Assignment 5 MAT 378

* Aditya Parey

Q1)

1. Code and the graph:

getwd()

ppm.data <- read.table("HW5Data.txt")

ppm.data

summary(ppm.data)

data = c(107.2, 48.0, 75.8, 46.0, 131.6, 57.0, 25.0, 127.4, 133.1, 33.8, 58.2, 94.8, 36.2, 104.3, 85.9)

##install.packages("cluster")

##library(cluster)

diss\_matrix <- dist(data)

#single linkage clustering #

agn.tox <- agnes(diss\_matrix, diss = FALSE, stand = FALSE, method = "single")

dend.ang.tox <- as.dendrogram(agn.tox)

plot(dend.ang.tox, xlab = "Data Points", main = "Single-linkage clustering")

Graph attached in 2nd question.   
  
The two positions that most similar are 0 and 40. We can conclude this as the dendrogram indicates that the separation between these two points is the shortest of all the distances that exist between any two points.2Two sites are more similar to one another the closer they are on the dendrogram.

1. If we were to draw an imaginary line at 15 ppm height through the graph, we find out that there are 4 clusters that intersect the 15-ppm height mark. Each branch that intersects the line represents a cluster.   
   Hence, 4 clusters.

A graph of a number

Description automatically generated with medium confidence

1. Complete linkage clustering

> #complete linkage clustering

> agn.completetox <- agnes(diss\_matrix, diss = FALSE, stand = FALSE, method = "complete")

> dend.ang.completetox <- as.dendrogram(agn.completetox)

> plot(dend.ang.completetox, xlab = "Data Points", main = "Complete-linkage clustering")

A graph of a clustering diagram

Description automatically generated with medium confidence

1. Farthest neighbor approach. The distance between the two most remote locations in each cluster is used to establish the distance between two clusters in the complete linkage technique.

Determine the distance between each point in the cluster and every other point in the cluster.

Locate the spots that are farthest apart.

In the cluster, these two spots are the farthest apart.

Cluster 1: {50, 100, 150}

Cluster 2: {200, 250}

Distance: 200 points, farthest distance between two different points in two clusters.

Q2)

1. The k-means result where k =3  
   A screenshot of a computer code

   Description automatically generated
2. Variable values for 3 different clusters

A screenshot of a computer code

Description automatically generated

1. Individual 12 cluster assigned to: 2

A close-up of a computer screen

Description automatically generated

1. Cluster 1: 8   
   Cluster 2: 3

Cluster 3: 9

A close-up of a white background

Description automatically generated

1. The code and the graph:

A screenshot of a computer code

Description automatically generated

A screen shot of a graph

Description automatically generated

Sum of squares between clusters is 122.2026

A close-up of a computer screen

Description automatically generated

The pseudo-F stat is 24.97237

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

1. The pseudo-F stat is 21.22555

A screenshot of a computer

Description automatically generated

Full code

#aparey@oswego.edu

getwd()

ppm.data <- read.table("HW5Data.txt")

ppm.data

summary(ppm.data)

data = c(107.2, 48.0, 75.8, 46.0, 131.6, 57.0, 25.0, 127.4, 133.1, 33.8, 58.2, 94.8, 36.2, 104.3, 85.9)

install.packages("cluster")

library(cluster)

diss\_matrix <- dist(data)

#single linkage clustering #

agn.tox <- agnes(diss\_matrix, diss = FALSE, stand = FALSE, method = "single")

dend.ang.tox <- as.dendrogram(agn.tox)

plot(dend.ang.tox, xlab = "Data Points", main = "Single-linkage clustering")

diss\_matrix <- dist(data)

#complete linkage clustering

agn.completetox <- agnes(diss\_matrix, diss = FALSE, stand = FALSE, method = "complete")

dend.ang.completetox <- as.dendrogram(agn.completetox)

plot(dend.ang.completetox, xlab = "Data Points", main = "Complete-linkage clustering")

# Q2 #

# adding 'kmeans' in the dataset

kmeans\_result <- kmeans(ppm.data[, c("x", "y")], centers = 3)

kmeans\_result

# Extract cluster centers

cluster\_centers <- kmeans\_result$centers

cluster\_centers

# cluster assignments for 12

cluster\_assignment <- kmeans\_result$cluster[12]

cluster\_assignment

# number of individuals in each cluster

table(kmeans\_result$cluster)

# Plot of data points with cluster assignments

plot(ppm.data$x, ppm.data$y, col = kmeans\_result$cluster, pch = 19)

# cluster centers to the plot

points(cluster\_centers[, 1], cluster\_centers[, 2], col = 1:3, pch = 3)

# sum of squares (tot.withinss) from kmeans result

betweenss <- kmeans\_result$between

betweenss

#pseudo F stat

k.mean.data = kmeans(ppm.data[, c("x", "y")], centers = 3)

k.mean.data

k.mean.data$totss

k.mean.data$withinss

k.mean.data$tot.withinss

k.mean.data$between

MSB\_k3 <- betweenss / (3 - 1)

MSE\_k3 <- kmeans\_result$tot.withinss / (20 - 3)

pseudo\_F\_k3 <- MSB\_k3 / MSE\_k3

print(pseudo\_F\_k3)

# Implement k-means clustering with k=4

kmeans\_k4 <- kmeans(ppm.data[, c("x", "y")], centers = 4)

kmeans\_k4

# Calculate pseudo-F statistic for k=4

MSB\_k4 <- kmeans\_result\_k4$betweenss / (4 - 1)

MSE\_k4 <- kmeans\_result\_k4$tot.withinss / (20 - 4)

pseudo\_F\_k4 <- MSB\_k4 / MSE\_k4

print(pseudo\_F\_k4)