HW4

Runye Hu A13410385
Yijie Fan A13485989
Xinyi He A13561164
Yifan Wu A14060535
Yiwen Cai A13530685

```
############## Load Data ##############
library(datasets)
data(iris)

#size=train size from each species
#Prior=relative sample size in train datap is prori
```

```
LDA <- function(train1,train2,train3, prior)
{
  #Divide data into train and test
  train_set = c(train1,train2,train3)
  train=iris[train_set,]
  test=iris[-train_set,]
  #Sample size
  n_setosa=length(train1)
  n_versicolor=length(train2)
  n_virginica=length(train3)
  ############## Calculate sample mean vectors ##############
  Mean_setosa=colMeans(iris[train1,1:4])
  Mean_versicolor=colMeans(iris[train2,1:4])
  Mean_virginica=colMeans(iris[train3,1:4])
  ############## Calculate pooled variance-covariance matrix ##############
  #Sample variance-covariance matrix for each species
  S_setosa=cov(iris[train1,1:4])
  S_versicolor=cov(iris[train2,1:4])
  S_virginica=cov(iris[train3,1:4])
  #Complete fomula
  S_pooled= ((n_setosa-1)*S_setosa+(n_versicolor-1)*S_versicolor+(n_virginica-1)*S_virgi
nica)/(n_setosa+n_versicolor+n_virginica-3)
  S_inv=solve(S_pooled)
  #Simple way
  #S_pooled=(S_setosa+S_versicolor+S_virginica)/3
  ############## Calculate alpha_i ##############
  alpha_setosa= -0.5* t(Mean_setosa) %*% S_inv %*% Mean_setosa + log(prior[1])
  alpha_versicolor= -0.5* t(Mean_versicolor) %*% S_inv %*% Mean_versicolor + log(prior[2
])
  alpha_virginica= -0.5* t(Mean_virginica) %*% S_inv %*% Mean_virginica + log(prior[3])
  ############## Calculate beta_i ##############
  beta_setosa=S_inv %*% Mean_setosa
  beta_versicolor=S_inv %*% Mean_versicolor
  beta_virginica=S_inv %*% Mean_virginica
  ############## Classification ##############
  prediction=c()
  d_setosa_vec=c()
  d_versicolor_vec=c()
  d_virginica_vec=c()
  label=c("setosa", "versicolor", "virginica")
  for(i in 1:nrow(test)){
    #Read an observation in test data
    x=t(test[i,1:4])
    #Calculate linear discriminant functions for each species
    d_setosa=alpha_setosa+ t(beta_setosa) %*% x
    d_versicolor=alpha_versicolor+ t(beta_versicolor) %*% x
    d_virginica=alpha_virginica+ t(beta_virginica) %*% x
    #Classify the observation to the species with highest function value
    d_vec=c(d_setosa, d_versicolor, d_virginica)
    prediction=append(prediction, label[which.max( d_vec )])
    d_setosa_vec=append(d_setosa_vec, d_setosa)
    d_versicolor_vec=append(d_versicolor_vec, d_versicolor)
    d_virginica_vec=append(d_virginica_vec, d_virginica)
```

```
  }
  #Combine the predicted results to the test dataset.
  test$prediction=prediction
  return (prediction)
}
```

Q1:

```
pred = LDA(c(1:40),c(51:90),c(101:140),c(0.8,0.1,0.1))
test_error = 1 - sum(iris[c(41:50, 91:100, 141:150),][5] == pred)/30
pred
```

```
##  [1] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [6] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [11] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [16] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [21] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [26] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
```

```
test_error
```

```
## [1] 0
```

```
pred = LDA(c(1:40),c(51:90),c(101:140), c(0.1,0.8,0.1))
test_error = 1 - sum(iris[c(41:50, 91:100, 141:150),][5] == pred)/30
pred
```

```
##  [1] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [6] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [11] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [16] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [21] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [26] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
```

```
test_error
```

```
## [1] 0
```

```
pred = LDA(c(1:40),c(51:90),c(101:140), c(0.1,0.1,0.8))
test_error = 1 - sum(iris[c(41:50, 91:100, 141:150),][5] == pred)/30
pred
```

```
##  [1] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [6] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [11] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [16] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [21] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [26] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
```

```
test_error
```

```
## [1] 0
```

Conclusion:
In the above chart, there are results of using priors (0.8,0.1,0.1);
using priors (0.1,0.8,0.1);using priors (0.1,0.1,0.8).
For all three sets of priors, we all get 100% correct.
So in this special case, we find that LDA methos is not sensitive to the choices of priors. However, according to the construction of the discriminant function, the choices of priors should affect our result, that is LDA method is sensitive to the choice of prior.

Q2:

```
pred_1 = LDA(c(1:30),c(51:80),c(101:130), c(1/3,1/3,1/3))
test_error1 = 1 - sum(iris[c(31:50, 81:100, 131:150),][5] == pred_1)/60
pred_1
```

```
##  [1] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [6] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [11] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [16] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [21] "versicolor" "versicolor" "versicolor" "virginica"  "versicolor"
## [26] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [31] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [36] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [41] "virginica"  "virginica"  "virginica"  "versicolor" "virginica"
## [46] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [51] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [56] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
```

```
test_error1
```

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

test error: 0.03333333

```
pred_2 = LDA(c(1:20),c(51:70),c(101:120), c(1/3,1/3,1/3))
test_error2 = 1 - sum(iris[c(21:50, 71:100, 121:150),][5] == pred_2)/90
pred_2
```

```
##  [1] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [6] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [11] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [16] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [21] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [26] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
## [31] "virginica"  "versicolor" "versicolor" "versicolor" "versicolor"
## [36] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [41] "versicolor" "versicolor" "versicolor" "virginica"  "versicolor"
## [46] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [51] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [56] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
## [61] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [66] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [71] "virginica"  "virginica"  "virginica"  "versicolor" "virginica"
## [76] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [81] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [86] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
```

```
test_error2
```

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

test error: 0.03333333

```
pred_3 = LDA(c(1:10),c(51:60),c(101:110), c(1/3,1/3,1/3))
test_error3 = 1 - sum(iris[c(11:50, 61:100, 111:150),][5] == pred_3)/120
pred_3
```

```
##   [1] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##   [6] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [11] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [16] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [21] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [26] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [31] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [36] "setosa"     "setosa"     "setosa"     "setosa"     "setosa"
##  [41] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
##  [46] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
##  [51] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
##  [56] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
##  [61] "versicolor" "versicolor" "versicolor" "virginica"  "versicolor"
##  [66] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
##  [71] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
##  [76] "versicolor" "versicolor" "versicolor" "versicolor" "versicolor"
##  [81] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
##  [86] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
##  [91] "virginica"  "virginica"  "virginica"  "versicolor" "virginica"
##  [96] "virginica"  "versicolor" "versicolor" "virginica"  "virginica"
## [101] "virginica"  "virginica"  "virginica"  "versicolor" "virginica"
## [106] "virginica"  "virginica"  "virginica"  "versicolor" "virginica"
## [111] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
## [116] "virginica"  "virginica"  "virginica"  "virginica"  "virginica"
```

```
test_error3
```

```
## [1] 0.05
```

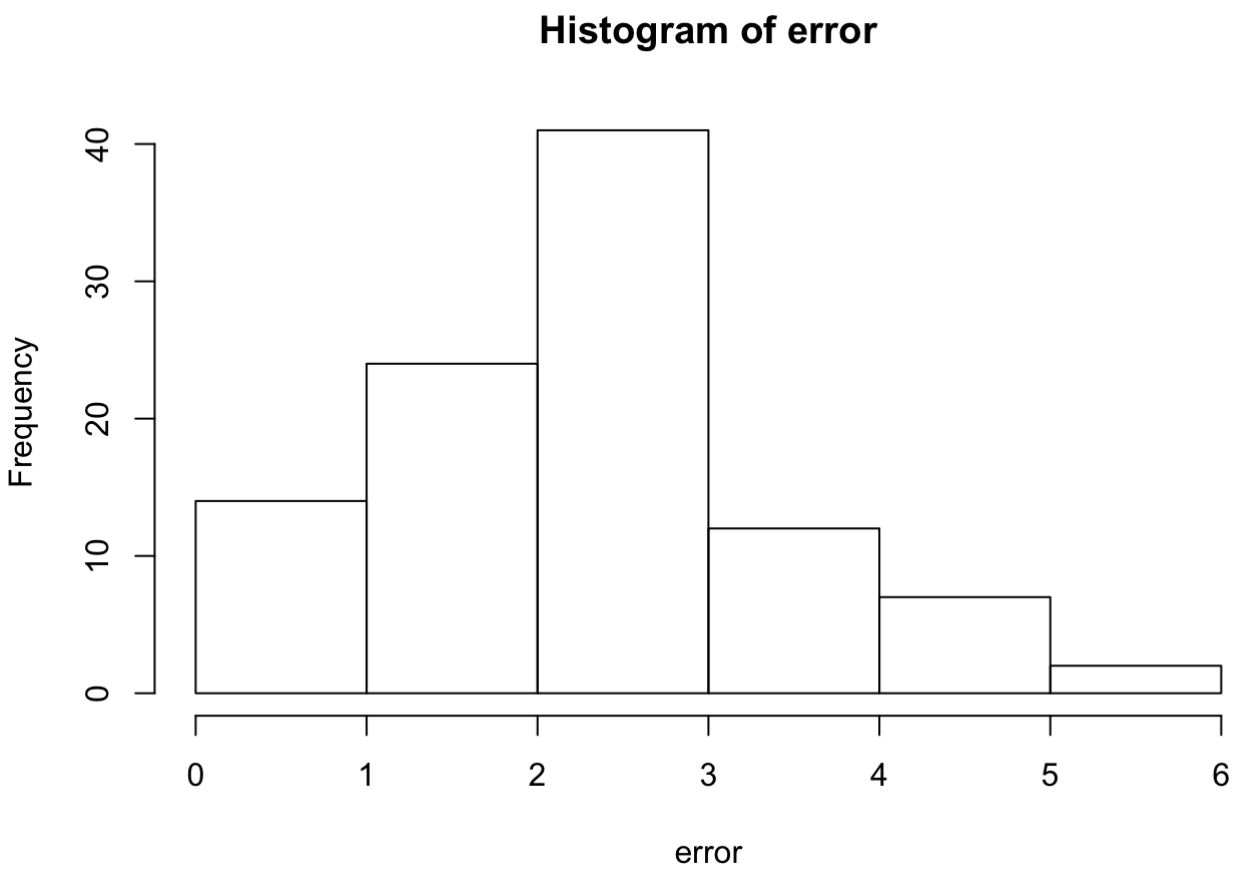
test error: 0.05

Q3:

```
error = c()
for (i in 1:100){
  s = sample(150,50);
  train1 = s[which( s <= 50)];
  train2 = s[which( 50< s & s <= 100 )];
  train3 = s[which( 100< s & s <= 150 )];
  pred = LDA(train1,train2,train3,c(length(train1)/50,length(train2)/50,length(train3)/5
0));
  error = c(error, sum(iris[-c(train1,train2,train3),][5] != pred))
}
error
```

```
##   [1] 3 4 2 5 3 3 3 3 1 3 1 4 5 1 2 2 1 4 2 3 3 5 3 4 0 3 3 3 3 1 4 2 2 3 2
##  [36] 2 3 3 2 3 2 1 3 2 3 3 3 0 3 4 3 3 5 3 4 4 5 4 1 0 3 2 4 3 2 3 2 6 3 3
##  [71] 2 5 2 2 3 3 3 1 2 2 3 4 3 1 1 2 3 6 3 3 2 5 3 1 3 4 2 2 3 2
```

```
hist(error)
```


Histogram of error
Frequency
0
10
20
30
40
0
1
2
3
4
5
6
error