STOR 767 Spring 2019 Hw3

Due on 02/18/2019 in Class

Zhenghan Fang

Remark. This homework focuses on linear classification methods.

Instruction.

- Submission of handwritten solution for the **Theoretical Part** of this homework is allowed.
- Please use **RMarkdown** to create a formatted report for the **Computational Part** of this homework.
- Some of the problems are selected or modified from the textbook [1].

Theoretical Part

1. (*15 pt*) (LDA and Least-Squared Rule, Textbook Ex. 4.2) Suppose we have training features $\boldsymbol{X}_i \in \mathbb{R}^p$, a two-class response with class sizes N_1, N_2, $N = N_1 + N_2$, coded as

$$Y_i = \begin{cases} -N/N_1, & \text{the } i\text{-th sample belongs to class 1} \\ N/N_2, & \text{the } i\text{-th sample belongs to class 2} \end{cases}$$

   (a) Derive from scratch that the LDA rule classifies $\boldsymbol{x} \in \mathbb{R}^p$ to class 2 if

$$\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1) > \frac{1}{2}\widehat{\boldsymbol{\mu}}_2^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_2 - \frac{1}{2}\widehat{\boldsymbol{\mu}}_1 \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1 + \log\left(\frac{N_1}{N}\right) - \log\left(\frac{N_2}{N}\right),$$

   and class 1 otherwise. Remember to declare $\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}$.

   Hint. Derive the Bayes rule under multivariate Gaussian with equal covariance assumption for class-conditional densities of $\boldsymbol{X}_i$'s, and plug in training sample estimates for unknown parameters.

   The LDA rule classifies $\boldsymbol{x}$ to class 2 if

$$\begin{aligned} & \log \frac{\Pr(G=1|X=x)}{\Pr(G=2|X=x)} < 0 \\ \iff & \log\frac{\pi_1}{\pi_2} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + x^T\Sigma^{-1}(\mu_1 - \mu_2) < 0 \\ \iff & x^T\Sigma^{-1}(\mu_2 - \mu_1) > \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\pi_1 - \log\pi_2 \end{aligned}$$

   Let $\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}, \frac{N_1}{N}, \frac{N_2}{N}$ be the estimation of $\mu_1, \mu_2, \Sigma, \pi_1, \pi_2$ respectively, then we get the expression in the question.

(b) Consider minimization of the least squares criterion

$$\min_{\beta_0\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{N}(Y_i-\beta_0-\boldsymbol{\beta}^T\boldsymbol{X}_i)^2.$$

Show that the solution $\widehat{\boldsymbol{\beta}}$ satisfies

$$\left((N-2)\widehat{\boldsymbol{\Sigma}}+\frac{N_1N_2}{N}\widehat{\boldsymbol{\Sigma}}_{\mathrm{B}}\right)\widehat{\boldsymbol{\beta}}=N(\widehat{\boldsymbol{\mu}}_2-\widehat{\boldsymbol{\mu}}_1)$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathrm{B}}:=(\widehat{\boldsymbol{\mu}}_2-\widehat{\boldsymbol{\mu}}_1)(\widehat{\boldsymbol{\mu}}_2-\widehat{\boldsymbol{\mu}}_1)^T$.

Hint. Solve $\widehat{\beta}_0(\boldsymbol{\beta})$ for fixed $\boldsymbol{\beta}$ first, then write down the normal equations for $\widehat{\boldsymbol{\beta}}$. Consider an ANOVA-type covariance matrix decomposition for $\boldsymbol{X}_i$'s: SS = SSW (within-class) + SSB (between-class). You might refer to Problem 2(b)(ii) for a more abstract development.

Answer:

Solve $\widehat{\beta}_0(\boldsymbol{\beta})$ for fixed $\boldsymbol{\beta}$.

$$\widehat{\beta}_0=-\boldsymbol{\beta}^T\frac{\sum_{i=1}^{N}\boldsymbol{X}_i}{N}$$

Write down the normal equations for $\widehat{\boldsymbol{\beta}}$.

$$\begin{aligned}&\left(\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)\widehat{\boldsymbol{\beta}}=\sum_{i=1}^{N}(Y_i-\beta_0)\boldsymbol{X}_i\\ \Longleftrightarrow\ &\left(\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)\widehat{\boldsymbol{\beta}}+\beta_0\sum_{i=1}^{N}\boldsymbol{X}_i=\sum_{i=1}^{N}Y_i\boldsymbol{X}_i\end{aligned}$$

Plug $\widehat{\beta}_0(\boldsymbol{\beta})$ in the above equation.

$$\begin{aligned}&\left(\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)\widehat{\boldsymbol{\beta}}-\left(\widehat{\boldsymbol{\beta}}^T\frac{\sum_{i=1}^{N}\boldsymbol{X}_i}{N}\right)\sum_{i=1}^{N}\boldsymbol{X}_i=\sum_{i=1}^{N}Y_i\boldsymbol{X}_i\\ \Longleftrightarrow\ &\left(\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^T-\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\sum_{i=1}^{N}\boldsymbol{X}_i^T\right)\widehat{\boldsymbol{\beta}}=\sum_{i=1}^{N}Y_i\boldsymbol{X}_i.\end{aligned}\tag{a1}$$

By definition,

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_1&=\frac{1}{N_1}\sum_{g_i=1}\boldsymbol{X}_i\\ \widehat{\boldsymbol{\mu}}_2&=\frac{1}{N_2}\sum_{g_i=2}\boldsymbol{X}_i\end{aligned}$$

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}&=\frac{1}{N-2}\left(\sum_{g_i=1}(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}_1)(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}_1)^T+\sum_{g_i=2}(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}_2)(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}_2)^T\right)\\ &=\frac{1}{N-2}\left(\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^T-N_1\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\mu}}_1^T-N_2\widehat{\boldsymbol{\mu}}_2\widehat{\boldsymbol{\mu}}_2^T\right)\\ \Longleftrightarrow\ &\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^T=(N-2)\widehat{\boldsymbol{\Sigma}}+N_1\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\mu}}_1^T+N_2\widehat{\boldsymbol{\mu}}_2\widehat{\boldsymbol{\mu}}_2^T\end{aligned}\tag{a2}$$

$$\sum_{i=1}^{N} \boldsymbol{X}_i = N_1\hat{\boldsymbol{\mu}}_1 + N_2\hat{\boldsymbol{\mu}}_2 \tag{a3}$$

$$\sum_{i=1}^{N} Y_i\boldsymbol{X}_i = \sum_{g_i=1} -N/N_1\boldsymbol{X}_i + \sum_{g_i=2} N/N_2\boldsymbol{X}_2 = N(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \tag{a4}$$

Plug (a2), (a3), (a4) into (a1), we get the expression in the question.

(c) Argue that

$$\hat{\boldsymbol{\beta}} \propto \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1), \tag{1}$$

i.e. the least squares coefficient is identical to the LDA coefficient up to a scalar multiple.

Plug $\hat{\boldsymbol{\Sigma}}_{\mathrm{B}} = (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^T$ into

$$\left((N-2)\hat{\boldsymbol{\Sigma}} + \frac{N_1N_2}{N}\hat{\boldsymbol{\Sigma}}_{\mathrm{B}}\right)\hat{\boldsymbol{\beta}} = N(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1),$$

we get

$$\begin{aligned}
&\left((N-2)\hat{\boldsymbol{\Sigma}} + \frac{N_1N_2}{N}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^T\right)\hat{\boldsymbol{\beta}} = N(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\
\iff &(N-2)\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}} = \left(N - \frac{N_1N_2}{N}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^T\hat{\boldsymbol{\beta}}\right)(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\
\iff &\hat{\boldsymbol{\beta}} = \frac{1}{N-2}\left(N - \frac{N_1N_2}{N}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^T\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\
\implies &\hat{\boldsymbol{\beta}} \propto \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)
\end{aligned}$$

(d) Argue that (1) holds invariantly for any (distinct) coding of Y_i's.

For any distinct coding of Y_i,

$$\sum_{i=1}^{N} Y_i\boldsymbol{X}_i = Y_1\sum_{g_i=1}\boldsymbol{X}_i + Y_2\sum_{g_i=2}\boldsymbol{X}_2 = Y_1N_1\hat{\boldsymbol{\mu}}_1 + Y_2N_2\hat{\boldsymbol{\mu}}_2.$$

Then $\hat{\boldsymbol{\beta}}$ satisfies

$$\left((N-2)\hat{\boldsymbol{\Sigma}} + \frac{N_1N_2}{N}\hat{\boldsymbol{\Sigma}}_{\mathrm{B}}\right)\hat{\boldsymbol{\beta}} = Y_1N_1\hat{\boldsymbol{\mu}}_1 + Y_2N_2\hat{\boldsymbol{\mu}}_2$$

(e) Find the least-square prediction function $\hat{f}_{\mathrm{LS}}$. Consider the least-square discriminant rule

$$\hat{d}_{\mathrm{LS}}(\boldsymbol{x}) = \begin{cases} \text{class 2}, & \hat{f}_{\mathrm{LS}}(\boldsymbol{x}) > 0 \\ \text{class 1}, & \text{otherwise} \end{cases}$$

Show that $\hat{d}_{\mathrm{LS}}$ is not the same as the LDA rules unless $N_1 = N_2$. If, alternatively, we use uniform class prior rather than plug-in estimates, then least-square discriminant rule is equivalent to the LDA.

Let

$$\widehat{\boldsymbol{\beta}} = \lambda \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1)$$

where $\lambda \in \mathbb{R}$ is a constant. Then,

$$\begin{aligned}
\widehat{\beta}_0 &= -\boldsymbol{\beta}^T \frac{\sum_{i=1}^N \boldsymbol{X}_i}{N} \\
&= -\boldsymbol{\beta}^T \frac{N_1\widehat{\boldsymbol{\mu}}_1 + N_2\widehat{\boldsymbol{\mu}}_2}{N} \\
&= -\left(\lambda \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1)\right)^T \frac{N_1\widehat{\boldsymbol{\mu}}_1 + N_2\widehat{\boldsymbol{\mu}}_2}{N} \\
&= -\frac{\lambda}{N}\left((N_1 - N_2)\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_2^T - N_1\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1^T + N_2\widehat{\boldsymbol{\mu}}_2\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_2^T\right)
\end{aligned}$$

Then

$$\begin{aligned}
\widehat{f}_{\mathrm{LS}} &= \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i \\
&= -\frac{\lambda}{N}\left((N_1 - N_2)\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_2^T - N_1\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1^T + N_2\widehat{\boldsymbol{\mu}}_2\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_2^T\right) + \lambda \boldsymbol{X}_i^T\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1)
\end{aligned}$$

Therefore, $\widehat{d}_{\mathrm{LS}}$ classifies $\boldsymbol{X}_i$ to class 2 if

$$\begin{aligned}
&\widehat{f}_{\mathrm{LS}} > 0 \\
\iff &\boldsymbol{X}_i^T\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1) > \frac{1}{N}\left((N_1 - N_2)\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_2^T - N_1\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1^T + N_2\widehat{\boldsymbol{\mu}}_2\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_2^T\right) \qquad \text{(a5)}
\end{aligned}$$

(a5) is not equivalent to the LDA rule (stated in question 1(a)) unless $N_1 = N_2$. Therefore, $\widehat{d}_{\mathrm{LS}}$ is not the same as the LDA rules unless $N_1 = N_2$.

2. (*20 pt*) (Fisher's Discriminant Analysis/Reduced Rank LDA, Textbook Ex. 4.1, 4.8) Let $\mathbf{X} = [X_{ij}]_{n\times p} \in \mathbb{R}^{n\times p}$ be a centered[1] data matrix with n observations (in rows) of p-dimensional features (in columns), $\mathbf{Y} = [Y_{ik}]_{n\times K} \in \{0,1\}^{n\times K}$ be an indicator matrix with 0-1 entries encoding memberships out of $K(\leqslant n)$ classes for observations, $n_1, n_2, \cdots, n_K \geqslant 1$ be class sizes.

   (a) Derive the class-centroid matrix $\mathbf{M} \in \mathbb{R}^{K\times p}$ in terms of $\mathbf{X}$ and $\mathbf{Y}$.

$$\mathbf{Y}^T\mathbf{Y} = \mathrm{diag}(n_1, n_2, \ldots, n_K)$$

$$\mathbf{M} = (\mathrm{diag}(n_1, n_2, \ldots, n_K))^{-1}\mathbf{Y}^T\mathbf{X} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}$$

   (b) Denote rows in $\mathbf{X}$ and $\mathbf{M}$ as $\{\boldsymbol{X}_{i\cdot}\}_{i=1}^n$ and $\{\widehat{\boldsymbol{\mu}}_k\}_{k=1}^K$ respectively. Define

$$\mathbf{S} := \sum_{i=1}^n \boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T, \quad \mathbf{S}_{\mathrm{B}} := \sum_{k=1}^K n_k\widehat{\boldsymbol{\mu}}_k\widehat{\boldsymbol{\mu}}_k^T, \quad \mathbf{S}_{\mathrm{W}} := \sum_{k=1}^K \sum_{Y_i=k} (\boldsymbol{X}_{i\cdot} - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{X}_{i\cdot} - \widehat{\boldsymbol{\mu}}_k)^T.$$

[1] That is, $\sum_{i=1}^n X_{ij} = 0$ for all $1 \leqslant j \leqslant p$.

(i) Write $\mathbf{S}, \mathbf{S}_{\mathrm{B}}, \mathbf{S}_{\mathrm{W}}$ in compact form in terms of $\mathbf{X}$ and $\mathbf{Y}$.

$$\mathbf{S} = \mathbf{X}^T\mathbf{X}$$

$$\mathbf{S}_{\mathrm{B}} = \mathbf{M}^T \operatorname{diag}(n_1, \ldots, n_K)\mathbf{M} = \mathbf{M}^T\mathbf{Y}^T\mathbf{Y}\mathbf{M} = \mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}$$

$$\begin{aligned}
\mathbf{S}_{\mathrm{W}} &= (\mathbf{X} - \mathbf{Y}\mathbf{M})^T(\mathbf{X} - \mathbf{Y}\mathbf{M}) \\
&= \mathbf{X}^T\mathbf{X} - \mathbf{M}^T\mathbf{Y}^T\mathbf{X} - \mathbf{X}^T\mathbf{Y}\mathbf{M} + \mathbf{M}^T\mathbf{Y}^T\mathbf{Y}\mathbf{M} \\
&= \mathbf{X}^T\mathbf{X} - 2\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X} + \mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X} \\
&= \mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}
\end{aligned}$$

(ii) Prove an ANOVA-type covariance matrix decomposition

$$\mathbf{S} = \mathbf{S}_{\mathrm{B}} + \mathbf{S}_{\mathrm{W}}.$$

Plug in the results from (i).

(c) Suppose $\mathbf{S}_{\mathrm{W}} \succ 0$. Show how to solve the generalized eigenvalue problem

$$\begin{aligned}
\max_{\boldsymbol{\phi}} \quad & \boldsymbol{\phi}^T\mathbf{S}_{\mathrm{B}}\boldsymbol{\phi} \\
\text{s.t.} \quad & \boldsymbol{\phi}^T\mathbf{S}_{\mathrm{W}}\boldsymbol{\phi} = 1
\end{aligned} \tag{2}$$

by transforming it to a standard eigenvalue problem.

There is an orthogonal matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{P}^T\mathbf{S}_{\mathrm{W}}\mathbf{P} = \mathbf{D}$. Let $\mathbf{S}_{\mathrm{W}}^{\frac{1}{2}} = \mathbf{P}\mathbf{D}^{\frac{1}{2}}\mathbf{P}^T$, $\mathbf{S}_{\mathrm{W}}^{-\frac{1}{2}} = \mathbf{P}\mathbf{D}^{-\frac{1}{2}}\mathbf{P}^T$, $\boldsymbol{\theta} = \mathbf{S}_{\mathrm{W}}^{\frac{1}{2}}\boldsymbol{\phi}$. Transform (2) to

$$\begin{aligned}
\max_{\boldsymbol{\theta}} \quad & \boldsymbol{\theta}^T\mathbf{S}_{\mathrm{W}}^{-\frac{1}{2}}\mathbf{S}_{\mathrm{B}}\mathbf{S}_{\mathrm{W}}^{-\frac{1}{2}}\boldsymbol{\theta} \\
\text{s.t.} \quad & \boldsymbol{\theta}^T\boldsymbol{\theta} = 1
\end{aligned}$$

(d) Suppose $\mathbf{S}_{\mathrm{W}} \succ 0$. Let $\{\widehat{\boldsymbol{\phi}}_j\}_{j=1}^p$ be Fisher's discriminant coordinates, *i.e.* the solutions to (2) with each one orthogonal in $\mathbf{S}_{\mathrm{W}}$[2] to the predecessors. For $1 \leqslant m \leqslant p$, define the (Fisher's) discriminant rules based on nearest centroids projected on $\{\widehat{\boldsymbol{\phi}}_j\}_{j=1}^m$

$$d_m(\boldsymbol{x}) = \operatorname*{argmin}_{1 \leqslant k \leqslant K} \sum_{j=1}^m \left(\widehat{\boldsymbol{\phi}}_j^T\boldsymbol{x} - \widehat{\boldsymbol{\phi}}_j^T\widehat{\boldsymbol{\mu}}_k\right)^2.$$

Now let $L := \dim\left(\mathbf{span}\{\widehat{\mu}_k\}_{k=1}^K\right) = \mathbf{rank}(\mathbf{S}_{\mathrm{B}}) \leqslant \min\{K-1, p\}$. Take the following steps to show the equivalence between LDA based on multivariate-Gaussian Bayes rule introduced in class and Fisher's discrimination.

(i) Using the fact that $\widehat{\boldsymbol{\phi}}_j^T\mathbf{S}_{\mathrm{W}}\widehat{\boldsymbol{\phi}}_{j'} = \mathbb{1}(j = j')$ $(1 \leqslant j, j' \leqslant p)$, show that

$$\mathbf{S}_{\mathrm{W}}^{-1} = \sum_{j=1}^p \widehat{\boldsymbol{\phi}}_j\widehat{\boldsymbol{\phi}}_j^T.$$

[2] We say $\boldsymbol{x}$ is orthogonal to $\boldsymbol{y}$ in $\mathbf{A}$ if $\boldsymbol{x}^T\mathbf{A}\boldsymbol{y} = 0$.

Let

$$\mathbf{A} = \begin{bmatrix} \widehat{\boldsymbol{\phi}}_1^T \\ \vdots \\ \widehat{\boldsymbol{\phi}}_p^T \end{bmatrix}$$

Then,

$$\begin{aligned} & \widehat{\boldsymbol{\phi}}_j^T \mathbf{S}_{\mathrm{W}} \widehat{\boldsymbol{\phi}}_{j'} = \mathbb{1}(j = j') \\ \Longrightarrow\ & \mathbf{A}\mathbf{S}_{\mathrm{W}}\mathbf{A}^T = \mathbf{I}_p \\ \Longrightarrow\ & \mathbf{A}\mathbf{S}_{\mathrm{W}} = (\mathbf{A}^T)^{-1} \\ \Longrightarrow\ & \mathbf{A}^T\mathbf{A}\mathbf{S}_{\mathrm{W}} = \mathbf{I}_p \\ \Longrightarrow\ & \mathbf{S}_{\mathrm{W}}^{-1} = \mathbf{A}^T\mathbf{A} = \sum_{j=1}^{p} \widehat{\boldsymbol{\phi}}_j \widehat{\boldsymbol{\phi}}_j^T \end{aligned}$$

(ii) Show that

$$\widehat{\boldsymbol{\phi}}_j^T \mathbf{S}_{\mathrm{B}} \widehat{\boldsymbol{\phi}}_j = 0. \quad (L+1 \leqslant j \leqslant p)$$

Because $\mathbf{S}_{\mathrm{W}}^{-\frac{1}{2}}$ has full rank, $\mathbf{S}_{\mathrm{W}}^{-\frac{1}{2}}\mathbf{S}_{\mathrm{B}}\mathbf{S}_{\mathrm{W}}^{-\frac{1}{2}}$ has rank of L. Therefore, $\mathbf{S}_{\mathrm{W}}^{-\frac{1}{2}}\mathbf{S}_{\mathrm{B}}\mathbf{S}_{\mathrm{W}}^{-\frac{1}{2}}$ has $p - L$ zero eigenvalues. Therefore, $\widehat{\boldsymbol{\phi}}_j^T \mathbf{S}_{\mathrm{B}} \widehat{\boldsymbol{\phi}}_j = 0$. $(L+1 \leqslant j \leqslant p)$.

(iii) Show the equivalence between (ii) and

$$\widehat{\boldsymbol{\phi}}_j^T \widehat{\boldsymbol{\mu}}_k = 0. \quad (L+1 \leqslant j \leqslant p,\ 1 \leqslant k \leqslant K)$$

$$\begin{aligned} & \widehat{\boldsymbol{\phi}}_j^T \mathbf{S}_{\mathrm{B}} \widehat{\boldsymbol{\phi}}_j = 0 \\ \Longleftrightarrow\ & \sum_{k=1}^{K} n_k \widehat{\boldsymbol{\phi}}_j^T \widehat{\boldsymbol{\mu}}_k \widehat{\boldsymbol{\mu}}_k^T \widehat{\boldsymbol{\phi}}_j = 0 \\ \Longleftrightarrow\ & \sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\phi}}_j^T \widehat{\boldsymbol{\mu}}_k)^2 = 0 \\ \Longleftrightarrow\ & \widehat{\boldsymbol{\phi}}_j^T \widehat{\boldsymbol{\mu}}_k = 0 \quad (1 \leqslant k \leqslant K) \end{aligned}$$

(iv) Show that $d_L, d_{L+1}, \cdots, d_p$ are equivalent to the multivariate-Gaussian-based LDA with uniform prior.

Let $d(\boldsymbol{x})$ be the estimation from multivariate-Gaussian-based LDA with uniform prior.

$$\begin{aligned} d(\boldsymbol{x}) &= \underset{1\leqslant k\leqslant K}{\operatorname{argmin}} (\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k) \\ &= \underset{1\leqslant k\leqslant K}{\operatorname{argmin}} \frac{1}{n-K} (\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T \mathbf{S}_{\mathrm{W}}^{-1} (\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k) \\ &= \underset{1\leqslant k\leqslant K}{\operatorname{argmin}} \sum_{j=1}^{p} (\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\phi}}_j \widehat{\boldsymbol{\phi}}_j^T (\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k) \\ &= \underset{1\leqslant k\leqslant K}{\operatorname{argmin}} \sum_{j=1}^{p} (\widehat{\boldsymbol{\phi}}_j^T \boldsymbol{x} - \widehat{\boldsymbol{\phi}}_j^T \widehat{\boldsymbol{\mu}}_k)^2 \\ &= d_p(\boldsymbol{x}). \end{aligned}$$

$$\begin{aligned} &\widehat{\boldsymbol{\phi}}_j^T \widehat{\boldsymbol{\mu}}_k = 0 \quad (L+1 \leqslant j \leqslant p,\ 1 \leqslant k \leqslant K) \\ \Longrightarrow\ & d_L(\boldsymbol{x}) = d_{L+1}(\boldsymbol{x}) = \cdots = d_p(\boldsymbol{x}). \end{aligned}$$

Therefore, $d_L, d_{L+1}, \cdots, d_p$ are equivalent to the multivariate-Gaussian-based LDA with uniform prior.

Remark (Reduced-Rank LDA). The scaling matrices for Fisher's discriminant rules are actually low-rank approximations[3] to that of multivariate-Gaussian LDA. In particular, $m = p$ recovers the scaling matrix of multivariate-Gaussian LDA as in (i), and the $m(\geqslant L)$-rank approximations don't lose information in the sense of discriminating centroids in the reduced space as in (iii). The equivalence of (ii) and (iii) reveals the rationale of finding directions to maximize between-class variances.

3. (*15 pt*) (Logistic Regression Recession, Textbook Ex. 4.5) Consider a two-class Logistic regression problem. We encode labels $Y_i \in \{\pm 1\}$.

   (a) Show that solving the maximum-likelihood estimate (MLE) for the intercept and slope parameters $(\beta_0, \boldsymbol{\beta})$ is equivalent to the following empirical risk minimization (ERM) problem[4]

   $$\min_{(\beta_0,\boldsymbol{\beta})} \frac{1}{n} \sum_{i=1}^{n} \log\left\{1 + \exp\left[-Y_i(\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta})\right]\right\}.$$

   Punchline. In standard generalized linear model (GLM) formulation[5] for the Logistic likelihood, Y_i's are coded as 0-1, but now we are coding it into ± 1.

   Let

   $$y_i = \begin{cases} 1 & Y_i = 1 \\ 0 & Y_i = -1 \end{cases}$$

[3] NOT under the usual spectral/Frobenius norm, but under a Mahalanobis-type Frobenius norm.
[4] We call $\ell(u, v) := \log(1 + e^{-uv})$ as the **Logistic loss**.
[5] See Section 4.4.1 at page 114 in [2, McCullagh and Nelder (1989)].

The maximum-likelihood estimate is

$$\begin{aligned}
&\max_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ y_i \log \frac{\exp\left[\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}\right]}{1+\exp\left[\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}\right]} + (1-y_i)\log \frac{1}{1+\exp\left[\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}\right]} \right\} \\
\iff &\max_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ y_i \log \frac{1}{1+\exp\left[-(\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta})\right]} + (1-y_i)\log \frac{1}{1+\exp\left[\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}\right]} \right\} \\
\iff &\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ y_i \log\left\{1+\exp\left[-(\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta})\right]\right\} + (1-y_i)\log\left\{1+\exp\left[\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}\right]\right\} \right\} \\
\iff &\min_{(\beta_0,\boldsymbol{\beta})} \sum_{i=1}^{n} \log\left\{1+\exp\left[-Y_i(\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta})\right]\right\}
\end{aligned}$$

(b) Argue that if the training data can be perfectly separated by a hyperplane (Figure 4.16 at page 134 in Textbook), or mathematically, there exists a hyperplane $H \subseteq \mathbb{R}^p$ with upper (*resp.* lower) halfspace H^+ (*resp.* H^-)[6] such that

$$\{\boldsymbol{X}_i : Y_i = +1 \ (1 \leqslant i \leqslant n)\} \subseteq H^+, \quad \{\boldsymbol{X}_i : Y_i = -1 \ (1 \leqslant i \leqslant n)\} \subseteq H^-,$$

then the optimal solution to Logistic regression is characterized by some unbounded directions[7]. That is, there exists a direction $(\alpha, \boldsymbol{\gamma}) \in \mathbb{R}^{p+1}$ along which $(\beta_0, \boldsymbol{\beta}) = c \times (\alpha, \boldsymbol{\gamma})$ will give objectives increasing/decreasing to an objective supremum/infimum as $c \to +\infty$.

Let $H = \alpha + \mathbf{X}^T\boldsymbol{\gamma}$. Then,

$$\alpha + \boldsymbol{X}_i^T\boldsymbol{\gamma} > 0, \quad \text{if } Y_i = +1$$

$$\alpha + \boldsymbol{X}_i^T\boldsymbol{\gamma} < 0, \quad \text{if } Y_i = -1$$

Therefore,

$$Y_i(\alpha + \boldsymbol{X}_i^T\boldsymbol{\gamma}) > 0, \quad 1 \leqslant i \leqslant n$$

Then,

$$\sum_{i=1}^{n} \log\left\{1+\exp\left[-Y_i(\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta})\right]\right\} = \sum_{i=1}^{n} \log\left\{1+\exp\left[-cY_i(\alpha + \boldsymbol{X}_i^T\boldsymbol{\gamma})\right]\right\} \to 0$$

as $c \to +\infty$.

(c) Describe the recession/unboundedness phenomenon in (b) when the number of classes is greater than two.

Denote the number of classes as K. Denote the class of the ith sample as G_i.

If for each k, $1 \leqslant k \leqslant K-1$, there exists a hyperplane $H \subseteq \mathbb{R}^p$ with upper (*resp.* lower) halfspace H^+ (*resp.* H^-) such that

$$\{\boldsymbol{X}_i : G_i = k \ (1 \leqslant i \leqslant n)\} \subseteq H^+, \quad \{\boldsymbol{X}_i : G_i \neq k \ (1 \leqslant i \leqslant n)\} \subseteq H^-,$$

[6] If we write a hyperplane into $H = \{\boldsymbol{x} \in \mathbb{R}^p : \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta} = 0\}$, then the halfspaces are $H^{+(-)} = \{\boldsymbol{x} \in \mathbb{R}^p : \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta} \geqslant (\leqslant) 0\}$.

[7] In convex analysis, we say such directions compose of the **recession cone** of the objective function. More general discussion of recession and unboundedness can be found in Section 8 in [3, Rockafellar (1970)].

then the optimal solution to Logistic regression is characterized by some unbounded directions. That is, there exists directions $(\alpha_k, \boldsymbol{\gamma}_k) \in \mathbb{R}^{p+1}$ $(1 \leqslant k \leqslant K-1)$ along which $(\beta_{k0}, \boldsymbol{\beta}_k) = c \times (\alpha_k, \boldsymbol{\gamma}_k)$ will give objectives increasing/decreasing to an objective supremum/infimum as $c \to +\infty$.

Computational Part

1. (*25 pt*) **South African Heart Disease Data**
   Perform LDA, QDA, and Logistic regression analysis on the *South African Heart Disease Data*[8] and compare their prediction performances. Find more information in Textbook Section 4.4.2. Remember to hold out a test set to perform prediction.

2. (*25 pt*) **Zip Code Digits Data**
   Classify between 3's and 8's from the *Zip Code Digits* data[9]. Build up the LDA, QDA, and Logistic regression models on the training dataset and compare their prediction performances on the test dataset.

References

[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer-Verlag, https://web.stanford.edu/~hastie/Papers/ESLII.pdf, second edition, 2009. 1

[2] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman and Hall, second edition, 1989. 7

[3] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 1970. 8

[8] Available at https://web.stanford.edu/~hastie/ElemStatLearn/datasets/SAheart.data.

[9] Available in https://web.stanford.edu/~hastie/ElemStatLearn/datasets/zip.train.gz (training dataset) and https://web.stanford.edu/~hastie/ElemStatLearn/datasets/zip.test.gz (test dataset).