ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

UNIT – 4

CLUSTERING

Dr. G. Ravi & Dr. S. Peerbasha
PG & Research Department of Computer Science
Jamal Mohamed College (Autonomous)
Trichy-620 020.

UNIT - 4

CLUSTERING

CLUSTERING:

- Clustering is also known as Cluster Analysis.
- It is the process of grouping objects with similar behaviour. Thus, objects in same cluster exhibit similar character.
- It is an important Unsupervised Learning Problem

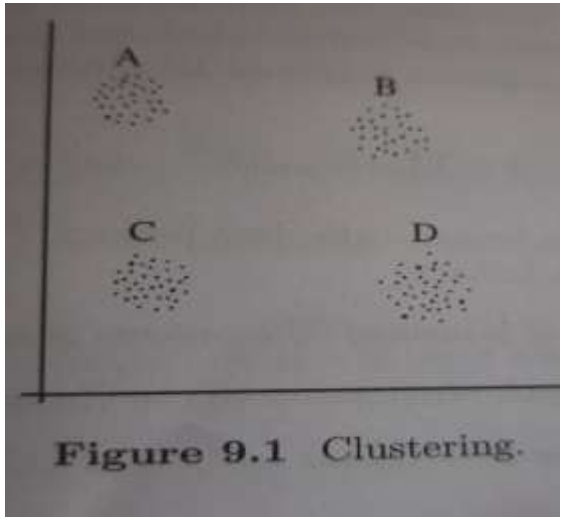


Figure 9.1 Clustering.

CLUSTERING ALGORITHMS

- Based on the cluster model, clustering algorithms can be classified as follows:-

1. **Partition clustering** (Here, data included in one cluster will not appear in another one) Ex:- K-Means algorithm
2. **Overlapping clustering** (Here, data may belong to more than one cluster depending on the degree of correlation. Ex:- Fuzzy C-Means algorithm
3. **Hierarchical clustering** (Here, the clusters which are close to one another can form a union) Ex:- Agglomerative algorithm
4. **Probabilistic clustering** (Here, the initial stage is set by treating each datum as separate clusters. Final clusters are formed after few iterations.) Ex:- Probabilistic D-Clustering algorithm

K-MEANS CLUSTERING / FORGY'S ALGORITHM

➔Unsupervised Learning algorithm

➔Developed by MacQueen in the year 1967.

➔Almost all the clustering problems can be solved using K-Means algorithm.

➔ It is an easy procedure which starts by defining the centroid of clusters.

Hint:- Centroid is nothing but the average location of all the points (or) balance point of set of points.

➔The objective function of K-means method is a Squared error function.

Algorithm: K-Means Clustering

Step-1: Let m1, m2, …… m_k be the set of k-means. The process is carried out in 2 steps

Step-2: Assignment step: Assign each observation to the cluster with the closest mean

Step-3: Update step: Calculate the new means to be the centroid of the observations in the cluster

Step-4: The algorithm seems to have converged when the assignments no longer change

Step-5: As k-means is a simple and basic algorithm, it has been taken up by problems in various domains

Facts about K-Means Clustering

➔ There is no efficient scheme that guarantees to find the optimal function for this clustering.

➔ The only advisable method is to run the algorithm several times so that we get an accurate result.

➔ The start decides the results. We can use various methods for selecting the initial centroid position such as the following:-

1. Select the centroid from the dataset in a random scheme.
2. Select the farthest n data points fro the mean of the dataset.
3. Select the nearest n data points from the mean of the dataset.
4. Select n data points that have the largest sum of pair-wise square distance.

K-Means Clustering Weakness

→ It is always the initial clustering that decides the fate of the result.

→ When the dataset if few, clustering can be inaccurate.

→ As we consider the variables to have the same weights, we never know which contributes more to the clustering process.

→ The outliers or noise can reduce the mathematical accuracy of averaging, which tries to pull the centroid away from its actual position.

Possible Solutions

→ Make the dataset into a big space; when the dataset increases, more accurate results will be produced.

→ By using median instead of mode, the outlier distributions can be prevented to an extent.

Applications of K-Means Clustering

→ Market Segmentation

→ Computer Vision

→ Geo-Statistics

→ Astronomy

→ Agriculture

→ Serves as a preprocessing step for other algorithms for creating a starting configuration.

WORKING WITH AN EXAMPLE

Object	1st attribute (X):	2nd attribute (Y):
Dataset A	1	1
Dataset B	2	1
Dataset C	4	3
Dataset D	5	4

Let us first represent it in the graph (Figure 9.2).

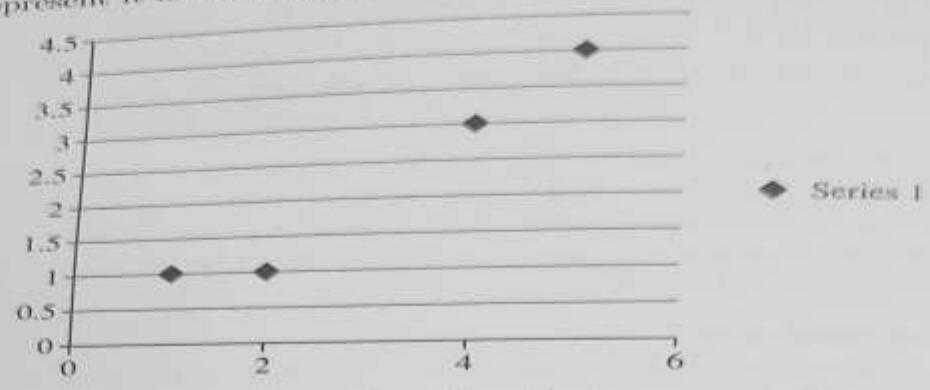


Figure 9.2 Example—Graphical representation.

The process starts by choosing the centroids. Let the datasets A and B be the first centroids, so the positions are $C_1 = (1, 1)$ and $C_2 = (2, 1)$.

Now, we need to calculate the distance between each dataset and the cluster centroid. Using the Euclidean distance formula, we get the first distance matrix. The distance matrix will have two rows, as we have taken two centroids; the values of each column will correspond to the distance of x coordinate and y coordinate. For example,

Distance of objects from centroid 1, $C_1 = (1, 1)$ is found as (with respect to x coordinate)

A from $c_1 = 0$

B from $c_1 = 1$

C from $c_1 = \sqrt{\{(4-1)^2 + (3-1)^2\}} = 3.61$

D from $c_1 = \sqrt{\{(5-1)^2 + (4-1)^2\}} = 5$

Distance of objects from centroid 2, $C_2 = (2, 1)$ is found as (with respect to y coordinate)

A from $c_2 = 1$

B from $c_2 = 0$

C from $c_2 = \sqrt{\{(4-2)^2 + (3-1)^2\}} = 2.83$

D from $c_2 = \sqrt{\{(5-2)^2 + (4-1)^2\}} = 4.24$

$$D^0 = \begin{Bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{Bmatrix}$$

c_1 = (1,1): group 1

c_1 = (2,1): group 2

Next, we need to assign each object based on the minimum distance. By doing so, dataset A will be grouped with the 1st centroid and all the others will be with the second group. The grouping can be well understood by checking the grouping matrix.

$$G^0 = \begin{Bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ A & B & C & D \end{Bmatrix}$$

c_1 = (1,1): group 1

c_1 = (2,1): group 2

Now, the process is repeated for the new set of centroids. We know that group 1 has only one element; hence, the centroid is the same, $C_1 = (1, 1)$. Group 2 has 3 elements; so the centroid will be $C_2 = ((2+4+5)/3, (1+3+4)/3) = (11/3, 8/3)$. The new centroid and the objects are shown in Figure 9.3.

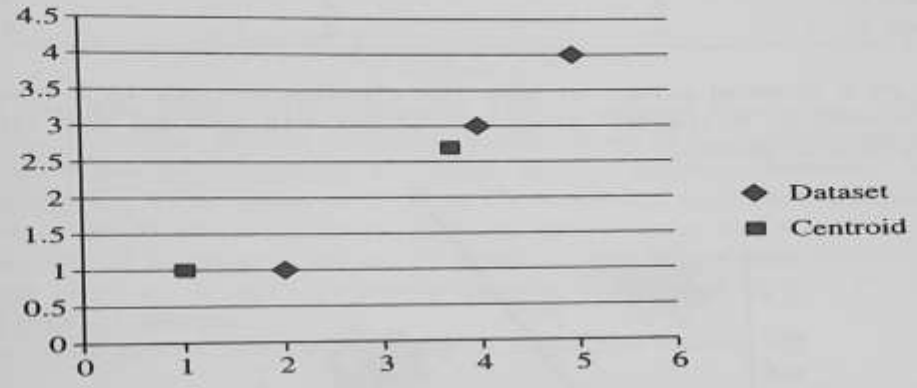


Figure 9.3 Graphical representation.

As we did before, the distance matrix is found out using the Euclidean distance formula, and the new distance matrix we get is

$$D^1 = \begin{Bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{Bmatrix}$$

c_1 = (1,1): group 1

$c_1 = (\frac{11}{3}, \frac{8}{3})$: group 2

The process of assigning the objects to a particular group is done by checking the distance. Now, objects A and B will be placed in group 1 and C and D in group 2.
The next iteration begins and new centroids are calculated: $C_1 = (1.5, 1)$ and $C_2 = (4.5, 3.5)$.

The new distance matrix is calculated and the distance matrix is

$$D^2 = \begin{Bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{Bmatrix}$$

$c_1 = (1\frac{1}{2},1)$: group 1

$c_1 = (4\frac{1}{2},3\frac{1}{2})$: group 2

While assigning objects to group, again datasets A and B are grouped under group 1, and datasets C and D under group 2. While comparing the grouping, we can see that the objects stick to the groups and have not changed. Thus, we can say that computation has reached a safe state and it need not require any more grouping. Here, the algorithm terminates. So, the final results and grouping are as shown in Table 9.2.

Table 9.2 Example—k-means

Object	1st attribute (X):	2nd attribute (Y):	Group number
Dataset A	1	1	1
Dataset B	2	1	1
Dataset C	4	3	2
Dataset D	5	4	3

So, let us look at a general image of how the clusters appear in k-means, i.e., each data object will be allocated to a specific cluster. There will not be any overlapping of data. Figure 9.4 shows such a clustering.

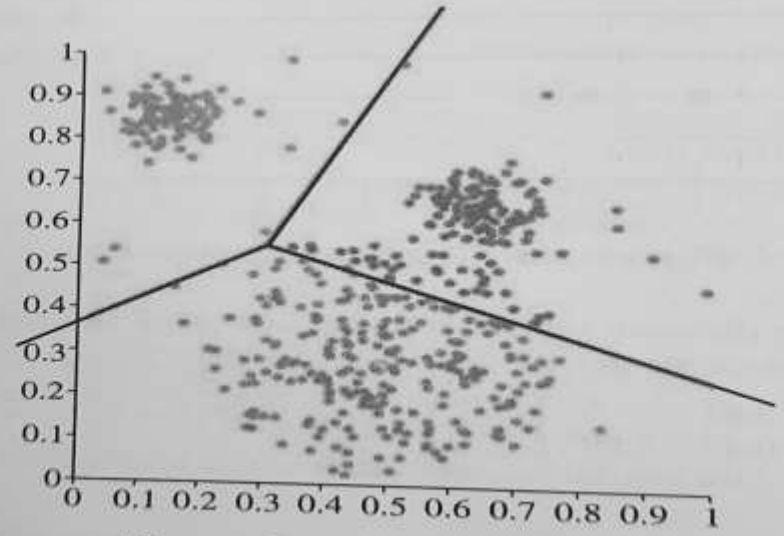


Figure 9.4 k-mean clusters.

FUZZY CLUSTERING / FUZZY C MEANS CLUSTERING (FCM)

➔ It is a type of Overlapping clustering where the clusters belong to a particular group depending on the membership grade.

➔ Data elements can be seen in more than one cluster.

➔ Developed by Jim Bezdek in the year 1981.

➔ This algorithm provides detail about the data in the boundary regions (edges) of various clusers.

ADVANTAGES OF FUZZY CLUSTERING

➔ Unsupervised learning algorithm

➔ Always converges

DISADVANTAGES OF FUZZY CLUSTERING

➔ Long computational time

➔ Sensitivity to the initial guess (speed, local minima)

➔ Sensitivity to noise

➔ One expects low (or even no) membership degree for outliers (Noisy points)

FUZZY C MEANS CLUSTERING ALGORITHM

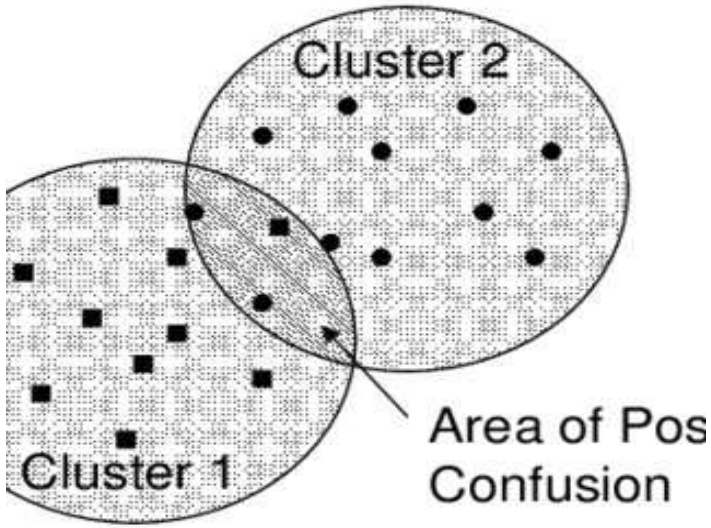
1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$

2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$



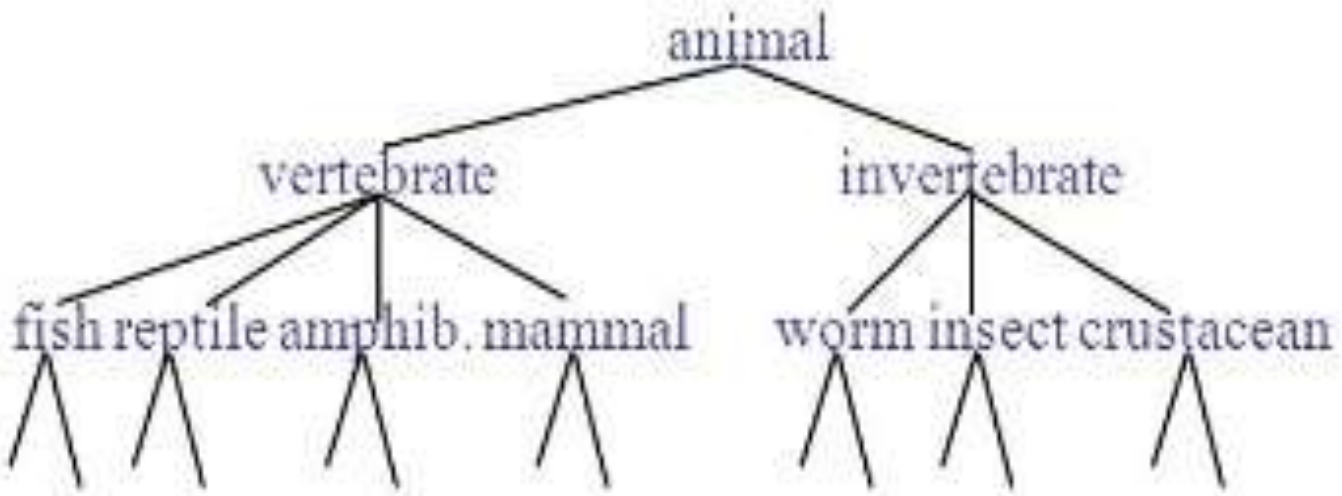
4. *If* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *then STOP; otherwise return to step 2.*

HIERARCHICAL CLUSTERING ALGORITHM

➔ It is a method of cluster analysis which builds a hierarchy of clusters.

➔ It always provides an output which is structured and more informative than the outputs we get from flat clustering methods.

➔ It does not required us to pre-specify the number of clusters.

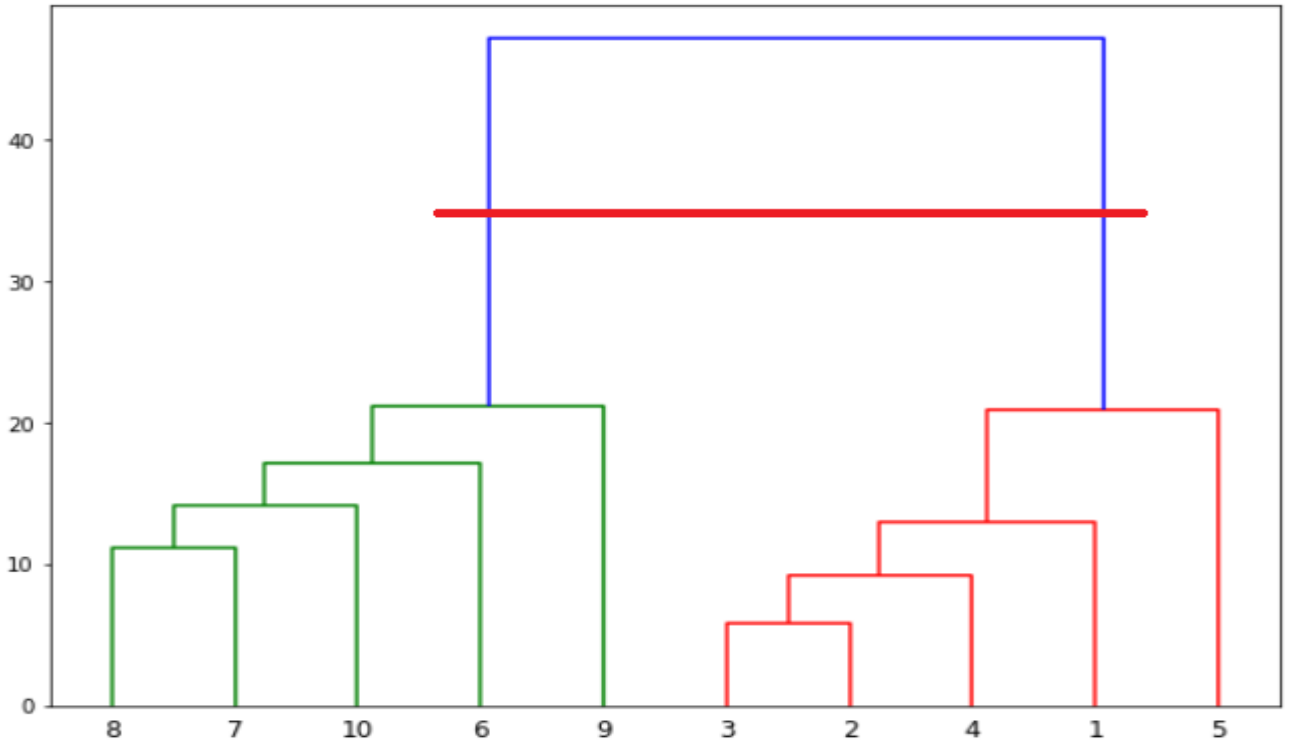


HIERARCHICAL CLUSTERING ALGORITHM

➔ With the help of dendogram, we can obtain the splitting point to determine the number of clusters.

➔ Here, in this example, the splitting point is taken as 2, and it cuts the total ten objects into two clusters.

➔ The first cluster has objects 8, 7, 10, 6, 9. The second cluster has objects 3, 2, 4, 1, 5.



HIERARCHICAL CLUSTERING EXAMPLE:-

➔ Consider six objects having two attributes each. Let the objects be A, B, C, D, E, and F. We can name the attributes as x1 and x2. Let us show the data and plot its graph.

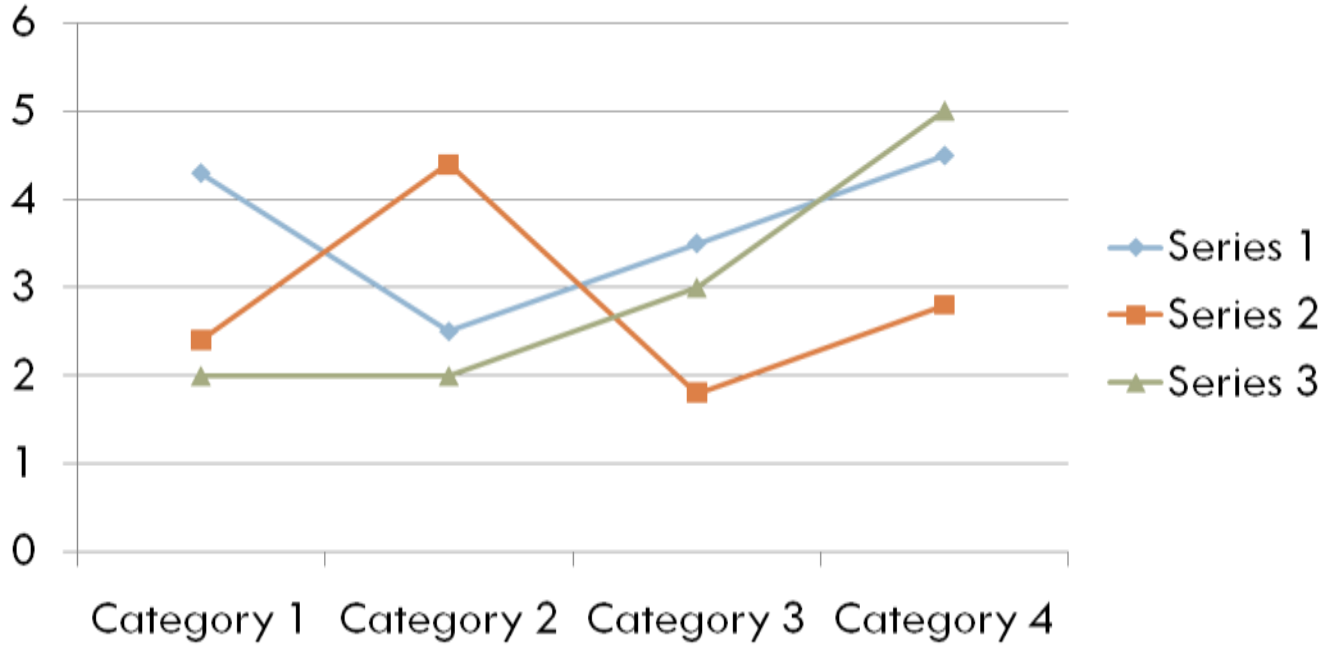
	X1	X2
A	1	1
B	2	2
C	3	4
D	4	5
E	5	5
F	6	2

HIERARCHICAL CLUSTERING EXAMPLE:-

➔ We should use the Euclidean distance formula to compute the distance between the objects.

➔ Euclidean distance is calculated as **the square root of the sum of the squared differences between the two vectors**

➔ The formula used to calculate the distance is:- $$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



HIERARCHICAL CLUSTERING EXAMPLE:-

➔ Thus, we obtain our distance matrix as given below:-

	A	B	C	D	E	F
A	0	1.41	3.6	5	5.65	5.09
B	1.41	0	2.24	3.6	4.24	4
C	3.6	2.24	0	1.41	2.24	3.6
D	5	3.6	1.41	0	1	3.6
E	5.65	4.24	2.24	1	0	3.16
F	5.09	4	3.6	3.6	3.16	0

HIERARCHICAL CLUSTERING EXAMPLE:-

➔ Our aim is to minimise the distance and make a single cluster. So, we combine the cluster having the distance 1 and make a new cluster. Also, we need to update the distance matrix table.

	A	B	C	D, E	F
A	0	1.41	3.6	5	5.09
B	**1.41**	0	2.24	3.6	4
C	3.6	2.24	0	1.41	3.6
D, E	5	3.6	**1.41**	0	3.6
F	5.09	4	3.6	3.6	0

HIERARCHICAL CLUSTERING EXAMPLE:-

➔ Again check for the least value, and combine the clusters. From the updated matrix, the shortest distance is 1.41. So, we need to combine the clusters A, B and (D,E).

	A, B, D, E	C	F
A, B, D, E	0	1.41	3.6
C	**1.41**	0	2.24
F	3.16	3.6	0

➔ Iteration continues and cluster C is joined with (A, B, D, E).

	A, B, D, E, C	F
A, B, D, E, C	0	3.16
F	3.16	0

➔ If we merge the remaining two clusters, we will get a single cluster and the algorithm exits and our computation is finished. We can summarize the process as follows.

1. Start with 6 clusters: A, B, C, D, E and F
2. Merge cluster D and E into (D,E) at distance 1.00
3. Merge cluster A, B and cluster (D,E) into (A,B,D,E) at distance 1.41
4. Merge cluster (A,B,D,E) and C into ((A,B,D,E),C) at distance 1.41
5. Merge cluster (A,B,C,D,E) and F into ((A,B,C,D,E),F) at distance 3.16 to get the single cluster

TWO APPROACHES OF HIERARCHICAL CLUSTERING

1. AGGLOMERATIVE AND DIVISIVE CLUSTERING (ADC)

2. HIERARCHICAL AGGLOMERATIVE CLUSTERING (HAC)

Agglomerative Clustering

- ➔ It is also known as Bottom-Up Approach
- ➔ Here, we start from the bottom and move to the top side. (i.e.,) We begin with individual objects and merge objects which are closely placed. The process is iterated until a single group is formed from it.

Divisive Clustering

- ➔ It is also known as Top-Down Approach.
- ➔ Here, we consider all the objects as a single cluster and then break down recursively until we get groups of single objects.

HIERARCHICAL AGGLOMERATIVE CLUSTERING

- ➔ Here, a similarity factor is used for determining the similarity of two instances.
- ➔ It starts by considering every separate instance as a cluster.
- ➔ It then joins the clustering having similar properties.
- ➔ The process repeats until there is a single cluster.
- ➔ The history of merging forms a binary tree or hierarchy.

The generalized HAC algorithm is as follows:-

1. Start with all instances in their own cluster.
2. Until there is one cluster.
    - Among the current clusters, determine the two clusters, Ci and Cj that are most similar.
    - Replace Ci and Cj with a single cluster Ci U Cj

Weakness of Agglomerative Clustering

- ➔ It does not scale well and for large datasets, time taken will be much more.
- ➔ It is a non-reversible process, i.e., it cannot undo the steps which are previously done.

THANK YOU