**Activity-17**

**NAME- Abhishek Srivastava**

**Regno.- 19BCE10071**

**Slot- E21+E22+E23**

**Faculty:- Dr. Nilamdhab Sir**

Q.

1. Using K-means clustering, cluster the following data into two clusters and show each step.

{3, 5, 10, 13, 4, 21, 31, 12, 26}.

Give your step-by-step computational analysis.

2. Formulate any four cluster scenarios to Calculate purity to measure the quality of each cluster.

3. Investigate the computational processes of the K-medoid algorithm with a suitable scenario.

1.

Solution: (Method-1)

Given: {3, 5, 10, 13, 4, 21, 31, 12, 26}

Step 1: Assign alternate value to each cluster randomly.

Step 2: k1= {3, 10, 4, 31, 26} Mean value= 14.8

k2= {5, 13, 21, 12} Mean value = 12.75

Step 3: Again, assign the values,

k1 = {21, 31, 26} Mean value = 26

k2= {3, 5, 10, 13, 4, 12} Mean value = 7.83

Step 4: Again, assign the values,

k1 = {21, 31, 26} Mean value = 26

k2= {3, 5, 10, 13, 4, 12} Mean value = 7.83

Step-2:

K1 cluster having cluster centroid C1 = 14.8

K2 cluster having Cluster centroid C2 = 12.75



Computation to move from Step-2 to Step-3.

In step-2, The clusters centroid values are as follows:

C1= 14.8 of K1 cluster

C2 = 12.75 of K2 cluster

Now Consider each data point from K1 and k2 clusters,

compute the distance from C1 and C2,

consider the minimum distance, and assign the respective data point to the cluster

k1 or k2.

Ex: for data point ‘2’: Min (│2-14.8 │, │2-12.75 │) = 10.75

so, data point ‘2’ assigns to cluster K2 having centroid C2[K2/C2].

(Method-2)

{3, 5, 10, 13, 4, 21, 31, 12, 26}

Step 1: Randomly assign the means: m1 = 4, m2 = 5

Step 2: Group the numbers close to mean m1 = 4 are grouped into cluster

k1 and m2 = 5 are grouped into cluster k2

Step 3: k1 = {3, 4}, k2 = {5, 10, 13, 21, 31, 12, 26}, m1= 3.5, m2 = 16.8

Step 4: k1 = {3, 4, 5}, k2 = {10, 13, 21, 31, 12, 26}, m1= 4 m2 = 18.8

Step 5: k1 = {3, 4, 5, 10}, k2= {13, 21, 31, 12, 26}, m1= 5.5, m2 = 20.6

Step 6: k1 = {3, 4, 5, 10, 12, 13}, k2 = {21, 31, 26}, m1 = 7.8, m2 = 26

Step 7: k1 = {3, 4, 5, 10, 12, 13}, k2 = {21, 31, 26}, m1 = 7.8, m2 = 26

Step 8: Stop. The clusters in step 6 and 7 are same.

Final answer: k1 = {3, 4, 5, 10, 12, 13} and k2 = {21, 31, 26}

2. Solution:

Assume that we cluster three category of data items (those colored with red, blue and green) into

three clusters.

Calculate purity to measure the quality of each cluster.

Cluster I: 5 red, 1 blue, 0 green

Cluster II: 1 green, 4 blue, 1 red



Cluster III: 2 red, 0 blue, 3 green

Cluster IV: 3 red, 2 blue, 1 green

Cluster I: Purity = 1/6 (max (5, 1, 0)) = 5/6 = 83%

Cluster II: Purity = 1/6 (max (1, 4, 1)) = 4/6 = 67%

Cluster III: Purity = 1/5 (max (2, 0, 3)) = 3/5 = 60%

Cluster IV: Purity = 1/6(max (3,2,1)) = 3/6= 50%

3.

K-Medoids (also called as Partitioning Around Medoid) algorithm:

A medoid can be deﬁned as the point in the cluster, whose dissimilarities with all

the other points in the cluster is minimum.

The dissimilarity of the medoid (Ci) and object (Pi) is calculated by using E = |Pi - Ci|

*The cost in K-Medoids algorithm is given as*

*.*

**Algorithm:**

*1. Initialize: select k random points out of the n data points as the medoids.*

*2. Associate each data point to the closest medoid by using any common distance*

*metric methods.*

*3. While the cost decreases:*

*For each medoid m, for each data o point which is not a medoid:*

*1. Swap m and o, associate each data point to the closest medoid,*

*recompute the cost.*

*2. If the total cost is more than that in the previous step, undo the swap.*



Let’s consider the following example:

If a graph is drawn using the above data points, we obtain the following:

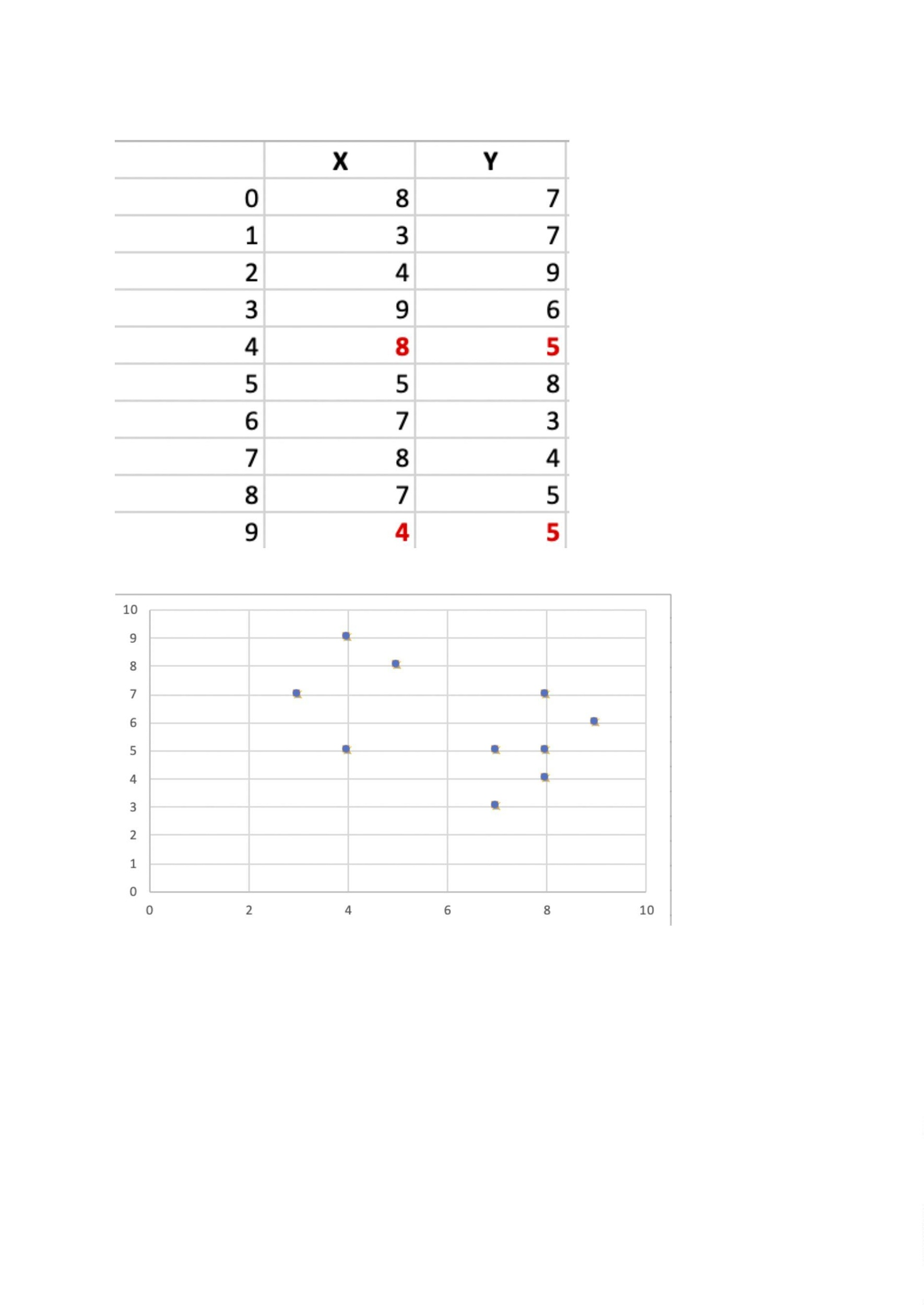
**Step 1:**

**Let the randomly selected 2 medoids, so select k = 2 and let C1 -(4, 5)** and **C2**

**-(8, 5)** are the two medoids.

**Step 2: Calculating cost.**

**The dissimilarity of each non-medoid point with the medoids is calculated and**



**tabulated:**

Each point is assigned to the cluster of that medoid whose dissimilarity is less.

The points 1, 2, 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2.

The Cost = (3 + 4 + 4) + (3 + 1 + 1 + 2 + 2) = 20

**Step 3: randomly select one non-medoid point and recalculate the cost.**

**Let the randomly selected point be (8, 4). The dissimilarity of each non-medoid**

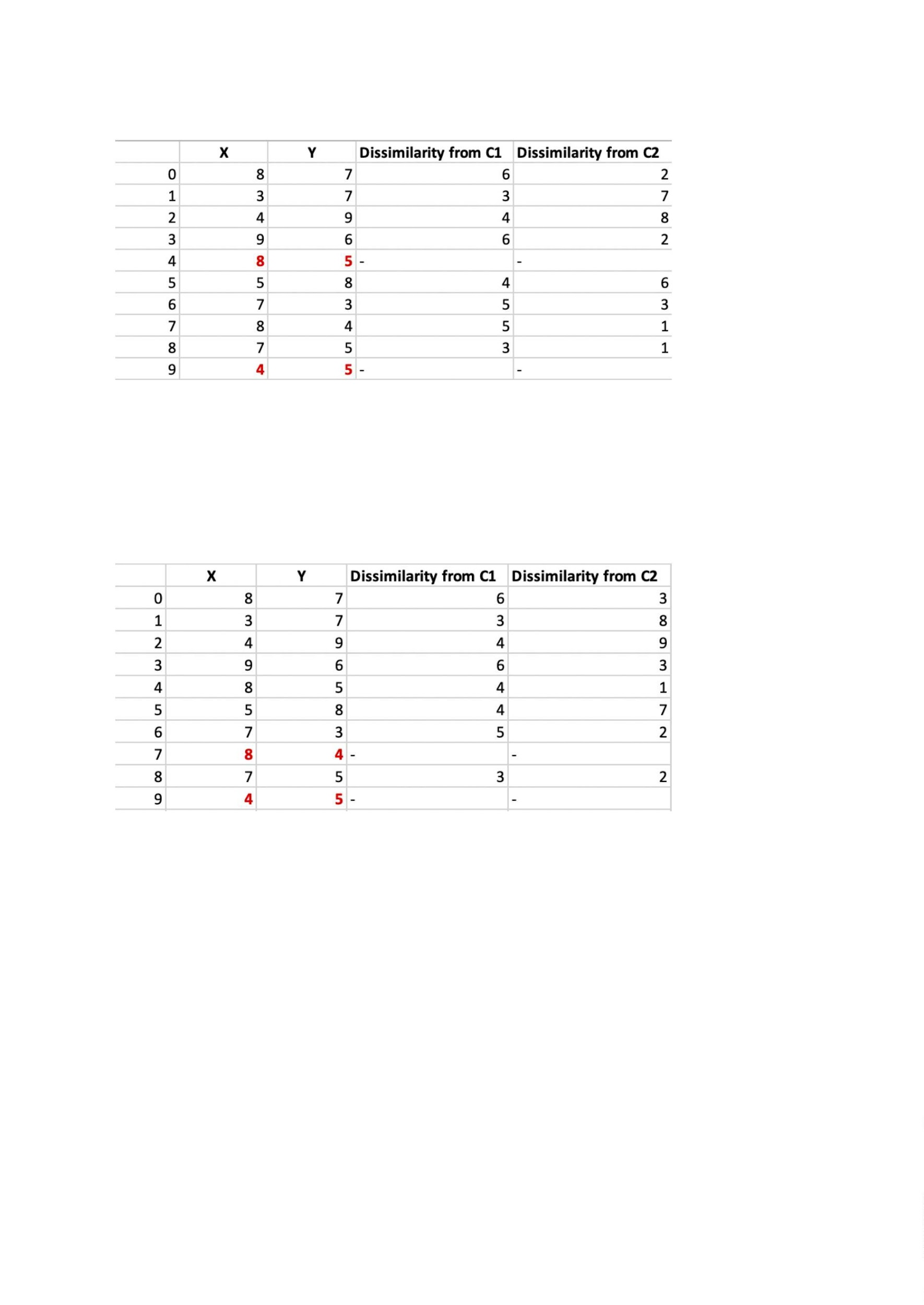
**point with the medoids –** C1 (4, 5) and C2 (8, 4) is calculated and tabulated.

Each point is assigned to that cluster whose dissimilarity is less. So, the points 1,

2, 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2.

The New cost = (3 + 4 + 4) + (2 + 2 + 1 + 3 + 3) = 22

Swap Cost = New Cost – Previous Cost = 22 – 20 and **2 >0**



As the swap cost is not less than zero, we undo the swap. Hence (3, 4) and (7, 4)

are the ﬁnal medoids. The clustering would be in the following way

The **time complexity** is

.

**Advantages:**

1. It is simple to understand and easy to implement.

2. K-Medoid Algorithm is fast and converges in a ﬁxed number of steps.

3. PAM is less sensitive to outliers than other partitioning algorithms.

**Disadvantages:**

4. The main disadvantage of K-Medoid algorithms is that it is not suitable for

clustering non-spherical (arbitrary shaped) groups of objects. This is

because it relies on minimizing the distances between the non-medoid

objects and the medoid (the cluster centre) – brieﬂy, it uses compactness

as clustering criteria instead of connectivity.

5. It may obtain different results for different runs on the same dataset

because the ﬁrst k medoids are chosen randomly.

