Math 680, Fall 2020
Homework 2 Due: Tuesday, 10/20/2020

1. Read the paper "Choosing between logistic regression and discriminant analysis?" by Press and Wilson (1978). Summarize your understanding in three sentences.

2. Assume that n data points are uniformly distributed in a p-dimensional unit ball centered at the origin. Let R_i be the distance from the origin to the ith data point, for $i = 1, \ldots, n$.

   (i) Derive the median of $R_{(1)}$, where $R_{(1)}$ is the minimal order statistics of R_i's.

   (ii) Compute the values of $R_{(1)}$ if $n = 500, p = 10$ and if $n = 500, p = 100$, respectively. What is your conclusion?

3. (**Linear Methods for Classification: Scenerio 1**)

   (a) Train the linear regression model, using the function "lm(y∼x)", with the training set generated in Question 4 of HW1. Report the training and testing errors.

   (b) Fit the LDA for the training data in Scenario 1. Report the training and testing errors.

   (c) Fit the logistic regression for the training data in Scenario 1. Report the training and testing errors.

   (d) Compare (a)(b)(c) with the Bayes rule in terms of their errors (you have done in Question 5 of HW 1), and write down your comments.

4. (**Two-Class Classification Problem: Scenerio 2**) (Textbook page 17) Generate a training set of $n = 200$ from a mixture data as follows.

step 1: Generate 10 points $\mu_k, k = 1, ..., 10$ from a bivariate Gaussian distribution $N((1, 0)^T, \mathbf{I})$. They will be used as means (centers) to generate the **Green** class for both training and test data.

step 2: Generate 10 points $\nu_k, k = 1, ..., 10$ from a bivariate Gaussian distribution $N((0, 1)^T, \mathbf{I})$. They will be used as means (centers) to generate the **Red** class.

step 3: For the **Green** class, generate 100 observations as follows: for each observation, randomly pick a μ_k with probability 1/10, and then generate a point from $N(\mu_k, \mathbf{I}/5)$.

step 4: For the **Red** class, generate 100 observations as follows: for each observation, randomly pick a ν_k with probability 1/10, and then generate a point from $N(\nu_k, \mathbf{I}/5)$.

(a) Use the following code to generate the training set:

```
library(MASS)

#generate ten centers, which are treated
as fixed parameters
Sig <- matrix(c(1,0,0,1),nrow=2)
```

```
seed_center <- 16
set.seed(seed_center)
center_green <- mvrnorm(n=10,c(1,0),Sig)
center_red <- mvrnorm(n=10,c(0,1),Sig)

##define a function "gendata2" first
gendata2 <-function(n,mu1,mu2,Sig1,Sig2,myseed)
{
set.seed(myseed)
mean1 <- mu1[sample(1:10,n,replace=T),]
mean2 <- mu2[sample(1:10,n,replace=T),]
green <- matrix(0,ncol=2,nrow=n)
red <- matrix(0,ncol=2,nrow=n)
for(i in 1:n){
green[i,] <- mvrnorm(1,mean1[i,],Sig1)
red[i,] <- mvrnorm(1,mean2[i,],Sig2)
}
x <- rbind(green,red)
return(x)
}

#generate the training set
seed_train <- 2000
ntrain <- 100
train2 <- gendata2(ntrain,center_green,center_red,
Sig/5, Sig/5,seed_train)
```

```
ytrain <- c(rep(1,ntrain),rep(0,ntrain))
```

(b) Draw the scatter plot of the training set, using different labels/colors for two classes.

(c) Generate a test set, with 500 observations from each class, using *set.seed(2014)*. The same center parameters are used in the training and test sets. Save the test set for future use.

Submit the scatter plot.

5. (**Linear Methods for Classification: Scenario 2**)

   (a) Train the linear regression model, using the function "lm(y$\sim$x)", with the training set generated in Question 4.

   (b) Add the linear decision boundary to the scatterplot.

   (c) Report the training and test errors for this linear classification rule.