Stat 6559 / Stat 4560
Assignment #2: Due Wednesday March 15 , 2023 in class

Question 1. Observations on two response variables are collected for two treatments. The observation vectors $[x_1, x_2]$ are Treatment 1: (3,3) (1,6), (2,3)
Treatment 2: (2,3), (5,1), (3,1), (2,3)
a) Calculate the $\mathbf{S_{pooled}}$
b) Test $H_0 : \mu_1 = \mu_2$ employing a two sample approach with $\alpha = 0.01$

Question 2. Generate a data set with two explanatory variables x_1 and x_2 from multinomial Normal distribution with covariance matrix $\sigma = c(1, .2, .2, 4)$ in two classes with mean for Class 0 is (3,7) and and Class 1 is (6,10). For the Class 0, generate 50 observations and for Class 1, 50 observations. While generating this data, use the set.seed("99"). Find the linear discriminant function weights. Plot the data with two colors and draw the discriminant function for classification. Also plot the 4 test data $(3.68, 5.65), (3.28, 5.20), (3.57, 8.82), (4.64, 7.98)$ and predict the test data. Use the R program also to predict the test data.

Question 3. This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.
(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?
(b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).
(e) Repeat (d) using LDA and QDA. Interpret the results.
(f) Which of these methods appears to provide the best results on this data?

Question 4. Construct the Hotelling T^2 charts for future observations using the a simulated data
Simulation Set-up
a) Use the set.seed("6559)
b) Generate 100 observations from bivariate normal distribution with $\mu = (2, 5)$ and covariance matrix - var(x1)=1; var(x2)=.5, cov(x1,x2)=0.3.
c) Estimate the classical estimators of mean and covariances
d) Generate 25 future observations, using bivariate normal distribution with $\mu = (2, 5)$ and covariance matrix - var(x1)=1; var(x2)=.5, cov(x1,x2)=0.3.
e) Draw three T^2 control chart for future observation using classical estimator and robust estimators of mean and covariance matrix. Draw your conclusions. g) Generate another 25 future observations, using bivariate normal distribution with $\mu = (2.4, 6)$ and covariance matrix - var(x1)=1; var(x2)=.5, cov(x1,x2)=0.3. and repeat (e).
f) Offer your comments. Compare your results with univariate charts for individual observations.

Question 5.(Bonus Question) Refer the class note on discriminant analysis and definition notations of $\mathbf{w}, \mathbf{B}\&\mathbf{S}$. Show that the $\mathbf{w}$ maximizing

$$\frac{w^T B w}{w^T S w}$$

satisfies

$$S^{-1} B w = \lambda w$$

(Hence, w is the eigenvector and λ is eigenvalue of $S^{-1}B$.)
Hint: Argue that we can maximize $w^T B w$ subject to $w^T S w = a$, where a is a constant. Then introduce a Lagrange multiplier for the constraint and differentiate with respect to elements of w.