Problem 1

a) $E[\mathbb{1}(f(x)\neq y)] = P_\beta(f(x)\neq y)$

$= E[E[\mathbb{1}(f(x)\neq y)|x]]$

$= E[[\mathbb{1}(f(x)\neq 1)]P(y=1|x) + [\mathbb{1}(f(x)\neq 2)]P(y=2|x) + \cdots +$
$[\mathbb{1}(f(x)\neq m)]P(y=m|x) + \cdots + [\mathbb{1}(f(x)\neq k)]P(y=k|x)]$

$= E[\sum_{y_i=1}^{k}[\mathbb{1}(f(x)\neq m)P(y=m|x)]]$

$= E[1-P(y=m|x)]$ when $f(x)=m$, $m\in[k]$

$\Rightarrow \arg\min_f E[\mathbb{1}(f(x)\neq y)] = \arg\max_{m\in[k]} P(y=m|x_i)$

P1

b) take $y_i = j \quad j \in [k]$

Given $P(y_i = j \mid x_i) = \dfrac{\exp(x_i^T \beta_j)}{\sum_{m=1}^{k} \exp(x_i^T \beta_m)}$ is the conditional probability that y_i taking on arbitrary values j

Then $P(y_i \mid x_i) = \dfrac{\exp(x_i^T \beta_{y_i})}{\sum_{m=1}^{k} \exp(x_i^T \beta_m)}$ is the conditional probability model of y_i

log-likelihood
$$\log L(\beta) = \log P(y_i \mid x_i)$$
$$= \log \left(\frac{\exp(x_i^T \beta_{y_i})}{\sum_{m=1}^{k} \exp(x_i^T \beta_m)} \right)$$
$$= x_i^T \beta_{y_i} - \log \sum_{m=1}^{k} \exp(x_i^T \beta_m)$$

c) decision rule :
$$\arg\max_{m \in [k]} P(y = m \mid x_i) = \arg\max_{m \in [k]} \frac{\exp(x_i^T \beta_m)}{\sum_{j=1}^{k} \exp(x_i^T \beta_j)}$$
$$= \arg\max_{m \in [k]} (x_i^T \beta_m)$$

because exp is a monotonically increasing function and $\sum_{j=1}^{k} \exp(x_i^T \beta_j)$ is a constant for example i.

Problem 1

d) Given $\mathbb{P}(y_i | x_i) = \prod_{l=1}^{k} p_l(x_i)^{\mathbb{1}(y_i = l)} = p_{y_i}(x_i)$

Likelihood $L = \prod_{i=1}^{n} P(y_i | x_i) = \prod_{i=1}^{n} p_{y_i}(x_i)$

Negative Log-likelihood $NLL = -\log L = \sum_{i=1}^{n} -\log p_{y_i}(x_i)$

Thus Negative log-likelihood loss of example i is $-\log p_{y_i}(x_i)$

Problem 2

a) if $y=1$ $\quad x = 1\times w + (1-1)\times v = w \sim N(\mu_1, \Sigma_1)$

$$f(x|y=1) = \frac{1}{(2\pi)^d \det(\Sigma_1)^{1/2}} \exp\left[-\frac{(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)}{2}\right]$$

---

b) if $y=0$ $\quad x = 0\times w + (1-0)\times v = v \sim N(\mu_0, \Sigma_0)$

$$f(x|y=0) = \frac{1}{(2\pi)^d \det(\Sigma_0))^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0)}{2}\right]$$

---

c) a) $P(y=1|x) = \dfrac{P(x|y=1)p(y=1)}{P(x|y=1)p(y=1) + P(x|y=0)p(y=0)}$

$$= \frac{f(x|y=1)\pi_1}{f(x|y=1)\pi_1 + f(x|y=0)(1-\pi_1)}$$

plug in $f(x|y=1)$ and $f(x|y=0)$ from a) and b)

$$\Rightarrow P(y=1|x) = \frac{(\det\Sigma_1)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right]\pi_1}{(\det\Sigma_1)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right]\pi_1 + (\det\Sigma_0)^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right](1-\pi_1)}$$

---

b) $\hat{f} = \arg\min_f \mathbb{1}(\eta(x) > 0.5)$

$= \arg\min_f \mathbb{1}[\eta(x) > 1-\eta(x)]$

$= \arg\min_f \mathbb{1}[P(y=1|x) > P(y=0|x)]$

$= \arg\min_f \mathbb{1}[P(y=1|x) > P(y=0|x)]$

solve $\quad P(y=1|x) > P(y=0|x)$

$$\frac{P(x|y=1)P(y=1)}{\cancel{P(x|y=1)P(y=1)+P(x|y=0)P(y=0)}} > \frac{P(x|y=0)p(y=0)}{\cancel{P(x|y=1)P(y=1)+P(x|y=0)P(y=0)}}$$

$$p(x|y=1)P(y=1) > P(x|y=0)P(y=0)$$

$$\left((2\pi)^{\frac{1}{2}}(\det \Sigma_1)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right]\right)\pi_1 > (2\pi)^{-\frac{1}{2}}(\det\Sigma_0)^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right](1-\pi_1)$$

multiply both sides of inequality by log

$$-\frac{1}{2}\log(\det\Sigma_1) - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \log\pi_1 > -\frac{1}{2}\log(\det\Sigma_0) - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_2) + \log(1-\pi_1)$$

$$-(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) > \log\frac{\det\Sigma_1}{\det\Sigma_0} + 2\log\frac{1-\pi_1}{\pi_1}$$

$$\text{left side} = -\left(x^T\Sigma^{-1}x - 2x^T\Sigma_1^{-1}\mu_1 + \mu_1^T\Sigma_1^{-1}\mu_1\right) + \left(x^T\Sigma_0^{-1}x - 2x^T\Sigma_0^{-1}\mu_0 + \mu_0^T\Sigma_0^{-1}\mu_0\right)$$

$$= x^T(-\Sigma_1^{-1}+\Sigma_0^{-1})x + 2x^T(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) - \mu_1^T\Sigma_1^{-1}\mu_1 + \mu_0^T\Sigma_0^{-1}\mu_0$$

$$= x^T(-\Sigma_1^{-1}+\Sigma_0^{-1})x + (2\Sigma_1^{-1}\mu_1 - 2\Sigma_0^{-1}\mu_0)^T x - \mu_1^T\Sigma_1^{-1}\mu_1 + \mu_0^T\Sigma_0^{-1}\mu_0$$

original inequality $\Rightarrow$ $x^T(-\Sigma_1^{-1}+\Sigma_0^{-1})x + (2\Sigma_1^{-1}\mu_1 - 2\Sigma_0^{-1}\mu_0)^T x > \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0 + \log\frac{\det\Sigma_1}{\det\Sigma_0} + 2\log\frac{1-\pi_1}{\pi_1}$

$$(2\Sigma_1^{-1}\mu_1 - 2\Sigma_0^{-1}\mu_0)^T x + x^T(-\Sigma_1^{-1}+\Sigma_0^{-1})x > \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0 + \log\frac{\det\Sigma_1}{\det\Sigma_0} + 2\log\frac{1-\pi_1}{\pi_1}$$

$$\Rightarrow \hat{f} = \mathbb{1}(v^T x + x^T A x > \tau)$$

where
$$\begin{cases} v = 2\Sigma_1^{-1}\mu_1 - 2\Sigma_0^{-1}\mu_0 \\ A = \Sigma_0^{-1} - \Sigma_1^{-1} \\ \tau = \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0 + \log\frac{\det\Sigma_1}{\det\Sigma_0} + 2\log\frac{1-\pi_1}{\pi_1} \end{cases}$$

Problem 2

c) c)

Given $\hat{f} = \mathbb{1}(v^T x + x^T A x > \tau)$ when $x \in R^p$

$$\begin{cases} v = 2\Sigma_1^{-1}\mu_1 - 2\Sigma_0^{-1}\mu_0 \\ A = \Sigma_0^{-1} - \Sigma_1^{-1} \\ \tau = \mu_1^T \Sigma_1^{-1}\mu_1 - \mu_0^T \Sigma_0^{-1}\mu_0 + \log\frac{\det\Sigma_1}{\det\Sigma_0} + 2\log\frac{1-\pi_1}{\pi_1} \end{cases}$$

when $p=1$, $x \in R$

Then $\hat{f} = \mathbb{1}(vx + Ax^2 > \tau) = \mathbb{1}\left(2\left(\frac{\mu_1}{\Sigma_1} - \frac{\mu_0}{\Sigma_0}\right)x + \left(\frac{1}{\Sigma_0} - \frac{1}{\Sigma_1}\right)x^2 > \tau\right)$

$$\begin{cases} v = 2\left(\frac{\mu_1}{\Sigma_1} - \frac{\mu_0}{\Sigma_0}\right) \\ A = \frac{1}{\Sigma_0} - \frac{1}{\Sigma_1} \\ \tau = \frac{\mu_1^2}{\Sigma_1} - \frac{\mu_0^2}{\Sigma_0} + \log\frac{\det\Sigma_1}{\det\Sigma_0} + 2\log\frac{1-\pi_1}{\pi_1} \end{cases}$$

This decision rule depends on inequality of function of x-quadratic

Thus it's called QDA

---

d) denote $\Sigma_1 = \Sigma_0 = \Sigma$

Then
$$\begin{cases} v = 2\Sigma^{-1}\mu_1 - 2\Sigma^{-1}\mu_0 = 2\Sigma^{-1}(\mu_1 - \mu_0) \\ A = 0 \\ \tau = \mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0 + \log\frac{\det\Sigma}{\det\Sigma} + 2\log\frac{1-\pi_1}{\pi_1} \\ \quad = \mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0 + 2\log\frac{1-\pi_1}{\pi_1} \end{cases}$$

Given $\hat{f} = \mathbb{1}(v^T x + x^T A x > \tau)$

$= \mathbb{1}(v^T x > \tau)$

$= \mathbb{1}(\gamma^T x > \varepsilon)$

where
$$\begin{cases} \gamma = v = 2\Sigma^{-1}(\mu_1 - \mu_0) \\ \varepsilon = \tau = \mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0 + 2\log\frac{1-\pi_1}{\pi_1} \end{cases}$$

Problem 2 c) e)

Given $\hat{f} = \mathbb{1}(\gamma^T x > \varepsilon)$

$$\begin{cases} \gamma = 2\Sigma^{-1}(\mu_1 - \mu_0) \\ \varepsilon = \mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0 + 2\log \frac{1-\pi_1}{\pi_1} \end{cases}$$

when $p = 1$, $x \in R$

Then $\hat{f} = \mathbb{1}(\gamma x > \varepsilon)$

where $$\begin{cases} \gamma = 2\Sigma^{-1}(\mu_1 - \mu_0) \\ \varepsilon = (\mu_1^2 - \mu_0^2)\Sigma^{-1} + 2\log\frac{1-\pi_1}{\pi_1} = \gamma \cdot \frac{1}{2}(\mu_1 + \mu_0) + 2\log\frac{1-\pi_1}{\pi_1} \end{cases} \quad \varepsilon \in R$$

This decision rule of input x being in a class y is purely a function of linear combination of known observations, in geometrical perspective, it's a functor of projection of $x \in R^d$ onto vector γ, an observation belongs to class y if its x is located on a certain side of a hyperplane (orthogonal to γ, location is defined by threshold ε)

Problem 3

Part 1 Show that $g^T v = 0$

Define vector $g = (x_i^T w) w$ and vector $e = x_i - g$

$$g^T v = [(x_i^T w) w]^T v$$

$$= (x_i^T w) w^T v$$

Given $\langle v, w \rangle = 0$

then $w^T v = 0$

then $g^T v = (x_i^T w) \cdot 0 = 0$

Thus for any vector $v \in H$, vector g is orthogonal to vector v

Part 2 Show that $e^T w = 0$

$$e^T w = (x_i - g)^T w$$

$$= (x_i^T - g^T) w$$

$$= x_i^T w - g^T w$$

$$= x_i^T w - [(x_i^T w) w]^T w$$

$$= x_i^T w - (x_i^T w) w^T w$$

$$= (x_i^T w)(1 - w^T w)$$

$$= (x_i^T w)(1 - \|w\|_2)$$

Since we assume $\|w\|_2 = 1$

Thus $e^T w = (x_i^T w) \cdot 0 = 0$

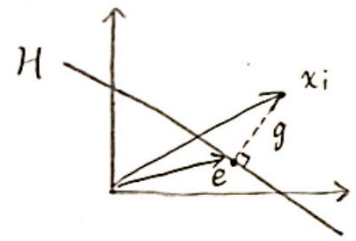
Given set $H = \{v \mid \langle v, w \rangle = 0\} = \{v \mid v^T w = 0\}$

Thus $e \in H$

Part 3 Show that $\|v - x_i\|_2^2 = \|v-e\|_2^2 + \|g\|_2^2$

$$\begin{aligned}
\text{left side} &= \|v - e - g\|_2^2 \\
&= \|(v-e) - g\|_2^2 \\
&= [(v-e)-g]^T [(v-e)-g] \\
&= [(v-e)^T - g^T] [(v-e)-g] \\
&= (v-e)^T(v-e) - (v-e)^T g - g^T(v-e) + g^T g \\
&= (v-e)^T(v-e) + g^T g \\
&= \|v-e\|_2^2 + \|g\|_2^2 \\
&= \text{right side}
\end{aligned}$$

Part 4



Since $e = x_i - g$ $\quad g \perp H$, $e \in H$

Then vector e is the projection of vector x_i onto hyperplane H

vector g is the minimum length from vector x_i to hyperplane H

vector g is parallel to vector w, so $g = \alpha w$ for some $\alpha \in R$

Given $e^T w = 0$, $e = x_i - g$

then $e^T w = (x_i - g)^T w = (x_i - \alpha w)^T w = x_i^T w - \alpha w^T w = 0$

Thus $\alpha = \dfrac{x_i^T w}{w^T w} = \dfrac{x_i^T w}{\|w\|_2^2}$

Thus $\|g\|_2^2 = \|\alpha w\|_2^2$

$$= (\alpha w)^T(\alpha w)$$

$$= \alpha^2 w^T w$$

$$= \frac{(x_i^T w)^2}{\|w\|_2^2} \cdot \|w\|_2^2$$

$$= (x_i^T w)^2$$

Thus $\delta_i^2 = \min\limits_{v|v \in H} \|v-e\|_2^2 + \|g\|_2^2$

$$= \min_{v|v \in H} \|v-e\|_2^2 + (x_i^T w)^2$$

when vector v = vector e, δ_i^2 gets the minimum

$$\delta_i^2 = (x_i^T w)^2$$

From the definition of margin of example i, we know

$$\delta_i = \frac{y_i \langle x_i, w \rangle}{\|w\|_2}$$

then $\delta_i^2 = \dfrac{y_i^2 (x_i^T w)^2}{\|w\|_2^2}$

Given $y_i \in \{-1, 1\}$ then $y_i = 1$

also we assume $\|w\|_2^2 = 1$

Then $\delta_i^2 = (x_i^T w)^2 = \min\limits_{v|v \in H} \|v-e\|_2^2 + \|g\|_2^2 = \min\limits_{v|v \in H} \|v - x_i\|_2^2$