STAT 445 Assignment 2

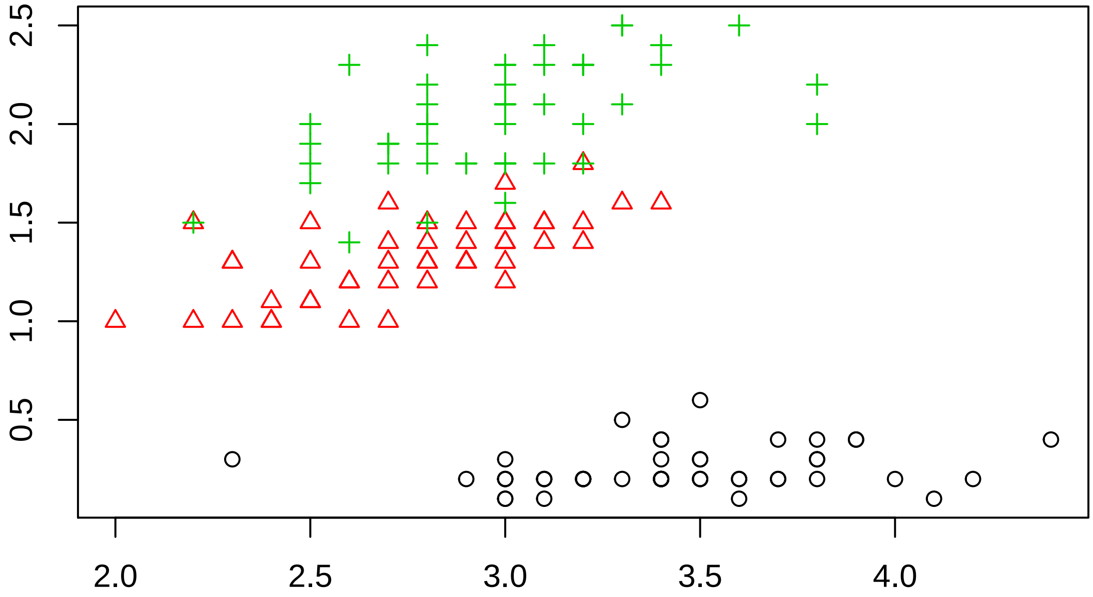
Question 11.27

11.27 a) Read in the dataset.

```
t11_5 <- read.table("T11-5.DAT.txt", header = F, col.names=c("sepal_length", "sepal_width", "petal_leng
t11_5$groups <- factor(t11_5$groups, labels = c("setosa", "versicolor", "virginica"))
summary(t11_5)
```

```
##   sepal_length    sepal_width     petal_length    petal_width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        groups  
##  setosa    :50  
##  versicolor:50  
##  virginica :50  
##                 
##                 
##                 
```

```
plot(as.matrix(t11_5[,c(2,4)]), xlab="", ylab="", col = as.integer(t11_5$groups), pch= as.integer(t11_5
```



Looking at the plot, it is not fair to assume a multivariate normal distribution as the cluster of data points for each group is not shaped elliptically. c)

```
library(MASS)
qda_model <- qda(groups ~ sepal_width + petal_width, data=t11_5, prior=c(1/3, 1/3, 1/3) )
qda_pred <- predict(qda_model, newdata = data.frame(sepal_width=3.5, petal_width=1.75))
xbar <- by(t11_5[,c(2,4)], t11_5$groups, colMeans )
n <- rep(50,3)
```

```r
S <- by(t11_5[,c(2,4)], t11_5$groups, var)
Spool <- (S[[1]]*(n[1]-1) + S[[2]]*(n[2]-1) + S[[3]]*(n[3]-1))/(sum(n-1))
prior <- c(1/3, 1/3, 1/3)
for (i in 1:3) {
  print(log(det(S[[i]])))
}
```

```
## [1] -6.496053
## [1] -6.140903
## [1] -5.189136
```

The above list all the

$$ln(|S_i|)$$

.

d)

```r
ceofMat <- matrix(0, 3, 3)

for(i in 1:nrow(ceofMat)){
ceofMat[i,] <- c(-0.5*t(xbar[[i]])%*%solve(Spool)%*%xbar[[i]],
                 t(xbar[[i]])%*%solve(Spool))
}
ceofMat
```

```
##            [,1]     [,2]      [,3]
## [1,] -58.99711 36.01791 -22.25685
## [2,] -37.73207 19.30501  16.58314
## [3,] -59.78197 15.49036  36.27622
```

The above is the matrix of coefficients for the linear discrimate formula.

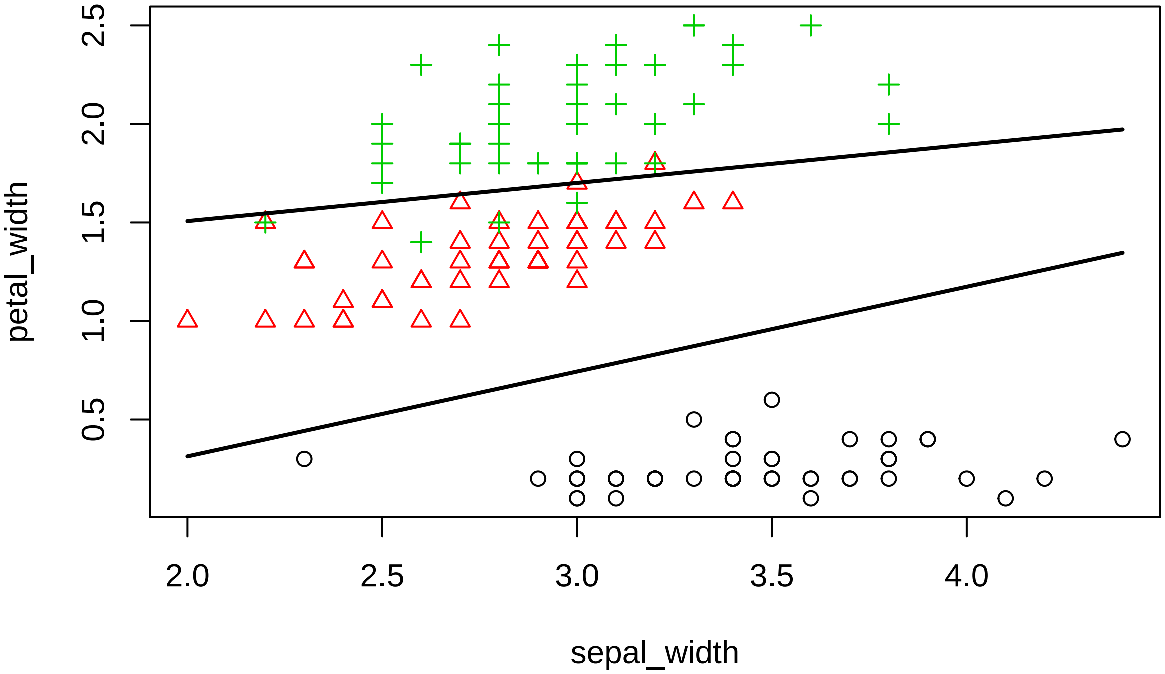
e) Functins ‘yval’, ‘dhat’, and ‘drawline’ are taken from ldaClassificationBoundary.rmd.

```r
d_matrix <- matrix(0, 3, 3)
x = c(3.5, 1.75)
for (i in 1:3) {
  for (j in 1:3)
    d_matrix[i,j] = t(xbar[[i]]-xbar[[j]])%*%solve(Spool)%*%x - 0.5*t(xbar[[i]]-xbar[[j]])%*%solve(Spool
}
calCeofMat <- function(data, prior){
ngroup <- length(unique(data$groups))
xbar <- by(data[,c(2,4)], data$groups, colMeans)
n <-by(data[,c(2,4)], data$groups, nrow)
S <- by(data[,c(2,4)], data$groups, var)
Spool <- matrix(0,2,2)
for(i in 1:ngroup)
  Spool <- Spool+S[[i]]*(n[i]-1)
Spool <- Spool/(sum(n-1))
ceofMat <- matrix(0, ngroup, 3)
for(i in 1:nrow(ceofMat)){
ceofMat[i,] <- c(-0.5*t(xbar[[i]])%*%solve(Spool)%*%xbar[[i]]+log(prior[i]),
                 t(xbar[[i]])%*%solve(Spool))
}
ceofMat
}
# coef includes intercept, coefficient for x1, coefficient for x2 (y)
# get the y value when we know the intercept, coefficients, and x
yval <- function(coef, x)
{
  -coef[1]/coef[3] - coef[2]*x/coef[3]
}

# the discriminant function
dhat <- function(coef, x1,x2)
{
  cbind(1,x1,x2)%*%coef
}


# draw the boundary between groups g1 and g2
drawline <- function(g1, g2, ngroup, ceofMat, x1){
y1  <- yval(ceofMat[g1,]-ceofMat[g2,], x1)
sel <- rep(TRUE,length(x1))
for(i in 1:ngroup)
{
  if(i!=g1 && i!=g2)
sel <- sel & (dhat(ceofMat[g1,], x1,y1) >= dhat(ceofMat[i,], x1,y1))
}
lines(x1[sel],y1[sel], lwd=2)
}

# draw all the boundaries between groups
drawBoundaries <- function(data, prior)
{
x1range <- range(data[,2])
x1points  <- seq(x1range[1], x1range[2], length.out = 200)
ceofMat <- calCeofMat(data, prior)
plot(data[,c(2,4)], col = as.integer(data$groups), pch = as.integer(data$groups))
ngroup <- length(unique(data$groups))
for(i in 1:(ngroup-1))
  for(j in (i+1):ngroup)
  drawline(i,j, ngroup,ceofMat,x1points)
}
drawBoundaries(t11_5, c(1/3, 1/3, 1/3))
```



e)

```
z <- lda(groups ~ sepal_width + petal_width, data=t11_5, prior=c(1/3, 1/3, 1/3))
conf_matrix <- table((predict(z, newdata=t11_5[,c(2,4)]))$class, t11_5$groups)
APER = (conf_matrix[2,3] + conf_matrix[3,2])/150
APER
```

```
## [1] 0.03333333
```

```
z1 <- lda(groups ~ sepal_width + petal_width, data=t11_5, prior=c(1/3, 1/3, 1/3), CV=T)
conf_matrix_h <- table(z1$class, t11_5$groups)
AER <- (conf_matrix_h[2,3] + conf_matrix_h[3,2])/150
AER
```

```
## [1] 0.04
```

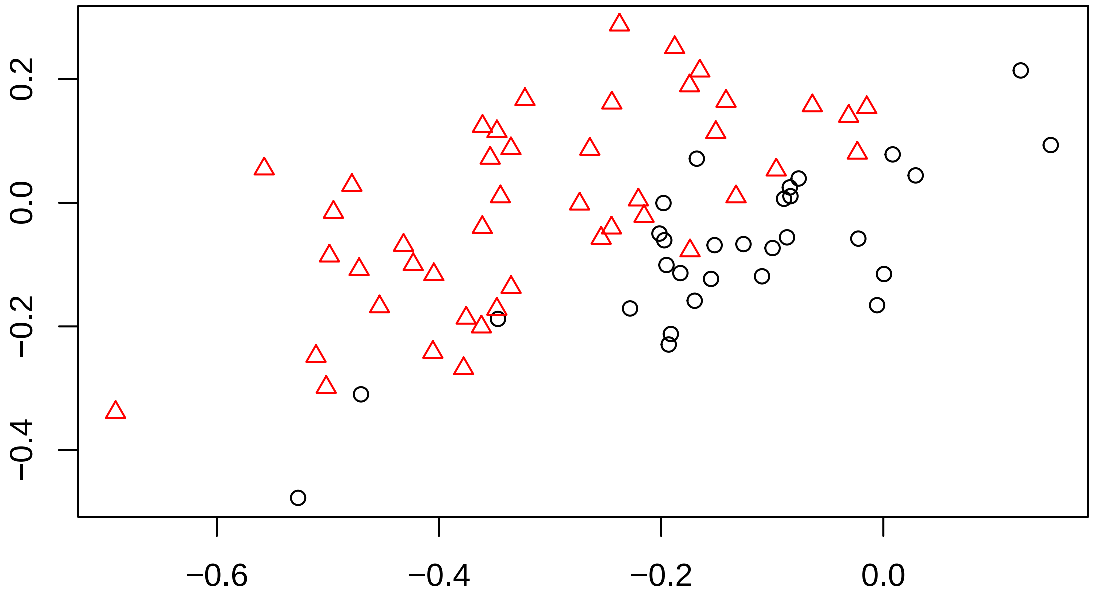
$$APER = \frac{1+4}{150} = \frac{1}{30} = 0.033$$

$$AER = \frac{2+4}{150} = \frac{6}{150} = 0.040$$

Question 11.32

a)

```
t11_8 <- read.table("T11-8.DAT.txt", header = F, col.names=c("group", "AFH_activity", "AFH_antigen"))
plot(as.matrix(t11_8[,2:3]), xlab="", ylab="", col = as.integer(t11_8$group), pch= as.integer(t11_8$grou
```



It is fair to assume a bivariate normal distribution because the data points is elliptical shape.

b)

```
xbar1 <- by(t11_8[,2:3], t11_8$group, colMeans)
n1 <-by(t11_8[,2:3], t11_8$group, nrow)
S1 <- by(t11_8[,2:3], t11_8$group, var)
Spool = (S1[[1]]*(n1[[1]]-1) + S1[[2]]*(n1[[2]]-1))/sum(n1-1)
m_hat <- 0.5*t(xbar1[[1]] - xbar1[[2]])%*%solve(Spool)%*%(xbar1[[1]]+xbar1[[2]])
cat(t(xbar1[[1]]-xbar1[[2]])%*%solve(Spool), -m_hat)
```

```
## 19.319 -17.12424 3.559472
```

```
zz <- lda(group ~ AFH_activity + AFH_antigen, data=t11_8, CV=T)

table(zz$class, t11_8$group)
```

```
##
##      1  2
##   1 26  7
##   2  4 38
```

We allocate x to population 1 if

$$19.319x_1 - 17.12424x_2 + 3.559472 \geq 0$$

Otherwise, to population 2.

$$AER = \frac{4+7}{75} = \frac{11}{75}$$

c)

```
new_cases <- matrix(c(-.112,-.279,-0.059,-.068,0.064,0.012,-.043, -.052, -.05, -.098, -.094, -.113, -.1

cbind(new_cases, rep(1, 10)) %*% t(cbind(t(xbar1[[1]]-xbar1[[2]])%*%solve(Spool), -m_hat))
```

```
##           [,1]
##  [1,] 6.173406
##  [2,] 3.584100
##  [3,] 4.590397
##  [4,] 3.619216
##  [5,] 4.271698
##  [6,] 3.678525
##  [7,] 3.632001
##  [8,] 3.980560
##  [9,] 1.043664
## [10,] 1.450639
```

The list is the 10 linear discrimate functions for the 10 cases. Since all prior probabilities for both populations are equal,

$$ln(\frac{p_i}{p_j}) = 0$$

. All cases have linear discriminate value greater than or equal to 0. We classify the 10 cases under population 1, noncarriers.

d)

```
cbind(new_cases, rep(1, 10)) %*% t(cbind(t(xbar1[[1]]-xbar1[[2]])%*%solve(Spool), -m_hat)) - rep(log(1/3
```

```
##             [,1]
##  [1,] 7.272019
##  [2,] 4.682712
##  [3,] 5.689010
##  [4,] 4.717828
##  [5,] 5.370310
##  [6,] 4.777138
##  [7,] 4.730614
##  [8,] 5.079172
##  [9,] 2.142276
## [10,] 2.549252
```

The above values are calculated from the formula given below. We classify x to population 1, noncarriers if

$$(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) - \ln(\frac{p_2}{p_1}) \geq 0$$

And population 2, otherwise. We classify the ten cases under population 1, noncarriers, because all the values are greater than 0.