1 Ex.4.1

To solve constrained maximization or minimization problems we want to use the idea of Lagrangian multipliers. Define the Lagrangian $\mathcal{L}$ as

$$\mathcal{L}(a;\lambda) = a^T B a + \lambda \left(a^T W a - 1\right).$$

Here λ is the Lagrange multiplier. Taking the a derivative of this expression and setting it equal to zeros gives

$$\frac{\partial \mathcal{L}(a;\lambda)}{\partial a} = 2Ba + \lambda(2Wa) = 0.$$

This last equation is equivalent to

$$Ba + \lambda W a = 0,$$

or multiplying by W^{-1} on both sides and moving the expression with B to the left hand side gives the

$$W^{-1} B a = \lambda a,$$

Notice this is a standard eigenvalue problem, in that the solution vectors a must be an eigenvector of the matrix $W^{-1}B$ and λ is its corresponding eigenvalue. From the form of the objective function we seek to maximize we would select a to be the eigenvector corresponding to the maximum eigenvalue.

2 Ex.4.2

2.1 (a)

Part (a): Under zero-one classification loss, for each class ω_k the Bayes' discriminant func-tions $\delta_k(x)$ take the following form

$$\delta_k(x) = \ln\left(p\left(x \mid \omega_k\right)\right) + \ln\left(\pi_k\right).$$

If our conditional density $p\left(x \mid \omega_k\right)$ is given by a multidimensional normal then its function form is given by

$$p\left(x \mid \omega_k\right) = \mathcal{N}\left(x; \mu_k, \Sigma_k\right) \equiv \frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right\}.$$

Taking the logarithm of this expression as required by Equation 91 we find

$$\ln\left(p\left(x \mid \omega_k\right)\right) = -\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\left|\Sigma_k\right|\right),$$

so that our discriminant function in the case when $p\left(x \mid \omega_k\right)$ is a multidimensional Gaussian is given by

$$\delta_k(x) = -\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\left|\Sigma_k\right|\right) + \ln\left(\pi_k\right).$$

We will now consider some specializations of this expression for various possible values of Σ_k and how these assumptions modify the expressions for $\delta_k(x)$. Since linear discriminant analysis (LDA) corresponds to the case of equal covariance matrices our decision boundaries (given by Equation 93), but with equal covariances ($\Sigma_k = \Sigma$). For decision purposes we can drop the two terms $-\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|)$ and use a discriminant $\delta_k(x)$ given by

$$\delta_k(x) = -\frac{1}{2}\left(x-\mu_k\right)^T \Sigma^{-1}\left(x-\mu_k\right) + \ln\left(\pi_k\right).$$

Expanding the quadratic in the above expression we get

$$\delta_k(x) = -\frac{1}{2}\left(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k\right) + \ln\left(\pi_k\right).$$

Since $x^T\Sigma^{-1}x$ is a common term with the same value in all discriminant functions we can drop it and just consider the discriminant given by

$$\delta_k(x) = \frac{1}{2}x^T\Sigma^{-1}\mu_k + \frac{1}{2}\mu_k^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln\left(\pi_k\right).$$

Since $x^T\Sigma^{-1}\mu_k$ is a scalar, its value is equal to the value of its transpose so

$$x^T\Sigma^{-1}\mu_k = \left(x^T\Sigma^{-1}\mu_k\right)^T = \mu_k^T\left(\Sigma^{-1}\right)^T x = \mu_k^T\Sigma^{-1}x,$$

since Σ^{-1} is symmetric. Thus the two linear terms in the above combine and we are left with

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln(\pi_k)\,.$$

Next we can estimate π_k from data using $\pi_i = \frac{N_i}{N}$ for $i = 1, 2$ and we pick class 2 as the classification outcome if $\delta_2(x) > \delta_1(x)$ (and class 1 otherwise). This inequality can be written as

$$x^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \ln\left(\frac{N_2}{N}\right) > x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \ln\left(\frac{N_1}{N}\right).$$

or moving all the x terms to one side

$$x^T\Sigma^{-1}(\mu_2 - \mu_1) > \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \ln\left(\frac{N_1}{N}\right) - \ln\left(\frac{N_2}{N}\right),$$

as we were to show.

2.2 (b)

Part (b): To minimize the expression $\sum_{i=1}^{N}\left(y_i - \beta_0 - \beta^T x_i\right)^2$ over $(\beta_0, \beta)'$ we know that the solution $\left(\hat\beta_0, \hat\beta\right)'$ must satisfy the normal equations which in this case is given by

$$X^T X \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = X^T\mathbf{y}.$$

Our normal equations have a block matrix $X^T X$ on the left-hand-side given by

$$\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{N_1} & x_{N_1+1} & x_{N_1+2} & \cdots & x_{N_1+N_2} \end{bmatrix} \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \\ 1 & x_{N_1}^T \\ 1 & x_{N_1+1}^T \\ 1 & x_{N_1+2}^T \\ \vdots & \\ 1 & x_{N_1+N_2}^T \end{bmatrix}.$$

When we take the product of these two matrices we find

$$\begin{bmatrix} N & \sum_{i=1}^N x_i^T \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i x_i^T \end{bmatrix}.$$

For the case where we code our response as $-\frac{N}{N_1}$ for the first class and $+\frac{N}{N_2}$ for the second class (where $N = N_1 + N_2$), the right-hand-side or $X^T y$ of the normal equations becomes

$$\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{N_1} & x_{N_1+1} & x_{N_1+2} & \cdots & x_{N_1+N_2} \end{bmatrix} \begin{bmatrix} -N/N_1 \\ -N/N_1 \\ \vdots \\ -N/N_1 \\ N/N_2 \\ N/N_2 \\ \vdots \\ N/N_2 \end{bmatrix}.$$

When we take the product of these two matrices we get

$$\left[\begin{array}{c} N_1 \left(-\frac{N}{N_1}\right) + N_2 \left(\frac{N}{N_2}\right) \\ \left(\sum_{i=1}^{N_1} x_i\right) \left(-\frac{N}{N_1}\right) + \left(\sum_{i=N_1+1}^{N} x_i\right) \left(\frac{N}{N_2}\right) \end{array} \right] = \left[\begin{array}{c} 0 \\ -N\mu_1 + N\mu_2 \end{array} \right].$$

Note that we can simplify the $(1,2)$ and the $(2,1)$ elements in the block coefficient matrix $X^T X$ in Equation 95 by introducing the class specific means (denoted by μ_1 and μ_2) as

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N_1} x_i + \sum_{i=N_1+1}^{N} x_i = N_1\mu_1 + N_2\mu_2,$$

Also if we pool all of the samples for this two class problem ($K = 2$) together we can estimate the pooled covariance matrix $\hat{\Sigma}$ (see the section in the book on linear discriminant analysis) as

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i_j i=k} (x_i - \mu_k)(x_i - \mu_k)^T .$$

When $K = 2$ this is

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{N-2} \left[\sum_{ig_i=1} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{\dot{x}_{\xi_i}=2} (x_i - \mu_2)(x_i - \mu_2)^T \right] \\ &= \frac{1}{N-2} \left[\sum_{\dot{x}_i=1} x_i x_i^T - N_1 \mu_1 \mu_1^T + \sum_{ig_i=1} x_i x_i^T - N_2 \mu_2 \mu_2^T \right]. \end{aligned}$$

From which we see that the sum $\sum_{i=1}^{N} x_i x_i^T$ found in the $(2,2)$ element in the matrix from Equation 95 can be written as

$$\sum_{i=1}^{N} x_i x_i^T = (N-2)\hat{\Sigma} + N_1 \mu_1 \mu_1^T + N_2 \mu_2 \mu_2^T .$$

Now that we have evaluated both sides of the normal equations we can write them down again as a linear system. We get

$$\left[\begin{array}{cc} N & N_1\mu_1^T + N_2\mu_2^T \\ N_1\mu_1 + N_2\mu_2 & (N-2)\Sigma + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T \end{array} \right] \left[\begin{array}{c} \beta_0 \\ \beta \end{array} \right] = \left[\begin{array}{c} 0 \\ -N\mu_1 + N\mu_2 \end{array} \right].$$

In more detail we can write out the first equation in the above system as

$$N\beta_0 + \left(N_1\mu_1^T + N_2\mu_2^T\right)\beta = 0,$$

or solving for β_0, in terms of β, we get

$$\beta_0 = \left(-\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T\right)\beta.$$

When we put this value of β_0 into the second equation in Equation 96 we find the total equation for β then looks like

$$(N_1\mu_1 + N_2\mu_2)\left(-\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T\right)\beta + \left((N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T\right)\beta = N(\mu_2 - \mu_1).$$

Consider the terms that are outer products of the vectors μ_i (namely terms like $\mu_i\mu_j^T$) we see that taken together they look like

$$
\begin{aligned}
\text{Outer Product Terms } &= -\frac{N_1^2}{N}\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T - \frac{N_2^2}{N}\mu_2\mu_2^T + N_1\mu_1\mu_2^T + N_2\mu_2\mu_2^T \\
&= \left(-\frac{N_1^2}{N} + N_1\right)\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + \left(-\frac{N_2^2}{N} + N_2\right)\mu_2\mu_2^T \\
&= \frac{N_1}{N}\left(-N_1 + N\right)\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + \frac{N_2}{N}\left(-N_2 + N\right)\mu_2\mu_2^T \\
&= \frac{N_1N_2}{N}\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + \frac{N_2N_1}{N}\mu_2\mu_2^T \\
&= \frac{N_1N_2}{N}\left(\mu_1\mu_1^T - 2\mu_1\mu_2 - \mu_2\mu_2\right) = \frac{N_1N_2}{N}\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^T.
\end{aligned}
$$

Here we have used the fact that $N_1 + N_2 = N$. If we introduce the matrix Σ_B as

$$
\widehat{\Sigma}_B \equiv \left(\mu_2 - \mu_1\right)\left(\mu_2 - \mu_1\right)^T,
$$

we get that the equation for β boks like

$$
\left[(N-2)\Sigma + \frac{N_1N_2}{N}\Sigma_B\right]\beta = N\left(\mu_2 - \mu_1\right),
$$

as we were to show.

2.3 (c)

Note that $\hat{\Sigma}_B\beta$ is $\left(\mu_2 - \mu_1\right)\left(\mu_2 - \mu_1\right)^T\beta$, and the product $\left(\mu_2 - \mu_1\right)^T\beta$ is a scalar. Therefore the vector direction of $\Sigma_B\beta$ is given by $\mu_2 - \mu_1$. Thus in Equation 99 as both the right-hand-side and the term $\frac{N_1N_2}{N}\Sigma_B$ are in the direction of $\mu_2 - \mu_1$ the solution β must be in the direction (i.e. proportional to) $\hat{\Sigma}^{-1}\left(\mu_2 - \mu_1\right)$.

2.4 (d)

This follows directly from (b) for $N \neq 0$.

2.5 (e)

Assuming the encoding of $-N/N_1$ and N/N_2, by (b) we have

$$
\begin{aligned}
\hat{\beta}_0 &= \frac{1}{N}\left(\sum_{i=1}^{N} y_i - \left(\sum_{i=1}^{N} x_i^T\right)\beta\right) \\
&= -\frac{1}{N}\left(\sum_{i=1}^{N} x_i^T\right)\hat{\beta} \\
&= -\frac{1}{N}\left(N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T\right)\hat{\beta}
\end{aligned}
$$

so that

$$
\hat{f}(x) = \hat{\beta}_0 + x^T\hat{\beta} = \left[x^T - \frac{1}{N}\left(N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T\right)\right]\hat{\beta}
$$

Since $\hat{\beta} \propto \hat{\Sigma}^{-1}\left(\hat{\mu}_2 - \hat{\mu}_1\right)$, there exists $\lambda > 0$ (up to a scalar constant, i.e., we can flip the classification sign if $\lambda < 0$) such that $\hat{\beta} = \lambda\hat{\Sigma}^{-1}\left(\hat{\mu}_2 - \hat{\mu}_1\right)$. Therefore, $\hat{f}(x) > 0$ is equivalent to

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

which is equivalent to LDA rule (1) when $N_1 = N_2$. When $N_1 \neq N_2, \log\left(N_2/N_1\right) \neq 0$ in (1) so they are not equivalent.

3 Ex.4.3

We start by introducing notations used in Chapter 3 . Let $x_i^T=(x_{i1},\ldots,x_{ip})\in R^{1\times p}, 1^T=(1,\ldots,1)\in R^{1\times p}, Y^T=(y_1,\ldots,y_N)\in R^{1\times N}$. $\beta^T=(\beta_1,\ldots,\beta_p)\in R^{1\times p}$. Let

$$\mathbf{X}=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{N1} & \cdots & x_{N_p}\end{pmatrix}=\begin{pmatrix}1 & x_1^T\\ 1 & x_2^T\\ \vdots & \vdots\\ 1 & x_N^T\end{pmatrix}\in R^{N\times(p+1)}$$

and

$$\mathbf{X}^T=\begin{pmatrix}1 & 1 & \cdots & 1\\ x_{11} & x_{21} & \cdots & x_{N1}\\ \vdots & \vdots & \ddots & \vdots\\ x_{1p} & x_{2p} & \cdots & x_{N_p}\end{pmatrix}=\begin{pmatrix}1 & 1 & \cdots & 1\\ x_1 & x_2 & \cdots & x_N\end{pmatrix}\in R^{(p+1)\times N}$$

We have

$$\hat{\boldsymbol{Y}}=\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}=\mathbf{X}\hat{B}\in\mathbf{R}^{N\times k}$$

and $\hat{y}=\hat{B}^T\boldsymbol{x}$ for a single training sample $\boldsymbol{x}$. We estimate the new parameters of the Gaussian distributions from transformed data, denoted by $\pi_k^{\text{new}},\hat{\mu}_k^{\text{new}}$ and $\hat{\Sigma}^{\text{new}}$, and link them back with $\pi_k,\hat{\mu}_k$ and $\hat{\Sigma}$ estimated from original training data.

First, $\pi_k^{\text{new}}=\pi_k$ for $k=1,\ldots,K$ since the training sample classification does not change. Second, by definition of $\hat{\mu}_k^{\text{new}}$, note again the training sample classification does not change, we have

$$\begin{aligned}\hat{\mu}_k^{\text{new}} &=\sum_{g_r=k}\frac{\hat{B}^T x_i}{N_k}\\ &=\hat{B}^T\sum_{g_s=k}\frac{x_i}{N_k}\\ &=\hat{B}^T\hat{\mu}_k\end{aligned}$$

Third, by definition of $\hat{\Sigma}^{\text{new}}$ and result above, we have

$$\begin{aligned}\hat{\Sigma}^{\text{new}} &=\sum_{k=1}^K\sum_{g_i=k}\left(\hat{B}^T x_i-\hat{\mu}_k^{\text{new}}\right)\left(\hat{B}^T x_i-\hat{\mu}_k^{\text{new}}\right)^T/(N-K)\\ &=\frac{1}{N-K}\sum_{k=1}^K\sum_{g_i=k}\hat{B}^T\left(x_i-\hat{\mu}_k\right)\left(x_i-\hat{\mu}_k\right)^T\hat{B}\\ &=\hat{B}^T\left[\frac{1}{N-K}\sum_{k=1}^K\sum_{g_i=k}\left(x_i-\hat{\mu}_k\right)\left(x_i-\hat{\mu}_k\right)^T\right]\hat{B}\\ &=\hat{B}^T\hat{\Sigma}\hat{B}.\end{aligned}$$

Therefore, the new linear discriminant function is

$$\begin{aligned}\delta_k^{\text{new}}(x) &=\left(\hat{B}^T x\right)^T\left(\hat{\Sigma}^{\text{new}}\right)^{-1}\hat{\mu}_k^{\text{new}}-\frac{1}{2}\left(\hat{\mu}_k^{\text{new}}\right)^T\left(\hat{\Sigma}^{\text{new}}\right)^{-1}\hat{\mu}_k^{\text{new}}+\log\pi_k^{\text{new}}\\ &=x^T\hat{B}(\hat{B})^{-1}(\hat{\Sigma})^{-1}\left(\hat{B}^T\right)^{-1}\hat{B}^T\hat{\mu}_k-\frac{1}{2}\hat{\mu}_k^T\hat{B}(\hat{B})^{-1}(\hat{\Sigma})^{-1}\left(\hat{B}^T\right)^{-1}\hat{B}^T\hat{\mu}_k+\log\pi_k\\ &=x^T\hat{\Sigma}^{-1}\hat{\mu}_k-\frac{1}{2}\hat{\mu}_k^T(\hat{\Sigma})^{-1}\hat{\mu}_k+\log\pi_k,\end{aligned}$$

which is identical to the discriminant function used in the original space.

4 Ex.4.6

4.1 (a)

By definition of separability, there exists β such that

$$\begin{aligned}\beta^T x_i > 0 \text{ for } y_i = 1\\ \beta^T x_i < 0 \text{ for } y_i = -1.\end{aligned}$$

Thus we have $y_i \beta^T x_i > 0$ for all x_i, thus for $y_i \beta^T z_i > 0$ for all z_i. Define

$$m := \min_i \left\| y_i \beta^T z_i \right\|$$

Thus, $y_i \left(\frac{1}{m}\beta^T\right) z_i \geq 1$. So there exists a $\beta_{\text{sep}} := \frac{1}{m}\beta$ such that $y_i \beta_{\text{sep}}^T z_i \geq 1 \forall i$.

4.2 (b)

We have

$$\begin{aligned}\left\|\beta_{\text{new}} - \beta_{\text{sep}}\right\|^2 &= \left\|\beta_{\text{old}} - \beta_{\text{sep}} + y_i z_i\right\|^2 \\ &= \left\|\beta_{\text{old}} - \beta_{\text{sep}}\right\|^2 + \left\|y_i z_i\right\|^2 + 2 y_i \left(\beta_{\text{old}} - \beta_{\text{sep}}\right)^T z_i \\ &= \left\|\beta_{\text{old}} - \beta_{\text{sep}}\right\|^2 + 1 + 2 y_i \beta_{\text{old}}^T z_i - 2 y_i \beta_{\text{sep}}^T z_i \\ &\leq \left\|\beta_{\text{old}} - \beta_{\text{sep}}\right\|^2 + 1 + 2 \cdot 0 - 2 \cdot 1 \\ &= \left\|\beta_{\text{old}} - \beta_{\text{sep}}\right\|^2 - 1.\end{aligned}$$