Homework 10

Yi Chen

4/9/2020

## Homework 10

### problem 1

#### question a: single-link clustering

library(readr)  
winter <- read\_csv("olympic dissim data.csv")  
head(winter)

## # A tibble: 6 x 13  
## name XctySki biathlon bobsled luge hockey speedSkt icedance figure snowbrd  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Xcty… 0 4.14 6.5 6.57 6.71 5.79 6.64 6.5 5.93  
## 2 biat… 4.14 0 6.14 6.14 6.86 5.86 7.36 6.64 6.57  
## 3 bobs… 6.5 6.14 0 3.57 6.5 5.71 6.71 6.93 6.29  
## 4 luge 6.57 6.14 3.57 0 6.86 6.14 6.86 6.21 5.86  
## 5 hock… 6.71 6.86 6.5 6.86 0 5.07 5.29 5.21 6.93  
## 6 spee… 5.79 5.86 5.71 6.14 5.07 0 5.07 5.29 6.71  
## # … with 3 more variables: skiJump <dbl>, slalom <dbl>, downhill <dbl>

x <-as.matrix(winter)  
sports <- x[,1]  
x <- x[,2:13]  
y <- matrix(as.numeric(x),12,12)  
rownames(y) <- sports  
colnames(y) <- sports  
spdist <- as.dist(y)  
spdist

## XctySki biathlon bobsled luge hockey speedSkt icedance figure snowbrd  
## biathlon 4.14   
## bobsled 6.50 6.14   
## luge 6.57 6.14 3.57   
## hockey 6.71 6.86 6.50 6.86   
## speedSkt 5.79 5.86 5.71 6.14 5.07   
## icedance 6.64 7.36 6.71 6.86 5.29 5.07   
## figure 6.50 6.64 6.93 6.21 5.21 5.29 3.93   
## snowbrd 5.93 6.57 6.29 5.86 6.93 6.71 5.93 6.79   
## skiJump 5.29 6.14 6.29 5.64 6.71 6.36 6.29 6.21 4.71  
## slalom 6.21 5.57 5.79 5.21 6.79 5.29 6.43 6.57 4.86  
## downhill 5.50 5.71 5.14 4.71 6.21 5.64 6.29 6.43 4.79  
## skiJump slalom  
## biathlon   
## bobsled   
## luge   
## hockey   
## speedSkt   
## icedance   
## figure   
## snowbrd   
## skiJump   
## slalom 4.64   
## downhill 5.00 4.33

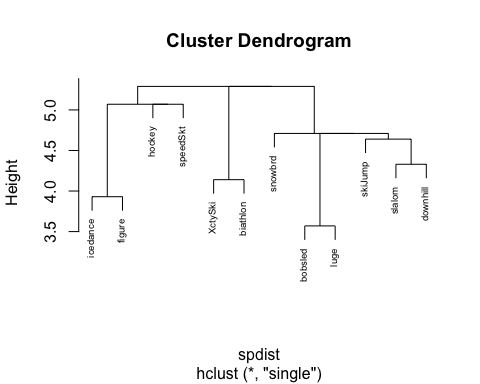
Single-link clustering is based on grouping clusters in bottom-up fashion (agglomerative clustering), at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other.

First step: based on spdist, the smallest distance 3.57 which is the distance between luge and bobsled. Thus, we cluster luge and bobsled. The corresponding first branch length will be . Also, we need to update the proximity matrix spdist into a new one called d2. The distance between the unclustered sports with the cluster will be updated by the *minimum distance* between the between each element of the first cluster and each of the remaining elements. For example, the distance between XctySki and the first cluster is:

The following step basically repeat the process until all the elements are clustered into a single cluster.

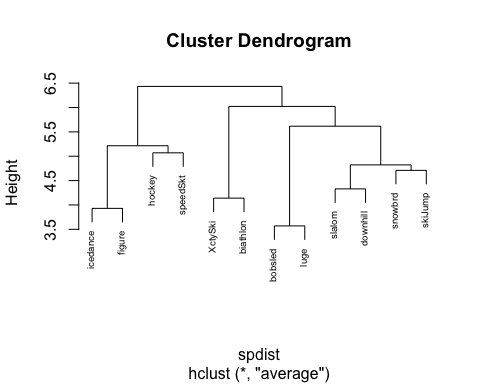
To do the visulization, I use the R function to repeat the step that I did in hand and show the resulting dendrogram.

single\_link\_HC <- hclust(d = spdist,method = 'single')  
plot(single\_link\_HC, cex = 0.6)



#### question b

single\_link\_HC <- hclust(d = spdist,method = 'average')  
plot(single\_link\_HC, cex = 0.6)



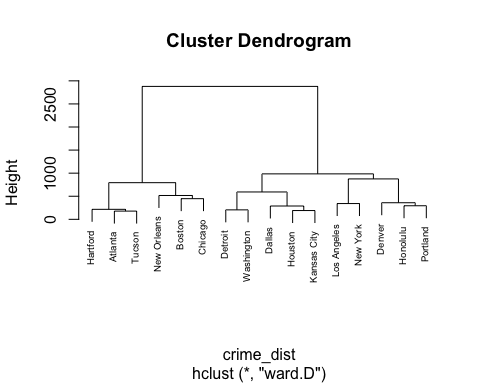
The average method give the result that are easier to interpreated. The average method combine the outdoor sports and indoor sports together. For the indoor sports, the sport emphasis more one the speed are clustered together and the sports emphasis more on the figure are cluster together. For the outdoor sports, the one with short distance are cluster together and long distance are clustered together.

### problem 2

Hartigan\_crime <- read.table("Hartigan\_crimes\_MLT.txt",sep = '\t',header = TRUE)  
rownames(Hartigan\_crime) <- Hartigan\_crime$city  
crime <- Hartigan\_crime[,2:ncol(Hartigan\_crime)]  
crime\_dist <- dist(crime)  
crime.HC.ward<- hclust(crime\_dist, method="ward")

## The "ward" method has been renamed to "ward.D"; note new "ward.D2"

plot(crime.HC.ward, cex = 0.6)



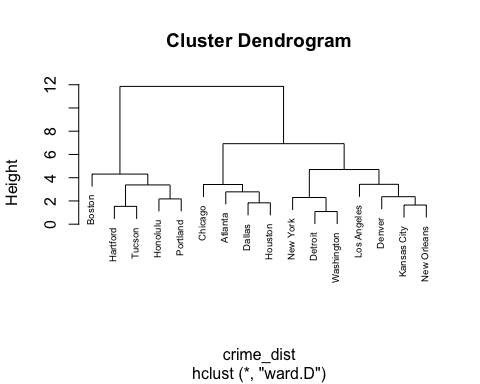
The states are clustered together by the space of crime. The states which have simialr number of crimes in general will be closer. For example, Honolulu and Portland are all high in burglary, larceny, and auto-theft, but low in murder, rape, and robbery. However, because the scale of different types of crime is very different, so the contribution each crime make for determine the distance between a pair of state are also different. Burglary, larceny, and auto-theft have a higher average number and bigger standard deviation, which will make more contribution in determine the hierarchical cluster than murder, rape, and robbery. Thus, we can see the overall differenec between the cluster is not very sigfinicant.

#### Optional question

crime <- scale(crime)  
crime\_dist <- dist(crime)  
crime.HC.ward<- hclust(crime\_dist, method="ward")

## The "ward" method has been renamed to "ward.D"; note new "ward.D2"

plot(crime.HC.ward, cex = 0.6)



As we mentioned above, the scale is very different for each crime. After normalized the data, we can see the distance between clusters are more clear since all the crime can make the contribution in determining the distance equally. The corresponding result shows a little differences. For example, Hartford this time is clustered with Tucson frist. But in general, the result is very similar.