Homework 9

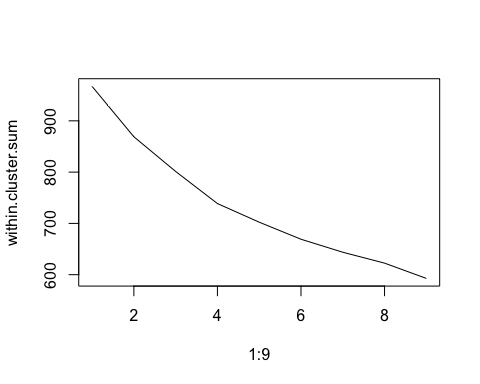
Yi Chen

3/27/2020

library(readr)  
CSNA\_MLT <- read\_csv("CSNA\_MLT.csv")  
x <- CSNA\_MLT[, 4:17]  
x <- as.matrix(x)

## question 1

within.cluster.sum <- c()  
for (k in 1:9){  
 kmean\_cluster <- kmeans(x = x,centers = k,iter.max = 100)  
 within.cluster.sum <- c(within.cluster.sum,sum(kmean\_cluster$withinss))  
}  
plot(1:9,within.cluster.sum,type = 'l')



I pick the default inital cluster seed. A random set of (distinct) rows in x is chosen as the initial centres.

## question 2

Based on the plot above, we can see that:

1. the sum of within cluster sum of square is always derease since it is a stepwise optimization and leads to a local minimum solution.
2. as the number of cluster approaching to the number of objectives of the input data, the sum of within cluster sum decrease to zero.
3. based on the marginal decrease of sun of within cluster sun of square, I think number of cluster should be 3 since after 3 the marginal decreas is not as obvious as 3.

# rerun the solution  
kmean\_cluster <- kmeans(x = x,centers = 3,iter.max = 100)  
(membership <- kmean\_cluster$cluster)

## [1] 2 1 3 2 3 3 1 1 2 1 2 3 2 3 2 1 1 3 2 1 1 1 1 3 2 2 2 1 3 3 1 2 3 2 3 1 3  
## [38] 2 3 1 1 1 3 2 1 1 3 3 2 1 1 2 3 3 1 3 1 3 1 3 2 2 2 2 3 2 2 3 2 3 3 3 2 3  
## [75] 1 2 1 1 1 1 1 3 1 1 3 1 1 3 3 3 1 1 3 3 3 3 1 1 2 2 2 1 2 3 2 2 2 3 3 1 3  
## [112] 2 1 1 3 1 1 3 1 3 1 2 1 2 2 3 3 1 3 2 1 3 3 3 2 3 3 2 3 3 3 2 1 2 3 3 1 1  
## [149] 1 2 3 3 1 1 2 2 2 2 3 1 3 3 1 2 3 2 1 2 3 3 3 2 2 3 3 1 1 2 3 2 3 2 2 3 2  
## [186] 2 1 2 3 1 1 2 2 1 1 1 1 2 2 3 2 2 3 1 3 3 1 2 1 1 2 3 2 1 1 2 2 3 1 2 1 3  
## [223] 2 1 1 1 1 1 3 1 3 1 1 2 2 3 1 1 1 1 2 3 1 3 2 2 2 1 1 1 3 1 3 1 1 2 2 3 1  
## [260] 1 1 1 2 3 1 3 2 2 2 2 3 2 3 2 3 1 1 1 1 1 3 2 3 2 1 3 1 3 2 3 1 3 3 2 1 3  
## [297] 1 3 1 3 3 2 2 1 3 3 2 3 3 3 1 1 3 2 3 1 2 1 1 1 1 3 3 1 3 2 1 3 1 2 1 2 3  
## [334] 3 1 3 2 3 3 1 2 3 2 1 3 3 1 3 2 2 3 2 1 1 3 3 3 3 3 2 2 3 3 2 3 1 3 2 3 1  
## [371] 3 3 2 2 1 2 1 3 2 2 1 2 2 2 2 1 3 1 1 2 1 3 2 2 2 2 2 1 3 1 2 3 3 3 2 3 3

## question 3

#### part A

# internal validity   
cluster\_mean <- matrix(NA, nrow = 14, ncol =3)  
cluster\_sd <- matrix(NA, nrow = 14, ncol =3)  
for (k in 1:3){  
 for (d in 1:14){  
 cluster\_mean[d,k] <- mean(x[which(membership==k),d])   
 cluster\_sd[d,k] <- sd(x[which(membership==k),d])   
 }  
}  
colnames(cluster\_mean) <- c("cluster 1","cluster 2","cluster 3")  
colnames(cluster\_sd) <- c("cluster 1","cluster 2","cluster 3")  
cluster\_mean

## cluster 1 cluster 2 cluster 3  
## [1,] 0.03623188 0.848 0.83333333  
## [2,] 0.65217391 0.800 0.84722222  
## [3,] 0.18115942 0.024 0.06944444  
## [4,] 0.27536232 0.248 0.48611111  
## [5,] 0.04347826 0.368 0.50000000  
## [6,] 0.10869565 0.496 0.32638889  
## [7,] 0.31884058 0.064 0.04166667  
## [8,] 0.10144928 0.272 0.31944444  
## [9,] 0.23188406 1.000 0.00000000  
## [10,] 0.09420290 0.096 0.20833333  
## [11,] 0.38405797 0.168 0.25694444  
## [12,] 0.10144928 0.056 0.04861111  
## [13,] 0.36231884 0.176 0.42361111  
## [14,] 0.10144928 0.056 0.11111111

cluster\_sd

## cluster 1 cluster 2 cluster 3  
## [1,] 0.1875474 0.3604656 0.3739788  
## [2,] 0.4780156 0.4016097 0.3610289  
## [3,] 0.3865533 0.1536649 0.2550957  
## [4,] 0.4483240 0.4335897 0.5015516  
## [5,] 0.2046740 0.4842020 0.5017452  
## [6,] 0.3123909 0.5019960 0.4705281  
## [7,] 0.4677248 0.2457379 0.2005238  
## [8,] 0.3030226 0.4467806 0.4678888  
## [9,] 0.4235728 0.0000000 0.0000000  
## [10,] 0.2931749 0.2957767 0.4075339  
## [11,] 0.4881436 0.3753708 0.4384736  
## [12,] 0.3030226 0.2308470 0.2158043  
## [13,] 0.4824213 0.3823526 0.4958550  
## [14,] 0.3030226 0.2308470 0.3153666

#### part b

cluster\_spec\_mean <- c()  
cluster\_spec\_sd <- c()  
for (k in 1:3){  
 cluster\_spec\_mean <- c(cluster\_spec\_mean,mean(CSNA\_MLT$spec[which(membership==k)]))  
 cluster\_spec\_sd <- c(cluster\_spec\_sd,sd(CSNA\_MLT$spec[which(membership==k)]))  
}  
cluster\_spec\_mean

## [1] 17.70290 17.64000 15.95139

cluster\_spec\_sd

## [1] 20.353593 8.047640 8.193746

## question 4

Based on the evidence of question 3, we can see the cluster well seperate in variable 13, which all of the cluster have significantly different center and small standard deviation. Similarly, variable 12, 14, and 3 also provide more information in clustering.

Variable spec als have certain level of explaination power in clustering. Cluster 1 and 2 have close cluster center and variability. But the cluster three have much higher value of center and more variability.