1. (25 points) Consider a binary classification problem that $\theta \in \{0, 1\}$ denotes the class label, $\boldsymbol{X}|(\theta = 0) \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $\boldsymbol{X}|(\theta = 1) \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, where N_p denotes the p-dimensional multivariate normal distribution. Suppose 0-1 loss is used, and the prior distribution of θ is $P(\theta = 0) = 1/2$ and $P(\theta = 1) = 1/2$.

   (i) (4 points) Derive the Bayes rule for a classifying a new observation $\boldsymbol{x} \in \mathcal{R}^p$.

   (ii) (4 points) Derive the misclassification rate R^* of the Bayes rule.

   (iii) (4 points) Let $\boldsymbol{X}_{0i}$ $(i = 1, \ldots, n_0)$ be independent and identically distributed (i.i.d) samples from the class of $\theta = 0$ and $\boldsymbol{X}_{1i}$ $(i = 1, \ldots, n_1)$ be i.i.d samples from the class of $\theta = 1$, and $\boldsymbol{X}_{0i}$ is independent of $\boldsymbol{X}_{1i}$. Derive the maximum likelihood estimators $(\widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}})$ of $(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$.

   (iv) (4 points) If we replace $(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ in the Bayes rule with $(\widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}})$, prove that the misclassification rate of the resulting rule, i.e., the probability of classifying $\boldsymbol{x}$ to a wrong class given the training data $\{\boldsymbol{X}_{0i}\}_{i=1}^{n_0}$ and $\{\boldsymbol{X}_{1i}\}_{i=1}^{n_1}$, is given by

$$\frac{1}{2}\Phi\left(\frac{\widehat{\boldsymbol{\delta}}^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}})}{\sqrt{\widehat{\boldsymbol{\delta}}^T \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\delta}}}}\right) + \frac{1}{2}\Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}})}{\sqrt{\widehat{\boldsymbol{\delta}}^T \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\delta}}}}\right),$$

   where $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}} = (\widehat{\boldsymbol{\mu}}_0 + \widehat{\boldsymbol{\mu}}_1)/2$.

   (v) (5 points) We propose another classification rule that assigns $\boldsymbol{x}$ to the class of $\theta = 0$ if and only if $\widehat{\boldsymbol{\beta}}^T(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}) \geq 0$, where $\widehat{\boldsymbol{\mu}} = (\widehat{\boldsymbol{\mu}}_0 + \widehat{\boldsymbol{\mu}}_1)/2$ and $\widehat{\boldsymbol{\beta}}$ solves the following problem

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{R}^p}{\operatorname{argmin}} \ \frac{1}{2}\boldsymbol{\beta}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta} - (\widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1)^T \boldsymbol{\beta} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

   Derive the Majorization-Minimization algorithm for solving $\widehat{\boldsymbol{\beta}}$. Give an explicit choice of step size and closed-form expressions on how iterations need to be done.

   (vi) (4 points) Let R_n denote the misclassification rate of the rule described in (v). Suppose we can show that $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ as $n \to \infty$. Using this result to prove $R_n \xrightarrow{P} R^*$.

You may use the following facts.

(a) The density of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $\{(2\pi)^p |\boldsymbol{\Sigma}|\}^{-1/2} \exp\{-(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})/2\}$;

(b) For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{M}$,

$$\frac{\partial \operatorname{tr}(\boldsymbol{A}\boldsymbol{M})}{\partial \boldsymbol{M}} = \frac{\partial \operatorname{tr}(\boldsymbol{M}\boldsymbol{A})}{\partial \boldsymbol{M}} = \boldsymbol{A}.$$

$$\frac{\partial \log |\boldsymbol{M}|}{\partial \boldsymbol{M}} = \boldsymbol{M}^{-1}.$$

Solution:

(i) Under 0-1 loss, the Bayes rule is the posterior mode. Therefore, the Bayes rule assigns $\boldsymbol{x}$ to $\theta = 0$ iff $f(\theta = 0|\boldsymbol{x}) > f(\theta = 1|\boldsymbol{x})$. That is equivalent as $-(1/2)(\boldsymbol{x} - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_0) > -(1/2)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)$. After some algebra, we find that the Bayes rule assigns $\boldsymbol{x}$ to $\theta = 0$ iff $\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) > 0$, where $\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)/2$.

(ii) The misclassification rate of Bayes rule is given by

$$
\begin{aligned}
R^* &= \frac{1}{2}P(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) > 0|\theta = 1) + \frac{1}{2}P(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \le 0|\theta = 0) \\
&= \frac{1}{2}P\left(\frac{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) - \boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu})}{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}} > -\frac{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu})}{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}}\right) \\
&\quad + \frac{1}{2}P\left(\frac{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) - \boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu})}{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}} \le -\frac{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu})}{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}}\right) \\
&= \frac{1}{2}P\left(Z > -\frac{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu})}{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}}\right) + \frac{1}{2}P\left(Z \le -\frac{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu})}{\sqrt{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}}\right) \\
&= \Phi\left(-\frac{1}{2}\sqrt{\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}\right),
\end{aligned}
$$

where $Z \sim N(0, 1)$ and $\Phi(\cdot)$ is the c.d.f of $N(0, 1)$.

(iii) The log-likelihood

$$
\begin{aligned}
\ell(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = &-\frac{1}{2}\sum_{i=1}^{n_0}(\boldsymbol{X}_{0i} - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{0i} - \boldsymbol{\mu}_0) - (n_0p/2)\log(2\pi) - \frac{n_0}{2}\log|\boldsymbol{\Sigma}| \\
&-\frac{1}{2}\sum_{i=1}^{n_1}(\boldsymbol{X}_{1i} - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{1i} - \boldsymbol{\mu}_1) - (n_1p/2)\log(2\pi) - \frac{n_1}{2}\log|\boldsymbol{\Sigma}|
\end{aligned}
$$

We set

$$
\frac{\partial\ell(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{\boldsymbol{\mu}_0} = \sum_{i=0}^{n_0}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{0i} - \boldsymbol{\mu}_0) = \boldsymbol{0}.
$$

Then, $\widehat{\boldsymbol{\mu}}_0 = \bar{\boldsymbol{X}}_0 = (1/n_0)\sum_{i=1}^{n_0}\boldsymbol{X}_{0i}$. Similarly, $\widehat{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{X}}_1 = (1/n_1)\sum_{i=1}^{n_1}\boldsymbol{X}_{1i}$. Note that,

$$
\begin{aligned}
\sum_{i=1}^{n_0}(\boldsymbol{X}_{0i} - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{0i} - \boldsymbol{\mu}_0) &= \sum_{i=1}^{n_0}(\boldsymbol{X}_{0i} - \bar{\boldsymbol{X}}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{0i} - \bar{\boldsymbol{X}}_0) + n_0(\bar{\boldsymbol{X}}_0 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}_0 - \boldsymbol{\mu}_0) \\
&= n_0\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{S}_0 + \boldsymbol{d}_0\boldsymbol{d}_0^T)\},
\end{aligned}
$$

where $\boldsymbol{S}_0 = (1/n_0)\sum_{i=1}^{n_0}(\boldsymbol{X}_{0i} - \bar{\boldsymbol{X}}_0)(\boldsymbol{X}_{0i} - \bar{\boldsymbol{X}}_0)^T$ and $\boldsymbol{d}_0 = \bar{\boldsymbol{X}}_0 - \boldsymbol{\mu}_0$. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, set

$$
\frac{\partial\ell(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}} = -\frac{1}{2}n_0(\boldsymbol{S}_0 + \boldsymbol{d}_0\boldsymbol{d}_0^T) + \frac{1}{2}n_0\boldsymbol{\Sigma} - \frac{1}{2}n_1(\boldsymbol{S}_1 + \boldsymbol{d}_1\boldsymbol{d}_1^T) + \frac{1}{2}n_1\boldsymbol{\Sigma} = \boldsymbol{0}.
$$

Insert $\widehat{\boldsymbol{\mu}}_0$ and $\widehat{\boldsymbol{\mu}}_1$ into the above equation, we have $\widehat{\boldsymbol{\Sigma}} = (1/n)(n_0\boldsymbol{S}_0 + n_1\boldsymbol{S}_1)$, where $n = n_0 + n_1$, $\boldsymbol{S}_1 = (1/n_1)\sum_{i=1}^{n_1}(\boldsymbol{X}_{1i} - \bar{\boldsymbol{X}}_1)(\boldsymbol{X}_{1i} - \bar{\boldsymbol{X}}_1)^T$.

(iv) Use similar argument as in (ii), the misclassification rate is given by

$$\frac{1}{2}\Phi\left(\frac{\widehat{\boldsymbol{\delta}}^T\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}})}{\sqrt{\widehat{\boldsymbol{\delta}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\delta}}}}\right)+\frac{1}{2}\Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^T\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_0-\widehat{\boldsymbol{\mu}})}{\sqrt{\widehat{\boldsymbol{\delta}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\delta}}}}\right),$$

where $\widehat{\boldsymbol{\delta}}=\widehat{\boldsymbol{\mu}}_0-\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}=(\widehat{\boldsymbol{\mu}}_0+\widehat{\boldsymbol{\mu}}_1)/2$.

(v) Suppose $\widetilde{\boldsymbol{\beta}}$ is the solution of $\boldsymbol{\beta}$ at the current iteration. Let $L(\boldsymbol{\beta})=(1/2)\boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}-\widehat{\boldsymbol{\delta}}^T\boldsymbol{\beta}$, we have

$$\begin{aligned}L(\boldsymbol{\beta})&=L(\widetilde{\boldsymbol{\beta}})+\{\nabla L(\widetilde{\boldsymbol{\beta}})\}^T(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})+\frac{1}{2}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})^T\nabla^2L(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})\\&=L(\widetilde{\boldsymbol{\beta}})+(\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\delta}})^T(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})+\frac{1}{2}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})^T\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})\\&\le L(\widetilde{\boldsymbol{\beta}})+(\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\delta}})^T(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})+\frac{c}{2}\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_2^2,\end{aligned}$$

where $c\ge\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})$ and $\bar{\boldsymbol{\beta}}$ is a vector on the line segment connecting $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$. Therefore, we solve

$$\widetilde{\boldsymbol{\beta}}^{(\text{new})}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \frac{c}{2}\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_2^2+(\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\delta}})^T(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})+\lambda\sum_{j=1}^p|\beta_j|.$$

The KKT condition is given by

$$\begin{aligned}c(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})_j+(\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\delta}})_j+\operatorname{sign}(\beta_j)=0,&\ \text{ for } \beta_j\ne 0;\\|c(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})_j+(\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\delta}})_j|<\lambda,&\ \text{ for } \beta_j=0.\end{aligned}$$

Hence, the solution is given by $\widetilde{\beta}_j^{(\text{new})}=s(\widetilde{\beta}_j-(1/c)(\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\delta}})_j,\lambda/c)$, where $s(x,\lambda)=\operatorname{sign}(x)(|x|-\lambda)_+$ is the soft-thresholding function. With stepsize being $1/c$ for any $c\ge\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})$, the algorithm is summarized as follows.

Step 1: Initialize $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^{(0)}$.

Step 2: At the kth iteration, let

$$\boldsymbol{\beta}^{(k)}=s\left(\boldsymbol{\beta}^{(k-1)}-\frac{1}{c}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^{(k-1)}-\widehat{\boldsymbol{\delta}}),\frac{\lambda}{c}\right)$$

where s is the soft-thresholding function defined above.

Update the gradient vector $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^{(k-1)}-\widehat{\boldsymbol{\delta}}$ with $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^{(k)}-\widehat{\boldsymbol{\delta}}$.

Step 3: Iterate until convergence.

(vi) Similar as in (iv), the misclassification rate

$$R_n=\frac{1}{2}\Phi\left(\frac{\widehat{\boldsymbol{\beta}}^T(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}})}{\sqrt{\widehat{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}}}\right)+\frac{1}{2}\Phi\left(-\frac{\widehat{\boldsymbol{\beta}}^T(\boldsymbol{\mu}_0-\widehat{\boldsymbol{\mu}})}{\sqrt{\widehat{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}}}\right).$$

By Law of Large Numbers, $\widehat{\boldsymbol{\mu}}\xrightarrow{P}\boldsymbol{\mu}$. This together with $\widehat{\boldsymbol{\beta}}\xrightarrow{P}\boldsymbol{\beta}$ imply $R_n\xrightarrow{P}R^*$.