MATH5470 Assigment 2

SUN Yuchang, 20034768R

1 Q1: ESL, 4.1

The generalized eigenvalue problem is

$$\max a^T Ba \tag{1}$$

$$\text{s.t.} a^T Wa = 1. \tag{2}$$

Its Laplacian form is

$$L(a;\lambda) = a^T Ba + \lambda(1 - a^T Wa). \tag{3}$$

Forcing the derivative equal to zero:

$$\frac{\partial L}{\partial a} = 2Ba - \lambda(2Wa) = 0. \tag{4}$$

And we will get

$$W^{-1}Ba = \lambda a. \tag{5}$$

Thus, a should be the eigenvector of $W^{-1}B$, and λ is the corresponding eigenvalue.

2 Q2: ESL, Q4.2

(a) We model each class density as Multivariate Gaussian

$$p(x|w_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)^{-1}\Sigma_K^{-1}(x-\mu_k)\right], k = 1, 2. \tag{6}$$

Note that in LDA we have $\Sigma_k = \Sigma, \forall k$. Then we have

$$\frac{p(x|w_2)}{p(x|w_1)} = \exp\left[-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right] \tag{7}$$

$$= \exp\left[x^T\Sigma^{-1}\mu_2 - x^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right]. \tag{8}$$

If $Pr(G = 2|X = x) > Pr(G = 1|X = x)$, it belongs to class 2. We do this by forcing

$$\log \frac{Pr(G = 2|X = x)}{Pr(G = 1|X = x)} = \log \frac{p(x|w_2)}{p(x|w_1)} + \log \frac{\pi_2}{\pi_1} \tag{9}$$

$$= x^T \Sigma^{-1}(\mu_2 - \mu_1) + \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + \log \frac{\pi_2}{\pi_1} \tag{10}$$

$$> 0, \tag{11}$$

which implies

$$x^T \Sigma^{-1}(\mu_2 - \mu_1) > \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \log \frac{N_1}{N} - \log \frac{N_2}{N}. \tag{12}$$

Otherwise, from Eq.11 we know it will be class 1.

(b) Let X_1 be the $N_1 \times p$ matrix of training set in class 1, and X_2 be the $N_2 \times p$ matrix of training set in class 2. Then the means are

$$\hat{\mu}_1 = \frac{1}{N} X_1^T \cdot 1_{N_1}; \tag{13}$$

$$\hat{\mu}_2 = \frac{1}{N} X_2^T \cdot 1_{N_2}. \tag{14}$$

Let $1_N = \begin{bmatrix} 1_{N_1} \\ 1_{N_2} \end{bmatrix}$, $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, and $y = \begin{bmatrix} -\frac{N}{N_1} 1_{N_1} \\ \frac{N}{N_2} 1_{N_2} \end{bmatrix}$. The solution $(\hat{\beta}_0, \hat{\beta})$ should satisfy

$$\begin{bmatrix} 1_N & X \end{bmatrix}^T \begin{bmatrix} 1_N & X \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 1_N & X \end{bmatrix}^T y. \tag{15}$$

Note that

$$1_N^T y = -1_{N_1}^T \frac{N_1}{N} 1_{N_1} + 1_{N_2}^T \frac{N_2}{N} 1_{N_2} \tag{16}$$

$$= 0; \tag{17}$$

and

$$\begin{bmatrix} 1_N & X \end{bmatrix}^T \begin{bmatrix} 1_N & X \end{bmatrix} = \begin{bmatrix} N & 1_N X \\ X^T 1_N & X^T X \end{bmatrix}. \tag{18}$$

Using Eq.18, Eq.15 becomes

$$\begin{bmatrix} N & 1_N X \\ X^T 1_N & X^T X \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 0 \\ X^T y \end{bmatrix}. \tag{19}$$

Then we can solve it and get the result:

$$\hat{\beta}_0 = 1\frac{1}{N} 1_N^T X \beta, \tag{20}$$

and

$$(-\frac{1}{N} X^T 1_N 1_N^T X + X^T X)\hat{\beta} = X^T y \tag{21}$$

$$= \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^T \begin{bmatrix} -\frac{N}{N_1} 1_{N_1} \\ \frac{N}{N_2} 1_{N_2} \end{bmatrix} \tag{22}$$

$$= N(\hat{\mu}_2 - \hat{\mu}_1) \tag{23}$$

$$= RHS. \tag{24}$$

Note that

$$X^T 1_N = N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2, \tag{25}$$

and thus

$$\frac{1}{N} X^T 1_N 1_N^T X = \frac{1}{N}(N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)(N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)^T \tag{26}$$

$$= \frac{1}{N}(N_1^2 \hat{\mu}_1 \hat{\mu}_1^T + 2N_1 N_2 \hat{\mu}_1 \hat{\mu}_2^T + N_2^2 \hat{\mu}_2 \hat{\mu}_2^T). \tag{27}$$

Also,

$$(N-2)\hat{\Sigma} = (X_1 - 1_{N_1}\hat{\mu}_1^T)^T (X_1 - 1_{N_1}\hat{\mu}_1^T) + (X_2 - 1_{N_2}\hat{\mu}_2^T)^T (X_2 - 1_{N_2}\hat{\mu}_2^T) \tag{28}$$

$$= X^T X - N_1 \hat{\mu}_1 \hat{\mu}_1^T - N_2 \hat{\mu}_2 \hat{\mu}_2^T. \tag{29}$$

The LHS of Eq.15 is

$$LHS = (-\frac{1}{N}X^T 1_N 1_N^T X + X^T X)\beta \tag{30}$$

$$= (-\frac{1}{N}(N_1^2\hat{\mu}_1\hat{\mu}_1^T + 2N_1N_2\hat{\mu}_1\hat{\mu}_2^T + N_2^2\hat{\mu}_2\hat{\mu}_2^T) + N_1\hat{\mu}_1\hat{\mu}_1^T + N_2\hat{\mu}_2\hat{\mu}_2^T + (N-2)\hat{\Sigma})\beta \tag{31}$$

$$= (-\frac{1}{N}(N_1N_2\hat{\mu}_1\hat{\mu}_1^T - 2N_1N_2\hat{\mu}_1\hat{\mu}_2^T + N_1N_2\hat{\mu}_2\hat{\mu}_2^T) + (N-2)\hat{\Sigma})\beta \tag{32}$$

$$= (\frac{N_1N_2}{N}(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T + (N-2)\hat{\Sigma})\beta \tag{33}$$

$$= (N\hat{\Sigma}_B + (N-2)\hat{\Sigma})\beta. \tag{34}$$

Then we complete the proof.

(c) Due to

$$\hat{\Sigma}_B\beta = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T\beta, \tag{35}$$

where $(\hat{\mu}_2 - \hat{\mu}_1)^T\beta$ is a scalar, we know that $\hat{\Sigma}_B\beta$ has the direction of $\hat{\mu}_2 - \hat{\mu}_1$.

From probelm (b), we know that

$$(N-2)\hat{\Sigma}\beta = N(\hat{\mu}_2 - \hat{\mu}_1) - \frac{N_1N_2}{N}\hat{\Sigma}_B\beta, \tag{36}$$

the first term and second term of RHS are both in the direction of $\hat{\mu}_2 - \hat{\mu}_1$.

Thus, by doing minor change, we see that

$$\hat{\beta} = \frac{1}{N-2}\hat{\Sigma}^{-1}(N(\hat{\mu}_2 - \hat{\mu}_1) - \frac{N_1N_2}{N}\hat{\Sigma}_B\beta) \tag{37}$$

$$\propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \tag{38}$$

This completes the proof.

(d) We code the values of class 1 and class 2 as k_1 and k_2 respectively. The normal equation Eq.**??** becomes

$$\begin{bmatrix} N & 1_N X \\ X^T 1_N & X^T X \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 1_N^T y \\ X^T y \end{bmatrix}, \tag{39}$$

where $1_N^T y = N_1k_1 + N_2k_2$ and $X^T y = N_1k_1\hat{\mu}_1 + N_2k_2\hat{\mu}_2$.

With similar operations in problem (b), Eq.24 becomes

$$(-\frac{1}{N}X^T 1_N 1_N^T X + X^T X)\beta = X^T y - \frac{1}{N}X^T 1_N 1_N^T y \tag{40}$$

$$= N_1 k_1 \hat{\mu}_1 + N_2 k_2 \hat{\mu}_2 - \frac{1}{N}(N_1\hat{\mu}_1 + N_1\hat{\mu}_1)(N_1 k_1 + N_2 k_2) \tag{41}$$

$$= \frac{N_1 N_2}{N}\hat{\mu}_2(k_2 - k_1) - \frac{N_1 N_2}{N}\hat{\mu}_1(k_2 - k_1) \tag{42}$$

$$= \frac{N_1 N_2}{N}(k_2 - k_1)(\hat{\mu}_2 - \hat{\mu}_1) \tag{43}$$

$$= RHS. \tag{44}$$

We notice that compared with problem (b), the RHS has a scalar factor, and the LHS stays the same. Therefore, the result holds.

(e) For $\hat{f} = \hat{\beta}_0 + \hat{\beta}^T x$, classifying to class 2 means

$$\hat{\beta}_0 + \hat{\beta}^T x > 0 \tag{45}$$

$$(x^T - \frac{1}{N}1_N^T X)\hat{\beta} > 0 \tag{46}$$

$$(x^T - (\frac{N_1}{N}\hat{\mu}_1 + \frac{N_2}{N}\hat{\mu}_2)^T)\hat{\beta} > 0 \tag{47}$$

Using $\hat{\beta} = \lambda\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$, where λ is a scalar, we have

$$\left(x^T - (\frac{N_1}{N}\hat{\mu}_1 + \frac{N_2}{N}\hat{\mu}_2)^T\right)\lambda\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > 0 \tag{48}$$

$$x^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > (\frac{N_1}{N}\hat{\mu}_1 + \frac{N_2}{N}\hat{\mu}_2)^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \tag{49}$$

When $N_1 = N_2$, above Eq.49 becomes

$$x^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \tag{50}$$

As for LDA rules, recall that in problem (a),

$$x^T\Sigma^{-1}(\mu_2-\mu_1) > \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\frac{N_1}{N} - \log\frac{N_2}{N} \tag{51}$$

$$x^T\Sigma^{-1}(\mu_2-\mu_1) > \frac{1}{2}(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1) + \log\frac{N_1}{N} - \log\frac{N_2}{N}. \tag{52}$$

When $N_1 = N_2$, above Eq.52 becomes

$$x^T\Sigma^{-1}(\mu_2-\mu_1) > \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1. \tag{53}$$

Generally, Eq.49 is not equal to Eq.52, implying different bounderies. But when $N_1 = N_2$, Eq.50 is the same as Eq.53, thus having the same rules.

3 Q3: ESL, Q3.17

Let the co-variance matrices for X and $\hat{Y}$ are respectively Σ and $\hat{\Sigma}$, and the k-class means for X and $\hat{Y}$ are respectively μ_k and $\hat{\mu}_k$. Also, $\hat{B} = (X^TX)^{-1}X^TY$.

The linear discriminant functions for X and $\hat{Y}$ are respectively

$$\delta_k = X^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k; \tag{54}$$

$$\delta_k' = \hat{Y}^T\Sigma^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + \log\pi_k. \tag{55}$$

I will show that LDA using $\hat{Y}$ is identical to LDA in the original space by proving that the first two terms are equal.

Note that

$$\hat{\mu}_k = \sum_{c_i=k}\frac{\hat{y}_i}{N_k} \tag{56}$$

$$= \sum_{c_i=k}\frac{\hat{B}^Tx_i}{N_k} \tag{57}$$

$$= \frac{B^TX^TY_k}{N_k}, \tag{58}$$

and

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{c_i=k} (\hat{y}_i - \hat{\mu}_k)(\hat{y}_i - \hat{\mu}_k)^T \tag{59}$$

$$= \frac{1}{N-K} \sum_{k=1}^{K} \sum_{c_i=k} (\hat{B}^T x_i - \hat{\mu}_k)(\hat{B}^T x_i - \hat{\mu}_k)^T \tag{60}$$

$$= \frac{1}{N-K} \sum_{k=1}^{K} \sum_{c_i=k} (\hat{B}^T x_i - \hat{B}^T \mu_k)(\hat{B}^T x_i - \hat{B}^T \mu_k)^T \tag{61}$$

$$= \hat{B}^T \Sigma \hat{B}, \tag{62}$$

where $c_i = k$ means the i-th sample's class label is k.

Firstly, we have

$$\hat{Y}^T \Sigma^{-1} \hat{\mu}_k = (XB)(B^T \Sigma B)^{-1} (\frac{B^T X^T Y_k}{N_k}) \tag{63}$$

and the concentration

$$\hat{Y}^T \Sigma^{-1} \hat{\mu} = (XB)(B^T \Sigma B)^{-1} (B^T X^T D^{-1}). \tag{64}$$

Let $D = diag(n_1, n_2, \ldots,_K)$. According to Eq.62, we have

$$\Sigma = \frac{1}{N-K} (X^T - X^T Y D^{-1} Y^T)(X^T - X^T Y D^{-1} Y^T)^T \tag{65}$$

$$= \frac{1}{N-K} X^T (I - Y D^{-1} Y^T)^2 X \tag{66}$$

$$= \frac{1}{N-K} X^T (I - Y D^{-1} Y^T) X. \tag{67}$$

Then

$$B^T \sigma B = \frac{1}{N-K} B^T X^T (I - Y D^{-1} Y^T) X B \tag{68}$$

$$= \frac{1}{N-K} B^T X^T Y - B^T X^T Y D^{-1} Y^T X B \tag{69}$$

$$= \frac{1}{N-K} (M - M D^{-1} M), \tag{70}$$

where $M = B^T X^T Y$ and $M^T = M$. Thus,

$$B(B^T \Sigma B)^{-1} B^T X^T Y = B(\frac{1}{N-K}(M - MD^{-1}M))^{-1} B^T X^T Y \tag{71}$$

$$= \Sigma^{-1}\Sigma(N-K)B(I - D^{-1}M)^{-1}M^{-1}B^T X^T Y \tag{72}$$

$$= \Sigma^{-1}(N-K)\Sigma B(I - D^{-1}M)^{-1}M^{-1}M \tag{73}$$

$$= \Sigma^{-1}X^T(I - YD^{-1}Y^T)XB(I - D^{-1}M)^{-1}M^{-1}M \tag{74}$$

$$= \Sigma^{-1}X^T Y(I - D^{-1}M)(I - D^{-1}M)^{-1} \tag{75}$$

$$= \Sigma^{-1}X^T Y. \tag{76}$$

Apply Eq.76 into Eq.64, we show that

$$\hat{Y}^T \Sigma^{-1} \hat{\mu} = X\Sigma^{-1}X^T Y D^{-1} = X\Sigma^{-1}\mu. \tag{77}$$

Secondly, we will prove the second term

$$\hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu} = (B^T \mu_k)^T \hat{\Sigma}^{-1}(B^T X^T Y D^{-1}) \tag{78}$$

$$= \mu_k^T B \hat{\Sigma}^{-1}(B^T X^T Y D^{-1}) \tag{79}$$

$$= \mu_k^T B(B^T \Sigma B)^{-1} B^T X^T Y D^{-1} \tag{80}$$

$$= \mu_k^T \Sigma^{-1} X^T Y D^{-1} \tag{81}$$

$$= \mu_k^T \Sigma^{-1} \mu. \tag{82}$$

Finally, with Eq.77 and Eq.82, we can show that LDA using $\hat{Y}$ is identical to LDA in the original space.

4 Q4: ESL, 4.9

My solution is written in Python on Colab. See GitHub link.
For the training data, the misclassification error is 0.0114, and for the test data, it is 0.5381.