Hierarchical Methods:-

In This approach, we Partition the date by grouping it into a Tree of clusters or a hierachy. Hierachical Representation of data objects is useful for data summarization and visualization. for Example, the manger of Human Resources can organize the employees into CEOs, Mangers, and Developers. These groups can be subdivided into subgroups. for Instance, The general group of developers Canbe further divided into subgroups of team lead and Trainee. This form a hierarchical Structure of employees that Canbe used to find the average salary of Mangers and team leads.

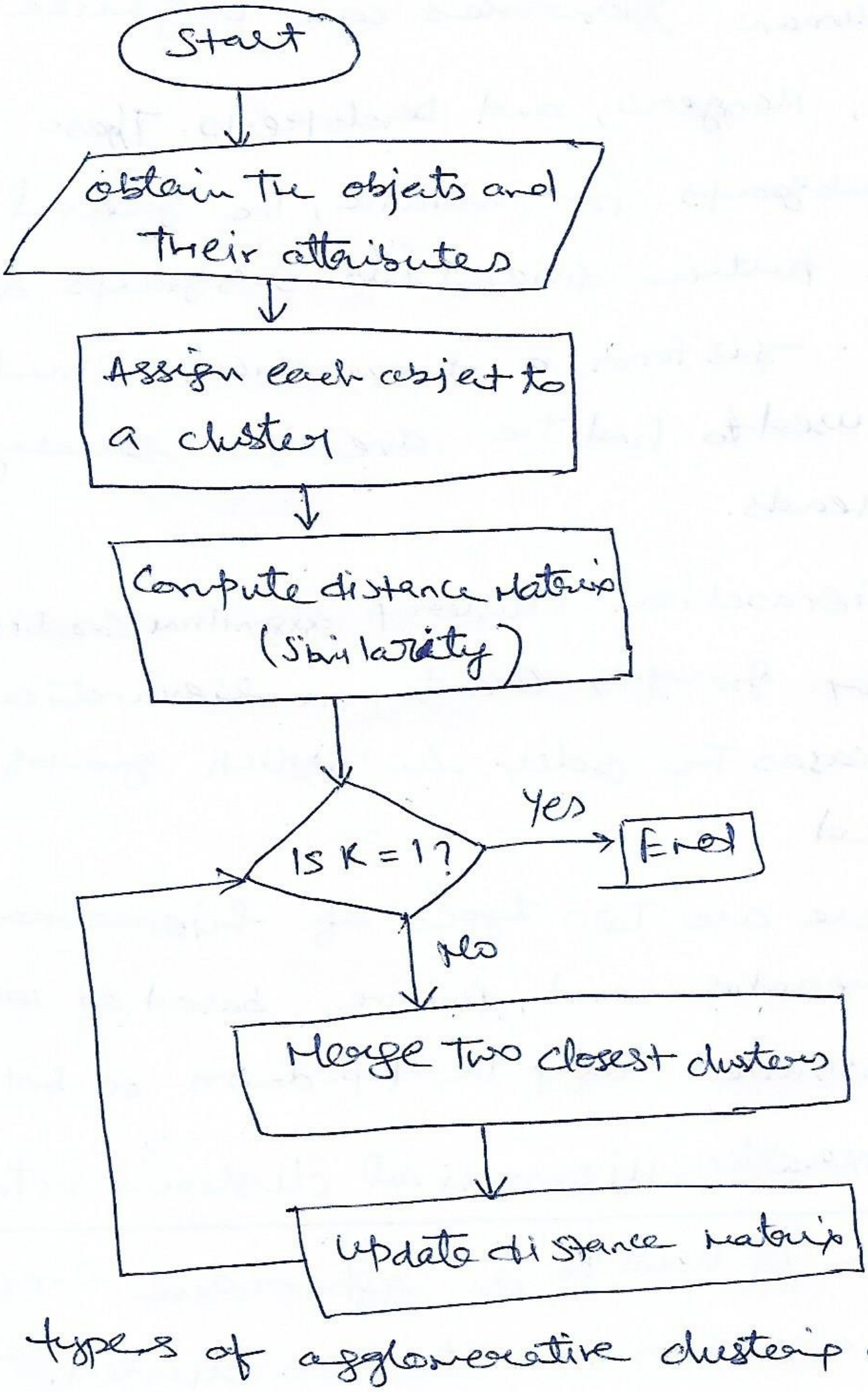
Hierarchical clustering algorithm Combines and divides existing groups, Creating a hierarchical Structure that Showcases The order in which groups are divided or merged

There are two types of hierachical clustering methods, agglomerative and divisive, based on whether the hierachy is constructed using the top-down or bottom-up approaches.

Agglomerative Hierarchical clustering method: → Bottom-up approach is used in an agglomerative hierachical clustering. Each objects forms its own cluster. The process Converges by iteratively merging into larger clusters, until all the objects are in a single cluster or a termination condition is reached. The single cluster becomes The hierachy root. During The Merging operation, Two Similar clusters are identified using distance measures and Combined to form a larger cluster contains at least one object.

An agglomerative method requires at the most n iterations.

Let K be the number of clusters. The process is illustrated using the flow chart in fig



There are three types of agglomerative clustering approaches based on the distance calculation.

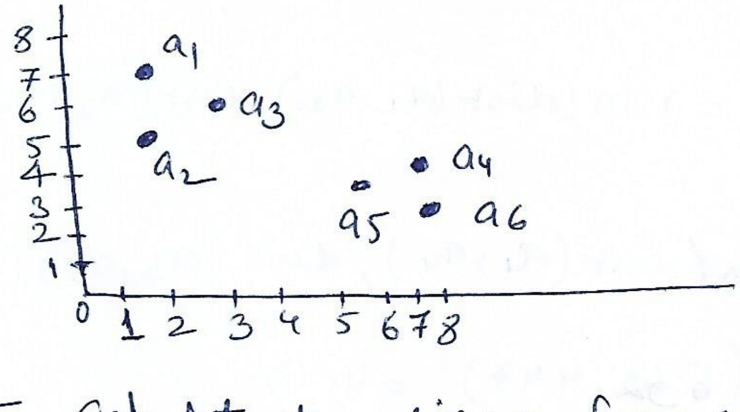
① Single link technique:- where the proximity of two clusters is identified using the minimum of the distance b/w the points belonging to two different clusters.

2. Complete link technique:- where the proximity of two clusters is identified using the maximum of the distance b/w the points belonging to two different clusters.

3. Average link technique:- where the proximity of two cluster is identified using the average distance b/w the points belonging to two different clusters.

Example:- Single link technique: Illustrate single link technique for clustering using the following dataset (DB) containing the data points {a1 (2,7), a2 (2,5), a3 (3,6), a4 (8,5), a5 (7,4), a6 (8,3)}. Use Euclidean distance m/s.

Soln:- Step ① Plot the objects in n-dimensional space (where n is the number of attributes). In our case, we have 2 attributes x and y. So, we plot the objects $a_1, a_2 \ldots a_6$ in two dimensional space.



Step2 :- Calculate the distance from each object (point) to all other objects using Euclidean distance measure to construct the following distance matrix.

Step3:- Identify the clusters with the shortest distance in the matrix and merge them together. Re-compute the distance matrix as those two clusters are now in a single cluster.

Here, we have three clusters with the shortest distance as 1.41. we will look at this, one by one. first we consider cluster a_1, a_3. So, we merge those clusters and recompute the distance Matrix.

	a_1	a_2	a_3	a_4	a_5	a_6
a_1	0					
a_2	2	0				
a_3	1.41	1.42	0			
a_4	6.32	6	5.1	0		
a_5	5.83	5.09	4.47	1.41	0	
a_6	7.21	6.32	5.83	2	1.41	0

with the single link method the proximity of two clusters is defined as the Minimum of the distance b/w the two points in the clusters.

Therefore the distance b/w (a_1, a_3) and a_2; (a_1, a_3) and a_4; (a_1, a_3) and a_5; (a_1, a_3) and a_6 will be calculated as follows:

$$Dist((a_1, a_3), a_2) = Min(dist(a_1, a_2), dist(a_3, a_2)) = Min(2, 1.41)$$
$$= 1.41$$

$$Dist((a_1, a_3), a_4) = Min(dist(a_1, a_4), dist(a_3, a_5))$$
$$= Min(\cancel{6.32}, 4.47) = 4.47$$

$$Dist((a_1, a_3), a_5) = Min(dist(a_1, a_5), dist(a_3, a_5))$$
$$= Min(5.83, 4.47) = 4.47$$

$$Dist((a_1, a_3), a_6) = Min(dist(a_1, a_6), dist(a_3, a_6))$$
$$= Min(7.21, 5.83)$$
$$= 5.83$$

The following is the resultant distance matrix

	a1,a3	a2	a4	a5	a6
a1,a3	0				
a2	1.41	0			
a4	~~4~~ 4.47	6	0		
a5	4.47	5.09	1.41	0	
a6	5.83	6.32	2	1.41	0

<u>Step 4</u> :- Repeat step 3 until all clusters are merged

Looking at the last distance matrix, we see that clusters (a_4, a_5), (a_5, a_6) and $((a_1, a_3), a_2)$ have the smallest distance as 1.41.

First we consider the cluster (a_4, a_5) and recompute the distance matrix. The distance b/w (a_4, a_5) and (a_1, a_3) would be calculated as follows:

$$dist(((a_4, a_5), (a_1, a_3)) = Min(d(a_4, a_1), dist(a_5, a_1),$$

$$di(a_4, a_3), di(a_5, a_3))$$

$$\Rightarrow Min(6.32, 5.83, 5.1, 4.47) \Rightarrow 4.47$$

The resultant distance matrix is reconstructed as follows:

	(a_1, a_3)	a_2	(a_4, a_5)	a_6
(a_1, a_3)	0			
a_2	1.41	0		
a_4, a_5	4.47	5.09	0	
a_6	5.83	6.32	1.41	0

Since, we have more clusters to merge, we continue to repeat step 3. Looking at the last distance matrix, we see that $((a_1, a_3), a_2)$, $((a_4, a_5), a_6)$ have the smallest distance as 1.41. Now we consider cluster $(a_1, a_3), a_2$ and calculate the distance b/w $((a_1, a_3), a_2)$ and other clusters. The resultant matrix is reconstructed as shown on right

	$(a_1, a_3), a_2$	a_4, a_5	a_6
$(a_1, a_3), a_2$	0		
(a_4, a_5)	4.47	0	
a_6	5.83	1.41	0

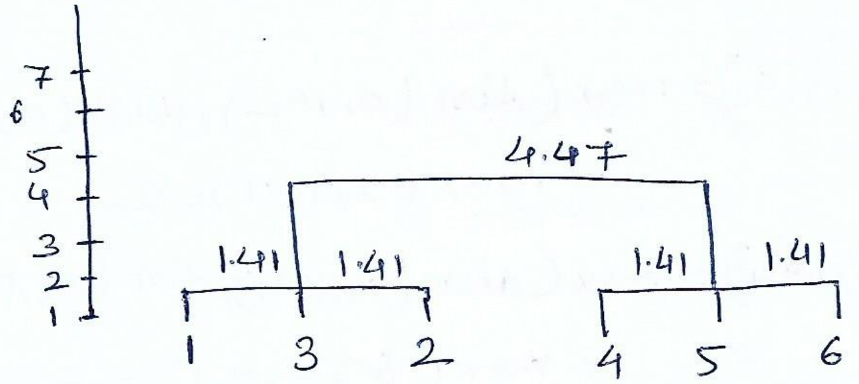
↓

In this step, we see the clusters (a_4, a_5) and a_6 have the smallest distance of 1.41; so we merge those two into a single cluster and re-compute the distance matrix.

	$(a_1, a_3), a_2$	$(a_4, a_5), a_6$
$(a_1, a_3), a_2$	0	
$(a_4, a_5), a_6$	4.47	0

Since we have one more cluster to merge, we continue to repeat step 3. Looking at the distance matrix, we see that the clusters $((a_4, a_5), a_6)$ and $((a_1, a_3), a_2)$ have the smallest distance 4.47. Hence they are merged and process stops with the root node of the cluster.

A Tree structure called Dendrogram is commonly used to represent the process of hierarchical clustering. It provides visualization of how objects are grouped together in an agglomerative method or partitioned in a divisive method at each step. The dendrogram for the single link method is shown in fig



Ex:- Complete link technique : for the same examples we also use here Euclidean distance measure

Sol^n:- step 1:- plot the objects in n-dimensional space (where n is the number of attributes). In our case, we have 2 attributes x and y. so, we plot the objects $a_1, a_2 \cdots a_6$ in 2-D space

Step 2:- calculate the distance from each object to all other points, using Euclidean distance M/S, and place the numbers in a distance ~~\#\#~~ matrix.

	a_1	a_2	a_3	a_4	a_5	a_6
a_1	0					
a_2	2	0				
a_3	1.41	1.42				
a_4	6.32	6	5.1	0		
a_5	5.83	5.09	4.47	1.41	0	
a_6	7.21	6.32	5.83	2	1.41	0

Step 3:- Identify the clusters with the shortest distance in the matrix, and merge them together. Recompute the distance matrix, as those two clusters are now in a single clusters.

Three set of clusters have The shortest distance of 1.41. we will consider This, one by one. we merge The cluster a_1, a_3 and Recompute The distance matrix. with The complete link Method, The proximity of two clusters is defined as The maximum of The distance b/w The two Points in The clusters. Therefore, The distance b/w (a_1, a_3) and a_2, (a_1, a_3) and a_4, (a_1, a_3) and a_5, (a_1, a_3) and a_6 will be calculated as follows:

$$\text{Dist}((a_1, a_3), a_2) = \max(\text{dist}(a_1, a_2), \text{dist}(a_3, a_2))$$
$$= \max(2, 1.41) = 2$$

$$\text{Dist}((a_1, a_3), a_4) = \max(\text{dist}(a_1, a_4), \text{dist}(a_3, a_4))$$
$$= \max(6.32, 5.1) = 6.32$$

$$\text{Dist}((a_1, a_3), a_5) = \max(\text{dist}(a_1, a_5), \text{dist}(a_3, a_5))$$
$$= \max(5.83, 4.47) = 5.83$$

$$\text{Dist}((a_1, a_3), a_6) = \max(\text{dist}(a_1, a_6), \text{dist}(a_3, a_6))$$
$$= \max(7.21, 5.83) = 7.21$$

The distance Matrix is Recomputed as given here

	a_1, a_3	a_2	a_4	a_5	a_6
a_1, a_3	0				
a_2	2	0			
a_4	6.32	6	0		
a_5	5.83	5.09	1.41	0	
a_6	7.21	6.32	2	1.41	0

step 4:- Repeat step 3 until all clusters are merged. so, looking at the last distance matrix, we see that cluster (a_4, a_5), (a_5, a_6) has The smallest distance of 1.41.

first, we consider the cluster (a_4, a_5) and re-compute the distance ~~measure~~ matrix. The distance b/w (a_4, a_5) and (a_1, a_3) would be calculated as follows:

$$dist((a_4, a_5), (a_1, a_3)) = max\ dist(a_4, a_1), dist(a_5, a_1),$$

$$dist(a_4, a_3), dist(a_5, a_3))$$

$$= max(6.32, 5.83, 5.1, 4.47) = 6.32$$

The resultant distance matrix is reconstructed on here.

	a_1, a_3	a_2	a_4, a_5	a_6
a_1, a_3	0			
a_2	1.41	0		
a_4, a_5	6.32	6	0	
a_6	5.83	6.32	2	0

Since, we have more clusters to merge, we continue to repeat step 3. in the last matrix, we see that (a_1, a_3) and a_2 as well as (a_4, a_5) and a_6 have the smallest distance of 1.41 and 2, respectively; Now we consider $((a_1, a_3), a_2)$ and recalculate the distance b/w $(a_1, a_3), a_2$ and other clusters. The resultant distance matrix is reconstructed on right.

After that we have merged cluster $(a_4, a_5), a_6$ b/c these have minimum distance.

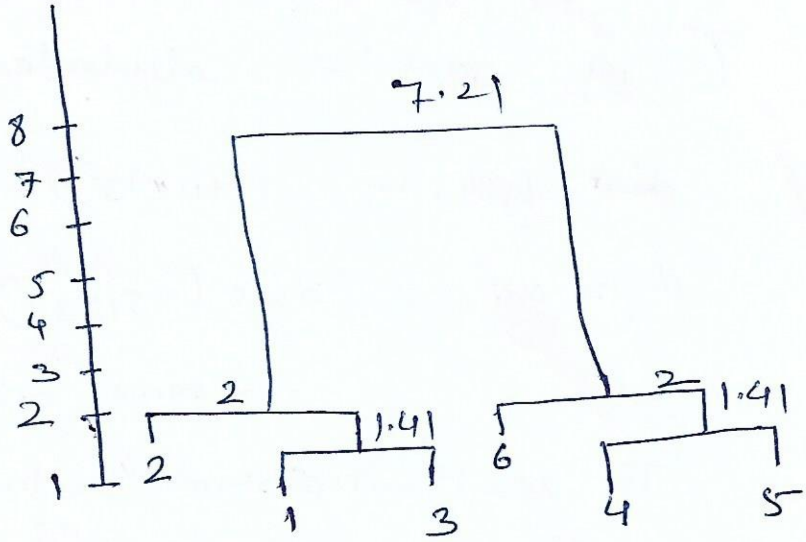
	$(a_1, a_3), a_2$	a_4, a_5	a_6
$(a_1, a_3), a_2$	0		
a_4, a_5	6.32	0	
a_6	7.21	2	0

and now we have re compute the distance from $(a_1, a_3) a_2$ to $(a_4, a_5), a_6$;

	$(a_1, a_3), a_2$	a_4, a_5, a_6
$(a_1, a_3), a_2$	0	
$(a_4, a_5), a_6$	7.21	0

Now we have to merge them in a single cluster. There is no need to re-compute the distance matrix, as there are no more clusters to merge.

The dendrogram for the complete link Method is shown in following figure



Divisive Hierarchical clustering Method :→ It start by placing all objects into one cluster at the root of the hierarchy, i.e. it follows a top-down approach. It then recursively partitions the clusters into smaller sub-clusters starting from the root.

We can stop the partitioning process when each cluster at the lowest level contains only one object, or the objects with in a cluster are sufficiently similar to each other. The divisive method is the exact opposite of the agglomerative method. In the divisive approach, the hierarchy is constructed using the split operation in contrast to that of merge operation in the agglomerative approach. Thus this approach is top-down strategy.

The process of divisive clustering is given as follows:

1. Start by placing all objects in one cluster
2. Repeat until all the clusters have a single object.
   a) Select a cluster with max inter-cluster distance to split.
   b) Replace the selected cluster with the sub-clusters.

In divisive Method, clusters can be split using minimum-spanning Tree, bisecting K-Means and Min-max cut techniques.

Splitting using Minimum Spanning Tree :→ This Process is as follows:

(1) Construct a Proximity graph to measure the distance b/w objects in The given set (D)

(2) Compute the minimum spanning Tree for The Proximity graph

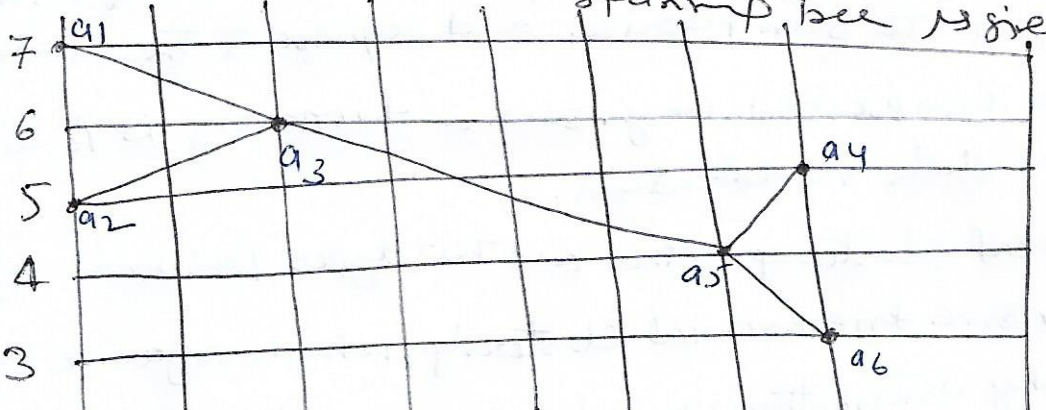
(3) Repeat

(4) Create a new cluster by breaking the line corresponding to The largest distance that represents objects that are the least similar.

(5) Continue until only singleton clusters remain

Ex:→ For The same example

Sol^n:- Construct the minimum spanning Tree for The given data Points based on The Proximity (distance) measure b/w The data Points. The Minimum spanning Tree is given here.



Break The largest line (a_3, a_5) to create two clusters (a_1, a_2, a_3) and (a_4, a_5, a_6)