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, you want to minimize

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

where \( \hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} \), letting \( \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} \) and \( \hat{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{bmatrix} \).

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 find the best choice for \( \hat{\beta} \), ie.,

\[ \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \]

Thus

\[ \mathbf{X}^T \hat{\mathbf{X}} \hat{\beta} = \hat{\mathbf{X}}^T \mathbf{y}. \] (1)

First consider the right hand side, \( \mathbf{X}^T \mathbf{y} \). 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 the right side of (1) becomes:

\[ \mathbf{X}^T \mathbf{y} = \begin{bmatrix} 1_N^T \\ \mathbf{x}_T \end{bmatrix} \mathbf{y} = \begin{bmatrix} 1_N^T \mathbf{y} \\ \mathbf{X}^T \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{X}^T \mathbf{y} \end{bmatrix}. \]

Meantime show

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

So
\[
\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 0 \\ N(\tilde{\mu}_2 - \tilde{\mu}_1) \end{bmatrix}.
\]

To calculate the left side of (1), you can write
\[
\mathbf{\bar{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^T \\ 1 & \vdots \\ 1 & \mathbf{x}_{N_1}^T \\ 1 & \mathbf{x}_{N_2}^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} \mathbf{\bar{X}}_1 \\ \mathbf{\bar{X}}_2 \end{bmatrix}.
\]

Let
\[
\mathbf{\bar{M}} = \begin{bmatrix} \frac{1}{N_1} \mathbf{1}_{N_1}^T \mathbf{\bar{X}}_1 \mathbf{1}_{N_1} \\ \frac{1}{N_2} \mathbf{1}_{N_2}^T \mathbf{\bar{X}}_2 \mathbf{1}_{N_2} \end{bmatrix} = [\mathbf{1}_N \mathbf{M}]
\]
i.e. \( \mathbf{M} \) is the matrix whose first \( N_1 \) rows are copies of \( \tilde{\mu}_1^T \), and whose last \( N_2 \) rows are copies of \( \tilde{\mu}_2^T \). Here \( \mathbf{1}_N \) is always a column vector of length \( N \) with all 1's.

Then show
\[
\mathbf{X}^T \mathbf{\bar{X}} = \begin{bmatrix} \mathbf{I}_N^T & \mathbf{X}_1 \mathbf{1}_N \end{bmatrix} [\mathbf{1}_N \mathbf{X}] = \begin{bmatrix} \mathbf{1}_N^T \mathbf{I}_N & \mathbf{1}_N^T \mathbf{X} \\ \mathbf{X}_1 \mathbf{1}_N & \mathbf{X}_1 \mathbf{X} \end{bmatrix} = \begin{bmatrix} \frac{N N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2} {N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2} \\ \mathbf{X}^T \mathbf{X} \end{bmatrix}
\]
Thus show
\[
\mathbf{X}^T \mathbf{\hat{X}} \mathbf{\hat{\beta}} = \begin{bmatrix} \frac{N}{N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2} & \frac{N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2} {N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2} \mathbf{X}^T \mathbf{X} \end{bmatrix} \begin{bmatrix} \tilde{\beta}_0 \\ \tilde{\beta} \end{bmatrix}
\]
\[
= \begin{bmatrix} \frac{N \tilde{\beta}_0 + (N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2) \tilde{\beta}} {N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2} \\ \tilde{\beta}_0 + \mathbf{X}^T \mathbf{\hat{X}} \tilde{\beta} \end{bmatrix}.
\]
Now from the relationship
\[
\mathbf{\hat{y}} = \mathbf{\hat{X}} \tilde{\beta} = \begin{bmatrix} \mathbf{1}_N & \mathbf{X} \end{bmatrix} \begin{bmatrix} \tilde{\beta}_0 \\ \tilde{\beta} \end{bmatrix} = \tilde{\beta}_0 \mathbf{1}_N + \mathbf{X} \tilde{\beta},
\]
if you average over the entries \( \hat{y}_i \) of \( \mathbf{\hat{y}} \), show
\[
0 = \mathbf{1}_N^T \mathbf{\hat{y}} = N \tilde{\beta}_0 + \mathbf{1}_N^T \mathbf{X} \tilde{\beta} = N \tilde{\beta}_0 + (N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2) \mathbf{\hat{\beta}},
\]
so
\[
0 = N \tilde{\beta}_0 + (N_1 \tilde{\mu}_1 + N_2 \tilde{\mu}_2) \mathbf{\hat{\beta}},
\]
\[
\hat{\beta}_0 = - \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T \hat{\beta},
\]

so now
\[
\hat{X}^T \hat{X} \hat{\beta} = \begin{bmatrix}
0 \\
- (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T \hat{\beta} + \hat{X}^T \hat{X} \hat{\beta}
\end{bmatrix}. \tag{3}
\]

You can write
\[
\hat{X}^T \hat{X} = (X - M)^T (X - M) + \hat{X}^T \hat{M} + M^T X - M^T M
\]

But show
\[
(X - M)^T (X - M) = (N - 2) \hat{\Sigma}
\]
\[
\hat{X}^T \hat{M} = \sum_{j=1}^{N_1} x_j \hat{\mu}_1^T + \sum_{j=N_1+1}^{N_2} x_j \hat{\mu}_2^T = N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T
\]
\[
M^T M = N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T.
\]

Thus
\[
\hat{X}^T \hat{X} = (N - 2) \hat{\Sigma} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T.
\]

So by (3) above, show
\[
\hat{X}^T \hat{X} \hat{\beta}
\]
\[
= \begin{bmatrix}
0 \\
- (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T + (N - 2) \hat{\Sigma} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T
\end{bmatrix} \hat{\beta}. \tag{4}
\]

Now show the bottom term coefficient is
\[
- (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T + (N - 2) \hat{\Sigma} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T
\]
\[
= (N - 2) \hat{\Sigma} + \left[ \frac{N_1 N_2}{N} \right] (\hat{\mu}_1 - \hat{\mu}_2) (\hat{\mu}_1 - \hat{\mu}_2)^T
\]
\[
= (N - 2) \hat{\Sigma} + \left[ \frac{N_1 N_2}{N} \right] \Sigma_B \tag{5}
\]

Now use (1), (2), (4) and (5).
(c) It follows that
\[ \hat{\Sigma} \hat{\beta} = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\beta} = [(\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\beta}](\hat{\mu}_2 - \hat{\mu}_1), \]
which is clearly in the direction of \((\hat{\mu}_2 - \hat{\mu}_1)\), since \([(\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\beta}]\) is a scalar (why?)

Finally from (4.56),
\[ \hat{\beta} = ((N - 2)\bar{\Sigma})^{-1} \left[ N(\hat{\mu}_2 - \hat{\mu}_1) - \frac{N_1N_2}{N}[(\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\beta}](\hat{\mu}_2 - \hat{\mu}_1) \right] \]
\[ = (\text{scalar}) \cdot \bar{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \quad (6) \]

(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 \(y\) has only entries \(-\frac{N}{N_1}\) and \(\frac{N}{N_2}\), then in their places the vector \(y^* = cy + d1_N\) will have \(a\) and \(b\) respectively.

We 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 \(\hat{\beta}\) is unchanged. First show this holds when each \(y_i\) is replaced by \(cy_i\) (what happens to \(\hat{\beta}\) and \(\beta_0\)?). Now show it holds when each \(y_i\) is replaced by \(y_i + d\) (what happens to \(\beta_0\)? Does \(\hat{\beta}\) change?). Finally show that this holds for general \(c, d\).

(e) Now you have \(\hat{\beta}\) and \(\beta_0\) and the regression function
\[ \hat{f}(x) = \beta_0 + \beta^T x. \]

From part (c), \(\hat{\beta} = k\bar{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)\) for some \(k\). Thus show from above that
\[ \beta_0 = -\left(\frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2\right)^T k\bar{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \]

Recall the group targets (\(y\)-values) on which we have trained the regression are:
\[ \text{class 1 : } y = -\frac{N}{N_1}; \quad \text{class 2 : } y = \frac{N}{N_2}. \]

For an input test vector \(x\), thus the predicted \(y\) will be in class 1 if \(f(x)\) 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
\[ f(x) > \frac{1}{2} \left( -\frac{N}{N_1} + \frac{N}{N_2} \right). \quad (7) \]
Show from above that the criterion for class 2 assignment is:

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

or

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

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