homework10

Yi Chen

4/30/2020

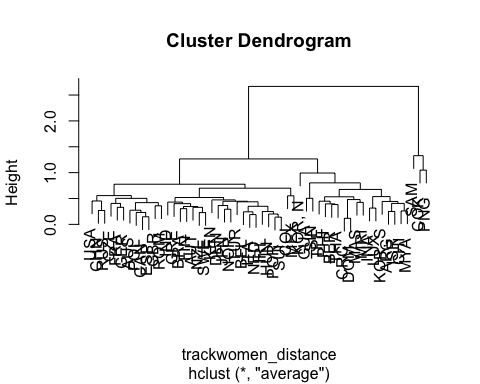
## Homework 10

## read the data  
trackwomen <- read.table("TrackWomen.dat", header = TRUE,sep = "\t")  
  
# convert the data into the same scale  
trackwomen$X100m <- 100/trackwomen$X100m  
trackwomen$X200m <- 200/trackwomen$X200m  
trackwomen$X400m <- 400/trackwomen$X400m  
trackwomen$X800m <- 800/(trackwomen$X800m\*60)  
trackwomen$X1500m <- 1500/(trackwomen$X1500m\*60)  
trackwomen$X3000m <- 3000/(trackwomen$X3000m\*60)  
trackwomen$Marathon <- 42195/(trackwomen$Marathon\*60)  
head(trackwomen)

## Country X100m X200m X400m X800m X1500m X3000m Marathon  
## 1 ARG 8.643042 8.718396 7.619048 6.504065 5.882353 5.440696 4.678353  
## 2 AUS 8.992806 8.996851 8.225375 6.734007 6.218905 5.793743 4.900355  
## 3 AUT 8.968610 8.810573 7.902015 6.872852 6.172840 5.694761 4.556203  
## 4 BEL 8.976661 8.896797 7.774538 6.768190 6.127451 5.668934 4.916113  
## 5 BER 8.726003 8.676790 7.504690 6.441224 5.827506 5.096840 4.037490  
## 6 BRA 8.952551 8.849558 7.902015 6.768190 5.995204 5.530973 4.770708

#### question 1

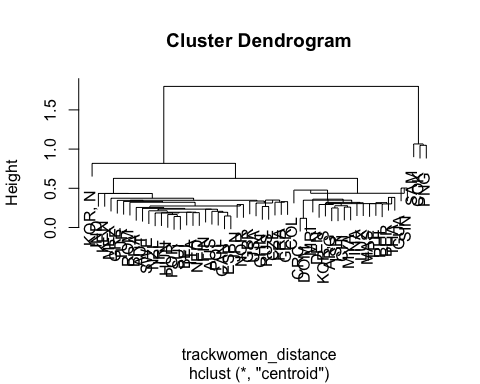
library(MASS)  
trackwomen\_distance <- dist(trackwomen[,2:8])  
av.clust = hclust(trackwomen\_distance, method = "average")  
plot(av.clust, labels = trackwomen$Country)



I will pick two clusters.

#### question 2

cen.clust = hclust(trackwomen\_distance, method = "centroid")  
plot(cen.clust, labels = trackwomen$Country)



cluster\_indicator <- cutree(cen.clust,2)  
  
sqrt(sum((apply((trackwomen[which(cluster\_indicator==1),2:8]), 2,mean)- apply((trackwomen[which(cluster\_indicator==2),2:8]), 2,mean))^2))

## [1] 2.555791

distance between the centroids of the two-cluster solution is 2.556.

#### question 3

cluster\_indicator <- cutree(av.clust,2)  
trackwomen$class <-cluster\_indicator  
lda\_data <- trackwomen[,2:9]  
ldamod <- lda(class ~ ., data= lda\_data)  
ldamod

## Call:  
## lda(class ~ ., data = lda\_data)  
##   
## Prior probabilities of groups:  
## 1 2   
## 0.94444444 0.05555556   
##   
## Group means:  
## X100m X200m X400m X800m X1500m X3000m Marathon  
## 1 8.854347 8.709740 7.757100 6.647154 6.043776 5.609671 4.692771  
## 2 8.141990 7.893767 6.946505 5.874250 5.070175 4.404213 3.387655  
##   
## Coefficients of linear discriminants:  
## LD1  
## X100m -0.2610821  
## X200m -2.1041894  
## X400m 1.6845920  
## X800m -1.1989193  
## X1500m 2.6845208  
## X3000m -1.8814474  
## Marathon -2.9213107

It has big positive value in x400m and x1500m and big negative value in x200m and marathon. I did not see clear interpreation of this result.

k2 <- kmeans(trackwomen[,2:8],centers = 2)  
cluster\_indicator <- k2$cluster  
  
trackwomen$class <-cluster\_indicator  
lda\_data <- trackwomen[,2:9]  
ldamod <- lda(class ~ ., data= lda\_data)  
ldamod

## Call:  
## lda(class ~ ., data = lda\_data)  
##   
## Prior probabilities of groups:  
## 1 2   
## 0.6296296 0.3703704   
##   
## Group means:  
## X100m X200m X400m X800m X1500m X3000m Marathon  
## 1 8.971297 8.851291 7.924231 6.758702 6.196201 5.781087 4.832331  
## 2 8.548680 8.346707 7.351390 6.341586 5.638613 5.137447 4.259750  
##   
## Coefficients of linear discriminants:  
## LD1  
## X100m 0.8200143  
## X200m 0.6227761  
## X400m -3.2158218  
## X800m 0.9317106  
## X1500m -4.2374394  
## X3000m 0.7398389  
## Marathon -0.5067418

The pattern of the result is similar but not the same. This time, x400m and x1500m are still the most contributing sports for determing the function. But the effects from marathon and x200m is small.

trackwomen <- read.table("TrackWomen.dat", header = TRUE,sep = "\t")  
k2 <- kmeans(trackwomen[,2:8],centers = 2)  
cluster\_indicator <- k2$cluster  
  
trackwomen$class <-cluster\_indicator  
lda\_data <- trackwomen[,2:9]  
ldamod <- lda(class ~ ., data= lda\_data)  
ldamod

## Call:  
## lda(class ~ ., data = lda\_data)  
##   
## Prior probabilities of groups:  
## 1 2   
## 0.8333333 0.1666667   
##   
## Group means:  
## X100m X200m X400m X800m X1500m X3000m Marathon  
## 1 11.26556 22.89667 51.34156 1.995333 4.109556 8.84800 147.9238  
## 2 11.81889 24.22778 55.22667 2.157778 4.588889 10.24444 182.0967  
##   
## Coefficients of linear discriminants:  
## LD1  
## X100m -0.06083260  
## X200m -0.84539345  
## X400m 0.21233566  
## X800m -1.40484495  
## X1500m 6.71672662  
## X3000m -1.63562215  
## Marathon 0.09215347

The patter is similar to the previous one, but not the same. This time the function depend more on x1500m and x300m than any other sports.