K-means algorithm

Problem

For the given dataset, consider the number of cluster as 3. Find the datapoints in these 3 clusters using K-means algorithm

$(x, y) = \{(2,4), (2,6), (5,6), (4,7), (8,3), (6,6), (5,2), (5,7), (6,3), (4,4)\}$

Use Euclidean distance as a distance function.

→ Iteration 01: (seed point) {Sometimes given}

Let's centroid be

Centroid → $C_1 \rightarrow (1,5)$

Centroid → $C_2 \rightarrow (4,1)$

Centroid → $C_3 \rightarrow (8,4)$

$C_1, C_2, C_3 \Rightarrow$ clusters

Euclidean distance.

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Calculate the distance of each point from each of the centroid of three clusters.

Given data points		Distance to centroid			cluster number
x	y	C_1 (1,5)	C_2 (4,1)	C_3 (8,4)	
2	4	1.41	3.61	6.00	C_1
2	6	1.41	5.39	6.32	C_1
5	6	4.12	5.10	3.61	C_3
4	7	3.61	6.00	5.00	C_1
8	3	7.28	4.47	1.00	C_3
6	6	5.10	5.39	2.83	C_3
5	2	5.00	1.41	3.61	C_2
5	7	4.47	6.08	4.24	C_3
6	3	5.39	2.83	2.24	C_3
4	4	3.16	3.00	4.00	C_2

Clusters after Iteration 1 are.

Cluster 1 → (2,4), (2,6), (4,7)
Cluster 2 → (5,2), (4,4)
Cluster 3 → (5,6), (8,3), (6,6), (5,7), (6,3)

Now us recompute the new cluster by determining the new centroid.
→ centroid can be computed by taking the mean of data points in that cluster.

Hence, Iteration 2

Centroid → C_1 = (2.66, 5.66)

$$\frac{2+2+4}{3} = \frac{8}{3} = 2.66$$

$$\frac{4+6+7}{3} = \frac{17}{3} = 5.66$$

Centroid → C_2 = (4.5, 3)

Centroid → C_3 = (6, 5)

Now determine the distance of all data points from new centroid.

x, y	distance to (2.66, 5.66) C_1	(4.5, 3) C_2	(6, 5) C_3	cluster number
2 4	1.79	2.69	4.12	C1
2 6	0.74	3.91	4.12	C1
5 6	2.36	3.04	1.41	C3
4 7	1.90	4.03	2.83	C1
8 3	5.87	3.5	2.83	C3
6 6	3.56	3.35	1	C3
5 2	4.34	1.12	3.16	C2
5 7	2.70	4.03	2.24	C3
6 3	4.27	1.5	2	C2
4 4	2.13	1.12	2.24	C2

Observe data point (6,3) is ~~changed~~ moved from cluster C_3 to C_2. Hence further centroid need to be calculated.

C_1 → (2,4) (2,6) (4,7)
C_2 → (5,2) (6,3) (4,4)
C_3 → (5,6) (8,3) (6,6), (5,7)

Iteration 3

New centroids will be,

Centroid $c_1 \rightarrow (2.66, 5.66)$

Centroid $c_2 \rightarrow (5, 3)$

Centroid $c_3 \rightarrow (6, 4.5)$

					cluster Number
2	4	1.79	2.16	4.12	c_1
2	6	0.74	2.24	4.03	c_1
5	6	2.36	3.00	1.12	c_3
4	7.0	1.90	4.12	2.50	c_1
8	3	5.97	5.00	3.20	(2)
6	6	3.36	3.12	0.50	c_3
5	2	4.34	1.00	3.64	c_2
9	7	2.70	4.00	1.80	c_3
6	3	4.27	1.00	2.50	c_2
4	4	2.13	1.41	2.50	c_2

Datapoint (8,3) is moved from cluster c_3 to c_2. Hence again new centroid need to be calculated.

Iteration 4:

Centroid → $c_1 = (2.66, 5.66)$

Centroid → $c_2 = (5.25, 3)$

Centroid → $c_3 = (5.33, 6.83)$

x	y	$(2.66, 5.66)$	$(5.25, 3)$	$(5.33, 6.83)$	Cluster number
2	4	1.74	3.88	4.06	C1
2	6	0.34	4.80	4.33	C1
5	6	2.31	3.09	0.47	C3
4	7	1.90	4.37	1.52	C3
8	3	5.92	3.05	4.04	C2
6	6	3.36	3.01	0.850	C3
5	2	4.34	1.02	4.34	C2
5	7	2.370	4.07	1.80 0.75	C3
6	3	4.27	0.75	3.40	C2
4	4	2.13	2.02	2.68	C2

$(5, 7)$

data point $(5, 7)$ is moved from C2 to C3.

∴ New centroid need to be recalculated.

Iteration 5

Centroid → $C_1 = (2, 5)$, Centroid$_2$ = $(5.7, 3)$

Centroid → $C_3 = (5, 6.4)$

x	y	$(2, 5)$	$(5.7, 3)$	$(5, 6.4)$	Cluster number
2	4	1.00	3.88	3.91	C1
2	6	1.00	4.80	3.04	C1
5	6	3.16	3.09	0.58	C3
4	7	2.83	4.37	1.12	C3
8	3	3.32	2.25	4.51	C2
6	6	4.12	3.01	1.12	C3
5	2	4.24	1.25	4.50	C2
5	7	3.61	4.07	0.90	C3
6	3	4.47	0.25	3.64	C2
4	4	2.24	2.02	2.69	C2

Now there is no data point which is moving from one cluster to another.

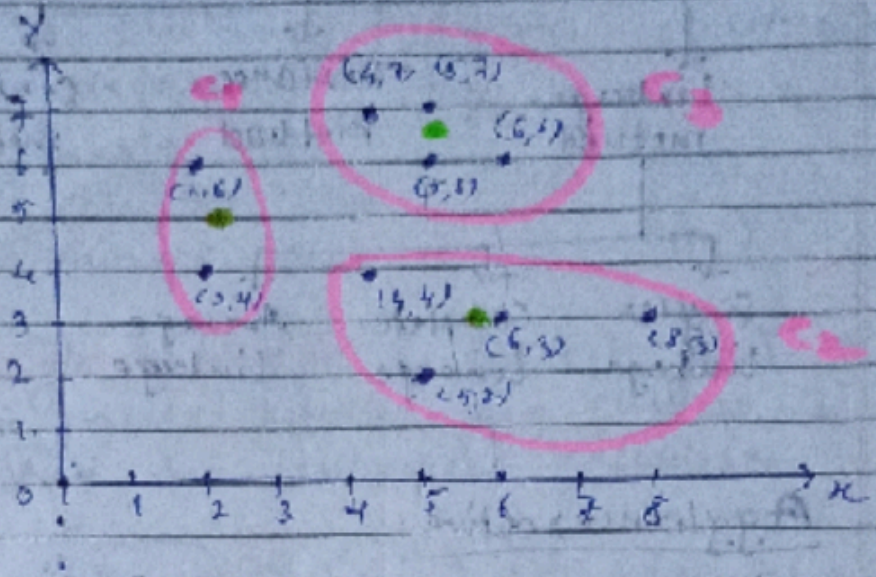
This is the

→ These are the final centroids of clusters whose data points are

$C_1 \rightarrow (3,4), (2,6)$ &

$C_2 \rightarrow (8,3), (5,2), (6,3), (4,4)$

$C_3 \rightarrow (4,6), (4,7), (6,1), (5,7)$



K-medoids algorithm

Manhattan distance between (x_1, y_1) & (x_2, y_2) is

$d(P,Q) = |P-Q|$

$= |x_2 - x_1| + |y_1 - y_2|$

→ It is very similar to k-means algorithm

→ Both k-means and k-medoids are partition based method and try to minimize the distance between data points and centroids

→ In k-medoids - data points are chosen as a centroid and uses manhattan distance to calculate the distance.

→ It is more robust to noise and outliers as compared to k-means.