Suggestions - Problem Set 3

4.2 (a) Show the discriminant condition from class or the text takes the form

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

as desired. We then replace the quantities $\mu_i,\ \Sigma_i$ by their estimates to get the proper form for this discriminant.

(b) Here using the output notations $y = -\frac{N}{N_1}$ and $y = \frac{N}{N_2}$ for classes 1 and 2 respectively, show you want to minimize

$$\sum_{i=1}^{N}(y_i-\beta_0-\beta^T\,\mathbf{x}_i)^2 = (\mathbf{y}-\tilde{\mathbf{X}}\tilde{\beta})^2,$$

where $\tilde{\beta} = \begin{bmatrix}\beta_0\\ \beta\end{bmatrix}$, letting $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^T\\ \mathbf{x}_2^T\\ \vdots\\ \mathbf{x}_N^T\end{bmatrix}$ and $\tilde{\mathbf{X}} = [\,\mathbf{1}_N\ \ \mathbf{X}] = \begin{bmatrix}1 & \mathbf{x}_1^T\\ \vdots & \vdots\\ 1 & \mathbf{x}_N^T\end{bmatrix}$.

In general vectors/matrices with a ~ on them can represent vectors augmented with 1's (and in some cases 0's).

Use the usual least squares to justify the best choice for $\tilde{\beta}$, ie.,

$$\widehat{\tilde{\beta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\,\tilde{\mathbf{X}}^T\,\mathbf{y}.$$

Thus

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\,\widehat{\tilde{\beta}} = \tilde{\mathbf{X}}^T\mathbf{y}. \qquad \textbf{(1)}$$

First consider the right hand side, $\tilde{\mathbf{X}}^T\mathbf{y}$. Show without loss you can arrange the data so the first N_1 examples $(\mathbf{x}_i, y_i)$ are in the first class and the last N_2 are in the second.

Thus show the right side of (1) becomes:

$$\tilde{\mathbf{X}}^T\,\mathbf{y} = \begin{bmatrix}\mathbf{1}_N^T\\ \mathbf{X}^T\end{bmatrix}\mathbf{y} = \begin{bmatrix}\mathbf{1}_N^T\,\mathbf{y}\\ \mathbf{X}^T\,\mathbf{y}\end{bmatrix} = \begin{bmatrix}0\\ \mathbf{X}^T\,\mathbf{y}\end{bmatrix}.$$

Meantime show

$$\mathbf{X}^T\mathbf{y} = -\frac{N}{N_1}\sum_{j=1}^{N_1}\mathbf{x}_i + \frac{N}{N_2}\sum_{j=N_1+1}^{N}\mathbf{x}_i = N(\widehat{\mu}_2-\widehat{\mu}_1)\,.$$

So

$$\tilde{\mathbf{X}}^T\,\mathbf{y} = \begin{bmatrix} 0 \\ N(\widehat{\mu}_2 - \widehat{\mu}_1) \end{bmatrix}. \tag{2}$$

To calculate the left side of (1), show you can write

$$\widetilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^T \\ 1 & \vdots \\ 1 & \mathbf{x}_{N_1}^T \\ 1 & \mathbf{x}_{N_1+1}^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}_1 \\ \tilde{\mathbf{X}}_2 \end{bmatrix}.$$

Let

$$\widetilde{\mathbf{M}} = \begin{bmatrix} \frac{1}{N_1}\,\mathbf{1}_{N_1}^T\tilde{\mathbf{X}}_1\,\mathbf{1}_{N_1} \\ \frac{1}{N_2}\mathbf{1}_{N_2}^T\tilde{\mathbf{X}}_2\,\mathbf{1}_{N_2} \end{bmatrix} = [\,\mathbf{1}_N \;\; \mathbf{M}\,]$$

i.e. $\mathbf{M}$ is the matrix whose first N_1 rows are copies of $\widehat{\mu}_1^T$, and whose last N_2 rows are copies of $\widehat{\mu}_2^T$. Here $\mathbf{1}_N$ is always a column vector of length N with all 1's.

Then show

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{1}_N^T \\ \mathbf{X}^T \end{bmatrix}[\mathbf{1}_N \;\; \mathbf{X}] = \begin{bmatrix} \mathbf{1}_N^T\mathbf{1}_N & \mathbf{1}_N^T\mathbf{X} \\ \mathbf{X}^T\mathbf{1}_N & \mathbf{X}^T\mathbf{X} \end{bmatrix} = \begin{bmatrix} N & N_1\widehat{\mu}_1^T + N_2\widehat{\mu}_2^T \\ N_1\widehat{\mu}_1 + N_2\widehat{\mu}_2 & \mathbf{X}^T\mathbf{X} \end{bmatrix}$$

Thus show

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\widehat{\tilde{\beta}} = \begin{bmatrix} N & N_1\widehat{\mu}_1^T + N_2\widehat{\mu}_2^T \\ N_1\widehat{\mu}_1 + N_2\widehat{\mu}_2 & \mathbf{X}^T\mathbf{X} \end{bmatrix}\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta} \end{bmatrix}$$

$$= \begin{bmatrix} N\widehat{\beta}_0 + (N_1\widehat{\mu}_1^T + N_2\widehat{\mu}_2^T)\widehat{\beta} \\ (N_1\widehat{\mu}_1 + N_2\widehat{\mu}_2)\widehat{\beta}_0 + \mathbf{X}^T\mathbf{X}\widehat{\beta} \end{bmatrix}.$$

Now from the relationship

$$\widehat{\mathbf{y}} = \tilde{\mathbf{X}}\,\widehat{\tilde{\beta}} = [\;\mathbf{1}_N \quad \mathbf{X}\;]\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta} \end{bmatrix} = \widehat{\beta}_0\,\mathbf{1}_N + \mathbf{X}\widehat{\beta},$$

if you average over the entries $\widehat{y}_i$ of $\widehat{\mathbf{y}}$, show

$$0 = \mathbf{1}_N^T\widehat{\mathbf{y}} = N\widehat{\beta}_0 + \mathbf{1}_N^T\mathbf{X}\widehat{\beta} = N\widehat{\beta}_0 + (N_1\widehat{\mu}_1 + N_2\widehat{\mu}_2)^T\,\widehat{\beta},$$

so

$$0 = N\,\widehat{\beta}_0 + (N_1\widehat{\mu}_1 + N_2\widehat{\mu}_2)^T\,\widehat{\beta}\,,$$

$$\widehat{\beta}_0 = -\Big(\frac{N_1}{N}\widehat{\mu}_1 + \frac{N_2}{N}\widehat{\mu}_2\Big)^T \widehat{\beta}\,,$$

so now

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\,\widehat{\tilde{\beta}} = \begin{bmatrix} 0 \\ -(N_1\widehat{\mu}_1 + N_2\widehat{\mu}_2)\Big(\frac{N_1}{N}\widehat{\mu}_1 + \frac{N_2}{N}\widehat{\mu}_2\Big)^T \widehat{\beta} + \mathbf{X}^T\mathbf{X}\,\widehat{\beta} \end{bmatrix}. \tag{3}$$

You can write

$$\mathbf{X}^T\mathbf{X} = (\mathbf{X}-\mathbf{M})^T(\mathbf{X}-\mathbf{M}) + \mathbf{X}^T\mathbf{M} + \mathbf{M}^T\mathbf{X} - \mathbf{M}^T\mathbf{M}$$

But show

$$(\mathbf{X}-\mathbf{M})^T(\mathbf{X}-\mathbf{M}) = (N-2)\widehat{\Sigma}$$

$$\mathbf{X}^T\mathbf{M} = \sum_{j=1}^{N_1}\mathbf{x}_j\widehat{\mu}_1^T + \sum_{j=N_1+1}^{N_2}\mathbf{x}_j\widehat{\mu}_2^T = N_1\widehat{\mu}_1\widehat{\mu}_1^T + N_2\widehat{\mu}_2\widehat{\mu}_2^T$$

$$\mathbf{M}^T\mathbf{M} = N_1\widehat{\mu}_1\widehat{\mu}_1^T + N_2\widehat{\mu}_2\widehat{\mu}_2^T.$$

Thus

$$\mathbf{X}^T\mathbf{X} = (N-2)\widehat{\Sigma} + N_1\widehat{\mu}_1\widehat{\mu}_1^T + N_2\widehat{\mu}_2\widehat{\mu}_2^T\,.$$

So by (3) above, show

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\;\widehat{\tilde{\beta}}$$

$$= \begin{bmatrix} 0 \\ \Big\{-(N_1\widehat{\mu}_1 + N_2\widehat{\mu}_2)\Big(\frac{N_1}{N}\widehat{\mu}_1 + \frac{N_2}{N}\widehat{\mu}_2\Big)^T + (N-2)\widehat{\Sigma} + N_1\widehat{\mu}_1\widehat{\mu}_1^T + N_2\widehat{\mu}_2\widehat{\mu}_2^T\Big\}\widehat{\beta} \end{bmatrix}. \tag{4}$$

Now show the bottom term coefficient is

$$-(N_1\widehat{\mu}_1 + N_2\widehat{\mu}_2)\Big(\frac{N_1}{N}\widehat{\mu}_1 + \frac{N_2}{N}\widehat{\mu}_2\Big)^T + (N-2)\widehat{\Sigma} + N_1\,\widehat{\mu}_1\,\widehat{\mu}_1^T + N_2\widehat{\mu}_2\widehat{\mu}_2^T$$

$$= (N-2)\widehat{\Sigma} + \Big[\frac{N_1N_2}{N}\Big](\widehat{\mu}_1 - \widehat{\mu}_2)(\widehat{\mu}_1 - \widehat{\mu}_2)^T$$

$$= (N-2)\widehat{\Sigma} + \Big[\frac{N_1N_2}{N}\Big]\Sigma_B \tag{5}$$

Now use (1), (2), (4) and (5).

(c) Show that it follows

$$\widehat{\Sigma}_B \widehat{\boldsymbol{\beta}} = (\widehat{\mu}_2 - \widehat{\mu}_1)[(\widehat{\mu}_2 - \widehat{\mu}_1)^T \widehat{\boldsymbol{\beta}}] = [(\widehat{\mu}_2 - \widehat{\mu}_1)^T \widehat{\boldsymbol{\beta}}](\widehat{\mu}_2 - \widehat{\mu}_1),$$

which is in the direction of $(\widehat{\mu}_2 - \widehat{\mu}_1)$, since $[(\widehat{\mu}_2 - \widehat{\mu}_1)^T \widehat{\boldsymbol{\beta}}]$ is a scalar (why?)

Finally from (4.56), show

$$\widehat{\beta} = ((N-2)\widehat{\Sigma})^{-1} \left[N(\widehat{\mu}_2 - \widehat{\mu}_1) - \frac{N_1 N_2}{N} [(\widehat{\mu}_2 - \widehat{\mu}_1)^T \widehat{\boldsymbol{\beta}}](\widehat{\mu}_2 - \widehat{\mu}_1) \right]$$

$$= (\text{scalar}) \cdot \widehat{\Sigma}^{-1} (\widehat{\mu}_2 - \widehat{\mu}_1). \qquad (6)$$

(d) Changing the coding for the two y values transforms the pair of numbers $-\frac{N}{N_1}$ and $\frac{N}{N_2}$ respectively into another pair a and b of possible y values. Show that there is a linear *scalar* transformation $y^* = cy + d = f(y)$ such that $f\left(-\frac{N}{N_1}\right) = a$ and $f\left(\frac{N}{N_2}\right) = b$. What are c and d? Now show that if $\mathbf{y}$ has only entries $-\frac{N}{N_1}$ and $\frac{N}{N_2}$, then in their places the vector $\mathbf{y}^* = c\mathbf{y} + d\mathbf{1}_N$ will have a and b respectively.

You wish to show that (6) above still holds. But show we are obtaining the minimizer of equation (4.55) with each y_i replaced by $cy_i + d$. Show that we only need to verify that the direction of $\widehat{\beta}$ is unchanged. First show this holds when each y_i is replaced by cy_i (what happens to $\widehat{\boldsymbol{\beta}}$ and β_0?). Now show it holds when each y_i is replaced by $y_i + d$ (what happens to $\beta_{0?}$ Does $\widehat{\boldsymbol{\beta}}$ change?). Finally show that this holds for general c, d.

(e) Now you have $\widehat{\beta}$ and β_0 and the regression function

$$\widehat{f}(\mathbf{x}) = \widehat{\beta}_0 + \widehat{\beta}^T \mathbf{x}\,.$$

From part (c), $\widehat{\beta} = k\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$ for some k. Thus show from above that

$$\widehat{\beta}_0 = -\left(\frac{N_1}{N}\widehat{\mu}_1 + \frac{N_2}{N}\widehat{\mu}_2\right)^T k\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1).$$

Recall the group targets (y-values) on which we have trained the regression are:

$$\text{class } 1: \ y = -\frac{N}{N_1}; \qquad \text{class2}: \ y = \frac{N}{N_2}\,.$$

For an input test vector $\mathbf{x}$, thus the predicted y will be in class 1 if $f(\mathbf{x})$ is closer to $-\frac{N}{N_1}$ than to $\frac{N}{N_2}$, and otherwise class 2. Show y should be assigned to class 2 if

$$f(\mathbf{x}) > \frac{1}{2}\left(\frac{-N}{N_1} + \frac{N}{N_2}\right)\,. \qquad (7)$$

Show from above that the criterion for class 2 assignment is:

$$f(\mathbf{x}) = \left[-\left(\frac{N_1}{N}\widehat{\mu}_1 + \frac{N_2}{N}\widehat{\mu}_2\right)^T + \mathbf{x}^T\right]k\widehat{\Sigma}^{-1}\left(\widehat{\mu}_2 - \widehat{\mu}_1\right) > \frac{1}{2}\left(\frac{-N}{N_1} + \frac{N}{N_2}\right)$$

or

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

Is this the same as the LDA criterion in (a)? Now assume $N_1 = N_2 = N/2$ - what happens then?