Unit-4

Cluster Analysis

- Clustering is the process of grouping the similar data points in the large dataset.
- Grouping is performed based on the similar characteristics.
- The final group of similar data points formed at the end is known as a cluster.
- The goal of clustering is to maximize the similarity within clusters and minimize the similarity between clusters.
- These clusters are then labeled and fed to machine learning models as training data.
- Hence, the process of analyzing a dataset to form clusters, is known as cluster analysis.
- And the algorithms used for cluster analysis are usually unsupervised algorithms.(non labeled data is provided)

Data Structures used in Cluster Analysis

- The working of clustering algorithms is simple:
  - + It takes a matrix which represents the dissimilarities b/w every pair of data points (known as dissimilarity matrix) and a data matrix which represents the data points with its features, as the inputs, and groups the pairs with less dissimilarity values to form a cluster.
  - + The dissimilarity matrix is built from a data matrix using a proximity measure(distance metric).
  - + Hence, Data Matrix and Dissimilarity Matrix are considered as the data structures used for clustering.

1) Data Matrix:

- It is a two dimensional matrix, whose rows represent the data points, and the columns represent a variable/feature/attribute of a data point.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$$

2) Dissimilarity Matrix:

+ It is a square matrix in which each element d(i, j) represents the dissimilarity/distance b/w data points (i) and (j) from the data matrix.
+ The diagonal elements d(i, i) or d(j, j) are 0, because it represents the same data point.

+ A dissimilarity value d(i, j) is calculated using a proximity measure(distance metric), such as, Euclidean Distance, Manhattan Distance, etc.
+ The Dissimilarity matrix formed at the end, and the data matrix, are fed to the clustering algorithm as input, which then forms clusters with the pairs having less dissimilarity values.

- The proximity measure(metric/formula to calculate the dissimilarity b/w two data points), is not the same for every data point. It depends on the type of attributes of a data point.
- Based on the type of value that can be stored in a column of a data matrix, the attribute types are classified into (4), they are:

1. Nominal Attributes
2. Binary Attributes
3. Ordinal Attributes
4. Numeric Attributes

1) Nominal Attributes:

- The attributes whose values represent names/categories/states, without any specific order/ranking, are nominal attributes.

Ex:

Attribute	Values
Colors	Black, Brown, Yellow
Birds	Peacock, Pigeon, Crow

- Proximity measure for nominal attributes is:

$$d(i, j) = \frac{p - m}{p},$$

2) Binary Attributes:

- The Attributes which have only two possible values/states, are binary attributes.
- These are of two types:
- Symmetric Attribute: When those two values are equally important.

Ex:

Attribute	Values
Gender	Male, Female

- Asymmetric Attribute: When those two values are not equally important(i.e., opposite values).

Ex:

Attribute	Values
Cancer	Yes, No
Result	pass, fail
Is employee	true, false

- Proximity Measure for Symmetric attributes is:

$$d(i, j) = \frac{r + s}{q + r + s + t}.$$

- Proximity Measure for Asymmetric attributes is:

$$d(i, j) = \frac{r + s}{q + r + s}.$$

3) Ordinal Attributes:

- Attributes whose values have a meaningful ranking/order/sequence, but the difference/intervals b/w them are not consistent/meaningful.

Ex:

Attributes	Values
Grade	A, B, C, D
Qualification	high school, inter, degree
Ratings	bad, good, satisfactory, best
Temp levels	low, medium, high

- Proximity measure, for ordinal attributes is:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}.$$

4) Numeric Attributes:

- The attributes which can hold numeric data (numbers), are numeric attributes.
- The values can be discrete or continuous.

Ex:

Attributes	Values
Brightness	25, 50, 30
Saturation	5.4, 26.7, 8.1

- Since the values are of numeric type, the dissimilarity value is termed as “distance”.
- Proximity measure for Numeric attributes are many, two of them are:

=>Euclidean Distance:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

=>Manhattan Distance:

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

Clustering Methods:

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method

1) Partitioning Method:

- The clustering method which forms clusters by initially assuming the partitioning of a given dataset into a specified no.of clusters and then iteratively assigning the data points to the nearest cluster (cluster from which the distance is minimum).

I.e, Initially, hypothetical random partitions/clusters are formed, and then the distances b/w every data point and every assumed cluster is calculated. Then, the data point which has the minimum distance to an assumed cluster is the new and actual assigned cluster for that data point. This goes on iteratively until a certain condition is met.

- For every iteration, the distance b/w data points and the cluster which the data point is supposed to be, keeps on decreasing.
- The major benefit of these methods is that they form non-overlapping clusters.
- There are many Algorithms for partitioning methods. Most used one is K-Means Algorithm.

K-Means Clustering Algorithm:

- A Clustering algorithm which follows the Partitioning method with the Euclidean Distance to create non-overlapping clusters from the given dataset.
- This Algorithm works by considering the centroid points of clusters.
- For every iteration the centroid keeps on changing and refining. And the algorithm terminates when centroids of previous iteration and current iteration are equal.

I.e, iteratively assigns the clusters to every datapoint, based on the distance from the centroids of clusters.

- Steps:

- Decide the no.of clusters to be created (k).
- Initialize any of (k) no.of data points as the centroids of (k) clusters.
- Assign each data point to the cluster with the nearest centroid. (use Euclidean Distance formula)
- Recalculate the centroids based on the mean of the points in each cluster.
- Repeat the steps (3-4) until the convergence of centroids.(centroids no longer change in iteration / previous iteration centroids and current iteration centroids are equal.)

- Hence, for every iteration, the centroid keeps on changing by averaging, and the distance b/w every data point and its suitable cluster keeps on decreasing.

Ex: (notes)

[(2) K - mediods]

2) Hierarchical method:

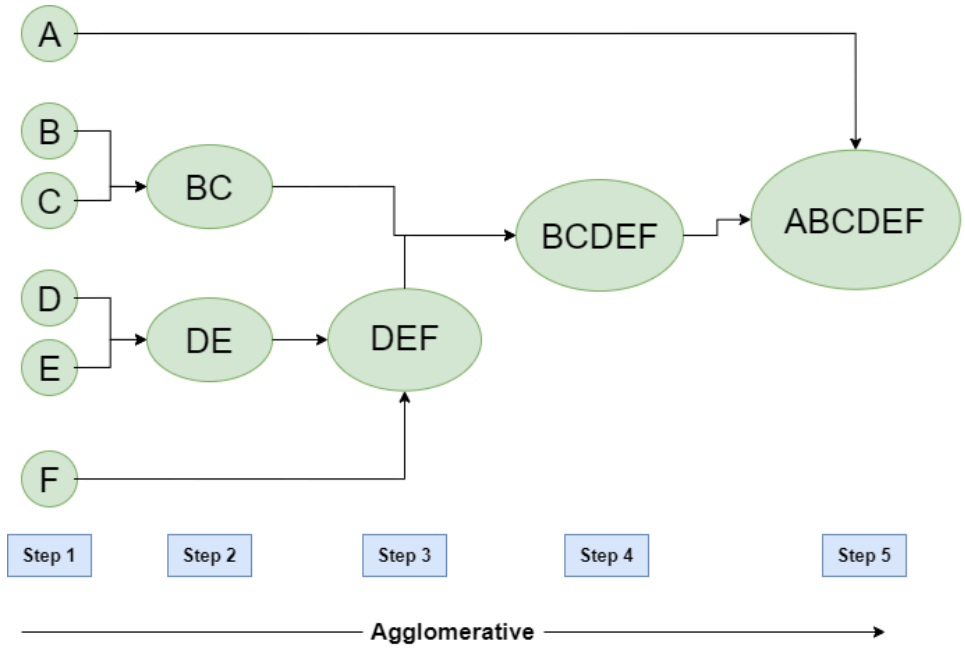
This method of clustering builds a hierarchy of clusters.

Clustering is performed by merging/dividing clusters and forming a tree-like structure, known as a dendrogram.

There are two approaches:

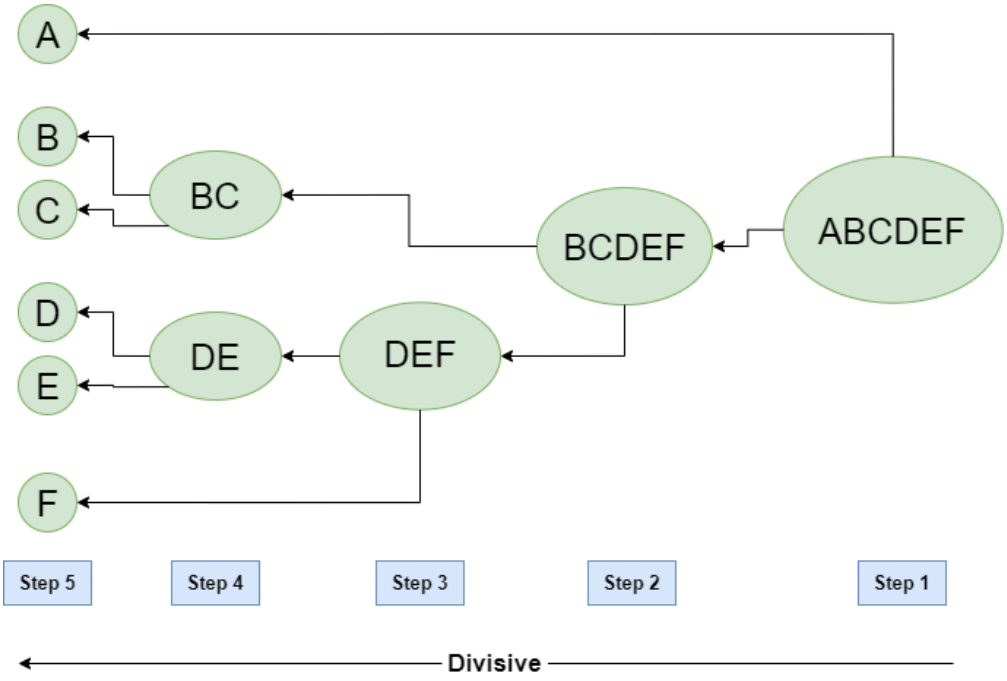
(a) Agglomerative (bottom-up)(merging):

- It starts with each data point as a single cluster and then merges the closest cluster/cluster with similar characteristics iteratively until only one cluster remains.
- Hence, this forms a single cluster for a given dataset.
- The root of the tree is the final cluster.



(b) Divisive (top-down)(splitting):

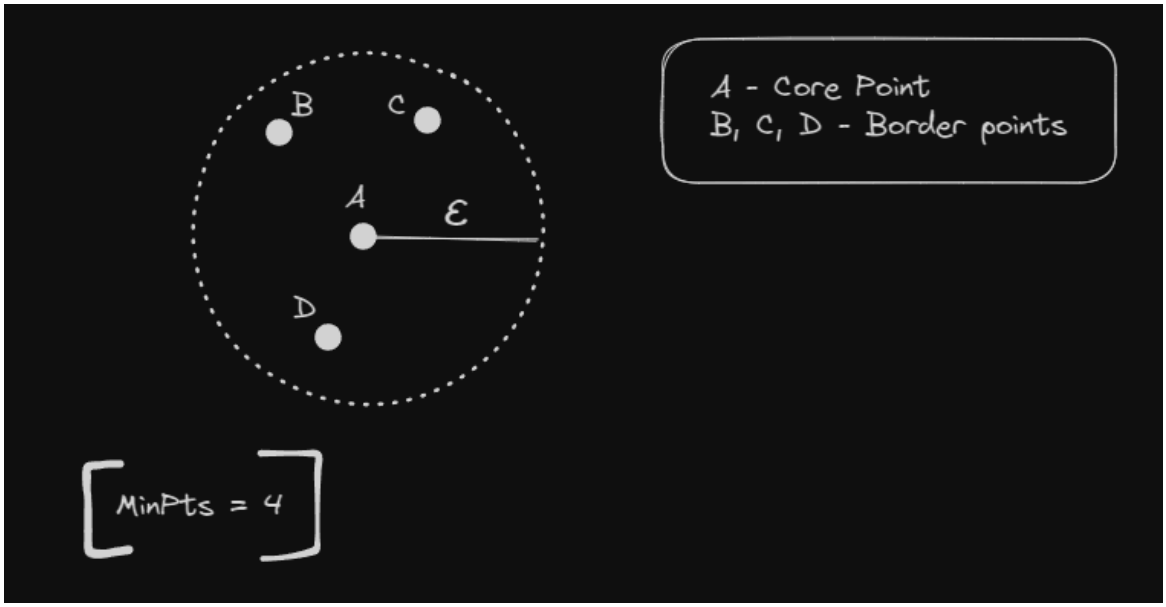
- It starts with all data points in a single cluster(dataset as a cluster) and then splits the cluster recursively until each data point is in its own cluster.
- Hence, this forms multiple separate clusters for a given data set.
- The terminal/leaf nodes of the tree are the final clusters.

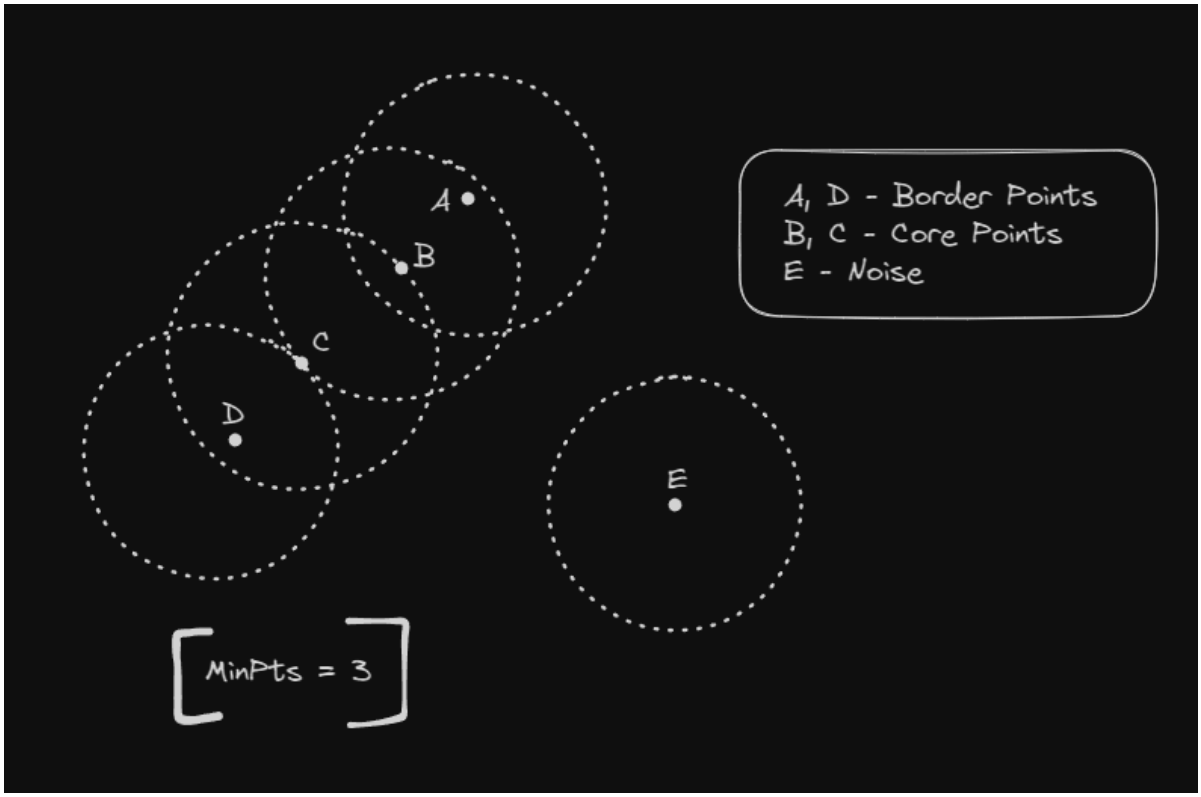


3) Density-based Methods:

- This method groups data points together based on their density up to the specified radius.
- A density is simply the no.of data points within a specified range.
- One popular density based clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise).
- some key points used in this method are:
  + Core Points: the central data points of clusters that have at least a specified no.of threshold neighboring points within the given radius(range). (n) no.of core points in a data set, represents (n) no.of clusters.
  + Border Points: the neighboring data points to a core point (or) the data points which are not core points but come under the radius of a core data point.
  + Noise: the data points which are neither core points nor the border points.
- Hence, the aim is to form clusters with core points as center,nearest data points as neighbors and eliminate noise data points.
- DBSCAN Algorithm:
  + Parameters used:

- Epsilon(ε): The radius from the core point, within which to search for neighboring points.
- MinPts: The minimum no.of data points required to form a dense region (core point)





+ Steps:
    1. Draw a table and determine the distances b/w every pair of data points of a given dataset.

2. Identify the neighboring points for every data point, based on the given (ε).
3. Identify core points based on the given threshold value of neighboring points (MinPts)
4. For non-core points, if it is a neighbor of any core points, it is said to be a border point. Else, it is a noise.
5. The resultant (n) core points are the centroids of (n) clusters and their corresponding border points are their neighbor points.

Ex Problem: [notes]

4) Gid-Based Method: