树的递归停止条件为：(1) 当前结点包含的样本属于同一类别，无需划分。(2) 当前的属性集为空，或是当前样本在所有属性上的取值相同，无法划分。(3) 当前结点的样本集合为空，不能划分。

假设对于训练得到的决策树存在结点，该结点中有无法划分的数据，即存在冲突数据，训练误差不为0，与原假设矛盾。

因此，对于不含冲突数据的训练集，必存在与训练集一致的决策树。

4.9 给定训练集D和属性a，令$\tilde{D}$表示D中在属性a上没有缺失值的样本子集。假设a有V个可取值$\{a^1, a^2, \cdots, a^V\}$，令$\tilde{D}^v$表示$\tilde{D}$上a属性取值为$a^v$的样本子集，$\tilde{D}_k$表示$\tilde{D}$中属于第k类$(k=1, \cdots, |Y|)$的样本子集。

显然有 $\tilde{D} = \bigcup_{k=1}^{|Y|} \tilde{D}_k$ $\tilde{D} = \bigcup_{v=1}^{V} \tilde{D}^v$

为每个样本赋予一个权重w_x

令 $\rho = \frac{\sum_{x \in \tilde{D}} w_x}{\sum_{x \in D} w_x}$ 表示在a上无缺失值的样本(权重)比例

$\tilde{p}_k = \frac{\sum_{x \in \tilde{D}_k} w_x}{\sum_{x \in \tilde{D}} w_x}$ 表示无缺失值样本中第k类权重比例

$\tilde{r}_v = \frac{\sum_{x \in \tilde{D}^v} w_x}{\sum_{x \in \tilde{D}} w_x}$ 表示无缺失值样本中a属性取a^v的样本权重比例

则由Gini指数定义：

$Gini(D) = 1 - \sum_{k=1}^{|Y|} \tilde{p}_k^2$ 即排除了所有随机抽取两样本后，其类别标记一致的情况的概率。

采用类似的符号表示，属性a的基尼指数定义为

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

(相当于原定义中的$\frac{|D^v|}{|D|}$换为了$\tilde{r}_v$)

补充题：$X=\{1,2,\cdots,k\}$　$P(X=k)=P_k$

熵 $H(P)=-\sum_k P_k \log_2 P_k$

写为规划问题，即有：

$$\max \quad -\sum_k P_k \log_2 P_k$$

$$s.t. \quad \sum_{i=1}^{k} P_i = 1$$

拉格朗日函数：$L(P,\lambda) = -\sum_{i=1}^{k} P_i \log_2 P_i + \lambda\left(\sum_{i=1}^{k} P_i - 1\right)$

求偏导：$\frac{\partial L(P,\lambda)}{\partial P_i} = -\log_2 P_i - \frac{1}{\ln 2} + \lambda$

$$\frac{\partial L(P,\lambda)}{\partial \lambda} = \sum_{i=1}^{k} P_i - 1 = 0.$$

令 $\frac{\partial L(P,\lambda)}{\partial P_i} = 0$，则 $P_i = 2^{\lambda - \frac{1}{\ln 2}}$

又∵ $\sum_{i=1}^{k} P_i - 1 = 0$

∴ $P_i = \frac{1}{k}$　　$\lambda = \log_2 \frac{1}{k} + \frac{1}{\ln 2} = -\log_2 k + \frac{1}{\ln 2}$

由此求出的 P_i 就是 $L(P,\lambda)$ 在条件 $\sum_{i=1}^{k} P_i = 1$ 下的可能的极值点。又因为这样的点只有一个，可直接确定这是所求的点。因此使熵最大的分布是均匀分布。

补充题 2.

(a). $P(\text{类别}=\text{"+"}) = \frac{1}{2}$

$P(\text{类别}=\text{"-"}) = \frac{1}{2}$.

$Ent(D) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$

(b) 以属性A来分析，有"T"、"F"两个取值.

$D^1 = \{1, 2, 3, 8\}$ D^1中有3个"+"，1个"-"

$D^2 = \{4, 5, 6, 7, 9, 10\}$ D^2中有2个"+"，4个"-"

$Ent(D^1) = -(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4})$

$= 2 - \frac{3}{4}\log_2 3$

$Ent(D^2) = -(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6})$

$= \log_2 3 - \frac{2}{3}$

$Gain(D, A) = Ent(D) - \sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v)$

$= 1 - \frac{2}{5}(2 - \frac{3}{4}\log_2 3) - \frac{3}{5}(\log_2 3 - \frac{2}{3})$

$= \frac{3}{5} - \frac{3}{10}\log_2 3$

以属性B来分析，有 "T"、"F" 两个取值.

$D^1 = \{1, 2, 5, 6, 9\}$ D^1中有2个"+"，3个"-"

$D^2 = \{3, 4, 7, 8, 10\}$ D^2中有3个"+"，2个"-"

$Ent(D^1) = -(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5})$

$= \log_2 5 - \frac{2}{5} - \frac{3}{5}\log_2 3$

$Ent(D^2) = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5})$

$= \log_2 5 - \frac{2}{5} - \frac{3}{5}\log_2 3$

$$Gain(D,B) = Ent(D) - \sum_{v=1}^{2} \frac{|D^v|}{|D|} Ent(D^v)$$

$$= 1 - (\log_2 5 - \frac{2}{5} - \frac{3}{5}\log_2 3)$$

$$= \frac{9}{5} + \frac{3}{5}\log_2 3 - \log_2 5.$$

(c) 统计C类，共8种取值，1.0、2.0、3.0、4.0、5.0、6.0、7.0、8.0

不妨取划分点分别为 1.5、2.5、3.5、4.5、5.5、6.5、7.5、 ...

① 划分点为 1.5

$D^1 = \{1\}$ 为"+"

$D^2 = \{2,3,4,5,6,7,8,9\}$ 4个为"+"，5个为"−"

$Ent(D^1) = -\log_2 1 = 0$

$Ent(D^2) = -(\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}) = -\frac{8}{9} - \frac{5}{9}\log_2 5 + 2\log_2 3$

$$Gain(D,C)_1 = Ent(D) - \sum_{v=1}^{2} \frac{|D^v|}{|D|} Ent(D^v)$$

$$= 1 - \frac{1}{10}\times 0 + \frac{9}{10}[\frac{8}{9} + \frac{5}{9}\log_2 5 - 2\log_2 3]$$

$$= \frac{9}{5} + \frac{1}{2}\log_2 5 - \frac{9}{5}\log_2 3$$

② 划分点为 2.5

$D^1 = \{1,10\}$ 2个均为"+"

$D^2 = \{2,3,4,5,6,7,8,9\}$ 3个为"+"，5个为"−"

$Ent(D^1) = 0$

$Ent(D^2) = -(\frac{3}{8}\log_2\frac{3}{8} + \frac{5}{8}\log_2\frac{5}{8}) = 3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5$

$$Gain(D,C)_2 = Ent(D) - \sum_{v=1}^{2} \frac{|D^v|}{|D|} Ent(D^v)$$

$$= 1 - \frac{4}{5}(3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5) = \frac{3}{10}\log_2 3 + \frac{1}{2}\log_2 5 - \frac{7}{5}$$

③ 划分点为 3.5

$D^1=\{1,6,10\}$　2个为"+"，1个为"−"

$D^2=\{2,3,4,5,7,8,9\}$　3个为"+"，4个为"−"

$Ent(D^1)=-[\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3}]=\log_2 3-\frac{2}{3}$

$Ent(D^2)=-[\frac{3}{7}\log_2\frac{3}{7}+\frac{4}{7}\log_2\frac{4}{7}]=\log_2 7-\frac{3}{7}\log_2 3-\frac{8}{7}$

$Gain(D,C)_3=Ent(D)-\sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v)$

$=1-\frac{3}{10}(\log_2 3-\frac{2}{3})-\frac{7}{10}(\log_2 7-\frac{3}{7}\log_2 3-\frac{8}{7})$

$=2-\frac{7}{10}\log_2 7$

④ 划分点为 4.5

$D^1=\{1,4,6,10\}$　3个为"+"，1个为"−"

$D^2=\{2,3,5,7,8,9\}$　2个为"+"，4个为"−"

$Ent(D^1)=-[\frac{3}{4}\log_2\frac{3}{4}+\frac{1}{4}\log_2\frac{1}{4}]=2-\frac{3}{4}\log_2 3$

$Ent(D^2)=-[\frac{2}{6}\log_2\frac{2}{6}+\frac{4}{6}\log_2\frac{4}{6}]=\log_2 3-\frac{2}{3}$

$Gain(D,C)_4=Ent(D)-\sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v)$

$=\frac{3}{5}-\frac{3}{10}\log_2 3$

⑤ 划分点为 5.5

$D^1=\{1,3,4,6,9,10\}$　3个为"+"，3个为"−"

$D^2=\{2,5,7,8\}$　2个为"+"，2个为"−"

$Ent(D^1)=-\log_2\frac{1}{2}=1$

$Ent(D^2)=-\log_2\frac{1}{2}=1$

$Gain(D,C)_5=Ent(D)-\sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v)=0$

⑥ 划分点为6.5

$D^1=\{1,2,3,4,6,9,10\}$ 4个为"+"，3个为"−"

$D^2=\{5,7,8\}$ 1个为"+"，2个为"−"

$Ent(D^1)=-[\frac{4}{7}\log_2\frac{4}{7}+\frac{3}{7}\log_2\frac{3}{7}]=\log_2 7-\frac{3}{7}\log_2 3-\frac{8}{7}$

$Ent(D^2)=-[\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}]=\log_2 3-\frac{2}{3}$

$Gain(D,C)_6=Ent(D)-\sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v)$

$=2-\frac{7}{10}\log_2 7$

⑦ 划分点为7.5

$D^1=\{1,2,3,4,5,6,8,9,10\}$ 5个为"+"，4个为"−"

$D^2=\{7\}$ 为"−"

$Ent(D^1)=-(\frac{4}{9}\log_2\frac{4}{9}+\frac{5}{9}\log_2\frac{5}{9})=-\frac{8}{9}-\frac{5}{9}\log_2 5+2\log_2 3$

$Ent(D^2)=0$

$Gain(D,C)_7=Ent(D)-\sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v)$

$=\frac{9}{5}+\frac{1}{2}\log_2 5-\frac{9}{5}\log_2 3$

(d) 对于属性A:

$Gini(D^1)=1-(\frac{3}{4})^2-(\frac{1}{4})^2=\frac{3}{8}$

$Gini(D^2)=1-(\frac{2}{3})^2-(\frac{1}{3})^2=\frac{4}{9}$

$Gini\_index(D,A)=\frac{3}{8}\times\frac{2}{5}+\frac{4}{9}\times\frac{3}{5}=\frac{5}{12}$

对于属性B:

$Gini(D^1) = 1-[(\frac{2}{5})^2+(\frac{3}{5})^2] = \frac{12}{25}$

$Gini(D^2) = 1-(\frac{2}{5})^2-(\frac{3}{5})^2 = \frac{12}{25}$

$Gini\_index(D,B) = \frac{12}{25}$

∵ $Gini\_index(D,A) < Gini\_index(D,B)$

∴ 属性B为最优划分.

(e) 计算可知. 对C的划分中信息增益最高的为以2.5为划分点

且 $Gain(D,C)_2 \approx 0.23645$

信息增益比 $Gain_R(D,C)_2 = \frac{-\frac{3}{10}\log_2\frac{3}{10}+\frac{1}{2}\log_2 5-\frac{7}{5}}{-\frac{1}{5}\log_2\frac{1}{5}-\frac{4}{5}\log_2\frac{4}{5}} \approx 0.3275$

$Gain_R(D,A) = \frac{\frac{3}{5}-\frac{3}{10}\log_2 3}{-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}} \approx 0.1282$

$Gain_R(D,B) = \frac{\frac{7}{5}+\frac{3}{5}\log_2 3-\log_2 5}{-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}} \approx 0.0291$

故 $Gain_R(D,C)_2 > Gain_R(D,A) > Gain_R(D,B)$

选择以C属性的情况为根结点.

决策树