Homework for DA with partial answers

1. This problem builds on the data simulation problem from the cluster analysis practice problems. Because we knew which observations belong in which cluster, we can now apply DA to the same data to better understand how well DA works.
	1. Simulate 100 observations each from two populations with the following distributions:

 and 

Set a seed number of 1181 before using the first distribution and a seed number of 1218 before using the second distribution. Plot the simulated data in a scatter plot where the plotting symbols correspond to the population. Based on this plot, how well do you think DA will separate out the observations into the correct populations?

> library(mvtnorm)

> set.seed(1181)

> pop1<-rmvnorm(n = 100, mean = c(0,0), sigma = matrix(data = c(1, 0, 0, 1), nrow =

 2, ncol = 2))

> head(pop1)

 [,1] [,2]

[1,] -0.2962019 -1.02229182

[2,] -0.2277220 2.09802300

[3,] 1.8393558 0.29361539

[4,] -0.7641729 -1.16671752

[5,] -0.2279988 0.05907115

[6,] -0.5433259 0.61151723

> set.seed(1218)

> pop2<-rmvnorm(n = 100, mean = c(6,0), sigma = matrix(data = c(1, 0, 0, 1), nrow =

 2, ncol = 2))

> head(pop2)

 [,1] [,2]

[1,] 6.005221 -0.264053889

[2,] 6.013727 0.594869118

[3,] 7.351746 0.003988692

[4,] 8.637141 0.778251779

[5,] 6.447020 0.211541760

[6,] 6.298675 -0.635608118

> dev.new()

> par(pty = "s")

> plot(x = pop1[,1], y = pop1[,2], xlab = "Variable #1", ylab = "Variable #2", pch

 = 1, col = "blue", xlim = c(-5, 10), ylim = c(-5, 10), panel.first = grid())

> points(x = pop2[,1], y = pop2[,2], xlab = "Variable #1", ylab = "Variable #2",

 pch = 2, col = "red")

> legend(x = 0, y = 8, legend = c("Pop. #1", "Pop. #2"), col = c("blue", "red"),

 pch = c(1, 2), bty = "n")



The two groups of points are fairly well separated so I would expect DA to find the correct populations.

* 1. Combine the two sets of data into one data frame and perform linear and quadratic DA. Examine how well DA differentiates between the two populations.

Both have 100% accuracy!

* 1. Repeat a) and b) with the variable #1 mean for population #2 being decreased from 6 to 5, 4, 3, and 2. Describe trends you see as the mean decreases.
	2. For the mean of 2 case given in c), construct a scatter plot showing the observations with their original and classified populations. Discuss which observations are misclassified.

Here’s my plot (smaller symbols denote the original populations and larger symbols denote the classifications):



One could plot a vertical line on it to show where the discriminant rule makes its decision. A plot of only the classifications can help you see this better.