MATH 4323, Data Science & Statistical Learning, Homework # 6.

DUE: November 22nd, at 11:59PM.

Instructions: Submit the solutions as a file (type it up and save as a *.pdf* or a *Word*-file) via UH Blackboard. Keep responses brief and to the point. For code & output: include only pieces that are of utmost relevance to the question.

Conceptual.

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

$$\begin{pmatrix} & 0.4 & 0.75 & 0.3 \\ 0.4 & & 0.5 & 0.8 \\ 0.75 & 0.5 & & 0.45 \\ 0.3 & 0.8 & 0.45 & \end{pmatrix}$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

(a) On the basis of this dissimilarity matrix, sketch the dendrogram (by hand is fine) that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

(b) Repeat (a), this time using single linkage clustering.

(c) Suppose that we cut the dendogram obtained in (a) such that two clusters result. Which observations are in each cluster?

(d) Suppose that we cut the dendogram obtained in (b) such that two clusters result. Which observations are in each cluster?

(e) It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

2. Here we work with the example introduced in lecture slides, where $n = 9$ observations are described by $p = 2$ predictors with the following dissimilarity matrix:

```
> round(dist(x),2)
     1    2    3    4    5    6    7    8
2 0.66
3 1.08 1.70
4 0.97 1.52 1.15
5 2.02 1.48 2.80 2.98
6 0.36 0.96 1.00 0.61 2.38
7 2.24 1.74 2.93 3.20 0.32 2.59
8 1.46 1.04 2.16 2.43 0.65 1.82 0.79
9 1.96 2.01 1.91 2.77 1.84 2.25 1.75 1.37
```

   We've already shown that clusters $\{5\}$ and $\{7\}$ will be merged first due to least dissimilarity ($dist(\{5\}, \{7\}) = 0.32$). Afterwards, we calculated the distances from new cluster $\{5, 7\}$ to all the other $n-2 = 7$ ”clusters” ($\{1\}, \{2\}, \{3\}, \{4\}, \{6\}, \{8\}, \{9\}$) by using **complete** linkage. Here, proceed to calculate those distances by using

   (a) **Single** linkage.

   (b) **Average** linkage.

   See the slide #9 from ”Unsupervised Learning. Hierarchical Clustering” lecture, for examples of calculations in case of **complete** linkage.

3. Suppose that for a particular data set, we perform hierarchical clustering using single linkage and using complete linkage. We obtain two dendrograms.

   (a) At a certain point on the single linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ fuse. On the complete linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

   (b) At a certain point on the single linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ fuse. On the complete linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

   Applied.

4. Consider the *USArrests* data. We will now perform hierarchical clustering on the states.

   (a) Using hierarchical clustering (Eucledian distance as the dissimilarity measure) with complete linkage, cluster the states. Provide the dendrogram.

   (b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters? Provide the cluster assignment output.

(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. Provide the dendrogram.

(d) Judging by the dendrogram, what appears to be a good number K of natural clusters? Cut the dendrogram at a height corresponding to that number K.

(e) Which states belong to which clusters from (d)? Provide the cluster assignment output. Describe which aspects unite the states within each of the clusters. E.g. "Cluster 1 contains states with low urban populations and low counts of murder, rape & assault."

(f) In your opinion, is there a reason to scale those variables before the inter-observation dissimilarities are computed? Why?

5. On the book website, *www.StatLearning.com*, there is a gene expression data set ($Ch10Ex11.csv$) that consists of 40 tissue samples with measurements on 1000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

   (a) Load in the data using $read.csv()$. You will need to select $header = F$.

   (b) Apply hierarchical clustering to the samples (using Eucledian distance), and plot the dendrogram, for the following linkages:

      i. Complete.

      ii. Single.

      iii. Average

   Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?