Unsupervised Learning Worked Example

K-Means Clustering Algorithm

Use the K-means algorithm and Euclidean distance to cluster the 5 data points given in Figure 1 into k = 2 clusters.

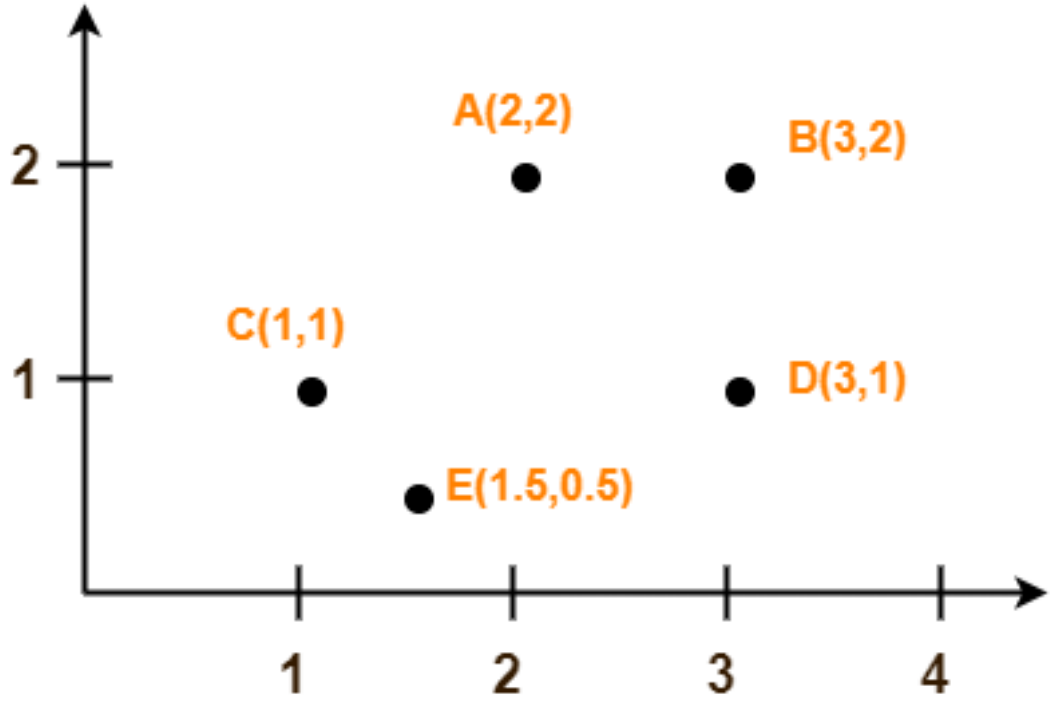


Figure 1: Dataset

Step 1 : Assume A(2, 2) and C(1, 1) are centers of the two clusters.
Step 2: Calculate the distance from each point to each cluster center.
Note: Check what are the formula to calculate the distance.

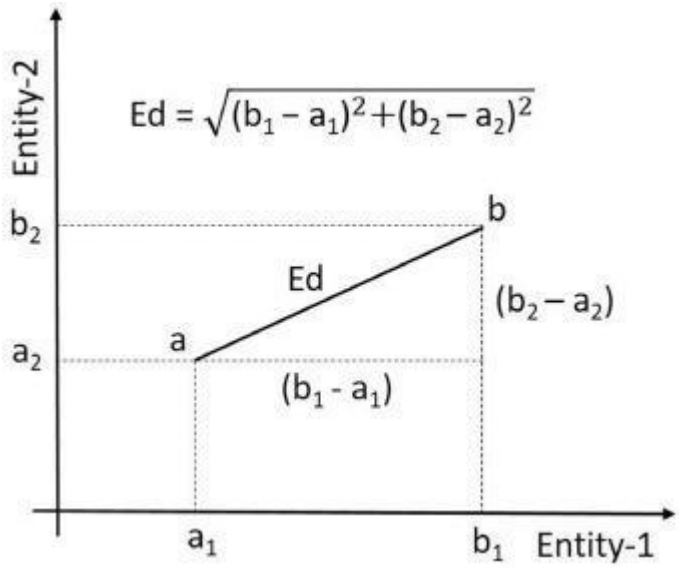


Figure 2: Euclidean distance

	Distance from Center 1 (2,2)	Distance from Center 2 (1,1)
A	0	1.41
B	1	2.24
C	1.41	0
D	1.41	2
E	1.58	0.71

E.g. of calculation:

P(A, C1)

$= \text{sqrt}\ [\ (x2 - x1)^2 + (y2 - y1)^2\ ]$

$= \text{sqrt}\ [\ (2 - 2)^2 + (2 - 2)^2\ ]$

$= \text{sqrt}\ [\ 0 + 0\ ]$

$= 0$

P(A, C2)

$= \text{sqrt}\ [\ (x2 - x1)^2 + (y2 - y1)^2\ ]$

$= \text{sqrt}\ [\ (1 - 2)^2 + (1 - 2)^2\ ]$

$= \text{sqrt}\ [\ 1 + 1\ ]$

$= \text{sqrt}\ [\ 2\ ]$

$= 1.41$

Step 3: Label all data points to closest cluster center.

	Distance from Center 1 (2,2)	Distance from Center 2 (1,1)	Point belongs to Cluster
A	0	1.41	C1
B	1	2.24	C1
C	1.41	0	C2
D	1.41	2	C1
E	1.58	0.71	C2

Step 4 : Recompute the center of newly formed clusters.
Note: The center of a cluster is computed by taking mean of all data points contained in that cluster.

	New Cluster Center
Center 1	2.67, 1.67
Center 2	1.25, 0.75

First Cluster: A, B, D
Second Cluster: C, E

For Cluster-01:

Center of Cluster-01

= ((2 + 3 + 3)/3, (2 + 2 + 1)/3)

= (2.67, 1.67)

For Cluster-02:

Center of Cluster-02

= ((1 + 1.5)/2, (1 + 0.5)/2)

= (1.25, 0.75)

This is completion of Iteration-01!!

Step 5 : Repeat Step 2 to 4 until stopping criteria is met.

i. Center of newly formed clusters do not change.
ii. Points remain present in the same cluster.
iii. Maximum number of iterations are reached.

Unsupervised Learning Worked Example

Hierarchical Agglomerative Clustering Algorithm

Question: Say there is **one dimensional** data set **{7,10,20,28,35}**, perform hierarchical clustering and plot the dendogram to visualize it.

- A = 7
- B = 10
- C = 20
- D = 28
- E = 35

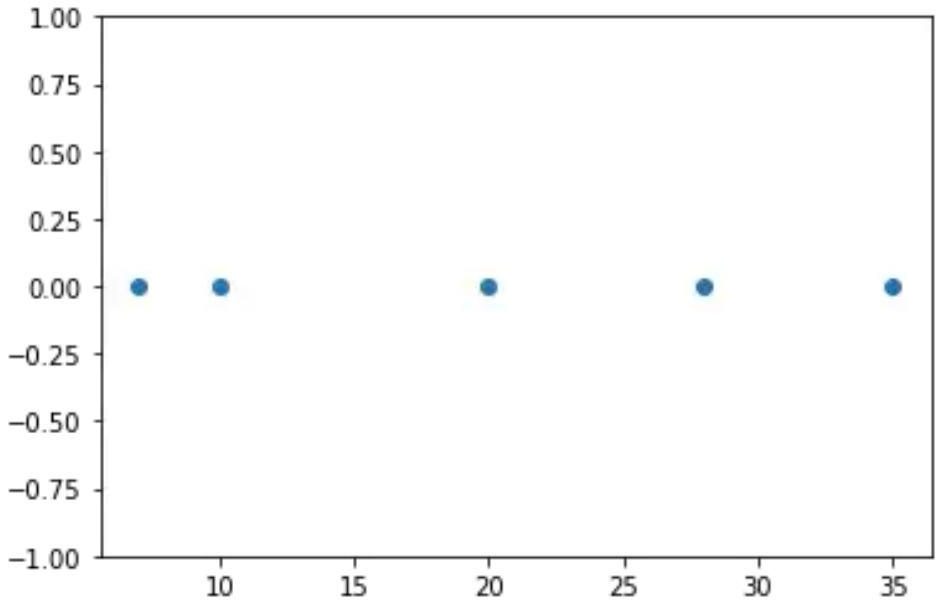
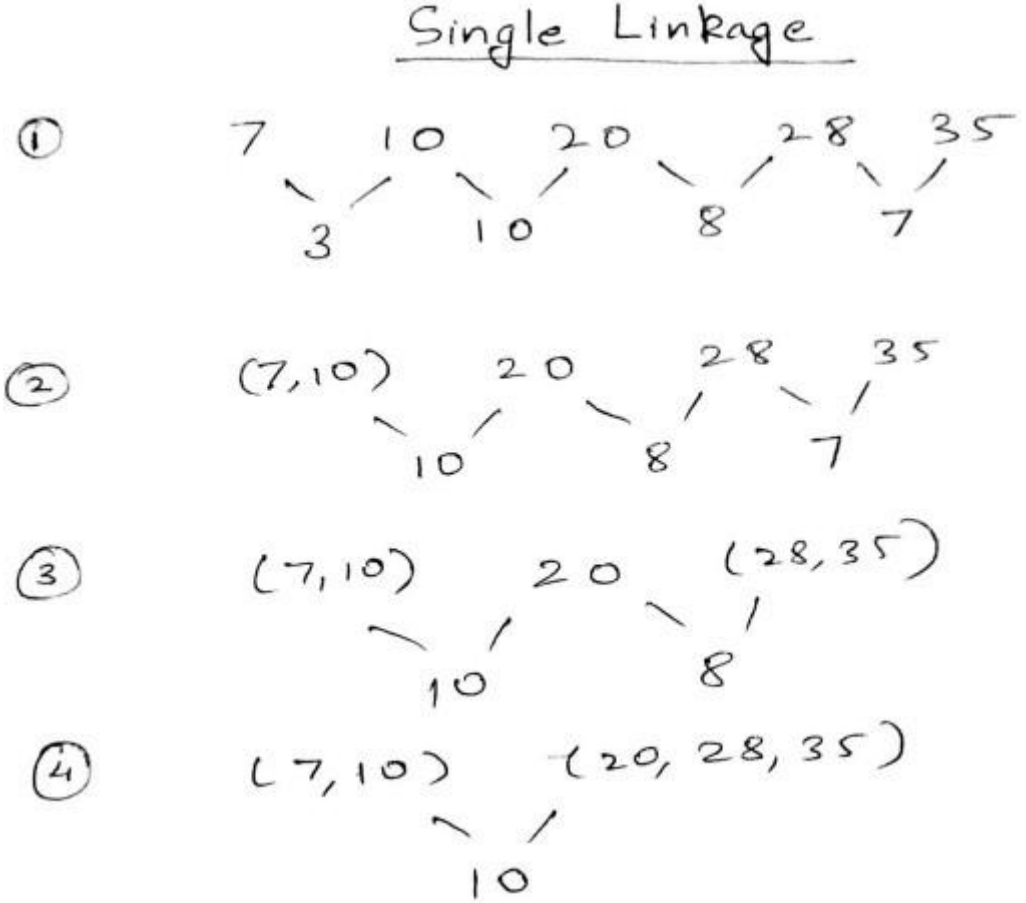
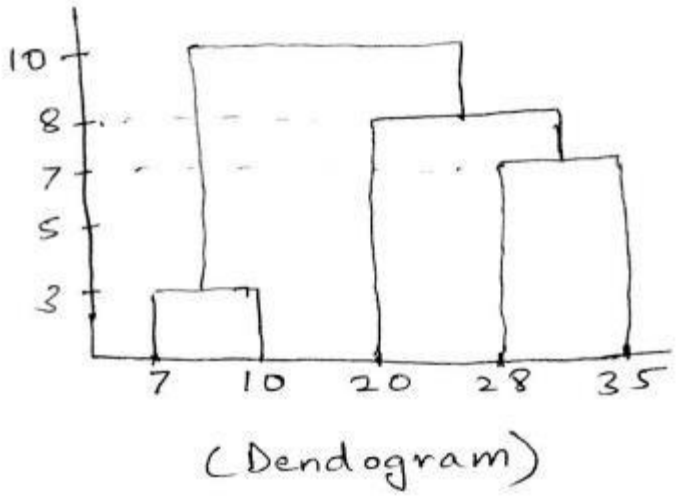


Figure 1: Visualize the data

IF Single Linkage is applied:



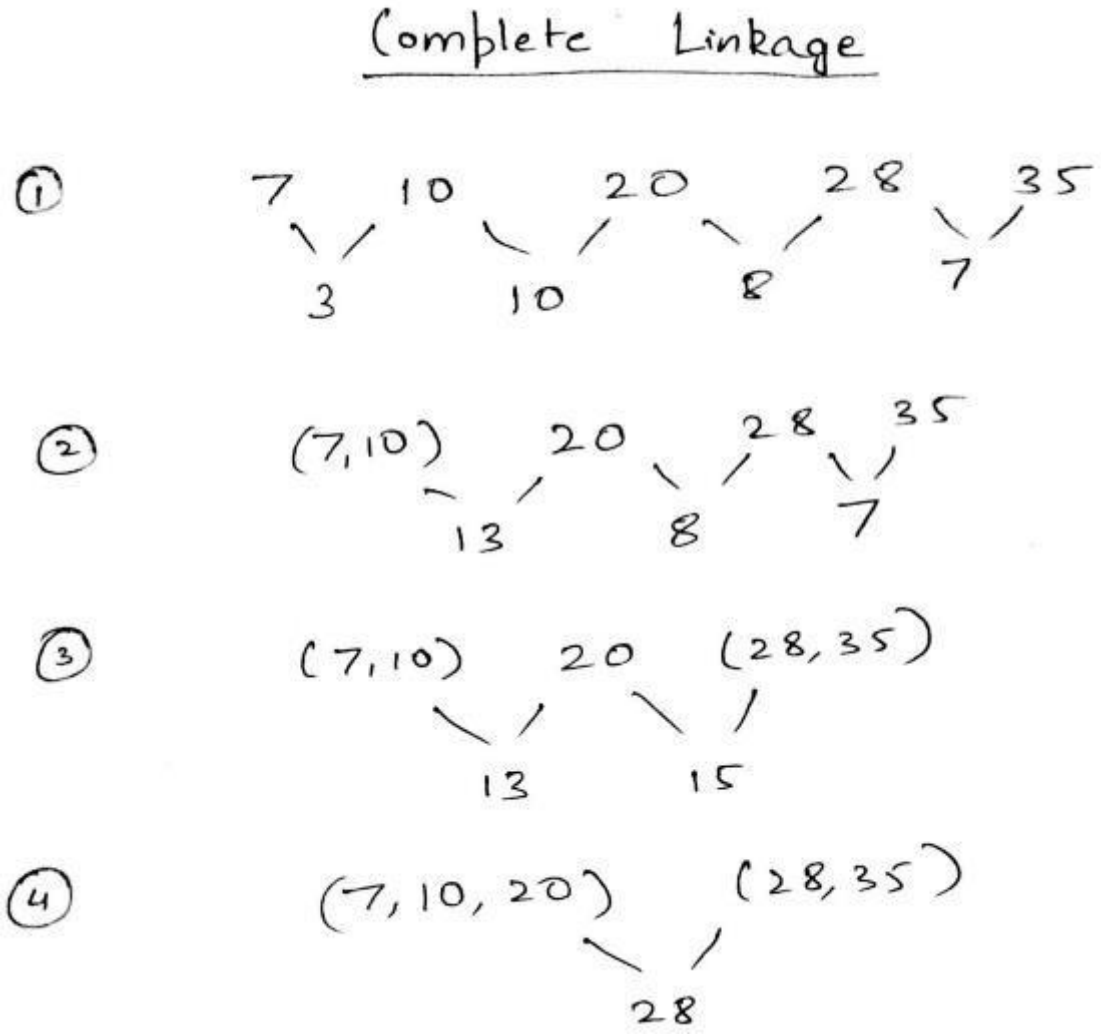


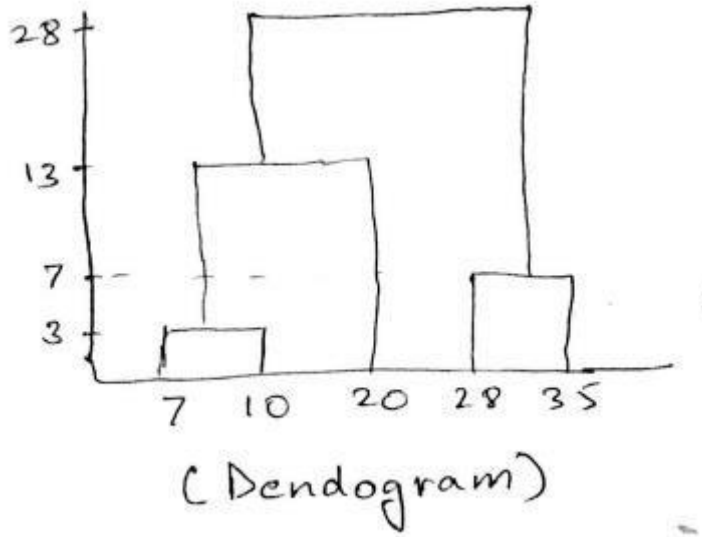
Final Answer:

Cluster 1 : (7,10) or (A, B)

Cluster 2 : (20,28,35) or (C, D, E)

IF Complete Linkage is applied:





Final Answer:

Cluster 1 : (7,10,20) or (A, B, C)

Cluster 2 : (28,35) or (D, E)