Hierarchical methods

* The hierarchical agglomerative clustering methods are most commonly used. The basic steps followed in this type of Hierarchical methods or general algorithm are.

1 * find the two closest objects + merge them into cluster

2 * find + merge the next two closest points, where a point is either an individual objects or a cluster of objects.

3 * If more than one cluster remain, again return to step 2.

+ Agglomerative Algorithm. :- * It follows bottom-up strategy

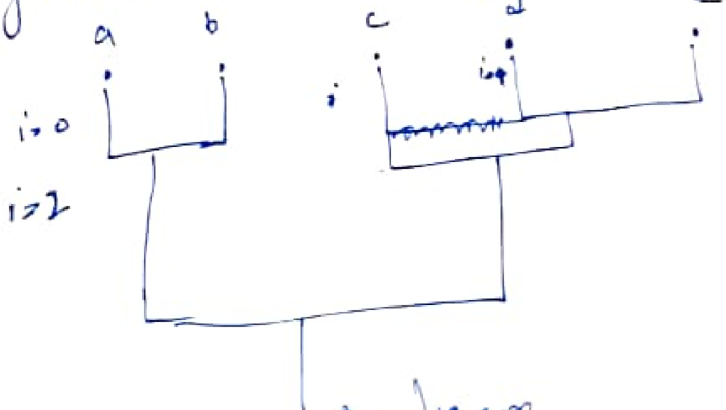
* According to some similarity measure (ED), the merging is done by choosing the closest clusters first.

* A dendogram, which is a tree like structure, which is used to represent hierarchical clustering.

* Individual objects are represented by leaf nodes + clusters are represented by root nodes.

This representation is known as dendogram as shown in below fig.

wrong fig



Dendogram

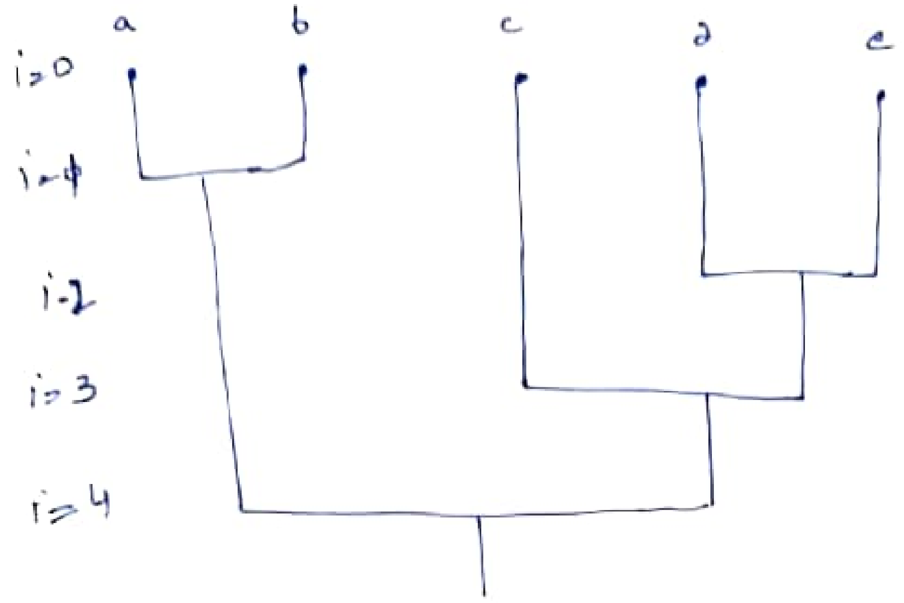


Fig 1. Dendogram

Distance measure / similarity measure.

This is called nearest-neighbour clustering algorithm.

Min dist : $dist_{min}(C_i, C_j) = \min_{p\in C_i, p'\in C_j}\{|p-p'|\}$

Max dist : $dist_{max}(C_i, C_j) = \max_{\underbrace{p\in C_i, p'\in C_j}_{\text{two clusters}}}\{|\overbrace{p-p'}^{\text{Two objects or Two points}}|\}$

Mean dist. : $dist_{mean}(C_i, C_j) = |m_i - m_j|$ $\{x\}$

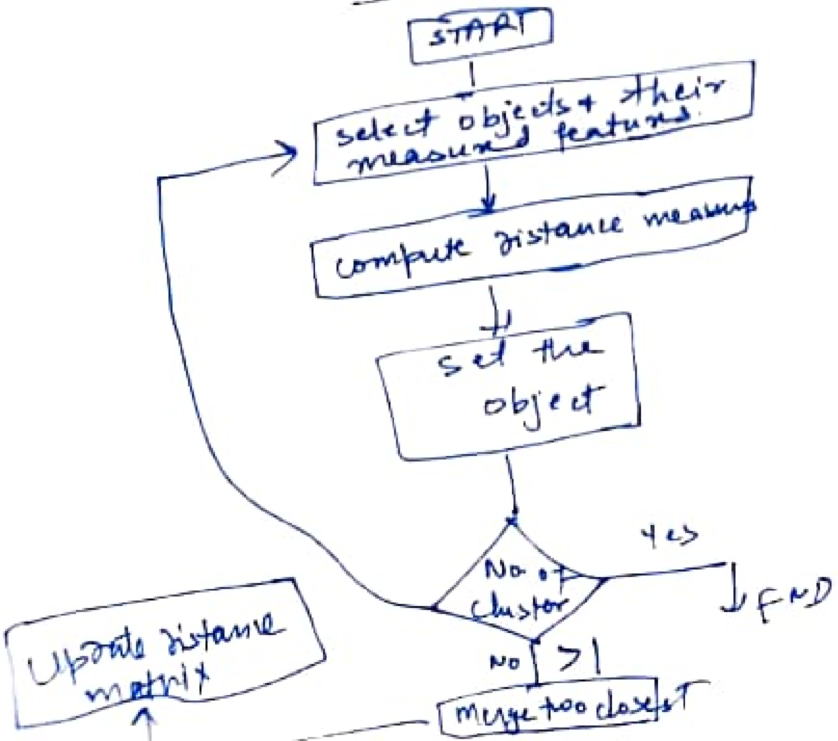
Avg dist : $dist_{Avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p\in C_i, p'\in C_j} |p-p'|$

* when an algorithm uses the min. distance $d_{min}(C_i, C_j)$ to measure the distance b/w clusters, it is called nearest-neighbor clustering algorithm.

* If the clustering process is terminated, when the distance b/w the nearest clusters exceeds user-defined threshold, it is called - Single linkage algorithm.

* Agglomerative hierarchical clustering algorithm with min. distance measure is called as minimum spanning tree algorithm

* An algorithm that uses the max. distance $d_{max}(C_i, C_j)$ to measure the distance b/w clusters is called farthest - neighbor clustering algo. If clustering is terminated when the max. distance exceeds a user defined threshold, it is called complete - linkage algorithm.



Agglomerative Alg. :- single link.

Find the clusters using single link technique. Use Euclidean distance as similarity measure & draw the dendogram.

Sample No.	X	Y
P_1	0.40	0.53
P_2	0.22	0.38
P_3	0.35	0.32
P_4	0.26	0.19
P_5	0.08	0.41
P_6	0.45	0.30

Distance matrix

$$d\left[(x,y),(a,b)\right] = \sqrt{(x-a)^2+(y-b)^2}$$

Euclidean distance $d(P_1, P_2) = \sqrt{(0.4-0.22)^2+(0.53-0.38)^2}$

$$= \sqrt{(0.18)^2+(0.15)^2}$$

$$= \sqrt{0.0324+0.0225}$$

$$= 0.23$$

Distance matrix :

$$\begin{array}{c|cccccc} & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ \hline P_1 & 0 & 0.23 & & & & \\ P_2 & 0.23 & 0 & & & & \\ P_3 & 0.22 & & 0 & & & \\ P_4 & 0.37 & & & 0 & & \\ P_5 & 0.34 & & & & 0 & \\ P_6 & 0.24 & & & & & 0 \end{array}$$

IIIly $d(P_1, P_3) = \sqrt{(0.4-0.35)^2+(0.53-0.32)^2}$

$= \sqrt{(0.05)^2+(0.21)^2} = \sqrt{0.0025+0.0441}$

$= 0.22$

$d(P_1, P_4) = \sqrt{(0.4-0.26)^2+(0.53-0.19)^2}$

$= \sqrt{(0.14)^2+(0.34)^2} = \sqrt{0.0196+0.115}$

$= 0.37.$

$d(P_1, P_5) = \sqrt{(0.4-0.08)^2+(0.53-0.41)^2}$

$= \sqrt{(0.32)^2+(0.12)^2}$

$= \sqrt{0.1024+0.0144} = 0.34.$

$d(P_1, P_6) = \sqrt{(0.4-0.45)^2+(0.53-0.30)^2}$

$= \sqrt{(0.05)^2+(0.23)^2} = 0.24.$

start For P_2.