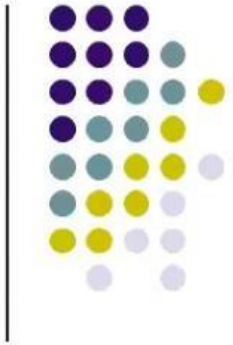
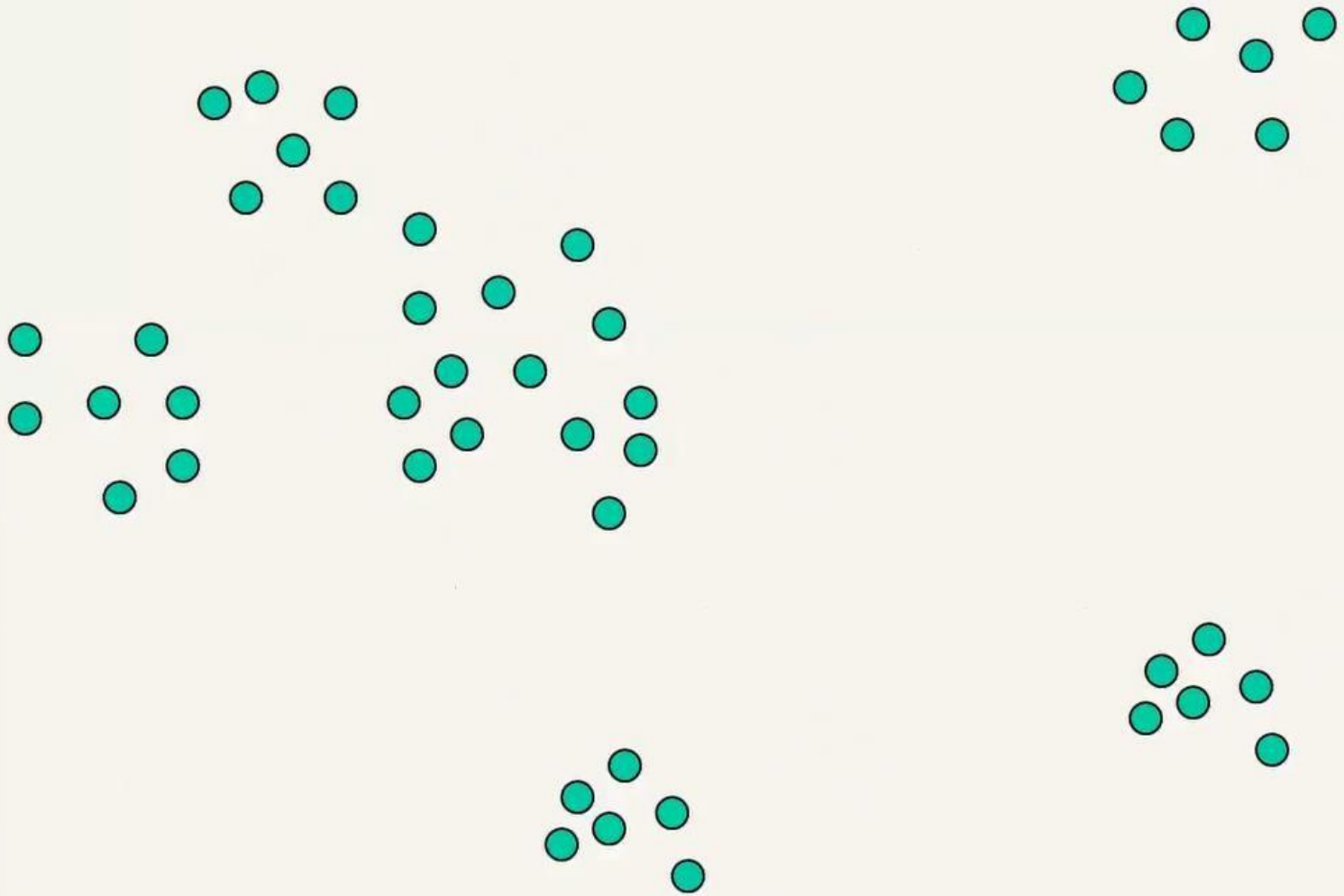
INTRODUCTION-
What is clustering?

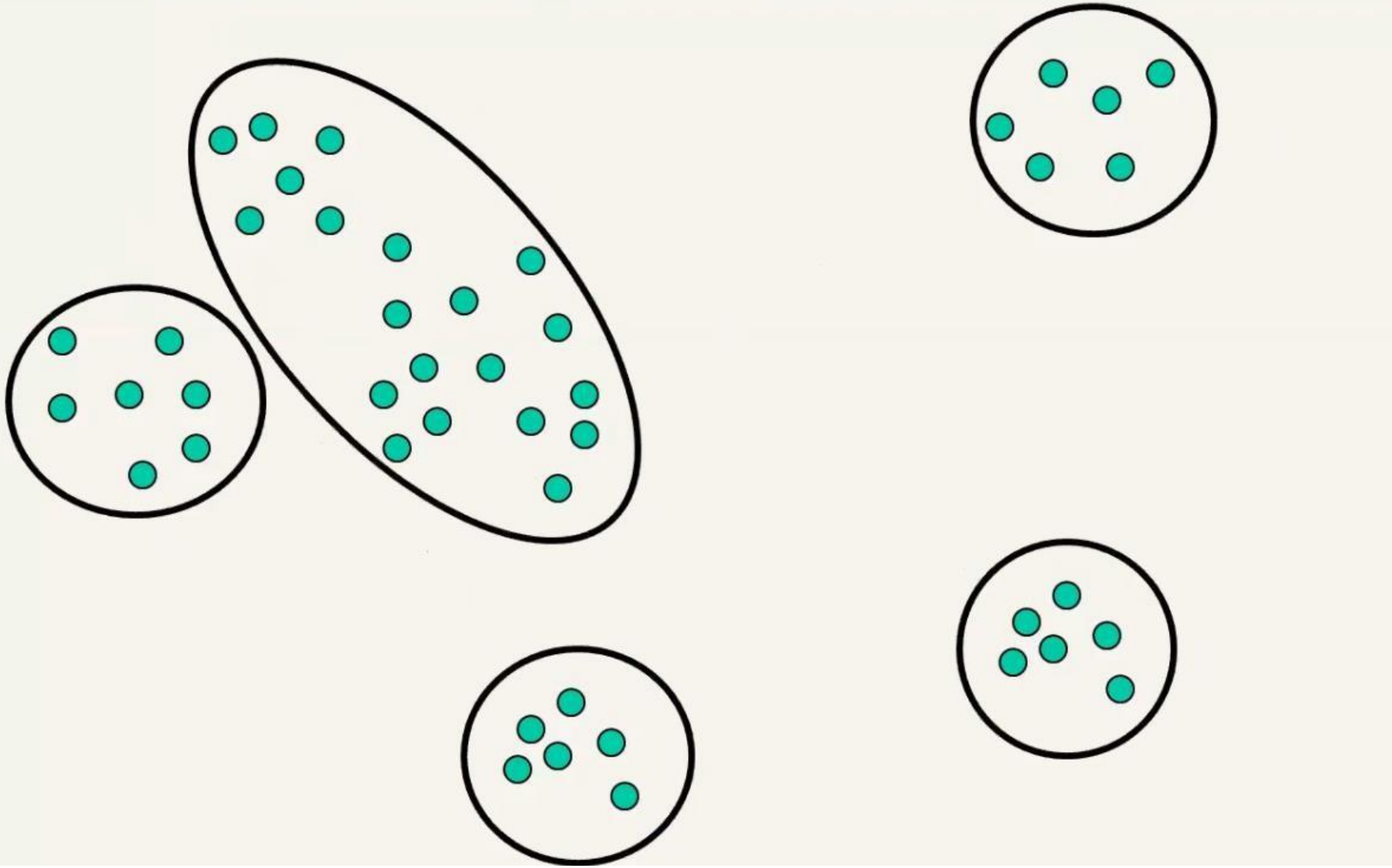


- **Clustering** is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

Clustering



Clustering



K-Means Clustering

Unsupervised Learning

K-means clustering is an unsupervised learning approach to cluster or group n objects into k clusters, where $k < n$

n datapoints

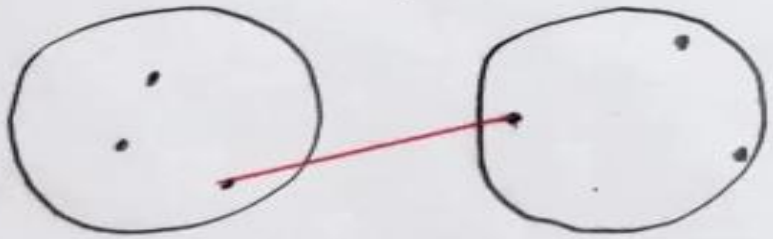
Maximum Clusters = n - 1

Minimum Clusters = 2

Measuring Similarity between Clusters [Slide 7]

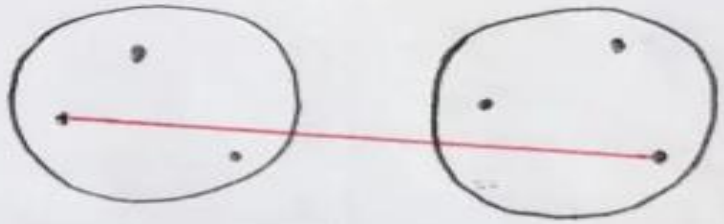
① Single Linkage

Distance between closest two datapoints



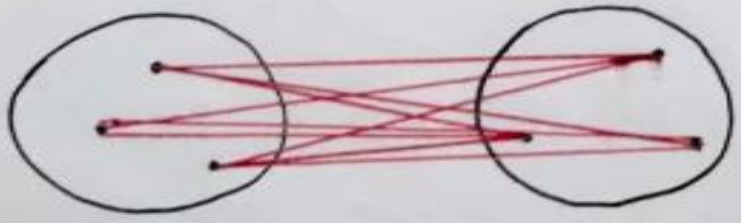
② Complete Linkage

Distance between farthest two datapoints



③ Average Linkage

Average of Distances between all datapoints

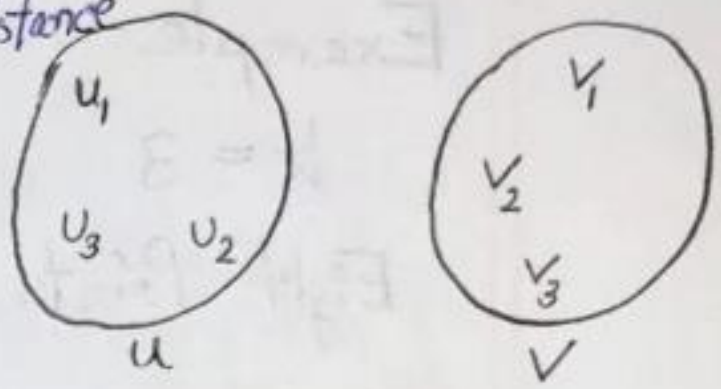


④ Average Group Linkage

Distance between centroids of each clusters

Euclidean Distance & Rectilinear/Manhattan Distance

Let u and v be two clusters/groups



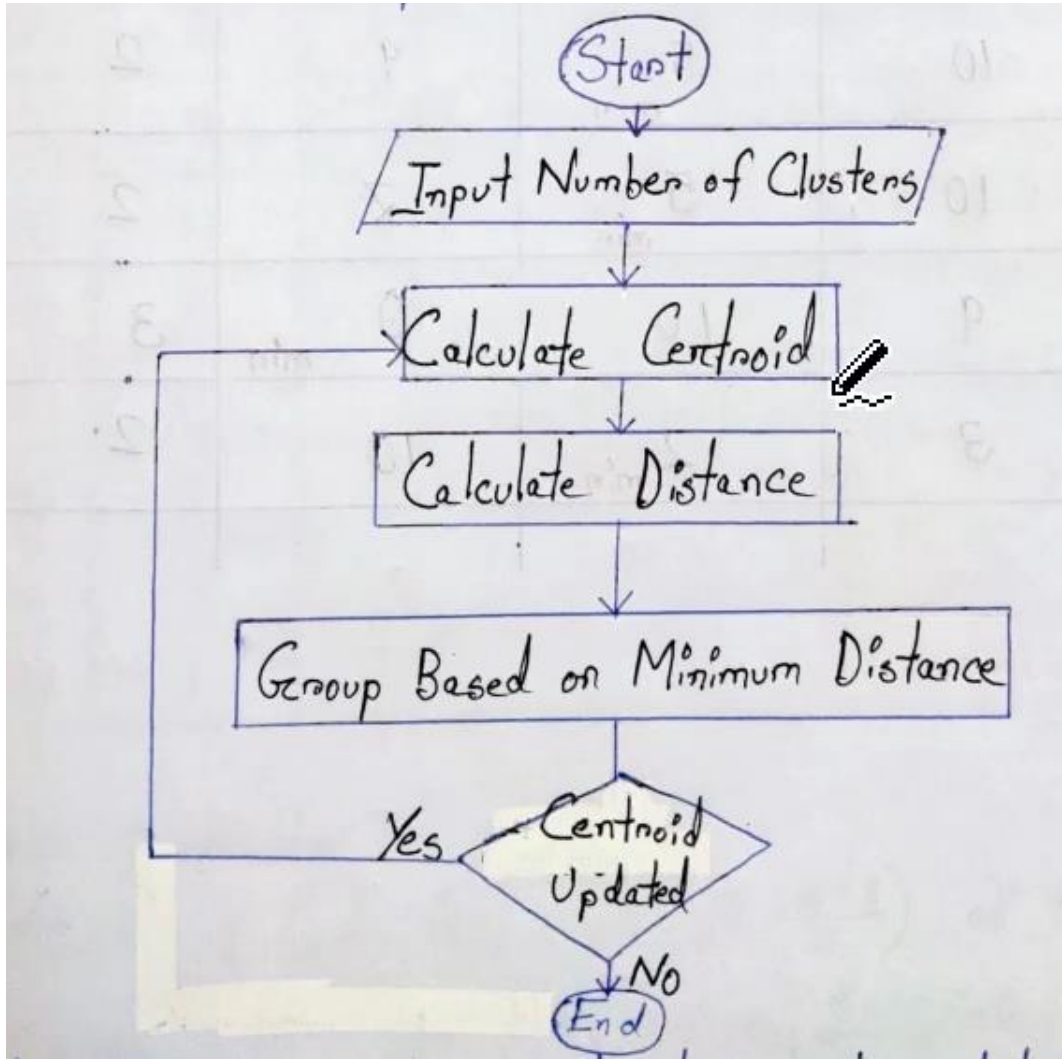
1 Euclidean Distance

$$d = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \ldots + (u_n - v_n)^2}$$

2 Rectilinear Distance / Manhattan Distance

$$d = |u_1 - v_1| + |u_2 - v_2| + \ldots + |u_n - v_n|$$

How the K-Mean Clustering algorithm works?



Example:

Cluster the following eight points (with (x,y) representing locations) into three clusters A1(2,10), A2(2,5) , A3(8,4), A4(5,8), A5(7.5), A6(6,4), A7(1,2) and A8(4,9)

Use Manhattan distance to solve it

N.B: If nothing is mentioned use Euclidean distance

Eight Points, A1(2,10) A2(2,5) A3(8,4) A4(5,8)

Iteration 1 A5(7,5) A6(6,4) A7(1,2) A8(4,9)

Using Rectilinear Distance

	C_1 (2,10)	C_2 (5,8)	C_3 (1,2)	Cluster
(2,10)	0 min	5	9	1
(2,5)	5	6	4 min	3
(8,4)	12	7 min	9	2
(5,8)	5	0 min	10	2
(7,5)	10	5 min	9	2
(6,4)	10	5 min	7	2
(1,2)	9	10	0 min	3
(4,9)	3	2 min	10	2

Cluster-1 only 1 datapoint

Cluster-2 (8,4) (5,8) (7,5) (6,1) (4,9)

$$\text{New Centroid} = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right)$$

$$= (6,6)$$

Cluster-3 (2,5) (1,2)

$$\text{New Centroid} = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1'5, 3'5)$$

Iteration 2

	C_1 (2,10)	C_2 (6,6)	C_3 (1.5,3.5)	Cluster
(2,10)	0	8	7	1
(2,5)	5	5	2	3
(8,4)	12	4	7	2
(5,8)	5	3	8	2
(7,5)	10	2	7	2
(6,4)	10	2	5	2
(1,2)	9	9	2	3
(4,9)	3	5	8	1 → Changed from Iteration-1

New Centroids

Cluster-1 $(2,10), (4,9) \rightarrow \left(\frac{2+4}{2}, \frac{10+9}{2}\right)$ or $(3, 9'5)$

Cluster-2 $(8,4), (5,8), (7,5), (6,4) \rightarrow \left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right)$ or $(6'5, 5'25)$

Cluster-3 $\rightarrow (1'5, 3'5)$ No change

Iteration-3

	C_1 (3,4.5)	C_2 (6.5,[illegible])	C_3 (1.5,3.5)	Clusters
(2,10)	1.5	9.25	7	1
(2,5)	5.5	4.75	2	3
(8,4)	10.5	2.75	7	2
(5,8)	3.5	4.25	8	1 → Changed from Iteration-2
(7,5)	8.5	0.75	7	2
(6,4)	8.5	1.75	5	2
(1,2)	9.5	8.75	2	3
(4,9)	1.5	6.25	8	1

New Centroids

Cluster-1 (2,10), (5,8), (4,9) $\longrightarrow \left(\frac{2+5+4}{3}, \frac{10+8+9}{3}\right)$ or (3·6, 9)

Cluster-2 (8,4), (7,5), (6,4) $\longrightarrow \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right)$ or (7, 4·33)

Cluster-3 $\longrightarrow$ (1·5, 3·5) No change

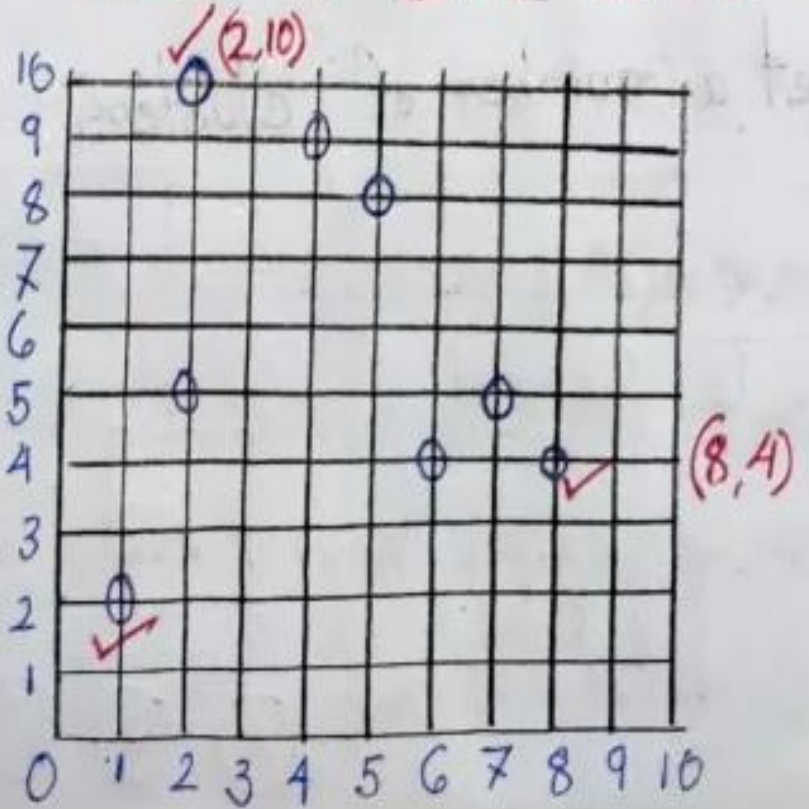
Iteration-4

	C_1 (3·6, 9)	C_2 (7, 4·3)	C_3 (1·5, 3·5)	Clusters
(2,10)	2·6	10·7	7	1
(2,5)	5·6	5·7	2	3
(8,4)	9·4	1·3	7	2
(5,8)	2·4	5·7	8	1
(7,5)	7·4	0·7	7	2
(6,4)	7·4	1·3	5	2
(1,2)	9·6	8·3	2	3
(4,9)	0·4	7·7	8	1

No change

clustering has converged

Trick: In order to solve with minimum iterations, take farthest points



Iteration 1:

Point	C1 (2,10)	C2 (1,2)	C3 (8,4)	Cluster
A1 (2,10)	**0(min)**	9	12	**1**
A2 (2,5)	5	**4(min)**	7	**2**
A3 (8,4)	12	9	**0(min)**	**3**
A4 (5,8)	**5(min)**	10	7	**1**
A5 (7,5)	10	9	**2(min)**	**3**
A6 (6,4)	10	7	**2(min)**	**3**
A7 (1,2)	9	**0(min)**	9	**2**
A8 (4,9)	**3(min)**	10	9	**1**

Cluster from Iteration 1

Cluster	Members
C1	A1 (2,10), A4 (5,8), A8 (4,9)
C2	A2 (2,5), A7 (1,2)
C3	A3 (8,4), A5 (7,5), A6 (6,4)

Compute New Centroids

Cluster 1 (A1, A4, A8)

$$x = (2 + 5 + 4)/3 = 11/3 = 3.67$$
$$y = (10 + 8 + 9)/3 = 27/3 = 9$$

New C1 = (3.67, 9)

Cluster 2 (A2, A7)

$$x = (2 + 1)/2 = 1.5$$
$$y = (5 + 2)/2 = 3.5$$

New C2 = (1.5, 3.5)

Cluster 3 (A3, A5, A6)

$$x = (8 + 7 + 6)/3 = 21/3 = 7$$
$$y = (4 + 5 + 4)/3 = 13/3 = 4.33$$

New C3 = (7, 4.33)

Iteration 2

Point	C1 (3.67, 9)	C2 (1.5, 3.5)	C3 (7, 4.33)	Cluster
A1 (2,10)	**2.67(min)**	7	10.67	**1**
A2 (2,5)	5.67	**2(min)**	5.33	**2**
A3 (8,4)	10.33	7	**0.67(min)**	**3**
A4 (5,8)	**1.67(min)**	8	5.67	**1**
A5 (7,5)	5.33	7	**0.33(min)**	**3**
A6 (6,4)	5.67	5	**1.33(min)**	**3**
A7 (1,2)	9.67	**2(min)**	8.33	**2**
A8 (4,9)	**0.33(min)**	8	8.67	**1**

Iteration 2 Clusters (same as Iteration 1)

Cluster	Members	Centroid
C1	A1, A4, A8	(3.67, 9)
C2	A2, A7	(1.5, 3.5)
C3	A3, A5, A6	(7, 4.33)

No points changed clusters → **clustering has converged**