5.2 Clearly posterior expected loss is Risk where

(a)

$$R(\hat{y}=0|x) = \lambda_{01} P(y=1|x) = \lambda_{01} P_1 \quad P_1 = P(y=1|x)$$

and $R(\hat{y}=0|x) = \lambda_{10} P(y=0|x) = \lambda_{10} P_0$
$= \lambda_{10}(1-P_1)$

So we will predict $\hat{y}=0$

if $R(\hat{y}=0|x) < R(\hat{y}=1|x)$

$$\lambda_{01} P_1 < \lambda_{10}(1-P_1)$$

$$\underline{\underline{P_1 < \frac{\lambda_{10}}{\lambda_{01}+\lambda_{10}} = \theta}}$$

(b) if $\dfrac{\lambda_{10}}{\lambda_{01}+\lambda_{10}} = 0.1 = \dfrac{1}{10} = \dfrac{1}{1+9}$

then $\lambda_{10} = 1$ and $\lambda_{01} = 9$ Note: (Not unique) clearly loss matrix will be

predicted $\hat{y}$	True y: 0	True y: 1
0	0	9
1	1	0

(Note any multiple of 1 and 9 will also give same threshold 0.1

5.3 (9) posterior expected loss/Risk

cost of rejecting is λ_r

cost of picking most probable class is

$j = \arg\max_{\ell} P(y=\ell|x)$ is

$\sum_{i \neq j} \lambda_s P(y=i|x)$ [cost of picking right class is 0]

so pick j if

$$\lambda_r \geq \sum_{i \neq j} \lambda_s P(y=i|x)$$

$$\frac{\lambda_r}{\lambda_s} \geq 1 - P(y=j|x)$$ [probability sum to one]

or $P(y=j|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$

otherwise choose reject.

<u>Note</u> if we decide to choose a class we has to choose

$$j = \arg\max_i P(y=i|x)$$

if we choose other class $K \neq j$ we will incur more cost.

ie cost of choosing K will be

$$\sum_{i \neq K} \lambda_s P(y=i|x) = \lambda_s (1 - P(y=K|x))$$

$$\geq \lambda_s (1 - P(y=j|x))$$

because $j = \arg\max_i P(y=i|x)$

(b) if $\frac{\lambda_r}{\lambda_s} = 0$ there is no cost of rejecting.

as $\frac{\lambda_r}{\lambda_s} \to 1$

$$P(y=j|x) \geq 1-1 \geq 0$$

cost of rejecting increases. Above inequality for most probable class is satisfied more and more. we always accept the most probable class.

Total 32 points

5.10

From section 5.7.2

Follow upto equation 5.114

Pick $\hat{y}=1$

if $$\frac{P(y=1|x)}{P(y=0|x)} > \frac{L_{FP}}{L_{FN}}$$

let $P(y=1|x) = P_1$

then $$\frac{P_1}{1-P_1} > \frac{1}{c} \qquad \because L_{FN} = c L_{FP}$$

$$cP_1 > 1 - P_1$$

$$\Rightarrow P_1 > \frac{1}{1+c}$$

ridge regression derivation

$$\arg\max_{W} \sum_{i=1}^{N} \log \mathcal{N}(y_i | w_0 + W^T x_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j | 0, \tau^2)$$

expression inside arg max will be

$$\sum_{i=1}^{N} \log \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{1}{2} \frac{(y_i - W^T X)^2}{\sigma^2}\right) + \sum_{j=1}^{D} \log \frac{1}{(2\pi)^{1/2}\tau} \exp\left(-\frac{1}{2} (w_i)^2 / \tau^2\right)$$

$$= \sum_{i=1}^{N} \left[\log \frac{1}{(2\pi)^{1/2}\sigma} - \frac{1}{2} \frac{(y_i - W^T X)^2}{\sigma^2}\right] + \sum_{i=1}^{D} \left[\log \frac{1}{(2\pi)^{1/2}\tau} - \frac{1}{2} \frac{w_i^2}{\tau^2}\right]$$

we can ignore constant term for maxmization

$$\simeq -\sum_{i=1}^{N} \frac{(y_i - W^T X)^2}{2\sigma^2} - \sum_{i=0}^{D} \frac{1}{2\tau^2} w_i^2$$

$$= \frac{1}{-2\sigma^2} \left[\sum_{i=1}^{N} (y_i - W^T X)^2 + \frac{\sigma^2}{\tau^2} \sum w_i^2\right]$$

Hence maximizing above objective is same as minimizing

$$= \frac{-1}{2\sigma^2} \left[\sum_{i=1}^{N} (y_i - W^T X)^2 + \lambda \|W\|_2^2\right] \quad \text{where } \lambda = \frac{\sigma^2}{\tau^2}$$

we can ignore +ve constant in front of whole objective as same W will minimize

$$\sum_{i=1}^{N} (y_i - W^T X)^2 + \lambda \|W\|_2^2$$

or

$$\arg\min_{W} \sum_{i=1}^{N} (y_i - W^T X)^2 + \lambda \|W\|_2^2$$

or

$$\arg\min_{W} \frac{1}{N} \sum_{i=1}^{N} (y_i - W^T X)^2 + \lambda \|W\|_2^2$$

Factor of $\frac{1}{N}$ doesn't affect the minimizer w.

7.2 X after encoding is

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}_{6\times 2}$$

$$y = \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}_{6\times 2}$$

compute

$$W = \left[X^T X\right]^{-1} X^T Y$$

using python numpy or matlab or any other matrix multiplication software you should get

$$W = \begin{bmatrix} -4/3 & -4/3 \\ 4/3 & 4/3 \end{bmatrix}$$

7.4 using equation 7.8 we have log likelihood is

$$l(\theta) = l(w, \sigma) = \frac{-1}{2\sigma^2} (y_i - w^T x_i)^2 - \frac{N}{2} \log (2\pi\sigma^2)$$

for ~~max~~ MLE estimate of σ^2 let take derivative of $l(w)$ with respect to σ

$$\frac{d\, l(w,\sigma)}{d\sigma} = \frac{-(-2)}{2\sigma^3} \sum_{i=1}^{N} (y_i - w^T x_i)^2 - \frac{N}{2} \frac{1}{2\pi\sigma^2} (\cancel{2\pi} 2\sigma)$$

At M.L.E estimate this derivative should be zero

$$\sum_{i=1}^{N} \left((y_i - w^T x_i)^2 \right) - N\sigma = 0$$

$$\Rightarrow \sigma = \sum_{i=1}^{N} \frac{(y_i - w^T x_i)^2}{N}$$

$$\hat{\sigma} = \sum_{i=1}^{N} \frac{(y_i - \hat{w}^T x_i)^2}{N}$$

We already know how to get MLE estimate of W.

8.3 (a) use 1D calculus to show

$$\frac{d\, \sigma(a)}{da} = \sigma(a) (1 - \sigma(a))$$

we know

(b) $NLL(w) = -\sum_{i=1}^{N} \left[y_i \log \mu_i + (1-y_i)\log(1-\mu_i) \right]$

where $[\mu_i = \sigma(w^T x_i)]$

Hence $\frac{d\, NLL(w)}{dw} = -\sum_{i=1}^{N} \left[\frac{y_i \cancel{\mu_i}(1-\mu_i) x_i}{\cancel{\mu_i}} + \frac{(1-y_i)(-1)\mu_i \cancel{(1-\mu_i)} x_i}{1-\mu_i} \right]$

$$= -\sum_{i=1}^{N} [y_i - \cancel{y_i\mu_i} - \mu_i + \cancel{y_i\mu_i}] x_i$$

$$= -\sum_{i=1}^{N} [y_i - \mu_i] x_i = \sum_{i=1}^{N} (\mu_i - y_i) x_i$$

$$= X^T (\mu - y)$$

where $X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}_{N \times d}$ $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}_{N \times 1}$ $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N \times 1}$

(c) H is positive definite if for non zero z

$$z^T X^T S X z > 0$$

clearly $z^T X^T S X z$

$$= (Xz)^T S (Xz)$$

as XZ is full rank

then $XZ = y \neq 0$ (zero vector)

then $y^T S y > 0$ as

S is positive definite

Infact $y^T S y$

$$= \sum_{i=1}^{n} \mu_i (1-\mu_i) y_i^2$$

7.9 We have from MLE estimate

$$\hat{\Sigma}_{xx} = \frac{1}{n} \sum_i (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{n} (X-\bar{X})^T (X-\bar{X})$$

$$\hat{\Sigma}_{yx} = \frac{1}{n} (y - \bar{y}1)^T (X - \bar{X})$$

From 4.3.1 we have

$$\underset{y}{\mu}_{1|\underset{x}{2}} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2)$$

or $E[y|x] = \bar{y} + \frac{1}{n} \underbrace{(y - \bar{y}1)^T}_{=y_c} \underbrace{(X-\bar{X})}_{=X_c} \left[\frac{1}{n} \underbrace{(X-\bar{X})^T}_{X_c} \underbrace{(X-\bar{X})}_{X_c}\right]^{-1} (x - \bar{x})$

$$= \bar{y} + y_c^T X_c [X_c^T X_c]^{-1} (x - \bar{x})$$

$$= \bar{y} + \underbrace{y_c^T X_c [X_c^T X_c]^{-1}}_{w^T} x - \underbrace{y_c^T X_c [X_c^T X_c]^{-1}}_{w^T} \bar{x}$$

or

$$E[Y|X] = \underbrace{\bar{y} - W^T x}_{w_0} + W^T X = w_0 + W^T X$$

where $W^T = Y_c^T X_c \left[X_c^T X_c\right]^{-1}$

or $W = \left[X_c^T X_c\right]^{-1} X_c^T Y_c$