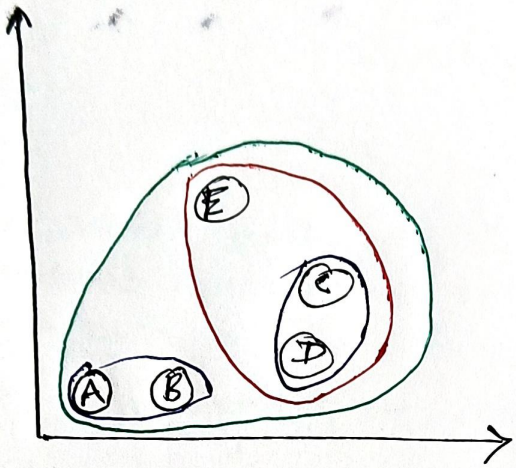
30-NOV-2022

Hierarchical Clustering

- Distance based approach

(A) (B) (C) (D) (E)



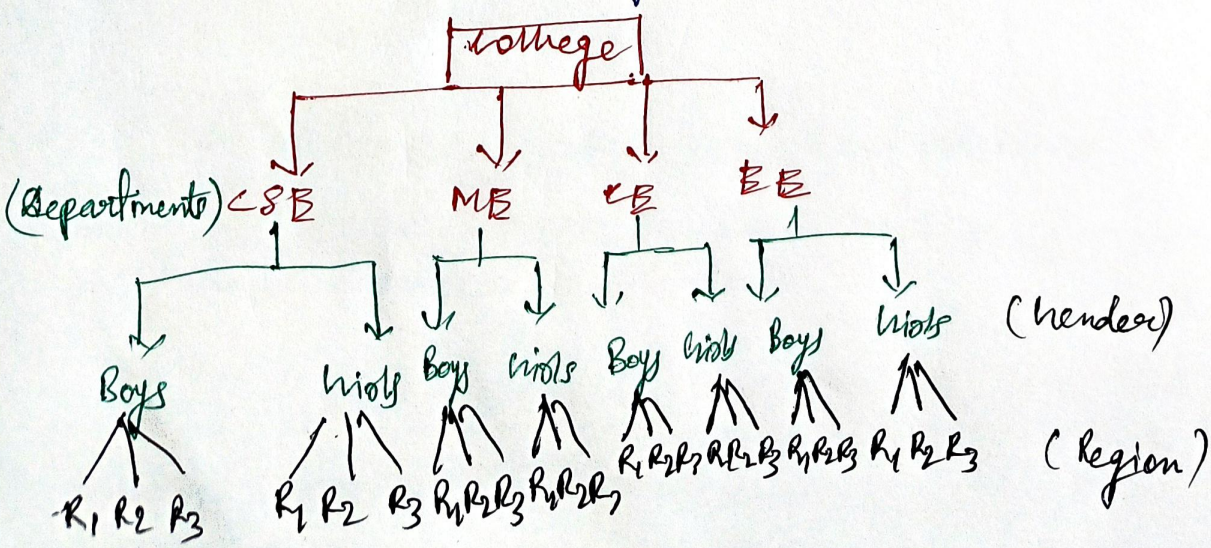
> A and B are near to each other, C and D are near to each other, E is near to CD cluster and can be converted into CDE cluster, and CDE is close AB cluster, so that final cluster will be CDEAB.

→ Mining each and every cluster.

Example.

College

> Consider college as a single cluster



→ Here, no restriction for no number of cluster. We can create as many as we want, whereas in K-Means we have to give number of cluster needed.

Q How can we represent Hierarchical clustering?

Ans Using "Dendogram" we can represent. It is a representation of Hierarchical clustering.

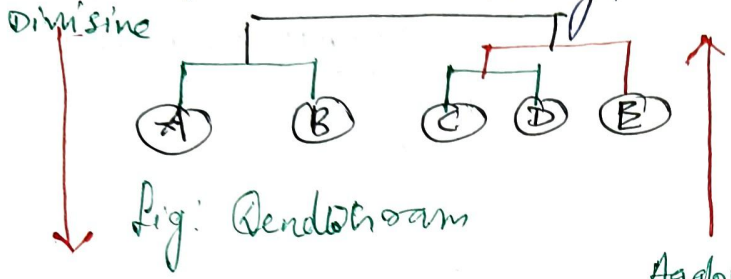


Fig: Dendrogram

AB and CD are close to each other as per previous graph and CD is close to E and CDE is close to AB.

Agglomerative
- Aggregating the data points
- Each point is cluster

Divisive
- Dividing the data points into different level.

→ In hierarchical clustering, we are considering all variations.

→ With the help of Point matrix, we can calculate distance.

Point Matrix

- 5 point (P_1) (P_2) (P_3) (P_4) (P_5)

→ Find distance b/w the points and create cluster

1st

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	9	0			
P_3	3	7	0		
P_4	6	5	9	0	
P_5	11	10	2		0

→ All the distances are assumed.

minimum distance = P_3 & P_5 = 2

2nd

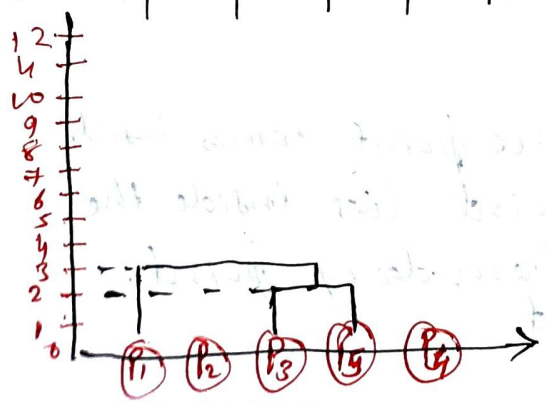
	P_1	P_2	$[P_3\ P_5]$	P_4
P_1	0			
P_2	9	0		
$[P_3\ P_5]$	3	7	0	
P_4	6	5	8	0

minimum distance = 3

$d(P_1, [P_3, P_5]) = [d((P_1, P_3), (P_1, P_5))]_{min} = (3, 11)_{min} = 3$

→ like that we can calculate for others

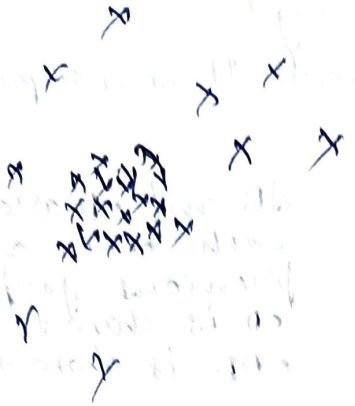
→ Similarly we need calculate for others points.



DBSCAN

- Density Based Spatial Clustering with application with noise.
- Density based approach

Dataset



DBSCAN can be used.

(Communication)

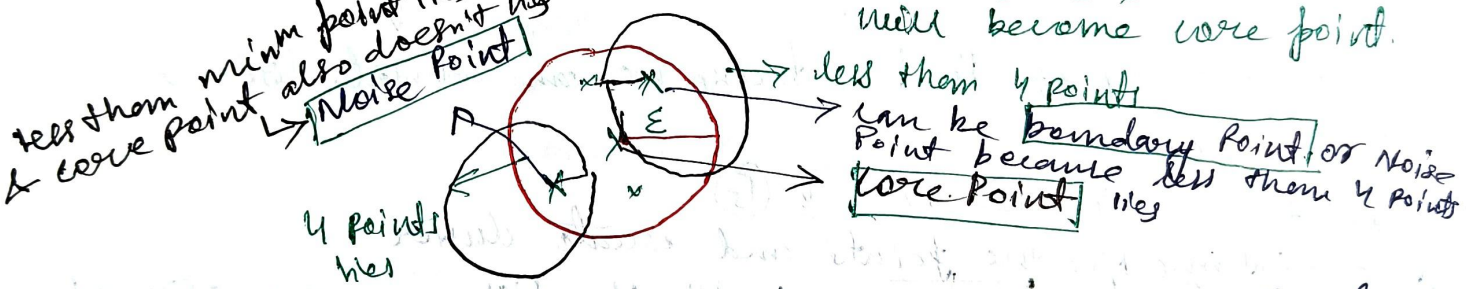
1. ~~absolute~~ epsilon (ε) distance
2. Core Point
3. Bordere Point
4. Noise Point
5. Minimum Point

→ Based on these 5 point, we have to make cluster.

Steps

Hyper Parameter { epsilon distance = ε
minimum point = 4

→ Make a circle with radius as ε and if 4 points ~~should~~ lies inside it then it will become core point.



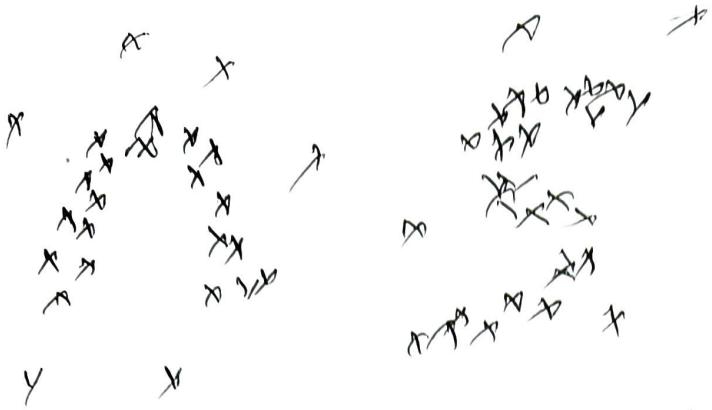
→ Pick one point and draw a circle with 'ε' as radius and if all the 'minimum point' lies inside it then that point become core point.

Noise Point

- It is a outlier, Not a core point nor a boundary point.

Boundary Point

- If we draw a circle, and core point comes inside and less than minimum point lies inside the circle then it will become boundary point.
- It is neighbour of the core point.



⇒ In this type of data K-Means and Hierarchical also clustering doesn't work, so we called here DBSCAN here.

> Epsilon is a unit distance.