

K-Medoids clustering Algorithm

K-Medoid

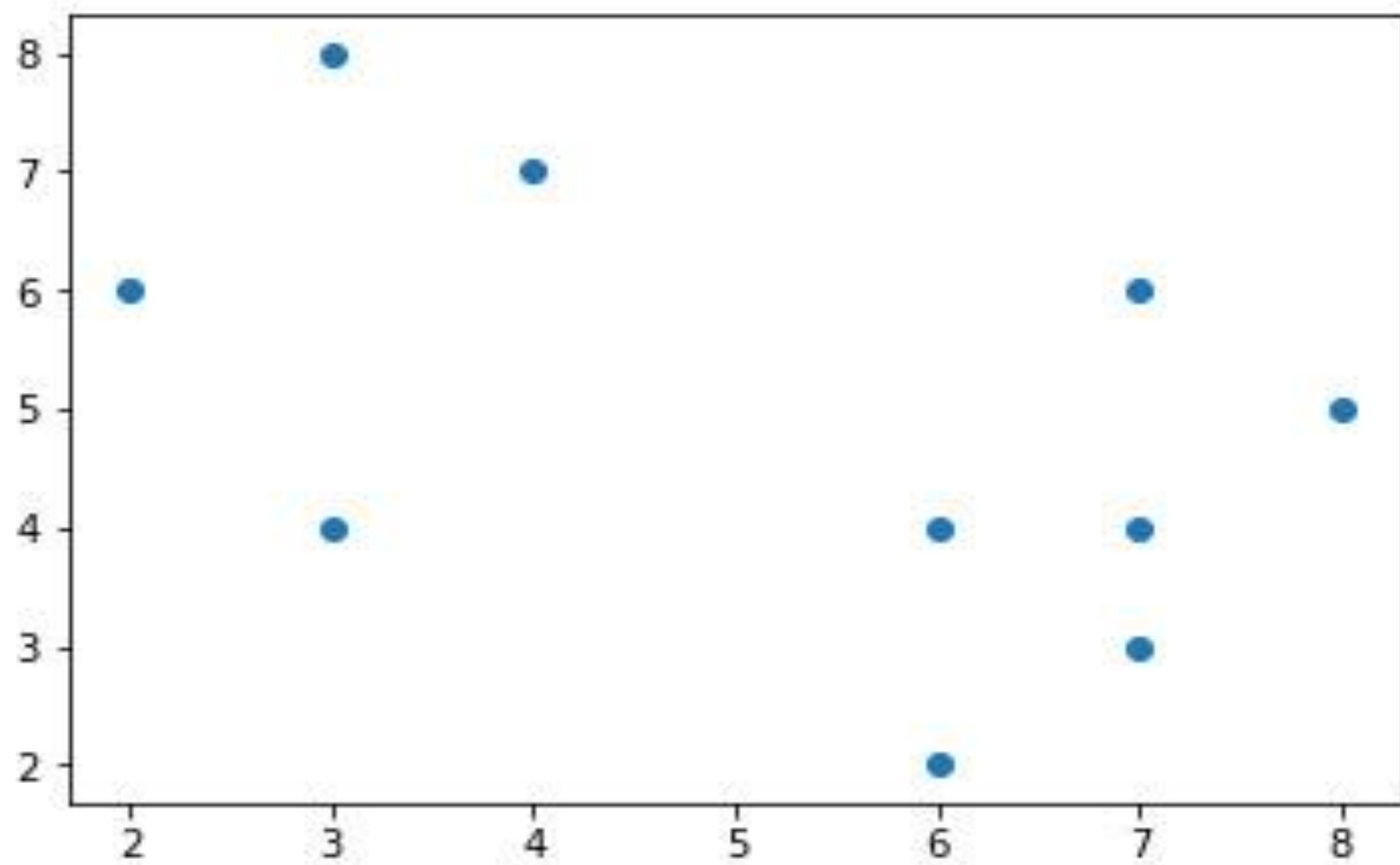
- K-Medoids (also called as Partitioning Around Medoid) algorithm
- A medoid can be defined as the point in the cluster, whose dissimilarities with all the other points in the cluster is minimum.
- The dissimilarity of the medoid(C_i) and object(P_i) is calculated by using $E = |P_i - C_i|$
- *The cost in K-Medoids algorithm is given as -*

$$c = \sum_{C_i} \sum_{P_i \in C_i} |P_i - C_i|$$

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 point o 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.*

	X	Y
0	7	6
1	2	6
2	3	8
3	8	5
4	7	4
5	4	7
6	6	2
7	7	3
8	6	4
9	3	4



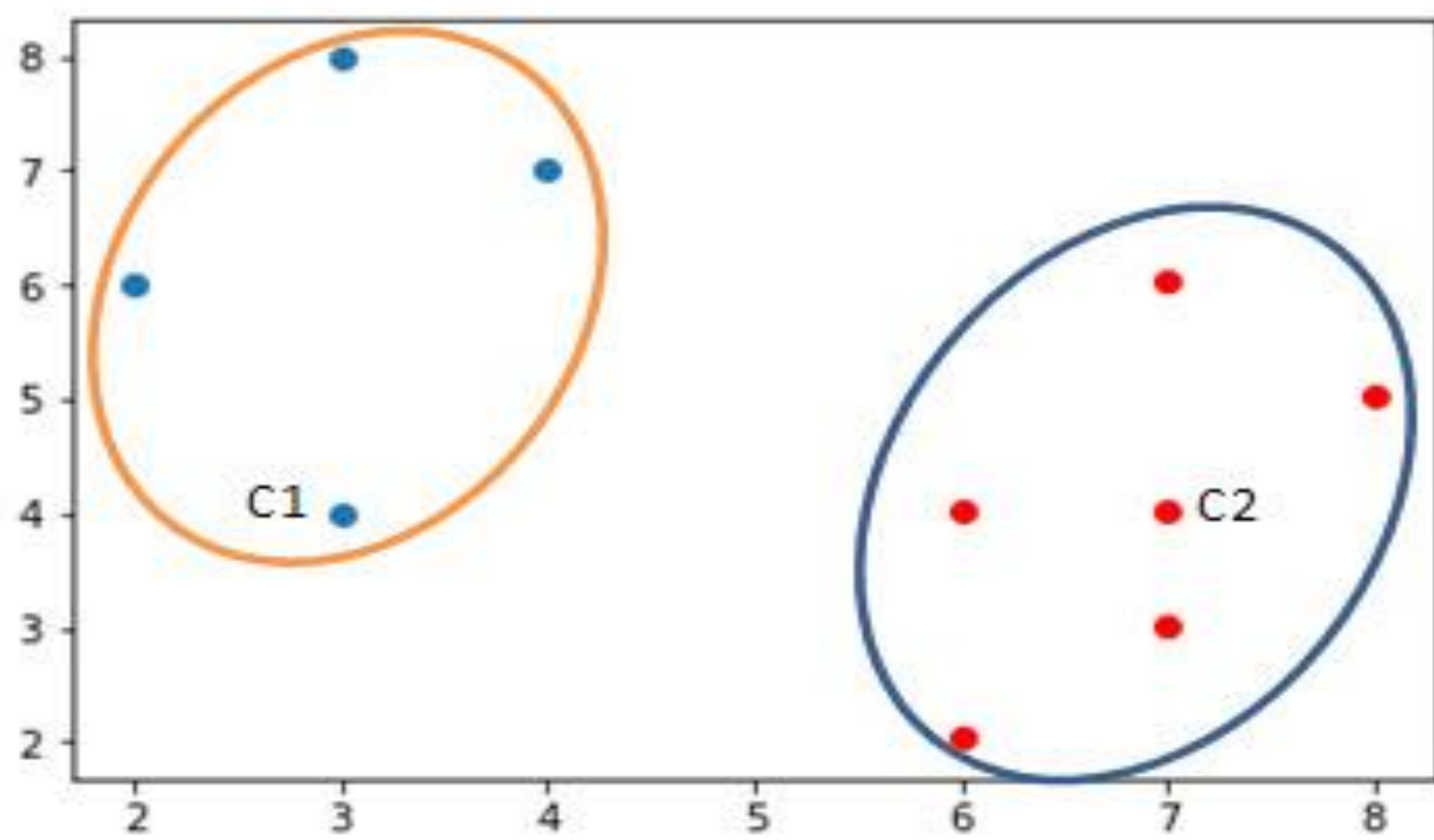
- Step 1: Let the randomly selected 2 medoids be C1 -(3, 4) and C2 -(7, 4).
- Step 2: The dissimilarity of each non-medoid point with the medoids is calculated and tabulated:

	X	Y	Dissimilarity From C1	Dissimilarity From C2
0	7	6	6	2
1	2	6	3	7
2	3	8	4	8
3	8	5	6	2
4	7	4	4	0
5	4	7	4	6
6	6	2	5	3
7	7	3	5	1
8	6	4	3	1
9	3	4	0	4

- 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 $C = (3 + 4 + 4) + (3 + 1 + 1 + 2 + 2) = 20$
- Step 3: Now randomly select one non-medoid point and recalculate the cost. Let the randomly selected point be (7, 3). The dissimilarity of each non-medoid point with the medoids – C1 (3, 4) and C2 (7, 3) is calculated and tabulated.

	X	Y	Dissimilarity From C1	Dissimilarity From C2
0	7	6	6	3
1	2	6	3	8
2	3	8	4	9
3	8	5	6	3
4	7	4	4	1
5	4	7	4	7
6	6	2	5	2
7	7	3	-	-
8	6	4	3	2
9	3	4	-	-

- 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 cost $C = (3 + 4 + 4) + (2 + 2 + 1 + 3 + 3)$
 $C = 22$
Swap Cost = Present Cost – Previous Cost
 $= 22 - 20 = 2 > 0$
- 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



Advantages:

- It is simple to understand and easy to implement.
- K-Medoid Algorithm is fast and converges in a fixed number of steps.
- PAM is less sensitive to outliers than other partitioning algorithms.

Disadvantages:

- 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 center) – briefly, it uses compactness as clustering criteria instead of connectivity.

?