K-Means Clustering – Solved Example

- Suppose that the data mining task is to cluster points into three clusters,
- where the points are
- *A*1(2, 10), *A*2(2, 5), *A*3(8, 4), *B*1(5, 8), *B*2(7, 5), *B*3(6, 4), *C*1(1, 2), *C*2(4, 9).
- The distance function is Euclidean distance.
- Suppose initially we assign *A*1, *B*1, and *C*1 as the center of each cluster, respectively.

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to			Cluster	New Cluster
A1	2	10					
A2	2	5					
A3	8	4					
B1	5	8					
B2	7	5					
B3	6	4					
C1	1	2					
C2	4	9					

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to			Cluster	New Cluster
			2 10	5 8	1 2		
A1	2	10	0.00	3.61	8.06	1	
A2	2	5	5.00	4.24	3.16	3	
A3	8	4	8.49	5.00	7.28	2	
B1	5	8	3.61	0.00	7.21	2	
B2	7	5	7.07	3.61	6.71	2	
B3	6	4	7.21	4.12	5.39	2	
C1	1	2	8.06	7.21	0.00	3	
C2	4	9	2.24	1.41	7.62	2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

New Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			2 10	5 8	1 2		
A1	2	10	0.00	3.61	8.06	1	
A2	2	5	5.00	4.24	3.16	3	
A3	8	4	8.49	5.00	7.28	2	
B1	5	8	3.61	0.00	7.21	2	
B2	7	5	7.07	3.61	6.71	2	
B3	6	4	7.21	4.12	5.39	2	
C1	1	2	8.06	7.21	0.00	3	
C2	4	9	2.24	1.41	7.62	2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			2 10	6 6	1.5 1.5		
A1	2	10	0.00	5.66	6.52	1	1
A2	2	5	5.00	4.12	1.58	3	3
A3	8	4	8.49	2.83	6.52	2	2
B1	5	8	3.61	2.24	5.70	2	2
B2	7	5	7.07	1.41	5.70	2	2
B3	6	4	7.21	2.00	4.53	2	2
C1	1	2	8.06	6.40	1.58	3	3
C2	4	9	2.24	3.61	6.04	2	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

New Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			2 10	6 6	1.5 1.5		
A1	2	10	0.00	5.66	6.52	1	
A2	2	5	5.00	4.12	1.58	3	
A3	8	4	8.49	2.83	6.52	2	
B1	5	8	3.61	2.24	5.70	2	
B2	7	5	7.07	1.41	5.70	2	
B3	6	4	7.21	2.00	4.53	2	
C1	1	2	8.06	6.40	1.58	3	
C2	4	9	2.24	3.61	6.04	1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

New Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			3, 9.5	6.5, 5.25	1.5, 3.5		
A1	2	10	1.12	6.54	6.52	1	
A2	2	5	4.61	4.51	1.58	3	
A3	8	4	7.43	1.95	6.52	2	
B1	5	8	2.50	3.13	5.70	1	
B2	7	5	6.02	0.56	5.70	2	
B3	6	4	6.26	1.35	4.53	2	
C1	1	2	7.76	6.39	1.58	3	
C2	4	9	1.12	4.51	6.04	1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			3.67, 9	7, 4.33	1.5, 3.5		
A1	2	10	1.94	7.56	6.52	1	1
A2	2	5	4.33	5.04	1.58	3	3
A3	8	4	6.62	1.05	6.52	2	2
B1	5	8	1.67	4.18	5.70	1	1
B2	7	5	5.21	0.67	5.70	2	2
B3	6	4	5.52	1.05	4.53	2	2
C1	1	2	7.49	6.44	1.58	3	3
C2	4	9	0.33	5.55	6.04	1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Agglomerative Hierarchical Clustering Solved Example

- Consider the following set of 6 one dimensional data points:
- 18, 22, 25, 42, 27, 43
- Apply the **agglomerative hierarchical clustering** algorithm to build the hierarchical clustering **dendogram**.
- Merge the clusters using **Min distance** and update the proximity matrix accordingly.
- Clearly show the **proximity matrix** corresponding to each iteration of the algorithm.

Agglomerative Hierarchical Clustering Solved Example

- Step – 1

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

Agglomerative Hierarchical Clustering Solved Example

- Step – 2

	18	22	25	27	42, 43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42, 43	24	20	17	15	0

Agglomerative Hierarchical Clustering Solved Example

- Step – 3

	18	22	25, 27	42, 43
18	0	4	7	24
22	4	0	3	20
25, 27	7	3	0	15
42, 43	24	20	15	0

Agglomerative Hierarchical Clustering Solved Example

- Step – 4

	18	22, 25, 27	42, 43
18	0	4	24
22, 25, 27	4	0	15
42, 43	24	15	0

Agglomerative Hierarchical Clustering Solved Example

- Step – 5

	18, 22, 25, 27	42, 43
18, 22, 25, 27	0	15
42, 43	15	0

Agglomerative Hierarchical Clustering Solved Example

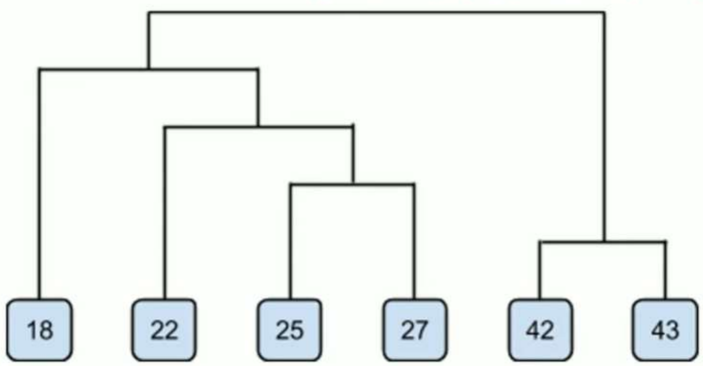
- Step – 6

	18, 22, 25, 27, 42, 43
18, 22, 25, 27, 42, 43	0

Agglomerative Hierarchical Clustering Solved Example

- **Dendrogram**

((42, 43), (((25, 27), 22), 18))



Clusters using a Single Link Technique Example - 1

Sample No.	X	Y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

Clusters using a Single Link Technique Example - 1

Step 1: Compute the distance matrix

- So we have to find the Euclidean distance between each and every points.
- Let $A(x_1, y_1)$ and $B(x_2, y_2)$ are two points.
- Then Euclidean distance between

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Clusters using a Single Link Technique Example - 1

Sample No.	X	Y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

$$d(p_1, p_2) = \sqrt{(0.22 - 0.40)^2 + (0.38 - 0.53)^2}$$
$$= 0.23$$

$$d(p_1, p_3) = \sqrt{(0.35 - 0.40)^2 + (0.32 - 0.53)^2}$$
$$= 0.22$$

$$d(p_2, p_3) = \sqrt{(0.35 - 0.22)^2 + (0.32 - 0.38)^2}$$
$$= 0.14$$

Clusters using a Single Lin

Sample No.	X	Y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

$$\begin{pmatrix} & P1 & P2 & P3 & P4 & P5 & P6 \\ P1 & 0 & & & & & \\ P2 & 0.23 & 0 & & & & \\ P3 & 0.22 & 0.14 & 0 & & & \\ P4 & 0.37 & 0.19 & 0.13 & 0 & & \\ P5 & 0.34 & 0.14 & 0.28 & 0.23 & 0 & \\ P6 & 0.24 & 0.24 & 0.10 & 0.22 & 0.39 & 0 \end{pmatrix}$$

Clusters using a Single Link Technique Example - 1

Step 2: Merging the two closest members.

- Here the **minimum value is 0.10** and hence we combine P3 and P6 (as 0.10 came in the P6 row and P3 column).
- Now, form clusters of elements corresponding to the minimum value and update the distance matrix.

Clusters using a Single Link Technique Example - 1

Now we will update the Distance Matrix:

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.14	0			
P4	0.37	0.19	0.13	0		
P5	0.34	0.14	0.28	0.23	0	
P6	0.24	0.24	0.10	0.22	0.39	0

	P1	P2	P3, P6	P4	P5
P1	0				
P2	0.23	0			
P3, P6	0.22	0.14	0		
P4	0.37	0.19	0.13	0	
P5	0.34	0.14	0.28	0.23	0

(P3, P6)

Clusters using a Single Link Technique Example - 1

Now we will update the Distance Matrix:

	P1	P2	P3, P6	P4	P5
P1	0				
P2	0.23	0			
P3, P6	0.22	0.14	0		
P4	0.37	0.19	0.13	0	
P5	0.34	0.14	0.28	0.23	0

	P1	P2	P3, P6, P4	P5
P1	0			
P2	0.23	0		
P3, P6, P4	0.22	0.14	0	
P5	0.34	0.14	0.28	0

{(P3, P6), P4}

Clusters using a Single Link Technique Example - 1

Now we will update the Distance Matrix:

$$\begin{array}{ccccc}
 & P1 & P2 & P3, P6, P4 & P5 \\
P1 & 0 & & & \\
P2 & 0.23 & 0 & & \\
P3, P6, P4 & 0.22 & 0.14 & 0 & \\
P5 & 0.34 & 0.14 & 0.28 & 0
\end{array}$$

$$\begin{array}{cccc}
 & P1 & P2, P5 & P3, P6, P4 \\
P1 & 0 & & \\
P2, P5 & 0.23 & 0 & \\
P3, P6, P4 & 0.22 & 0.14 & 0
\end{array}$$

{(P3, P6), P4} and **(P2, P5)**

Clusters using a Single Link Technique Example - 1

Now we will update the Distance Matrix:

$$\begin{array}{cccc}
 & P1 & P2, P5 & P3, P6, P4 \\
P1 & 0 & & \\
P2, P5 & 0.23 & 0 & \\
P3, P6, P4 & 0.22 & 0.14 & 0
\end{array}$$

$$\begin{array}{ccc}
 & P1 & P2, P5, P3, P6, P4 \\
P1 & 0 & \\
P2, P5, P3, P6, P4 & 0.22 & 0
\end{array}$$

[{(P3, P6), P4}, (P2, P5)]

Clusters using a Single Link Technique Example - 1

Now we will update the Distance Matrix:

$$\begin{pmatrix} & P1 & P2,P5 & P3,P6,P4 \\ P1 & 0 & & \\ P2,P5 & 0.23 & 0 & \\ P3,P6,P4 & 0.22 & 0.14 & 0 \end{pmatrix}$$

$$\begin{pmatrix} & P1 & P2,P5,P3,P6,P4 \\ P1 & 0 & \\ P2,P5,P3,P6,P4 & 0.22 & 0 \end{pmatrix}$$

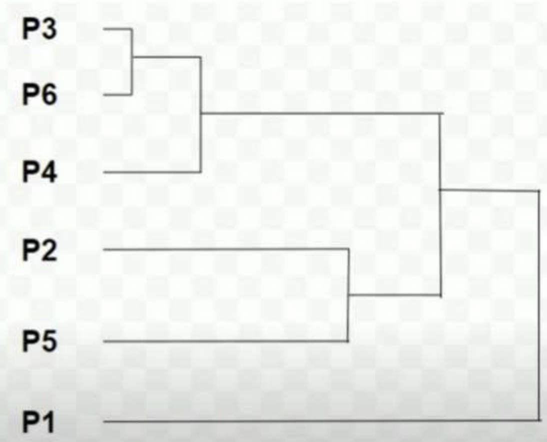
[{(P3, P6), P4}, (P2, P5)]

[{(P3, P6), P4}, (P2, P5)] , P1

Clusters using a Single Link Technique Example - 1

So now we have reached to the solution, the dendrogram for those question will be as follows:

[{(P3, P6), P4}, (P2, P5)] , P1



Dendogram of the cluster formed