**1. (10 points) [K-means Clustering] [John Wesley Hostetter (Designed) & Chengyuan**

**Liu (Graded)] Using K-means clustering and Euclidean distance, cluster the 11 data**

**points in Figure 1 into three clusters. We assume that the initial seeds are at points E,**

**F, and J (in yellow). Answer the following questions:**

**(a) (4 points) Run the K-means algorithm for one round. Calculate the coordinates**

**of the new centroids. What are the new clusters? Show your work in the first**

**subgraph in Figure 2.**

*Find the complete version of the code snippets attached to the document in the file HW5\_1+2.py in the GitHub repository engr-ALDA-fall2021-H12*

**ROUND 1:**

Centroids: E=(-8,0), F=(-2,-1), J=(0,-2)

**Euclidean distance between each centroid to a given point**

**Table, Excel

Description automatically generated**

**CLUSTERS AFTER ROUND 1**

C1 = (A,B,C,E,H,K)

C2: (F,G,I)

C3: (D,J)

**ROUND 2:**

Centroids: C1 = (-7.5,-0.666), C2 = (-3.333,-0.333), C3 = (0, -1.5)

**Euclidean distance between each centroid to a given point**

**Graphical user interface, table, Excel

Description automatically generated**

**CLUSTERS AFTER ROUND 2**

C1: (A,B,E,H,K)

C2: (C,F,G,I)

C3: (D,J)

**ROUND 3:**

Centroids: C1 = (-8,-1.4). C2 = (-3.75,0.5), C3 = (0, -1.5)

**Euclidean distance between each centroid to a given point**

**Table, Excel

Description automatically generated**

**CLUSTERS AFTER ROUND 3**

C1: (A,E,H,K)

C2: (B,C,G,I)

C3: (D,F,J)

**ROUND 4:**

Centroids: C1 = (-8.5,-2), C2 = (-4.75,1), C3 = (-0.666, -1.333)

**Euclidean distance between each centroid to a given point**

Table, Excel

Description automatically generated

**CLUSTERS AFTER ROUND 4:**

C1: (A,E,H,K)

C2: (B,C,G,I)

C3: (D,F,J)

**ROUND 5:**

Centroids:C1 = (-8.5,-2), C2 = (-4.75,1), C3 = (-0.666, -1.333)

**Euclidean distance between each centroid to a given point**

Table, Excel

Description automatically generated

**CLUSTERS AFTER ROUND 5:**

C1: (A,E,H,K)

C2: (B,C,G,I)

C3: (D,F,J)

**(b) (6 points) How many rounds are needed for the K-means clustering algorithm to converge? Draw the resulting clusters and new centroid at the end of each round (including the first round) in the Figure 2. Indicate the coordinates alongside corresponding centroids. Add new graphics if needed; Stop when the algorithm**

**converges and clearly label on the graph where the algorithm converges.**

**Diagram

Description automatically generated**

**ROUND 1:**

Centroids: E = (-8,0), F = (-2,-1), J = (0,-2)

**ROUND 2:**

Centroids: C1 = (-7.5,-0.666), C2 = (-3.333,-0.333), C3 = (0, -1.5)

**ROUND 3:**

Centroids: C1 = (-8,-1.4). C2 = (-3.75,0.5), C3 = (0, -1.5)

**ROUND 4:**

Centroids: C1 = (-8.5,-2), C2 = (-4.75,1), C3 = (-0.666, -1.333)

**ROUND 5:**

Centroids:C1 = (-8.5,-2), C2 = (-4.75,1), C3 = (-0.666, -1.333)

**2. (15 points) [Hierarchical Clustering] [John Wesley Hostetter(Designed) & Chengyuan** **Liu (Graded)] We will use the same dataset as in Question 1 for the following problem. The Euclidean Distance matrix between each pair of the data points is listed in the figure below:**

**(a) (8 points) Perform single and complete link hierarchical clustering. Show your results by drawing corresponding dendrogram. The dendrogram should clearly show the order and the height in which the clusters are merged. In case of a tie please resolve in alphabetical order of the points' labels. NO PARTIAL CREDIT.**

**LIST OF COORDINATES**

**Graphical user interface, application

Description automatically generated**

from sklearn.metrics.pairwise import euclidean\_distances

mat = euclidean\_distances(df)

euc\_matrix = pd.DataFrame(mat, columns = df.index)

euc\_matrix.to\_csv('euc\_mat.csv')

**SINGLE LINK HIERARCHICAL CLUSTERING**

* Single link hierarchical clustering is performed by considering the smallest distance in the distance matrix and clustering both the points, say A, B associated with that distance.
* We recalculate the distance from other points to the two points in consideration by taking the minimum distance to either one of the points. For E.g. If point C has distance 4 to A and distance 3 to B, we ignore the larger of the two distances and keep the smaller.
* We perform this operation until we are left with a single cluster

**Note: The Euclidean distance marked in red gives us the smallest distance and the points we merge.**

**ROUND 1:**

**Table

Description automatically generated**

**ROUND 2:**

**Graphical user interface, application, table, Excel

Description automatically generated**

**ROUND 3:**

**Table

Description automatically generated**

**ROUND 4:**

**A screenshot of a computer

Description automatically generated with medium confidence**

**ROUND 5:**

**A screenshot of a computer

Description automatically generated with low confidence**

**ROUND 6:**

**A screenshot of a computer

Description automatically generated with medium confidence**

**ROUND 7:**

**Table

Description automatically generated**

**ROUND 8:**

**Table

Description automatically generated**

**ROUND 9:**

**Table

Description automatically generated**

**ROUND 10:**

**Graphical user interface, text, application

Description automatically generated**

**DENDOGRAM FOR SINGLE LINK MATRIX**

**Diagram

Description automatically generated**

**COMPLETE LINK HIERARCHICAL CLUSTERING**

* Complete link hierarchical clustering is performed by considering the smallest distance in the distance matrix and clustering both the points, say A, B associated with that distance.
* We recalculate the distance from other points to the two points in consideration by taking the maximum distance to either one of the points. For E.g. If point C has distance 4 to A and distance 3 to B, we ignore the smaller of the two distances and keep the larger one.
* We perform this operation until we are left with a single cluster

**Note: The Euclidean distance marked in red gives us the smallest distance and the points we merge.**

**ROUND 1:**

**Table

Description automatically generated**

**ROUND 2:**

**Table

Description automatically generated**

**ROUND 3:**

**Table

Description automatically generated**

**ROUND 4:**

**Table, Excel

Description automatically generated**

**ROUND 5:**

**Table

Description automatically generated**

**ROUND 6:**

**Graphical user interface, application, table, Excel

Description automatically generated**

**ROUND 7:**

**Table

Description automatically generated**

**ROUND 8:**

**Table

Description automatically generated**

**ROUND 9:**

**Table

Description automatically generated**

**ROUND 10:**

**Graphical user interface, text, application

Description automatically generated**

**DENDOGRAM FOR COMPLETE LINK MATRIX**

**Chart

Description automatically generated**

**(b) (4 points) Using Sum of Squared Error (SSE) and assuming there are three clusters, which of the single link and complete link hierarchical clustering will yield better results? Justify your answer.**

**SINGLE LINK HIERARCHICAL CLUSTERING**

We have obtained three clusters,

Cluster 1:ABCEGHK,

Cluster 2: I

Cluster 3: DFJ

**Chart

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**COMPLETE LINK HIERARCHICAL CLUSTERING**

We have obtained three clusters,

Cluster 1:AEHIK

Cluster 2: BCG

Cluster 3: DFJ

**A picture containing diagram

Description automatically generated**

**Text

Description automatically generated**

From the script above, we notice that complete link hierarchical clustering has a lesser SSE rate of **38.53** compared to single link’s **70.76**.

**(c) (3 points) Compare the clusters from 2(b) with the clusters found using K-means in Question 1 by calculating their corresponding Sum of Squared Errors (SSE)s. According to their SSE results, which is better: K-means or hierarchical clustering?**

**Text

Description automatically generated**

Although hierarchical clustering is supposed to perform better on smaller datasets, we observe that K-Means clustering performs better in this case with an SSE of **35.0833 < 38.533 (Hierarchical Complete Clustering).** Thus, K-Means is the better suited for this dataset.