STAT 534: Homework 6

Fall 2021

Due: Friday, November 12

1. Suppose that we have four observations, for which we compute a dissimilarity matrix, given by:

$$\begin{bmatrix} & 0.2 & 0.7 & 0.5 \\ 0.2 & & 0.3 & 0.6 \\ 0.7 & 0.3 & & 0.4 \\ 0.5 & 0.6 & 0.4 & \end{bmatrix}.$$

   For instance, the dissimilarity (distance) between the first and second observations is 0.2, and the dissimilarity (distance) between the second and fourth observations is 0.6.

   (a) Which two points we need to cluster together first? (5 points)

   (b) Using complete (MAX) linkage, write the dissimilarity matrix after the first step in (a). You should get a 3×3 matrix. Based on this matrix, what should be the next step? (5 points)

   (c) Continue (b) to sketch the dendrogram using complete (MAX) linkage and get two clusters from it. (5 points)

   (d) Repeat (b) using single (MIN) linkage. What should be the next step then? (You do not need to complete the dendrogram here.) (5 points)

2. Carry out and interpret a clustering of vehicles from another manufacturer using the hierarchical clustering in the first course example. (Hint: you can find all the manufacture names in the `Mfr Name` column.) (10 points)

3. (a) Re-fit the k-means algorithm on the `BigCities` data of the second course example with a different value of k (i.e., not six). Experiment with at least two different values of k and report on the sensitivity of the algorithm to changes in this parameter. (10 points)
   (b) Project the `world_cities` coordinates using the Gall-Peters projection (see below) and run the k-means algorithm again. (you can try one of your k's in (a) or $k = 6$). Would you expect to obtain different results? Verify your guess by showing the clustering results. (10 points)

```
big_cities <- world_cities %>%
arrange(desc(population)) %>%
head(4000) %>%
transmute(x=pi*longitude/180/sqrt(2), y=sqrt(2)*sin(pi*latitude/180))
glimpse(big_cities)
```