ESI 4606 Analytics I - Foundations of Data Science
Homework 7 (Last Graded Homework)
Due: November 30st (11:00AM), 2022

Problem 1 (1.5 points)

Recall example for ID3 in lecture 14 (slide 12). Answer the following questions.

(a) If splitting variable for root node is "Temperature", what is the corresponding information gain, i.e., Gain(R,"Temp")?

(b) If splitting variable for root node is "Humidity", what is the corresponding information gain, i.e., Gain(R,"Hum")?

(c) If splitting variable for root node is "Wind", what is the corresponding information gain, i.e., Gain(R,"Wind")?

Note: **To get full points, include intermediate steps.**

Problem 2 (2 points)

In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. 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.

(b) At the initialization step, assign observations 1 and 2 as cluster 1 and observations 3,4,5 and 6 as cluster 2. Plot the observations and color the observations according to the their cluster labels.

(c) Compute the centroid for each cluster.

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

(e) Repeat (c) and (d) until the answers obtained stop changing.

(f) Based on the final clustering results, plot the observations and color the observations according to the their cluster labels.

Note: **To get full points, include intermediate steps. Only for plots, you can use R.**

Problem 3 (1.5 points)

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

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

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 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 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?

Note: **To get full points, include intermediate steps.**