ML2024 Fall Assignment 1

Wei-Chen Chang, R12227118

Due: Sep, 27 2024

Problem 1

(a) (i) Let $\mathbf{x} = \begin{pmatrix} x_1 & ... & x_n \end{pmatrix}^\top$ and $\mathbf{a} = \begin{pmatrix} a_1 & ... & a_n \end{pmatrix}^\top$, and

$$\begin{aligned}
\frac{\partial \|\mathbf{x}-\mathbf{a}\|_2}{\partial x_i} &= \frac{\partial}{\partial x_i}[(\mathbf{x}-\mathbf{a})^\top(\mathbf{x}-\mathbf{a})]^{1/2} \\
&= \frac{\partial}{\partial x_i}[\sum_{i=1}^{n}(x_i-a_i)^2]^{1/2} \\
&= \frac{1}{2}\frac{1}{\|\mathbf{x}-\mathbf{a}\|_2}\cdot 2(x_i-a_i) \\
&= \frac{(x_i-a_i)}{\|\mathbf{x}-\mathbf{a}\|_2}
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\partial \|\mathbf{x}-\mathbf{a}\|_2}{\partial \mathbf{x}} &= \begin{pmatrix} \frac{\partial}{\partial x_1}\|\mathbf{x}-\mathbf{a}\|_2 & ... & \frac{\partial}{\partial x_n}\|\mathbf{x}-\mathbf{a}\|_2 \end{pmatrix}^\top \\
&= \frac{1}{\|\mathbf{x}-\mathbf{a}\|_2}\begin{pmatrix} x_1-a_1 & ... & x_n-a_n \end{pmatrix}^\top \\
&= \frac{\mathbf{x}-\mathbf{a}}{\|\mathbf{x}-\mathbf{a}\|_2}.
\end{aligned}$$

(ii) Note that

$$\frac{\partial \mathbf{a}^\top\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial \mathbf{a}^\top\mathbf{X}\mathbf{b}}{\partial x_{11}} & ... & \frac{\partial \mathbf{a}^\top\mathbf{X}\mathbf{b}}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{a}^\top\mathbf{X}\mathbf{b}}{\partial x_{m1}} & ... & \frac{\partial \mathbf{a}^\top\mathbf{X}\mathbf{b}}{\partial x_{mn}} \end{pmatrix}$$

Observe that $\mathbf{a}^\top\mathbf{X}\mathbf{b} = a_1\mathbf{X_1}^\top\mathbf{b} + ... + a_m\mathbf{X_m}^\top\mathbf{b}$, where $\mathbf{X_i}$ is the i-th column of $\mathbf{X}$, and $\mathbf{X_i}^\top\mathbf{b} = x_{i1}b_1 + ... + x_{in}b_n$.
Thus,

$$\frac{\partial \mathbf{a}^\top\mathbf{X}\mathbf{b}}{\partial x_{ij}} = a_ib_j$$

and

$$\frac{\partial \mathbf{a}^\top\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \begin{pmatrix} a_1b_1 & ... & a_1b_n \\ \vdots & \ddots & \vdots \\ a_mb_1 & ... & a_mb_n \end{pmatrix} = \mathbf{a}\mathbf{b}^\top$$

(b) By cofactor expansion $\det(\mathbf{X}) = \sum_{j=1}^{n} x_{ij} C_{ij}$. So,

$$\begin{aligned}
\frac{\partial \det(\mathbf{X})}{\partial x_{ij}} &= \frac{\partial \sum_{k=1}^{n} x_{ik} C_{ik}}{\partial x_{ij}} \\
&= \sum_{k=1}^{n} \frac{\partial x_{ik} C_{ik}}{\partial x_{ij}} \\
&= \sum_{k=1}^{n} \frac{\partial x_{ik}}{\partial x_{ij}} C_{ik} + \sum_{k=1}^{n} \frac{\partial C_{ik}}{\partial x_{ij}} x_{ik} \\
&= C_{ij} + \sum_{k=1}^{n} 0 \cdot x_{ik}
\end{aligned}$$

Note that x_{ij} is not included in expansion of C_{ik}, so $\frac{\partial C_{ik}}{\partial x_{ij}} = 0$.
Thus,

$$\begin{aligned}
\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} &= \begin{pmatrix} C_{11} & \ldots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \ldots & C_{nn} \end{pmatrix} \\
&= \operatorname{adj}^{\top}(\mathbf{X}) \\
&= \det(\mathbf{X})(\mathbf{X}^{-1})^{\top} \quad (\text{Since } \mathbf{X}\operatorname{adj}(\mathbf{X}) = \det(\mathbf{X}))
\end{aligned}$$

(c) By cofactor expansion, $\det(\mathbf{A}) = \sum_{j=1}^{m} a_{ij} C_{ij}$. So,

$$\begin{aligned}
\frac{\partial \log(\det(\mathbf{A}))}{\partial a_{ij}} &= \frac{\partial}{\partial a_{ij}} \log\left(\sum_{j=1}^{m} a_{ij} C_{ij}\right) \\
&= \frac{1}{\sum_{k=1}^{m} a_{ik} C_{ik}} \cdot \frac{\partial}{\partial a_{ij}} \sum_{k=1}^{m} a_{ik} C_{ik} \\
&= \frac{1}{\det(\mathbf{A})} C_{ij} \\
&= \mathbf{e_j}^{\top} \mathbf{A^{-1}} \mathbf{e_i}
\end{aligned}$$

Note that $\mathbf{A^{-1}} = \frac{1}{\det(\mathbf{A})}\operatorname{adj}(\mathbf{A}) = \frac{1}{\det(\mathbf{A})} C^{\top}$. Thus, $\frac{1}{\det(\mathbf{A})} C_{ij}$ is the j-i-th element of $\mathbf{A}^{-1}$, that is, $\mathbf{e_j}^{\top} \mathbf{A^{-1}} \mathbf{e_i}$.

Problem 2

(a) (i) Let $A_1 : \{i | y_i = C_1\}, A_2 : \{j | y_j = C_2\}$ be the sets of indices which label is C_1/C_2, and the number of elements in each set be $|A_1| = N_1, |A_2| = N_2$, where $N_1 + N_2 = N$.

The likelihood function $L(\theta)$ can be expressed as:

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{N} \mathbb{P}_\theta(X = \mathbf{x_i}, Y = y_i) \\
&= \prod_{i\in A_1} \mathbb{P}_\theta(X = \mathbf{x_i} | Y = C_1) \prod_{j\in A_2} \mathbb{P}_\theta(X = \mathbf{x_j} | Y = C_2) \\
&= \prod_{i\in A_1} \pi_1 f_{\boldsymbol{\mu}_1,\Sigma_1}(\mathbf{x_i}) \prod_{j\in A_2} \pi_2 f_{\boldsymbol{\mu}_2,\Sigma_2}(\mathbf{x_j}) \\
&= \pi_1^{N_1}\pi_2^{N_2}(2\pi)^{-\frac{Nd}{2}}|\Sigma_1|^{-\frac{N_1}{2}}|\Sigma_2|^{-\frac{N_2}{2}} \\
&\quad \times \exp\left\{ -\frac{1}{2}\left[\sum_{i\in A_1} (\mathbf{x_i} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1} (\mathbf{x_i} - \boldsymbol{\mu}_1) + \sum_{j\in A_2} (\mathbf{x_j} - \boldsymbol{\mu}_2)^\top \Sigma_2^{-1} (\mathbf{x_j} - \boldsymbol{\mu}_2) \right] \right\}
\end{aligned}
$$

(ii) let $\ell(\theta) = -\log L(\theta)$. It is equivalent to minimize $\ell(\theta)$ and to maximize $L(\theta)$.

$$
\begin{aligned}
\ell(\theta) &= -\log L(\theta) \\
&= -N_1 \log \pi_1 - N_2 \log \pi_2 + \frac{Nd}{2}\log(2\pi) + \frac{N_1}{2}\log|\Sigma_1| + \frac{N_2}{2}\log|\Sigma_2| \\
&\quad + \frac{1}{2}\left[\sum_{i\in A_1} (\mathbf{x_i} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1} (\mathbf{x_i} - \boldsymbol{\mu}_1) + \sum_{j\in A_2} (\mathbf{x_j} - \boldsymbol{\mu}_2)^\top \Sigma_2^{-1} (\mathbf{x_j} - \boldsymbol{\mu}_2) \right]
\end{aligned}
$$

To minimize $\ell(\theta)$, we need to find its first derivative with respect to all parameters.

For π_1, π_2, since that $\sum_{k\in\{1,2\}} \int_X \pi_k f_{\mu_k,\Sigma_k}(x) = 1 = \pi_1 + \pi_2, \pi_2 = 1 - \pi_2$, we have:

$$
\begin{aligned}
&\frac{\partial \ell(\theta)}{\partial \pi_1} = -N_1 \frac{1}{\pi_1} - N_2 \frac{-1}{(1-\pi_1)} \\
\text{Set} \quad &\frac{\partial \ell(\theta)}{\partial \pi_1} = 0, \\
\implies \; &N_2 \pi_1 = (1-\pi_1) N_1 \\
\implies \quad &\pi_1^* = \frac{N_1}{N_1 + N_2} = \frac{N_1}{N}, \quad \text{and} \quad \pi_2^* = 1 - \pi_1^* = \frac{N_2}{N}
\end{aligned}
$$

Note that $\frac{\partial^2 \ell(\theta)}{\partial \pi_1^2} > 0$, thus π_1^*, π_2^* are critical points to attain minimum.

For the derivative with respect to $\boldsymbol{\mu}_1$:

$$
\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \boldsymbol{\mu}_1} &= \frac{1}{2}\sum_{i\in A_1}\frac{\partial}{\partial \boldsymbol{\mu}_1}(\mathbf{x_i}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x_i}-\boldsymbol{\mu}_1) \\
&= \frac{1}{2}\sum_{i\in A_1}\frac{\partial \boldsymbol{\mu}_1^\top \Sigma_1^{-1}\boldsymbol{\mu}_1}{\partial \boldsymbol{\mu}_1} - 2\frac{\partial \mathbf{x_i}^\top \Sigma_1^{-1}\boldsymbol{\mu}_1}{\partial \boldsymbol{\mu}_1} \\
&= \frac{1}{2}\sum_{i\in A_1} 2\Sigma_1^{-1}\boldsymbol{\mu}_1 - 2\Sigma_1^{-1}\mathbf{x_i} \\
&= \Sigma_1^{-1}\left(N_1\boldsymbol{\mu}_1 - \sum_{i\in A_1}\mathbf{x_i}\right). \\
\text{Set}\quad \frac{\partial \ell(\theta)}{\partial \boldsymbol{\mu}_1} &= \Sigma_1^{-1}\left(N_1\boldsymbol{\mu}_1 - \sum_{i\in A_1}\mathbf{x_i}\right) = 0 \\
\implies \boldsymbol{\mu}_1^* &= \frac{\sum_{i\in A_1}\mathbf{x_i}}{N_1} = \bar{\mathbf{x}}_1
\end{aligned}
$$

Note that $\frac{\partial^2 \ell(\theta)}{\partial \boldsymbol{\mu}_1^2} = N_1\Sigma_1^{-1}$, Since Σ_1^{-1} is positive semi-definite, $\boldsymbol{\mu}_1^*$ is a critical point that yields minimum. Similarly, we can get:

$$
\boldsymbol{\mu}_2^* = \frac{\sum_{i\in A_2}\mathbf{x_i}}{N_2} = \bar{\mathbf{x}}_2
$$

For Σ_1, consider:

$$
\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \Sigma_1} &= \frac{N_1}{2}(\Sigma_1^{-1})^\top + \frac{1}{2}\sum_{i\in A_1}\frac{\partial (\mathbf{x_i}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x_i}-\boldsymbol{\mu}_1)}{\partial \Sigma_1^{-1}}\frac{\partial \Sigma_1^{-1}}{\partial \Sigma_1} \\
&= \frac{N_1}{2}\Sigma_1^{-1} + \frac{1}{2}\sum_{i\in A_1}(\mathbf{x_i}-\boldsymbol{\mu}_1)(\mathbf{x_i}-\boldsymbol{\mu}_1)^\top \cdot \left(-\Sigma_1^{-1}\frac{\partial \Sigma_1}{\partial \Sigma_1}\Sigma_1^{-1}\right) \\
&= \frac{N_1}{2}\Sigma_1^{-1} - \frac{\left(\Sigma_1^{-1}\right)^2}{2}\sum_{i\in A_1}(\mathbf{x_i}-\boldsymbol{\mu}_1)(\mathbf{x_i}-\boldsymbol{\mu}_1)^\top \\
\text{Set}\quad \frac{\partial \ell(\theta)}{\partial \Sigma_1} &= 0, \\
\implies N_1 &= \Sigma_1^{-1}\sum_{i\in A_1}(\mathbf{x_i}-\boldsymbol{\mu}_1)(\mathbf{x_i}-\boldsymbol{\mu}_1)^\top \\
\implies \Sigma_1^* &= \frac{1}{N_1}\sum_{i\in A_1}(\mathbf{x_i}-\boldsymbol{\mu}_1)(\mathbf{x_i}-\boldsymbol{\mu}_1)^\top
\end{aligned}
$$

Similarly, $\Sigma_2^* = \frac{1}{N_2}\sum_{i\in A_2}(\mathbf{x_i}-\boldsymbol{\mu}_2)(\mathbf{x_i}-\boldsymbol{\mu}_2)^\top$

(iii) $\mathbb{P}_\theta(Y = C_1|X = \mathbf{x})$ is the probability of a data point is labelled C_1 (or generated from

$\mathbb{P}_\theta[X = \mathbf{x}, Y = C_1]$), given its feature is $\mathbf{x}$.

$$
\begin{aligned}
\mathbb{P}_\theta(Y = C_1 | X = \mathbf{x}) &= \frac{\mathbb{P}_\theta(X = \mathbf{x}, Y = C_1)}{\sum_{k \in \{1,2\}} \mathbb{P}_\theta(X = \mathbf{x}, Y = C_k)} \\
&= \frac{\pi_1 f_{\boldsymbol{\mu}_1, \Sigma_1}(\mathbf{x})}{\pi_1 f_{\boldsymbol{\mu}_1, \Sigma_1}(\mathbf{x}) + \pi_2 f_{\boldsymbol{\mu}_2, \Sigma_2}(\mathbf{x})} \\
&= \pi_1 |\Sigma_1|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right] \\
&\quad \times \left\{\pi_1 |\Sigma_1|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right]\right. \\
&\qquad \left. + \pi_2 |\Sigma_2|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \Sigma_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right]\right\}^{-1}
\end{aligned}
$$

And $\mathbb{P}_\theta(X = \mathbf{x} | Y = C_1)$ is the probability of a data point having feature $\mathbf{x}$, given it's labelled C_1.

$$
\begin{aligned}
\mathbb{P}_\theta(X = \mathbf{x} | Y = C_1) &= \frac{\mathbb{P}_\theta(X = \mathbf{x}, Y = C_1)}{\mathbb{P}_\theta(Y = C_1)} \\
&= \frac{\mathbb{P}_\theta(X = \mathbf{x}, Y = C_1)}{\int_X \mathbb{P}_\theta(X = \mathbf{x}, Y = C_1)} \\
&= \frac{\pi_1 f_{\mu_1, \Sigma_1}(\mathbf{x})}{\int_X \pi_1 f_{\mu_1, \Sigma_1}(\mathbf{x})} \\
&= \frac{\pi_1 f_{\mu_1, \Sigma_1}(\mathbf{x})}{\pi_1 \times 1} \\
&= f_{\mu_1, \Sigma_1}(\mathbf{x})
\end{aligned}
$$

(iv) From result of (iii), one can divide both the numerator and the denominator by $\pi_1 \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right]$. So,

$$
\begin{aligned}
\mathbb{P}_\theta(Y = C_1 | X = \mathbf{x}) &= \frac{1}{1 + \frac{\pi_2}{\pi_1}\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)^{\frac{-1}{2}} \frac{\exp\left\{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_2)^\top \Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]\right\}}{\exp\left\{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]\right\}}} \\
&= (1 + \exp(-z))^{-1} = \sigma(z)
\end{aligned}
$$

where $z = \log\frac{\pi_1}{\pi_2} - \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|} - \frac{1}{2}\left[(\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)^\top \Sigma_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right]$, or,

$$
\log\frac{\pi_1}{\pi_2} - \frac{1}{2}\left[\mathbf{x}^\top(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x} - 2(\boldsymbol{\mu}_1^\top \Sigma_1^{-1} - \boldsymbol{\mu}_2^\top \Sigma_2^{-1})\mathbf{x} + \boldsymbol{\mu}_1^\top \Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \Sigma_2^{-1}\boldsymbol{\mu}_2 + \log\frac{|\Sigma_1|}{|\Sigma_2|}\right].
$$

(b) (i)

$$
\begin{aligned}
L(\vartheta) &= \prod_{i=1}^{N} \mathbb{P}_\vartheta(X = \mathbf{x_i}, Y = y_i) \\
&= \pi_1^{N_1} \pi_2^{N_2} (2\pi)^{-\frac{Nd}{2}} |\Sigma|^{-\frac{N}{2}} \\
&\quad \times \exp\left\{ -\frac{1}{2} \left[\sum_{i\in A_1} (\mathbf{x_i} - \boldsymbol{\mu}_1)^\top \Sigma^{-1} (\mathbf{x_i} - \boldsymbol{\mu}_1) + \sum_{j\in A_2} (\mathbf{x_j} - \boldsymbol{\mu}_2)^\top \Sigma^{-1} (\mathbf{x_j} - \boldsymbol{\mu}_2) \right] \right\}
\end{aligned}
$$

(ii) The log-likelihood function becomes:

$$
\begin{aligned}
\ell(\vartheta) &= -N_1 \log \pi_1 - N_2 \log \pi_2 + \frac{Nd}{2} \log(2\pi) + \frac{N}{2} \log |\Sigma| \\
&\quad + \frac{1}{2} \left[\sum_{i\in A_1} (\mathbf{x_i} - \boldsymbol{\mu}_1)^\top \Sigma^{-1} (\mathbf{x_i} - \boldsymbol{\mu}_1) + \sum_{j\in A_2} (\mathbf{x_j} - \boldsymbol{\mu}_2)^\top \Sigma^{-1} (\mathbf{x_j} - \boldsymbol{\mu}_2) \right]
\end{aligned}
$$

$\pi_1^*, \pi_2^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*$ are same as those in (a). The arguments are similar thus omitted here. For Σ, consider:

$$
\begin{aligned}
\frac{\partial \ell(\vartheta)}{\partial \Sigma} &= \frac{N}{2} (\Sigma^{-1})^\top \\
&\quad + \frac{1}{2} \left[\sum_{i\in A_1} \frac{\partial (\mathbf{x_i} - \boldsymbol{\mu}_1)^\top \Sigma^{-1} (\mathbf{x_i} - \boldsymbol{\mu}_1)}{\partial \Sigma} + \sum_{j\in A_2} \frac{\partial (\mathbf{x_j} - \boldsymbol{\mu}_2)^\top \Sigma^{-1} (\mathbf{x_j} - \boldsymbol{\mu}_2)}{\partial \Sigma} \right] \\
&= \frac{N}{2} \Sigma^{-1} - \frac{(\Sigma^{-1})^2}{2} \left[\sum_{i\in A_1} (\mathbf{x_i} - \boldsymbol{\mu}_1)(\mathbf{x_i} - \boldsymbol{\mu}_1)^\top + \sum_{j\in A_2} (\mathbf{x_j} - \boldsymbol{\mu}_2)(\mathbf{x_j} - \boldsymbol{\mu}_2)^\top \right]
\end{aligned}
$$

Set $\quad \dfrac{\partial \ell(\vartheta)}{\partial \Sigma} = 0,$

$$
\implies N_1 = \Sigma^{-1} \left[\sum_{i\in A_1} (\mathbf{x_i} - \boldsymbol{\mu}_1)(\mathbf{x_i} - \boldsymbol{\mu}_1)^\top + \sum_{j\in A_2} (\mathbf{x_j} - \boldsymbol{\mu}_2)(\mathbf{x_j} - \boldsymbol{\mu}_2)^\top \right]
$$

$$
\implies \Sigma^* = \frac{1}{N} \left[\sum_{i\in A_1} (\mathbf{x_i} - \boldsymbol{\mu}_1)(\mathbf{x_i} - \boldsymbol{\mu}_1)^\top + \sum_{j\in A_2} (\mathbf{x_j} - \boldsymbol{\mu}_2)(\mathbf{x_j} - \boldsymbol{\mu}_2)^\top \right]
$$

(iii)

$$
\begin{aligned}
\mathbb{P}_\vartheta(Y = C_1 | X = \mathbf{x}) &= \pi_1 |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) \right] \\
&\quad \times \left\{ \pi_1 |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) \right] \right. \\
&\quad \left. + \pi_2 |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_2)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \right] \right\}^{-1}
\end{aligned}
$$

And,

$$\mathbb{P}_\vartheta(X = \mathbf{x}|Y = C_1) = \frac{\pi_1 f_{\mu_1,\Sigma}(\mathbf{x})}{\pi_1 \times 1} = f_{\mu_1,\Sigma}(\mathbf{x})$$

(iv) Set $\Sigma_1 = \Sigma_2 = \Sigma$, let $z' = \log\frac{\pi_1}{\pi_2} - \frac{1}{2}\left[-2(\boldsymbol{\mu}_1^\top - \boldsymbol{\mu}_2^\top)\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_1^\top\Sigma^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top\Sigma^{-1}\boldsymbol{\mu}_2\right]$, and

$$\mathbb{P}_\vartheta(Y = C_1|X = \mathbf{x}) = \sigma(z') = (1 + \exp(-z'))^{-1}$$

Problem 3

(a)

$$\begin{aligned}
L(\boldsymbol{\theta}) &= \sum_i \kappa_i (y_i - \mathbf{X_i}\theta)^2 + \lambda \sum_j w_j^2 \\
&= (\mathbf{y} - \mathbf{X}\theta)^\top \mathbf{K}(\mathbf{y} - \mathbf{X}\theta) + \boldsymbol{\theta}^\top \boldsymbol{(\lambda I)\theta} \\
&= \mathbf{y}^\top\mathbf{K}\mathbf{y} - 2\boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{K}\boldsymbol{y} + \boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{K}\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{(\lambda I)\theta}
\end{aligned}$$

Let $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that

$$\begin{aligned}
L(\boldsymbol{\theta}) &= (\theta - \theta^*)^\top(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})(\theta - \theta^*) + \Delta \\
&= \boldsymbol{\theta}^\top\boldsymbol{(X^\top KX + \lambda I)\theta} - 2\boldsymbol{\theta}^\top\boldsymbol{(X^\top KX + \lambda I)\theta^*} + {\theta^*}^\top(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})\theta^* + \Delta.
\end{aligned}$$

To maintain the equivalence, $\boldsymbol{\theta}^\top\mathbf{X}^\top\mathbf{K}\mathbf{y}$ should equal to $\boldsymbol{\theta}^\top(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\theta}^*$.
That is, $\mathbf{X}^\top\mathbf{K}\mathbf{y} = (\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\theta}^*$, or

$$\boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{K}\mathbf{y}.$$

Also,

$$\begin{aligned}
\Delta &= \mathbf{y}^\top\mathbf{K}\mathbf{y} - {\boldsymbol{\theta}^*}^\top(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\theta}^* \\
&= \mathbf{y}^\top\mathbf{K}\mathbf{y} - \left[(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{K}\mathbf{y}\right]^\top(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{K}\mathbf{y} \\
&= \mathbf{y}^\top\mathbf{K}\mathbf{y} - \mathbf{y}^\top\mathbf{K}\mathbf{X}(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{K}\mathbf{y}.
\end{aligned}$$

Plug in Δ, we have

$$\begin{aligned}
L(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top&(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&+ \left[\mathbf{y}^\top\mathbf{K}\mathbf{y} - \mathbf{y}^\top\mathbf{K}\mathbf{X}(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{K}\mathbf{y}\right].
\end{aligned}$$

Because the last term is irrelevant to θ, it suffices to show that $(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})$ is positive semi-definite to prove the minimum happens at $\boldsymbol{\theta} = \boldsymbol{\theta^*}$.

For $\mathbf{v} \in \mathbb{R}^d$, one can see that

$$\mathbf{v}^\top(\mathbf{X}^\top\mathbf{K}\mathbf{X} + \lambda\mathbf{I})\mathbf{v} = (\mathbf{K^{1/2}}\mathbf{X}\mathbf{v})^\top\mathbf{K^{1/2}}\mathbf{X}\mathbf{v} + \lambda\mathbf{v}^\top\mathbf{v} \geq 0,$$

where $\mathbf{K^{1/2}} = \begin{pmatrix} \kappa_1^{1/2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \kappa_n^{1/2} \end{pmatrix}$.

Thus, the minimum of $L(\boldsymbol{\theta})$ happens at $\boldsymbol{\theta}^* = (\mathbf{X}^\top \mathbf{K}\mathbf{X} + \lambda \mathbf{I})^{-\mathbf{1}}\mathbf{X}^\top \mathbf{K}\mathbf{y}$.

(b)

$$\begin{aligned}
L(\boldsymbol{\theta}) &= \sum_i \kappa_i (y_i - \mathbf{X_i}\boldsymbol{\theta})^2 + \lambda \sum_j w_j \\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top \mathbf{K}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \mathbf{w}^\top \lambda I \mathbf{w} \\
&= \mathbf{y}^\top \mathbf{K}\mathbf{y} - 2\boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{K}\mathbf{y} + \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{K}\mathbf{X}\boldsymbol{\theta} + \mathbf{w}^\top \lambda I \mathbf{w} \\
&= \mathbf{y}^\top \mathbf{K}\mathbf{y} - 2[\mathbf{w}^\top\ b][\tilde{\mathbf{X}}\quad \mathbf{e}]^\top (\mathbf{K}\mathbf{y}) + [\mathbf{w}^\top\quad b][\tilde{\mathbf{X}}\quad \mathbf{e}]^\top \mathbf{K}[\tilde{\mathbf{X}}\quad \mathbf{e}][\mathbf{w}^\top\quad b]^\top + \mathbf{w}^\top \lambda I \mathbf{w} \\
&= \mathbf{y}^\top \mathbf{K}\mathbf{y} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{K}\mathbf{y} + \mathbf{w}^\top \tilde{\mathbf{X}}^\top \mathbf{K}\tilde{\mathbf{X}}\mathbf{w} - 2b(\mathbf{e}^\top \mathbf{K}\mathbf{y} - \mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}}\mathbf{w}) + b^2\mathbf{e}^\top \mathbf{K}\mathbf{e} + \mathbf{w}^\top \lambda I \mathbf{w}
\end{aligned}$$

Note that $L(\boldsymbol{\theta})$ is a quadratic form of b and $\mathbf{e}^\top \mathbf{K}\mathbf{e} = \sum_i \kappa_i = \mathrm{Tr}(\mathbf{K})$. Thus $L(\boldsymbol{\theta})$ can be rewrite as:

$$\begin{aligned}
L(\boldsymbol{\theta}) &= \mathrm{Tr}(\mathbf{K})b^2 - 2b(\mathbf{e}^\top \mathbf{K}\mathbf{y} - \mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}}\mathbf{w}) + \Delta \\
&= \mathrm{Tr}(\mathbf{K})\left[b - \frac{1}{\mathrm{Tr}(\mathbf{K})}(\mathbf{e}^\top \mathbf{K}\mathbf{y} - \mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}}\mathbf{w})\right]^2 - \frac{1}{\mathrm{Tr}(\mathbf{K})}(\mathbf{e}^\top \mathbf{K}\mathbf{y} - \mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}}\mathbf{w})^2 + \Delta \\
&= \mathrm{Tr}(\mathbf{K})(b - b^*)^2 - \frac{1}{\mathrm{Tr}(\mathbf{K})}(\mathbf{y}^\top \mathbf{K}\mathbf{e} - \mathbf{w}^\top \tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e})(\mathbf{e}^\top \mathbf{K}\mathbf{y} - \mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}}\mathbf{w}) + \Delta,
\end{aligned}$$

where $\Delta = \mathbf{y}^\top \mathbf{K}\mathbf{y} + \mathbf{w}^\top (\tilde{\mathbf{X}}^\top \mathbf{K}\tilde{\mathbf{X}} + \lambda I)\mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{K}\mathbf{y}$.

Because $\mathrm{Tr}(\mathbf{K}) > 0$, $L(b|\mathbf{w})$ has minimum when $b = b^* = \frac{1}{\mathrm{Tr}(\mathbf{K})}(\mathbf{e}^\top \mathbf{K}\mathbf{y} - \mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}}\mathbf{w})$, now consider $L(\boldsymbol{\theta})$ as a function of $\mathbf{w}$:

$$\begin{aligned}
L(\mathbf{w}) = \mathrm{Tr}(\mathbf{K})(b - b^*)^2 &- \frac{1}{\mathrm{Tr}(\mathbf{K})}(\mathbf{e}^\top \mathbf{K}\mathbf{y} - \mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}}\mathbf{w})^2 \\
&+ \mathbf{y}^\top \mathbf{K}\mathbf{y} + \mathbf{w}^\top (\tilde{\mathbf{X}}^\top \mathbf{K}\tilde{\mathbf{X}} + \lambda I)\mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{K}\mathbf{y} \\
= \mathbf{w}^\top &\left(\tilde{\mathbf{X}}^\top \mathbf{K}\tilde{\mathbf{X}} + \lambda I - \frac{1}{\mathrm{Tr}(\mathbf{K})}(\tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}})\right)\mathbf{w} \\
&- 2\mathbf{w}^\top \left(\tilde{\mathbf{X}}^\top \mathbf{K} - \frac{\mathbf{1}}{\mathrm{Tr}(\mathbf{K})}(\tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K})\right)\mathbf{y} \\
&+ \mathbf{y}^\top \left(\mathbf{K} - \frac{1}{\mathrm{Tr}(\mathbf{K})}\mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K}\right)\mathbf{y} + \mathrm{Tr}(\mathbf{K})(b - b^*)^2 \\
= \mathbf{w}^\top &\mathbf{A}\mathbf{w} - 2\mathbf{w}^\top \mathbf{B}\mathbf{y} + C,
\end{aligned}$$

where $\mathbf{A} = \tilde{\mathbf{X}}^\top \mathbf{K}\tilde{\mathbf{X}} + \lambda I - \frac{1}{\mathrm{Tr}(\mathbf{K})}(\tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}})$, $\mathbf{B} = \tilde{\mathbf{X}}^\top \mathbf{K} - \frac{\mathbf{1}}{\mathrm{Tr}(\mathbf{K})}(\tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K})$, and $C = \mathbf{y}^\top \left(\mathbf{K} - \frac{1}{\mathrm{Tr}(\mathbf{K})}\mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K}\right)\mathbf{y}$.

Let $\mathbf{w}^* \in \mathbb{R}^d$ such that $L(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{A}(\mathbf{w} - \mathbf{w}^*) + D$,

$$\begin{aligned}
L(\mathbf{w}) &= \mathbf{w}^\top \mathbf{A}\mathbf{w} - 2\mathbf{w}^\top \mathbf{A}\mathbf{w}^* + {\mathbf{w}^*}^\top \mathbf{A}\mathbf{w}^* + D \\
&= \mathbf{w}^\top \mathbf{A}\mathbf{w} - 2\mathbf{w}^\top \mathbf{B}\mathbf{y} + C
\end{aligned}$$

To maintain the equivalence, $\mathbf{w}^\top \mathbf{B}\mathbf{y}$ should equal to $\mathbf{w}^\top \mathbf{A}\mathbf{w}^*$, that is,

$$\begin{aligned}
\mathbf{w}^* &= \mathbf{A}^{-\mathbf{1}}\mathbf{B}\mathbf{y} \\
&= \left(\tilde{\mathbf{X}}^\top \mathbf{K}\tilde{\mathbf{X}} + \lambda I - \frac{1}{\mathrm{Tr}(\mathbf{K})}(\tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}})\right)^{-1} \left(\tilde{\mathbf{X}}^\top \mathbf{K} - \frac{\mathbf{1}}{\mathrm{Tr}(\mathbf{K})}(\tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K})\right)\mathbf{y} \\
&= \left(\tilde{\mathbf{X}}^\top \mathbf{K}\tilde{\mathbf{X}} + \lambda I - \frac{1}{\mathrm{Tr}(\mathbf{K})}(\tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}})\right)^{-1} \tilde{\mathbf{X}}^\top \mathbf{K}\left(\mathbf{y} - \frac{\mathbf{1}}{\mathrm{Tr}(\mathbf{K})}\mathbf{e}\mathbf{e}^\top \mathbf{K}\mathbf{y}\right).
\end{aligned}$$

And $D = C - \mathbf{w}^{*\top}\mathbf{A}\mathbf{w}^*$, which is irrelevant to $\mathbf{w}$.

Lastly, we need to show that $\mathbf{A}$ is positive semi-definite to show $\mathbf{w}*$ attain the minimum. Note that $\mathbf{A}$ can be rewrite as $\tilde{\mathbf{X}}^\top(\mathbf{K} - \frac{\mathbf{1}}{\mathrm{Tr}(\mathbf{K})}\mathbf{K}\mathbf{e}\mathbf{e}^\top\mathbf{K})\tilde{\mathbf{X}} + \lambda I$, so it suffice to show that $\mathbf{v}^\top(\mathbf{K} - \frac{1}{\mathrm{Tr}(\mathbf{K})}\mathbf{K}\mathbf{e}\mathbf{e}^\top\mathbf{K})\mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^n$ to show positive semi-definite:

$$\begin{aligned}
\mathbf{v}^\top(\mathbf{K} - \frac{1}{\mathrm{Tr}(\mathbf{K})}\mathbf{K}\mathbf{e}\mathbf{e}^\top\mathbf{K})\mathbf{v} &= \mathbf{v}^\top\mathbf{K}\mathbf{v} - \frac{1}{\mathrm{Tr}\mathbf{K})}\mathbf{v}^\top\mathbf{K}\mathbf{e}\mathbf{e}^\top\mathbf{K}\mathbf{v} \\
&= \sum_{i=1}^n \kappa_i v_i^2 - \frac{1}{\sum_{i=1}^n \kappa_i}\left(\sum_{i=1}^n \kappa_i v_i\right)^2
\end{aligned}$$

Let a X be a r.v. which has the density $f_X(v_i) = \frac{\kappa_i}{\sum_{i=1}^n \kappa_i}$. Thus,

$$\begin{aligned}
\sum_{i=1}^n \kappa_i v_i^2 - \frac{1}{\sum_{i=1}^n \kappa_i}\left(\sum_{i=1}^n \kappa_i v_i\right)^2 &= \frac{1}{\sum_{i=1}^n \kappa_i}\left[\mathbb{E}(X^2) - \mathbb{E}^2(X)\right] \\
&= \frac{1}{\sum_{i=1}^n \kappa_i} Var(X) \geq 0
\end{aligned}$$

Thus $\mathbf{A}$ is positive semi-definite, and $L(\boldsymbol{\theta})$ yields the minimum when $\mathbf{w} = \mathbf{w}^*, b = b^*$, where:

$$\begin{aligned}
\mathbf{w}^* &= \left(\tilde{\mathbf{X}}^\top \mathbf{K}\tilde{\mathbf{X}} + \lambda I - \frac{1}{\mathrm{Tr}(\mathbf{K})}(\tilde{\mathbf{X}}^\top \mathbf{K}\mathbf{e}\mathbf{e}^\top \mathbf{K}\tilde{\mathbf{X}})\right)^{-1} \tilde{\mathbf{X}}^\top \mathbf{K}\left(\mathbf{y} - \frac{\mathbf{1}}{\mathrm{Tr}(\mathbf{K})}\mathbf{e}\mathbf{e}^\top \mathbf{K}\mathbf{y}\right), \\
b^* &= \frac{1}{\mathrm{Tr}(\mathbf{K})}(\mathbf{e}^\top\mathbf{K}\mathbf{y} - \mathbf{e}^\top\mathbf{K}\tilde{\mathbf{X}}\mathbf{w}^*).
\end{aligned}$$

Problem 4

$$
\begin{aligned}
\tilde{L}_{ss}(\mathbf{w}, b) &= \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x_i}+\eta_\mathbf{i}) - y_i)^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[(f_{\mathbf{w},b}(\mathbf{x_i}+\eta_\mathbf{i}) - y_i)^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\mathbf{w}^\top(\mathbf{x_i}+\eta_\mathbf{i}) + b - y_i\right)^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left\{\left[(f_{\mathbf{w},b}(\mathbf{x_i}) - y_i) + \mathbf{w}^\top\eta_\mathbf{i}\right]^2\right\} \\
&= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[(f_{\mathbf{w},b}(\mathbf{x_i}) - y_i)^2 + 2(f_{\mathbf{w},b}(\mathbf{x_i}) - y_i)\mathbf{w}^\top\eta_\mathbf{i} + (\mathbf{w}^\top\eta_\mathbf{i})^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x_i}) - y_i)^2 + 2(f_{\mathbf{w},b}(\mathbf{x_i}) - y_i)\cdot\mathbf{w}^\top\mathbb{E}\left[\eta_\mathbf{i}\right] + \mathbb{E}\left[(\mathbf{w}^\top\eta_\mathbf{i})^2\right]
\end{aligned}
$$

Note that $\mathbb{E}\left[\eta_i\right] = \mathbf{0}$ and $\mathbb{E}\left[\eta_{i,j}\eta_{i',j'}\right] = \begin{cases}\sigma^2, \text{ if } i = i' \text{ and } j = j' \\ 0, \text{ otherwise.}\end{cases}$

So, $\mathbb{E}\left[\eta_i\eta_i^\top\right] = \sigma^2 I$. And,

$$
\begin{aligned}
\mathbb{E}\left[(\mathbf{w}^\top\eta_\mathbf{i})^2\right] &= \mathbb{E}\left[\mathbf{w}^\top\eta_\mathbf{i}\eta_\mathbf{i}^\top\mathbf{w}\right] \\
&= \mathbf{w}^\top\mathbb{E}\left[\eta_\mathbf{i}\eta_\mathbf{i}^\top\right]\mathbf{w} \\
&= \mathbf{w}^\top\left(\sigma^2\mathbf{I}\right)\mathbf{w} \\
&= \sigma^2\mathbf{w}^\top\mathbf{w} = \sigma^2\|\mathbf{w}\|^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\tilde{L}_{ss}(\mathbf{w}, b) &= \frac{1}{2N}\sum_{i=1}^{N}\left[(f_{\mathbf{w},b}(\mathbf{x_i}) - y_i)^2 + \sigma^2\|\mathbf{w}\|^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x_i}) - y_i)^2 + \frac{1}{2N}\cdot N\sigma^2\|\mathbf{w}\|^2 \\
&= \frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x_i}) - y_i)^2 + \frac{\sigma^2}{2}\|\mathbf{w}\|^2
\end{aligned}
$$

This shows the equivalence.

Problem 5

(a)

$$
\ell^{(n)}(\mathbf{w}) = \log\left(1 + \exp(-y_n(\mathbf{w}^\top\mathbf{x_n}))\right).
$$

Let $L(\mathbf{w}) = \mathbb{1}\{\text{sign}(\mathbf{w}^\top \mathbf{x_n}) \neq y_n\}$. When $\text{sign}(\mathbf{w}^\top \mathbf{x_n}) = y_n$, $-y_n(\mathbf{w}^\top \mathbf{x_n}) \leq 0$, it's obvious to see $\ell^{(n)}(\mathbf{w}) \geq 0 = L(\mathbf{w})$, and

$$\frac{1}{\log 2}\ell^{(n)}(\mathbf{w}) \geq L(\mathbf{w}) = 0.$$

When $\text{sign}(\mathbf{w}^\top \mathbf{x_n}) \neq y_n, -y_n(\mathbf{w}^\top \mathbf{x_n}) \geq 0 \implies \exp(-y_n(\mathbf{w}^\top \mathbf{x_n})) \geq 1$. Thus,

$$\ell^{(n)}(\mathbf{w}) \geq \log(1+1) = \log(2) \implies \frac{1}{\log 2}\ell^{(n)}(\mathbf{w}) \geq 1 = L(\mathbf{w})$$

So $\frac{1}{\log 2}\ell^{(n)}(\mathbf{w})$ is an upper bound of $\mathbb{1}\{\text{sign}(\mathbf{w}^\top \mathbf{x_n}) \neq y_n\}$.

(b) Let $\mathbf{w} = \begin{pmatrix} w_1 & ... & w_m \end{pmatrix}^\top$, $\mathbf{x_n} = \begin{pmatrix} x_{n1} & ... & x_{nm} \end{pmatrix}^\top$,

$$\begin{aligned}
\frac{\partial \ell^{(n)}(\mathbf{w})}{\partial w_i} &= \frac{\exp(-y_n(\mathbf{w_n^\top x_n}))}{1+\exp(-y_n(\mathbf{w_n^\top x_n}))}(-y_n x_{ni}) \\
&= \frac{1}{1+\exp(y_n(\mathbf{w_n^\top x_n}))}(-y_n x_{ni}). \\
\nabla \ell^{(n)}(\mathbf{w}) &= \frac{\partial \ell^{(n)}(\mathbf{w})}{\partial \mathbf{w}} \\
&= \frac{-y_n}{1+\exp(y_n(\mathbf{w_n^\top x_n}))}\begin{pmatrix} x_{n1} & \dots & x_{nm} \end{pmatrix}^\top \\
&= \frac{-y_n}{1+\exp(y_n(\mathbf{w_n^\top x_n}))}\mathbf{x_n} \\
&= \begin{cases} -(1+\exp(\mathbf{w_n^\top x_n}))^{-1}\mathbf{x_n}, & \text{if } y_n = +1; \\ (1+\exp(-\mathbf{w_n^\top x_n}))^{-1}\mathbf{x_n}, & \text{if } y_n = -1. \end{cases}
\end{aligned}$$

(c) First notice that:

$$\ell^{(n)}(\mathbf{w}) = \log\Big(1+\exp(-y_n(\mathbf{w}^\top \mathbf{x_n}))\Big) = \begin{cases} \log\big(\exp(-(\mathbf{w}^\top \mathbf{x_n})+1\big) & \text{if } y_n = +1 \\ \log\big(\exp(\mathbf{w}^\top \mathbf{x_n})+1\big) & \text{if } y_n = -1 \end{cases}$$

Next, Let $z_n = \frac{1}{2}\mathbf{w}^\top \mathbf{x}_n$, and note that $1 + \tanh(z) = \frac{2\exp(2z)}{\exp(2z)+1}$, $1 - \tanh(z) = \frac{2}{\exp(2z)+1}$

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= -\sum_{n=1}^{d}\left(\frac{1+y_n}{2}\log\frac{1+\tanh(z_n)}{2} + \frac{1-y_n}{2}\log\frac{1-\tanh(z_n)}{2}\right) \\
&= -\sum_{n=1}^{d}\left(\frac{1+y_n}{2}\log\left(\frac{e^{2z_n}}{e^{2z_n}+1}\right) + \frac{1-y_n}{2}\log\left(\frac{1}{e^{2z_n}+1}\right)\right) \\
&= -\sum_{n=1}^{d}\left[\frac{1+y_n}{2}\left(\mathbf{w}^\top\mathbf{x}_n - \log(e^{2z_n}+1)\right) + \frac{1-y_n}{2}\left(-\log\left(e^{2z_n}+1\right)\right)\right] \\
&= -\sum_{n=1}^{d}\left[\frac{1+y_n}{2}\mathbf{w}^\top\mathbf{x}_n - \log(e^{\mathbf{w}^\top\mathbf{x}_n}+1)\right] \\
&= \sum_{n=1}^{d}\mathcal{L}^{(n)}(\mathbf{w}).
\end{aligned}$$

Where

$$\mathcal{L}^{(n)}(\mathbf{w}) = \begin{cases} -\mathbf{w}^\top\mathbf{x_n} + \log(\exp(\mathbf{w}^\top\mathbf{x_n})+1) & \text{if } y_n = +1, \\ \log(\exp(\mathbf{w}^\top\mathbf{x_n})+1) & \text{if } y_n = -1. \end{cases}$$

Note that

$$\begin{aligned}
-\mathbf{w}^\top\mathbf{x_n} + \log(\exp(\mathbf{w}^\top\mathbf{x_n})+1) &= -\mathbf{w}^\top\mathbf{x_n} + \log(\exp(\mathbf{w}^\top\mathbf{x_n})+1) \\
&\quad - \log(\exp(\mathbf{w}^\top\mathbf{x_n}) + \log(\exp(\mathbf{w}^\top\mathbf{x_n}) \\
&= \log(\exp(-\mathbf{w}^\top\mathbf{x_n})+1)
\end{aligned}$$

Thus $\mathcal{L}^{(n)}(\mathbf{w}) = \ell^{(n)}(\mathbf{w})$.

Since $\frac{1}{d}$ is irrelevant to $\mathbf{w}$, minimizing $\mathcal{L}(\mathbf{w})$ is equivalent to minimize $\ell(\mathbf{w})$.