DS 303 Homework 11
Due: Dec. 06, 2021 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Dendrogram

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

$$\begin{pmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4. & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{pmatrix}$$

For example, the dissimilarity between the 1st and 2nd observations is 0.3.

a. Using the dissimilarity matrix, sketch the dendrogram that would result from carrying out hierarchical clustering on these four observations using complete linkage. Be sure to indicate on the dendrogram the height at which each fusion occurs.

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

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

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

Continue on next page.

Problem 2: K-means

In this problem, we will implement K-means clustering "manually" in R. Let $K = 2$. The observations are as follows:

Obs	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

a. Plot the observations. Show the plot here.

b. Randomly assign a cluster label to each observation. You can use the `sample()` function in R to do this. Report the cluster labels for each observation.

c. Compute the centroid for each cluster. Report those values here.

d. Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

e. Repeat (c) and (d) until the answers obtained stop changing. Report the centroids and cluster labels for the first two iterations.

f. In your plot from (a), color the observations according to the cluster labels obtained. Show that plot here.

Problem 3: Hierarchical Clustering

We'll use the `USArrests` data set for this problem; it is part of the `ISLR2` library. We will perform hierarchical clustering on the states.

a. Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

b. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

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

d. What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the features be scaled before Euclidean distance is computed? Provide a justification for your answer.

Continue on next page.

Problem 4: K-means for NC160 Data

We will use the NCI60 cancer cell line microarray data from our in-class example. This is also part of the ISLR2 library. Carry out the following pre-processing:

```
nci.labs <- NCI60$labs
nci.data <- NCI60$data
```

a. Should we scale our features, which are gene expressions, in this setting? Justify your answer. If you decide to scale the features, do so.

b. Implement K-means clustering on the (possibly scaled) data. Experiment with $K = 2$ and $K = 4$. Report the total within-cluster sum of squares for both $K = 2$ and $K = 4$.

c. Implement hierarchical clustering with complete linkage and Euclidean distance on the (possibly scaled) data. Cut the dendrogram to obtain 2 clusters. How does this compare to the K-means results we obtained in part (b) for $K = 2$? Report a confusion matrix to compare the results.

d. From your results in part (c), cut the dendrogram to obtain 4 clusters. How does this compare to the K-means results we obtained in part (b) for $K = 4$? Report a confusion matrix to compare the results.