Clustering

- Method of identifying similar groups of data in a dataset.

- Independent variables having similar features are identified through clustering.

- Types of clustering

1) Hard clustering :- Each data point either belongs to a cluster completely or not.

2) Soft clustering :- Each data point can belong to more than one cluster

- Types of clustering algo

1) Centroid model : The similarity is determined by closeness of data point to the centroid of clusters. eg - Kmeans algo.

2) Connectivity model : Data points closer in data space exhibit more similarity to each other than data points away eg - Hierarchial clustering

3) Distribution model : How probable is that data points in the cluster belong to same distribution

4) Density model : It isolates various diff intensity regions and form clusters on basis of it

- Types of clustering

1) Partitioning

→ Cluster data set into set of groups
→ K-value - no of clusters to be formed
→ centroid based method.
→ So, if k=3, 3 centroids are formed and euclidean distance of a point from every centroid is seen. Whichever is min, it is assigned to that
→ Pre specify no of clusters

2) Hierarchial

→ Set of nesting clustering (A cluster is clustered into another cluster)
→ Organized by representation tree called dendogram

3) Exclusive (non overlapping)

→ assign each data point to a single group.

4) Non exclusive (overlapping)

→ we can have data points in more than one cluster

5) Fuzzy clustering

→ Object have a membership weight between 0 & 1

6) Probalistic clustering
→ We determine the prob that a data point belong to specific cluster

- Types of Clusters

1) Well separated

→ The data points are clustered on the basis of distance.

→ A set of objects in which each object is closer to the object in same cluster than object in diff cluster

2) Prototype based

→ A set of object such that an object in a cluster is closer to prototype that defines the cluster

→ Continuous attributes - prototype - centroid
Categorical attributes - " - most representative point (medoid)

3) Graph based

4) Density based

5) Shared property
A cluster in which set of objects share some property.

→ K-means Algo-

Steps -

1) Choose K number of random data points as initial centroids.

2) Repeat till cluster center stabilize:-

(i) Allocate each point in dataset to nearest of k^m centroid

(ii) Compute centroid for cluster using all points in cluster

Advantages :-

(i) Simple & easy to understand

(ii) Efficient algo.

Disadvantages

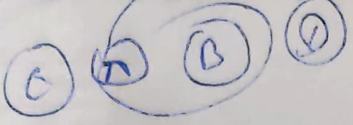
(i) We need to pre specify value of K.

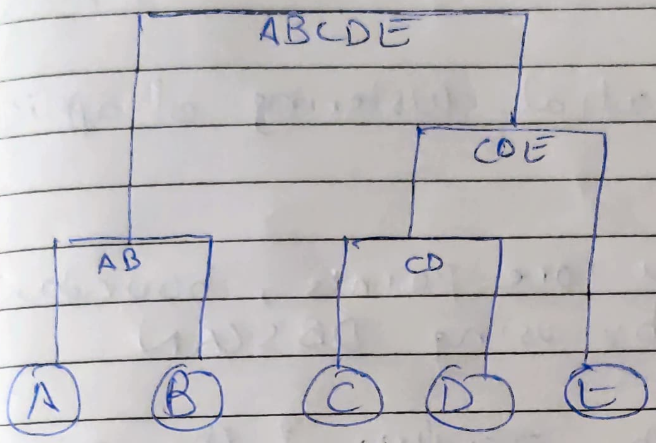
(ii) Process of finding clusters may not converge.

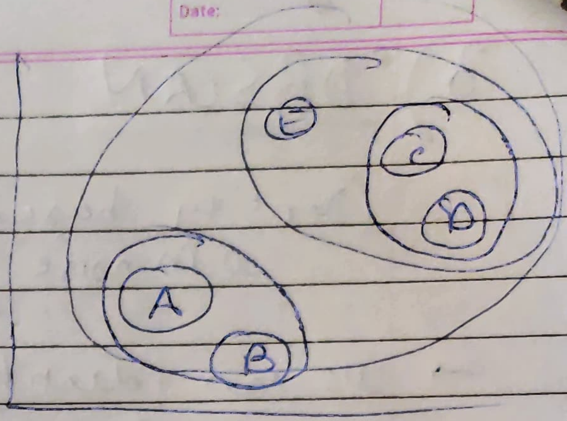
→ Hierarchial clustern

1) Agglomerative : (Bottom up approach)

→ First, we form the smaller clusters and then form the larger clusters.

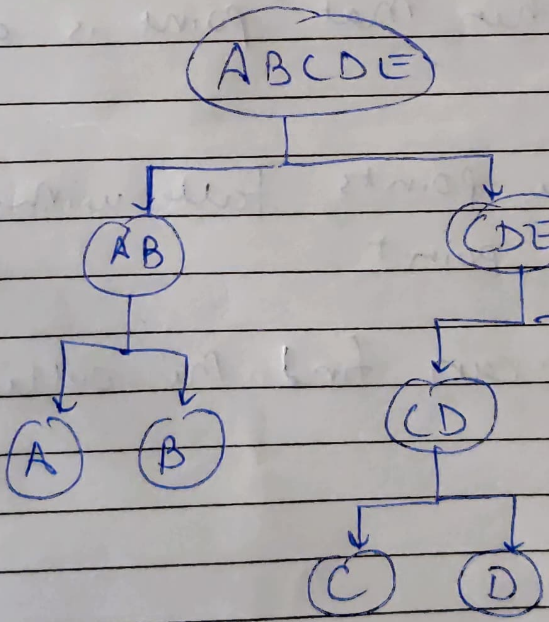






2) Divisive

→ Start wim one big cluster containing all data points & then we divide into smaller clusters



3) DBSCAN

(Density based spatial clustering of application with noise)

→ We can identify the core points, boundary points & noise by using DBSCAN.

→ We are given the radius & the minpoints

→ We draw a circle by keeping every point as centre. If no of min points fall within that circle, that point is a core point.

→ If any one core point fall within that circle, then that point is a boundary point

→ If none of the points fall within the circle, it is a noise point

→ So, this algo can find the outliers also.