Multivariate Analysis Assignment #4 Due 5/29/2018

<u>Perform the calculations in calculator, SAS IML or other Matrix software.</u>

1. (10%) Textbook 11.1

11.1. Consider the two data sets

$$\mathbf{X}_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{X}_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$$

for which

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

and

$$\mathbf{S}_{\text{pooled}} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Calculate the linear discriminant function in (11-19).

(b) Classify the observation $\mathbf{x}_0' = [2 \quad 7]$ as population π_1 or population π_2, using Rule (11-18) with equal priors and equal costs.

2. (10%) Textbook 11.4

11.4. A researcher wants to determine a procedure for discriminating between two multivariate populations. The researcher has enough data available to estimate the density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ associated with populations π_1 and π_2, respectively. Let $c(2|1) = 50$ (this is the cost of assigning items as π_2, given that π_1 is true) and $c(1|2) = 100$.

In addition, it is known that about 20% of all possible items (for which the measurements $\mathbf{x}$ can be recorded) belong to π_2.

(a) Give the minimum ECM rule (in general form) for assigning a new item to one of the two populations.

(b) Measurements recorded on a new item yield the density values $f_1(\mathbf{x}) = .3$ and $f_2(\mathbf{x}) = .5$. Given the preceding information, assign this item to population π_1 or population π_2.

3. (10%) Textbook 11.19(b)(c) Perform the calculations by textbook formula and SAS outputs

(b) Using the calculations in Part a, compute Fisher's linear discriminant function, and use it to classify the sample observations according to Rule (11-25). Verify that the confusion matrix given in Example 11.7 is correct.

(c) Classify the sample observations on the basis of smallest squared distance $D_i^2(\mathbf{x})$ of the observations from the group means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$. [See (11-54).] Compare the results with those in Part b. Comment.

4. (70%) Textbook 11.32

11.32 Data on homophilia A carriers, similar to those used in example 11.3, are listed in Table 11.8 on page 664. Using these data,

(a) (10%) Plot the bivariate plot of the data for two groups. Do bivariate normal distributions appear to be valid by visual inspection? Perform the following tests using SAS commands: pool=test (for testing common variance) manova (for testing if means are equal) Assume equal prior probabilities for (b) to (e)

(b) (10%) Construct linear discriminant function (textbook formula & SAS) Estimate the error rate using the holdout procedure (SAS output) What are the misclassified observations? (SAS output)

(c) (5%) Classify $\mathbf{x}'_0 = [-0.112, -0.279]$ using the discriminant functions in (b) (textbook formula)

(d) (10%) Classify $\mathbf{x}'_0 = [-0.112, -0.279]$ based on posterior probabilities (SAS output). Use Matlab to verify the posterior probabilities (textbook formula & SAS output such as pooled covariance matrix and group means)

(e) (5%) Calculate linear discriminant score of $\mathbf{x}'_0 = [-0.112, -0.279]$ (textbook formula & SAS output) Assume the prior probabilities of carriers (group 2) is 0.25 and that of noncarriers (group 1) is 0.75 for (f) to (i)

(f) (5%) Construct linear discriminant function. (textbook formula & SAS) Estimate the error rate using the holdout procedure. (SAS outputs) What are the misclassified observations? (SAS output)

(g) (5%) Classify $\mathbf{x}'_0 = [-0.112, -0.279]$ using the discriminant functions in (f) (textbook formula)

(h) (5%) Classify $\mathbf{x}'_0 = [-0.112, -0.279]$ based on posterior probabilities (SAS outputs). Use Matlab to verify the posterior probabilities (textbook formula & SAS output such as pooled covariance matrix and group means)

(i) (5%) Calculate linear discriminant score of $\mathbf{x}'_0 = [-0.112, -0.279]$ (textbook formula)

(j) (10%) Plot the discriminant functions constructed in (b) and (f) onto the bivariate plot constructed in (a). Show the point of $\mathbf{x}'_0 = [-0.112, -0.279]$. How do the changes of prior probabilities move the discriminant functions?