Stat223 Homework

Zihao Huang

April 10, 2018

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#STAT 223
#Zihao Huang
#Date 4/1/2018

#1.
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.4.3
```

```
usstates<- read_excel("C:/Users/simon/Desktop/STAT223/USStates(1).xlsx")
states<-usstates
obama<-subset(usstates, ObamaMcCain=="O", select=c(Smokers,PhysicalActivity,Obese,College,NonWhite))
mccain<-subset(usstates, ObamaMcCain=="M", select=c(Smokers,PhysicalActivity,Obese,College,NonWhite))
s1<-var(obama)
s2<-var(mccain)
n1<-nrow(obama)
n2<-nrow(mccain)
meany1<- apply(obama,2,mean)
meany2<- apply(mccain,2,mean)
meandiff<-(meany1-meany2)
p<-5
k<-2
sp<-((n1-1)*s1+(n2-1)*s2)/(n1+n2-2)
##a)
(T2<-(n1*n2/(n1+n2))*t(meandiff)%*%solve(sp)%*%(meandiff))
```

```
##          [,1]
## [1,] 43.21259
```

```
###T2 is 43.21259
p*(n1+n2-2)/(n1+n2-p-1)*qf(.95,p, n1+n2-p-1)
```

```
## [1] 13.2384
```

```
###critical value is 13.2384
###Since T2>13.2384, we reject the H0 and believe the people who voted Obama
###are different from others who vote McCain.
##b)
(a<-solve(sp)%*%(meany1-meany2))
```

```
##                          [,1]
## Smokers           0.50241515
```

```
## PhysicalActivity  0.20437562
## Obese            -0.41147569
## College           0.35304935
## NonWhite          0.04870765
```

```
(a.star<-diag(sp)*a)
```

```
##                       [,1]
## Smokers           4.721579
## PhysicalActivity  2.608150
## Obese            -2.674104
## College           6.522556
## NonWhite         10.191821
```

```
###The order is Nonwhite,College,smokers,obese,PhysicalActivity.
##c)
t.save<-rep(0,5)
for (i in 1:5){
(t.save[i]<-(meany1[i]-meany2[i])/sqrt((n1+n2)/(n1*n2)*sp[i,i]))
}
rbind(c("Smokers","PhysicalActivity","Obese","College","NonWhite"),round(t.save,4))
```

```
##      [,1]      [,2]               [,3]      [,4]      [,5]
## [1,] "Smokers" "PhysicalActivity" "Obese"   "College" "NonWhite"
## [2,] "-2.0003" "2.6975"           "-4.6777" "5.2215"  "0.8408"
```

```
round((p.value =2*pt(-abs(t.save), n1+n2-2)),4)
```

```
## [1] 0.0511 0.0096 0.0000 0.0000 0.4047
```

```
###The rank by T-statistics is College,Obese,PhysicalActivity,Smokers,Nonwhite
###Only Smoker and Nonwhite fail to reject H0.
##d)
full.lamb<-summary(manova(cbind(Smokers,PhysicalActivity,Obese,College,NonWhite)
                          ~as.factor(ObamaMcCain),data=usstates), test="Wilks")$stats[1,2]
partial.lambda <- rep(-1,p)

partial.lambda[1] <- summary(manova(cbind(PhysicalActivity,Obese,College,NonWhite)
                                    ~as.factor(ObamaMcCain),data=usstates), test="Wilks")$stats[1,2]
partial.lambda[2] <- summary(manova(cbind(Smokers,Obese,College,NonWhite)
                                    ~as.factor(ObamaMcCain),data=usstates), test="Wilks")$stats[1,2]
partial.lambda[3] <- summary(manova(cbind(Smokers,PhysicalActivity,College,NonWhite)
                                    ~as.factor(ObamaMcCain),data=usstates), test="Wilks")$stats[1,2]
partial.lambda[4] <- summary(manova(cbind(Smokers,PhysicalActivity,Obese,NonWhite)
                                    ~as.factor(ObamaMcCain),data=usstates), test="Wilks")$stats[1,2]
partial.lambda[5] <- summary(manova(cbind(Smokers,PhysicalActivity,Obese,College)
                                    ~as.factor(ObamaMcCain),data=usstates), test="Wilks")$stats[1,2]
partial.lambda<-full.lamb/partial.lambda
partial.F <- (n1+n2-k-p+1)/(k-1)*(1-partial.lambda)/partial.lambda
(pl<-rbind(c("Smokers","PhysicalActivity","Obese","College","NonWhite"),round(partial.F,4)))
```

```
##      [,1]      [,2]               [,3]     [,4]      [,5]
## [1,] "Smokers" "PhysicalActivity" "Obese"  "College" "NonWhite"
## [2,] "6.0135"  "1.3262"           "3.4342" "9.2453"  "2.1072"
```

```
###The rank by partial lambda is College,Smokers,Obese,NonWhite,PhysicalActivity,
##e)
```

```
###Conclusion:They are totally different. But for part b) and part d), the last places are both
###Nonwhite. For part c) and d), the 1st places are both College.
####part b: Nonwhite,College,smokers,obese,PhysicalActivity.
####part c: College,Obese,PhysicalActivity,Smokers,Nonwhite.
####part d: College,Smokers,Obese,NonWhite,PhysicalActivity.
##f)
z1 <- as.matrix(obama)%*%a
z2 <- as.matrix(mccain)%*%a
#boxplot(c(z1,z2)~usstates$ObamaMcCain)
t.test(z1,z2)
```

```
##
##  Welch Two Sample t-test
##
## data:  z1 and z2
## t = 6.72, df = 47.802, p-value = 2.014e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.457948 4.557083
## sample estimates:
## mean of x mean of y
##  29.88270  26.37518
```

```
###The t-statistics is 6.72, which is larger than any t value shown in part c).
###And reject the H0 and conclude voters for Obama are different from McCain's.

#2
##a)
library(MASS)
m1 <- lda(ObamaMcCain~Smokers+PhysicalActivity+Obese+College+NonWhite, data=usstates, prior=rep(1,k)/k)
pred2 <- predict(m1)$class #Predicting each state
pre<-data.frame(states$ObamaMcCain,pred2)#Comparing our predictions and R predictions
table(pre)
```

```
##                   pred2
## states.ObamaMcCain  M  O
##                  M 19  3
##                  O  7 21
```

```
1-sum(diag(table(pre)))/sum(table(pre))
```

```
## [1] 0.2
```

```
###Apparent error rate for LDA is 0.2.
##b)
library(class)
```

```
## Warning: package 'class' was built under R version 3.4.4
```

```
###K = sqrt(N/k) = sqrt(50/2) = 5
m3 <- knn(train=states[,12:16], test=states[,12:16], cl = states$ObamaMcCain, k=5)
(tab.knn <- table(ObamaMcCain = usstates$ObamaMcCain, Predicted = m3))
```

```
##            Predicted
## ObamaMcCain  M  O
##           M 18  4
##           O  6 22
```

```r
1-sum(diag(tab.knn))/sum(tab.knn)
```

```
## [1] 0.2
```

```r
###Apparent error rate for Knn is 0.2.
##c)
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.4.4
```

```r
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.4.4
```

```r
mct <- rpart(ObamaMcCain~Smokers+PhysicalActivity+Obese+College+NonWhite, data=usstates, method="class")
rpart.plot(mct, main= "votes", type=0,extra=101)
####Note: the plot for classification tree will be a bit behind from here,
####it supposes to be an issue for Knit the pdf.
pct <- predict(mct, states[,12:16], type="class")
(tab.ct <- table(Region = usstates$ObamaMcCain, Predicted = pct))
```

```
##       Predicted
## Region  M  O
##      M 21  1
##      O  7 21
```

```r
1-sum(diag(tab.ct))/sum(tab.ct)
```

```
## [1] 0.16
```

```r
###Apparent error rate for Classification tree is 0.16.
##d)
m.cv <- lda(ObamaMcCain~Smokers+PhysicalActivity+Obese+College+NonWhite, data=usstates,
            prior=rep(1,2)/2, CV=T)
(tab.ld.cv <- table(Region = usstates$ObamaMcCain, Predicted = m.cv$class))
```

```
##       Predicted
## Region  M  O
##      M 17  5
##      O  9 19
```

```r
(error.cv <- mean(usstates$ObamaMcCain != m.cv$class) )
```

```
## [1] 0.28
```

```r
###Cross validated error rate with LDA is 0.28
knncv <- knn.cv(train=usstates[,12:16], cl = usstates$ObamaMcCain, k=5)
(tab.knncv <- table(Vote=usstates$ObamaMcCain, Predicted = knncv))
```

```
##     Predicted
## Vote  M  O
##    M 13  9
##    O 10 18
```

```r
(error.cv <- mean(usstates$ObamaMcCain != knncv) )
```

```
## [1] 0.38
```

```r
###Cross validated error rate with KNN is 0.38.
pred.ct.cv <- rep(0,50)
for (i in 1:50){
```

```
  m.ct.cv <- rpart(ObamaMcCain~Smokers+PhysicalActivity+Obese+College+NonWhite, data=usstates[-i,], meth
  pred.ct.cv[i] <- predict(m.ct.cv, usstates[i,12:16], type="class")
}
obmc<-rep(2,50)
obmc[which(usstates$ObamaMcCain=="M")]<-1
(tab.ctcv<-table(Actual=obmc,pred=pred.ct.cv))
```

```
##       pred
## Actual  1  2
##      1 21  1
##      2  9 19
```

```
(error.cv <- mean(obmc!= pred.ct.cv) )
```

```
## [1] 0.2
```

```
###Cross validated error rate with Classification tree is 0.2.
##e)
m.e <- lda(ObamaMcCain~Smokers+PhysicalActivity+Obese+College+NonWhite, data=usstates,
           prior=rep(1,2)/2,CV=T)
sum(usstates$ElectoralVotes)
```

```
## [1] 535
```

```
sum(usstates$ElectoralVotes[which(m.e$class=="O")])
```

```
## [1] 247
```

```
###McCain wins. McCain had 288 votes,  Obama had 247 votes.

#3.
iri <- read.table("C:\\Users\\simon\\Desktop\\STAT223\\iris.txt",header=T)
N<-nrow(iri)
k<-3
p<-4
##a)
### There are 2 possible discriminant functions.
##b)
m3.1 <- manova(cbind(Sepal.Length,Sepal.Width,Petal.Length,Petal.Width)~as.factor(Species),data=iri)
H <- summary(m3.1)$SS[[1]]
E <- summary(m3.1)$SS[[2]]
(e.vals <- Re(round(eigen(solve(E)%*%H)$values,digits=5)))
```

```
## [1] 32.19193  0.28539  0.00000  0.00000
```

```
(e.vecs <- Re(round(eigen(solve(E)%*%H)$vectors,digits=5)))
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,]  0.20874 -0.00653  0.65786 -0.77854
## [2,]  0.38620 -0.58661  0.00881  0.41628
## [3,] -0.55401  0.25256  0.07274  0.42978
## [4,] -0.70735 -0.76945 -0.74957 -0.18941
```

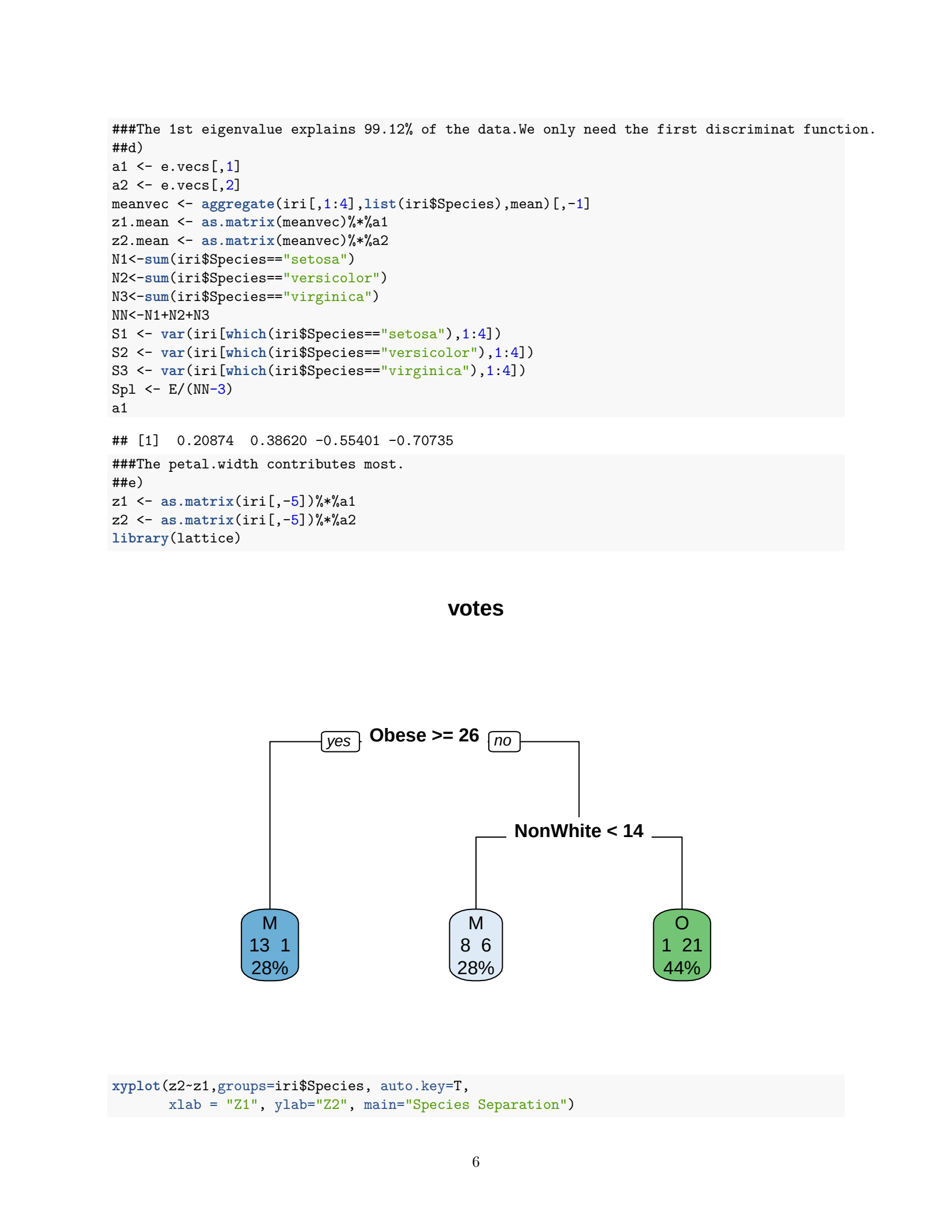
```
###The non-zero eigenvalues are 32.19 and 0.29
###And the non-zero eigenvectors are first two columns of eigenvectors
##c)
(e.vals[1]/sum(e.vals))
```

```
## [1] 0.9912126
```

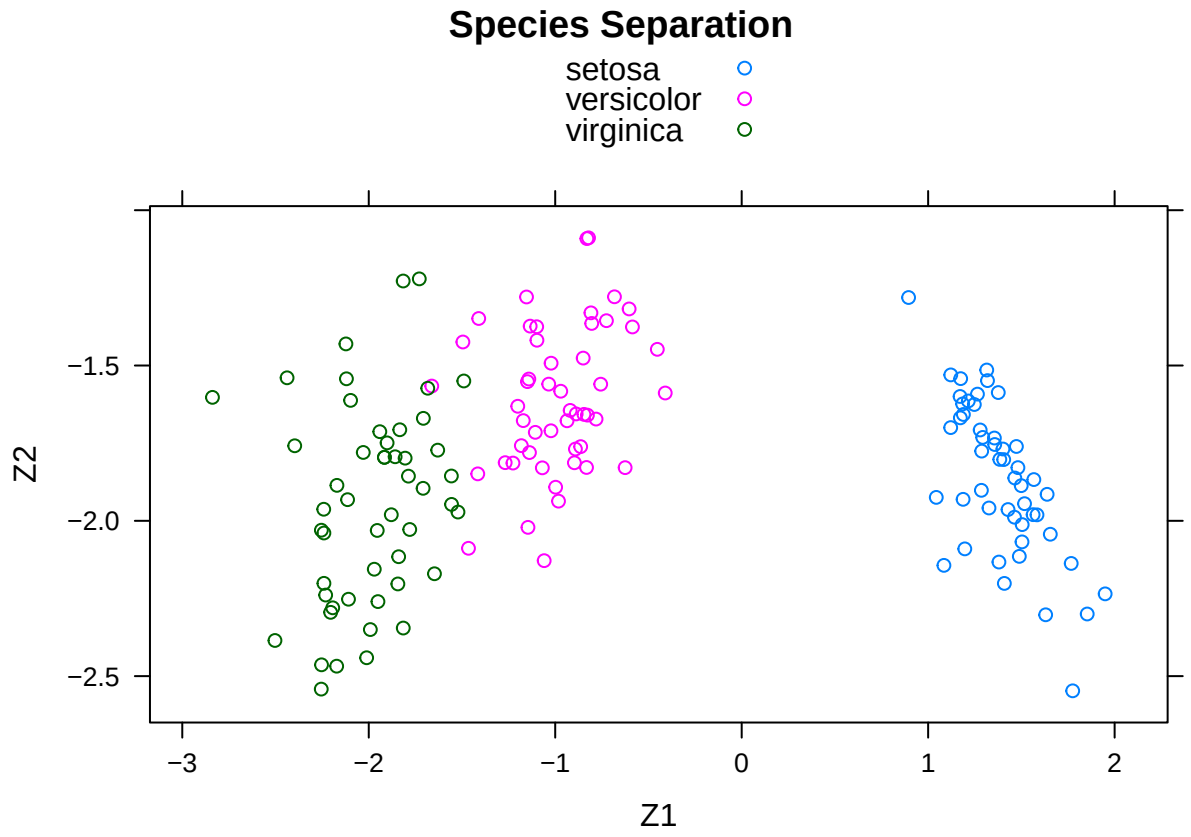
```
###The 1st eigenvalue explains 99.12% of the data.We only need the first discriminat function.
##d)
a1 <- e.vecs[,1]
a2 <- e.vecs[,2]
meanvec <- aggregate(iri[,1:4],list(iri$Species),mean)[,-1]
z1.mean <- as.matrix(meanvec)%*%a1
z2.mean <- as.matrix(meanvec)%*%a2
N1<-sum(iri$Species=="setosa")
N2<-sum(iri$Species=="versicolor")
N3<-sum(iri$Species=="virginica")
NN<-N1+N2+N3
S1 <- var(iri[which(iri$Species=="setosa"),1:4])
S2 <- var(iri[which(iri$Species=="versicolor"),1:4])
S3 <- var(iri[which(iri$Species=="virginica"),1:4])
Spl <- E/(NN-3)
a1
```

```
## [1]  0.20874  0.38620 -0.55401 -0.70735
```

```
###The petal.width contributes most.
##e)
z1 <- as.matrix(iri[,-5])%*%a1
z2 <- as.matrix(iri[,-5])%*%a2
library(lattice)
```




```
xyplot(z2~z1,groups=iri$Species, auto.key=T,
       xlab = "Z1", ylab="Z2", main="Species Separation")
```




```
###The first discriminant function is better. The scatter plot with transformed data
###shows the 1st discriminant function seperates the data well while 2nd is not as
###as better as previous function.
###I agree with the answer from part c)
t(a1)%*%H%*%a1/(t(a1)%*%E%*%a1)
```

```
##          [,1]
## [1,] 32.19193
```

```
t(a2)%*%H%*%a2/(t(a2)%*%E%*%a2)
```

```
##          [,1]
## [1,] 0.285391
```

```
##f)
m.full <- manova(cbind(Sepal.Length,Sepal.Width,Petal.Length,Petal.Width)~as.factor(Species),data=iri)
full.lambda <- summary(m.full, test="Wilks")$stats[1,2]
partial.lambda <- rep(-1,p)
partial.lambda[1] <- summary(manova(cbind(Sepal.Width,Petal.Length,Petal.Width)~as.factor(Species),data=
partial.lambda[2] <- summary(manova(cbind(Sepal.Length,Petal.Length,Petal.Width)~as.factor(Species),data
partial.lambda[3] <- summary(manova(cbind(Sepal.Length,Sepal.Width,Petal.Width)~as.factor(Species),data=
partial.lambda[4] <- summary(manova(cbind(Sepal.Length,Sepal.Width,Petal.Length)~as.factor(Species),data
###partial lambas are 0.02497554 0.03057958 0.03502453 0.03154590
lambda.ratio <- full.lambda/partial.lambda
partial.F <- (N-k-p+1)/(k-1)*(1-lambda.ratio)/lambda.ratio
rbind(colnames(iri)[1:4],partial.F)
```

```
##          [,1]              [,2]              [,3]
```

```
##           "Sepal.Length"     "Sepal.Width"      "Petal.Length"
## partial.F "4.72115209042432" "21.9359280889553" "35.5901748494336"
##           [,4]
##           "Petal.Width"
## partial.F "24.9043331921546"
```

```
### The order from Partial F             is Petal.Length, Petal.width, Sepal.width, Sepal.Length
### The order from standard coefficients is Petal.Length, Petal.Width, Sepal.Width, Sepal.Length
###The order is the same.
##g)
qf(.95,k-1,NN-k-p+1)
```

```
## [1] 3.058928
```

```
###The critical value is 3.0589
partial.F > qf(.95,k-1,N-k-p+1)
```

```
## [1] TRUE TRUE TRUE TRUE
```

```
###all the variables have partial F value>3.0589, they all reject the H0, and they are significant
###to 0.

#4.
##a) lda
library(MASS)
m1 <- lda(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data=iri, prior=rep(1,k)/k)
pred2 <- predict(m1)$class #Predicting each state
pre<-data.frame(iri$Species,pred2)#Comparing our predictions and R predictions
table(pre)
```

```
##             pred2
## iri.Species  setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         2
##   virginica       0          1        49
```

```
1-sum(diag(table(pre)))/sum(table(pre))
```

```
## [1] 0.02
```

```
###Apparent Error rate by LDA is 0.02
##b)
library(class)
###k = sqrt(N/k) = sqrt(150/3) = 7
m3 <- knn(train=iri[,1:4], test=iri[,1:4], cl = iri$Species, k=7)
(tab.knn <- table(Species = iri$Species, Predicted = m3))
```

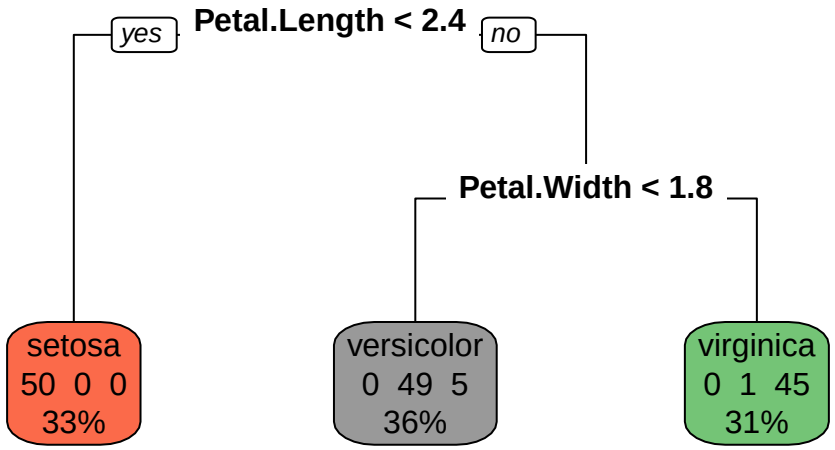
```
##             Predicted
## Species      setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         47         3
##   virginica       0          1        49
```

```
1-sum(diag(tab.knn))/sum(tab.knn)
```

```
## [1] 0.02666667
```

```
###Apparent Error rate by KNN is 0.027
##c)
```

```
library(rpart)
library(rpart.plot)
mct <- rpart(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data=iri, method="class")
rpart.plot(mct, main= "Classification for species", type=0,extra=101)
```




```
pct <- predict(mct, iri[,1:4], type="class")
(tab.ct <- table(Region = iri$Species, Predicted = pct))
```

```
##             Predicted
## Region       setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         49         1
##   virginica       0          5        45
```

```
1-sum(diag(tab.ct))/sum(tab.ct)
```

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

```
###Apparent Error rate by Classification tree is 0.04
##d)

m2 <- lda(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data=iri,
          prior=rep(1,k)/k, CV=T)
table(data.frame(iri$Species,m2$class))
```

```
##             m2.class
## iri.Species  setosa versicolor virginica
##   setosa         50          0         0
```

```
##   versicolor      0         48         2
##   virginica       0          1        49
```

```
(error.cv <- mean(iri$Species != m2$class) )
```

```
## [1] 0.02
```

```
###The cross validated error rate by LDA is 0.02
m4 <- knn.cv(iri[,1:4], cl = iri$Species, k=7)
(tab.knncv <- table(Species=iri$Species, Predicted = m4))
```

```
##             Predicted
## Species      setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         46         4
##   virginica       0          1        49
```

```
1-sum(diag(tab.knncv))/sum(tab.knncv)
```

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

```
###The cross validated error rate by KNN is 0.033
pred.ct.cv <- rep(0,N)
for (i in 1:N){
  m.ct.cv <- rpart(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data=iri[-i,], method="cla
  pred.ct.cv[i] <- predict(m.ct.cv,iri[i,1:4], type="class")
}
(tab.ct.cv <- table(Species=iri$Species, Predicted = pred.ct.cv))
```

```
##             Predicted
## Species       1  2  3
##   setosa     50  0  0
##   versicolor  0 45  5
##   virginica   0  5 45
```

```
1-sum(diag(tab.ct.cv))/sum(tab.ct.cv)
```

```
## [1] 0.06666667
```