DATA WAREHOUSING & MINING

Unit-5: Cluster Analysis

Parameshwar R Hegde, PhD

Basic Concepts

- **Clustering**
  - In clustering, a group of different data objects is classified as similar objects
  - It is an unsupervised machine learning-based algorithm that acts on unlabelled data
  - One group means a cluster of data
  - Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data
  - After the classification of data into various groups, a label is assigned to the group

Basic Concepts

- **Cluster Analysis**
  - In Data Mining means that to find out the group of objects which are similar to each other in the group but are different from the object in other groups
  - In this, we first partition the set of data into groups based on data similarity and then assign the labels to the groups
  - **Requirements**
    - **Scalability** – We need highly scalable clustering algorithms to deal with large databases
    - **Ability to deal with different kinds of attributes** – Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical and binary data

Basic Concepts

- **Cluster Analysis**
  - **Requirements**
    - **Discovery of clusters with attribute shape** – The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes
    - **High dimensionality** – The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space
    - **Ability to deal with noisy data** – Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters

Basic Concepts

- **Cluster Analysis - Applications**
  - Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing
  - Help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns
  - In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations
  - Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location

Clustering Approaches

- **Clustering methods can be classified into the following categories**
  i. Partitioning Method
  ii. Hierarchical Method
  iii. Density-based Method

Clustering Approaches

- **Partitioning method**
  - This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data
  - Its the data analysts to specify the number of clusters that has to be generated for the clustering methods
  - In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region
  - There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Mediods), CLARA algorithm (Clustering Large Applications) etc

Clustering Approaches

- **Partitioning method – K-Means algorithm**
  - The given data points are grouped into k-clusters, based on the similarity of the data points
  - K-Mean algorithms, where each cluster is represented by the centre of gravity of the cluster
  - The K means algorithm takes the input parameter K from the user and partitions the dataset containing N objects into K clusters so that resulting similarity among the data objects inside the group (intracluster) is high but the similarity of data objects with the data objects from outside the cluster is low (intercluster)
  - The similarity of the cluster is determined with respect to the mean value of the cluster

Clustering Approaches

- **Partitioning method – K-Means algorithm**
  - It is a type of square error algorithm
  - At the start randomly k objects from the dataset are chosen in which each of the objects represents a cluster mean(centre)
  - For the rest of the data objects, they are assigned to the nearest cluster based on their distance from the cluster mean
  - The new mean of each of the cluster is then calculated with the added data objects

Clustering Approaches

- **Partitioning method – K-Means algorithm**
  - **Step 1:** Randomly select k cluster centers, $v_1, v_2, \ldots\ldots, v_n$
  - **Step 2:** Calculate the distance between each data points a_j and each cluster center v_i
  - **Step 3:** Assign each data points a_j to the cluster center v_i for which the distance $a_j - v_i$ is minimum
  - **Step 4:** Recalculate each cluster center by taking the average of cluster's data points
  - **Step 5:** Repeat from step 2 to step 4 until the recalculated cluster centers are same as previous or no re-assignment of data points happened

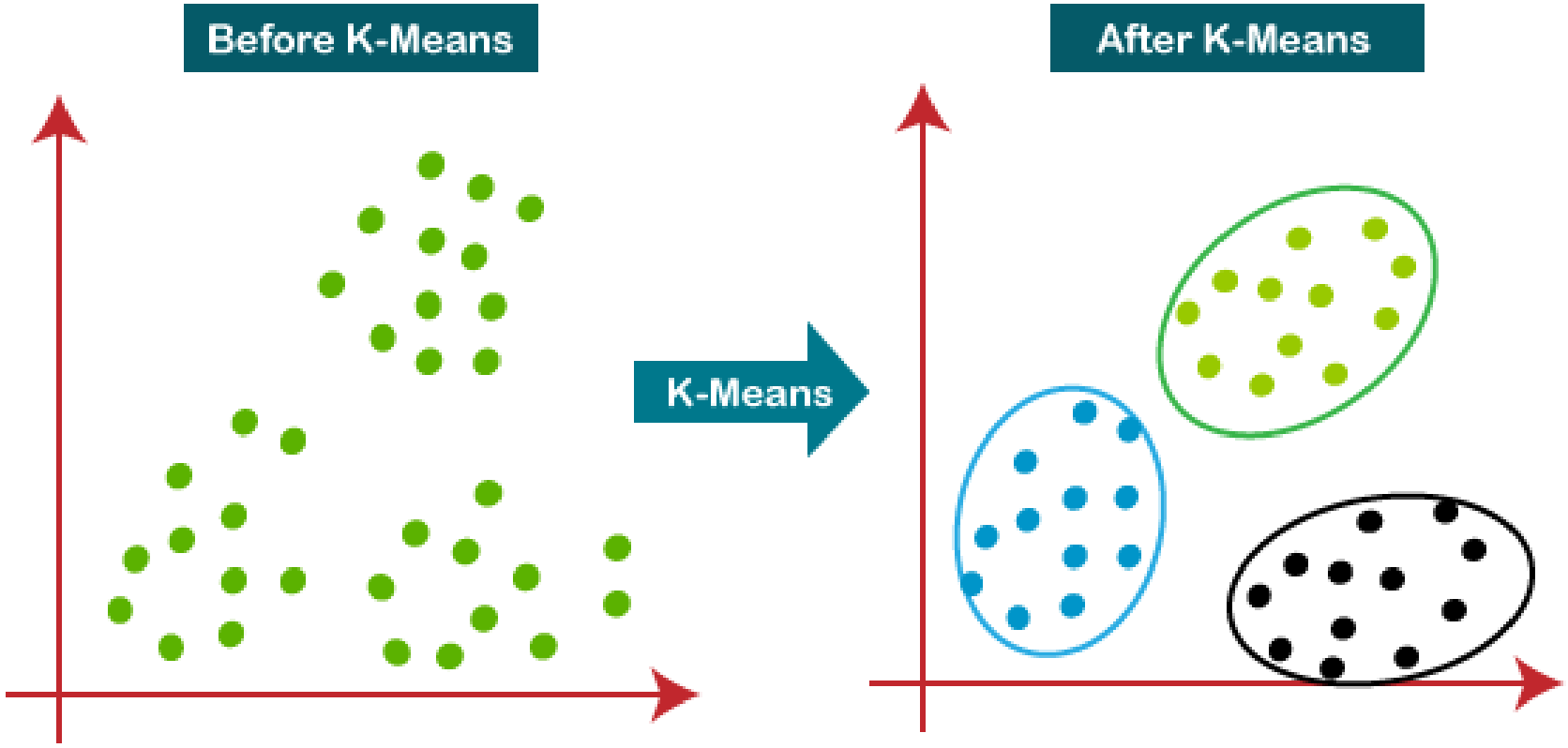
Clustering Approaches

- **Partitioning method – K-Means algorithm**
  - Distance between data points
    - We assume that each data points is a n-dimensional vector
    - The distance between two data points, x=(x_1, x_2, ……, x_n) and y= (y_1, y_2, ……, y_n) is defined by using Euclidean distance equation as

$$x - y = \sum_{i=1}^{n} \sqrt{(xi - yi)^2}$$

Clustering Approaches

- **Partitioning method – K-Means algorithm**



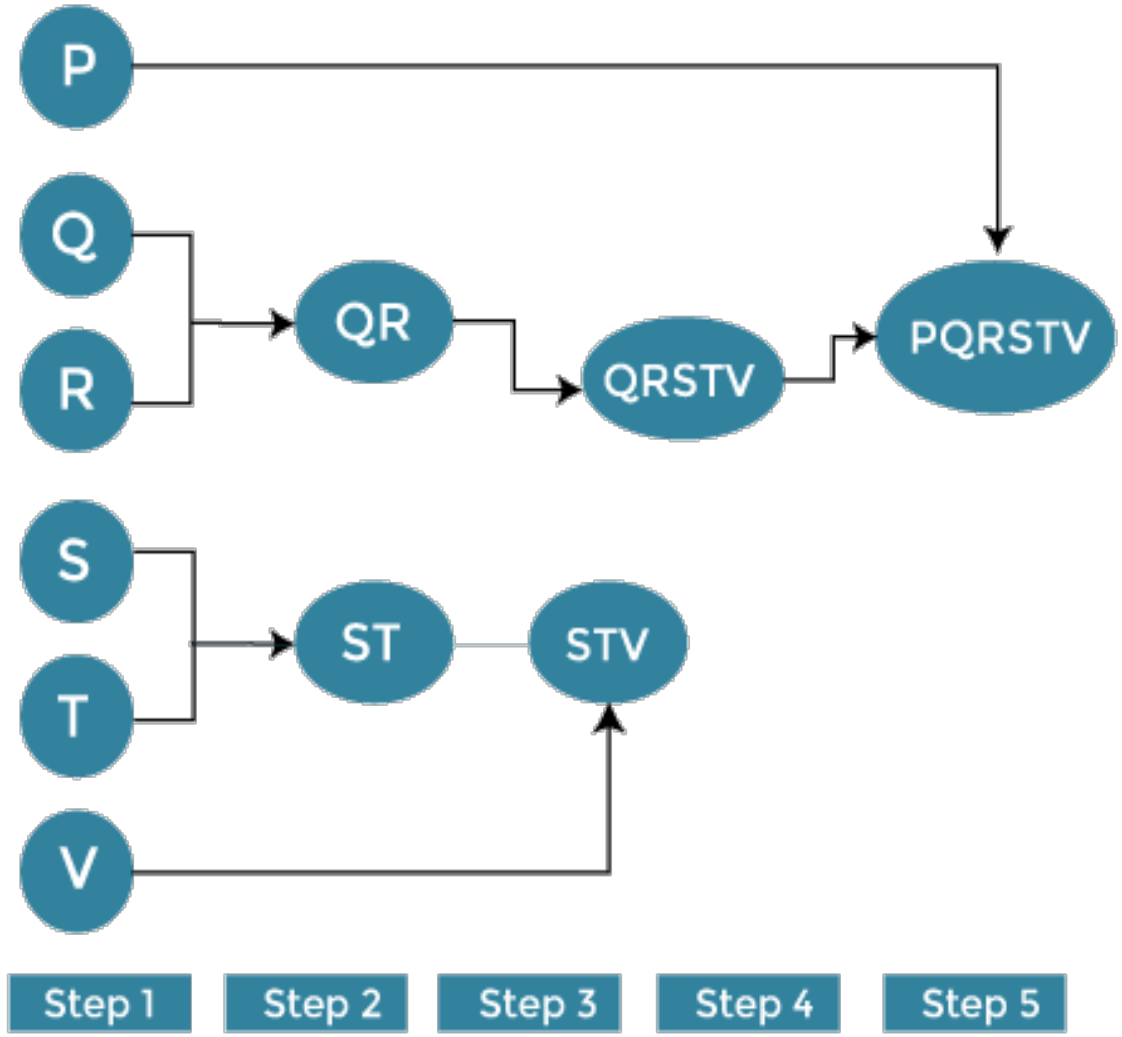
https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Clustering Approaches

- **Hierarchical method**
  - Is a method of clustering which seeks to build a heirarchy of clusters in a given dataset
  - It works via grouping data into a tree of clusters
  - Hierarchical clustering starts by treating each data points as an individual cluster
  - The endpoint refers to a different set of clusters, where each cluster is different from the other cluster, and the objects within each cluster are the same as one another

Clustering Approaches

- **Hierarchical method**
  - At the lowest level, each data point refers to a cluster. i.e, each cluster contains a single observation
  - At the highest level, there is only one cluster containing all the data
  - The decision regarding whether two clusters are to be merged or not is based on the “measure of dissimilarity between the clusters”

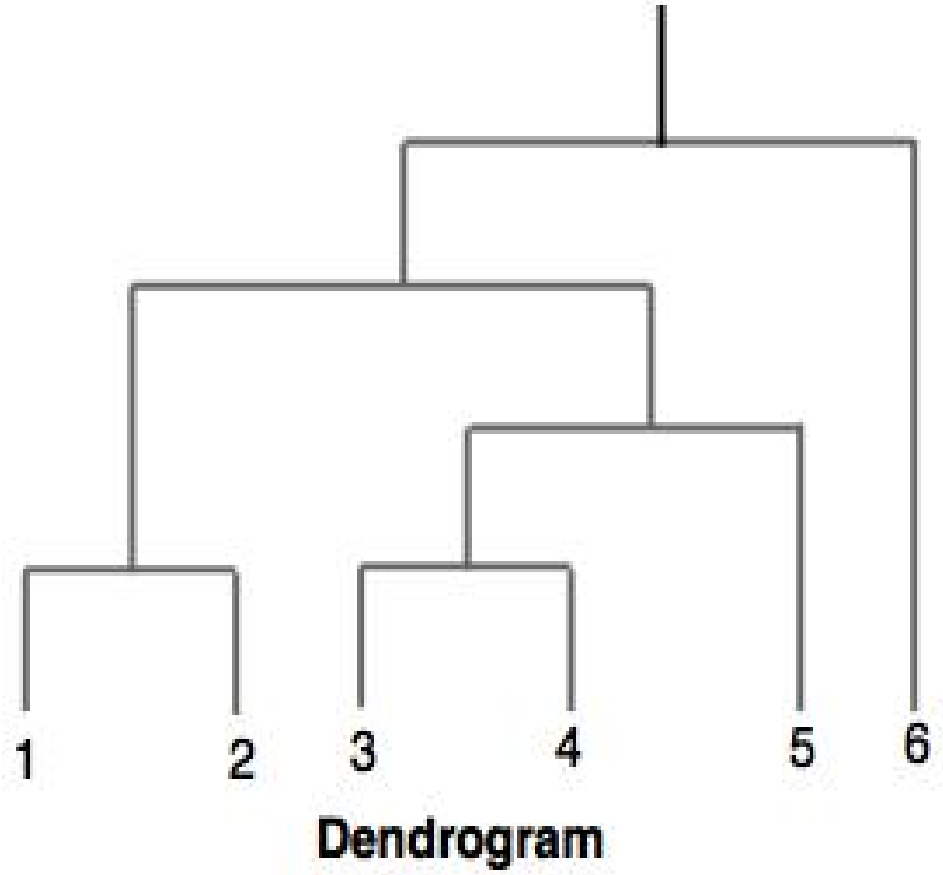


Clustering Approaches

- **Hierarchical method**
  - **Step 1:** Consider each alphabet (P, Q, R, S, T, V) as an individual cluster and find the distance between the individual cluster from all other clusters
  - **Step 2:** Now, merge the comparable clusters in a single cluster. Let's say cluster Q and Cluster R are similar to each other so that we can merge them in the second step. Finally, we get the clusters [(P), (QR), (ST), (V)]
  - **Step 3:** Here, we recalculate the proximity as per the algorithm and combine the two closest clusters [(ST), (V)] together to form new clusters as [(P), (QR), (STV)]
  - **Step 4:** Repeat the same process. The clusters STV and PQ are comparable and combined together to form a new cluster. Now we have [(P), (QQRSTV)]
  - **Step 5:** Finally, the remaining two clusters are merged together to form a single cluster [(PQRSTV)]

Clustering Approaches

- **Hierarchical method - Dendrogram**
  - It is a tree structure diagram which illustrates hierarchical clustering techniques
  - Each level shows clusters for that level
  - If there are ‘n’ data points, then we will get ‘n-1’ hierarchy of clusters



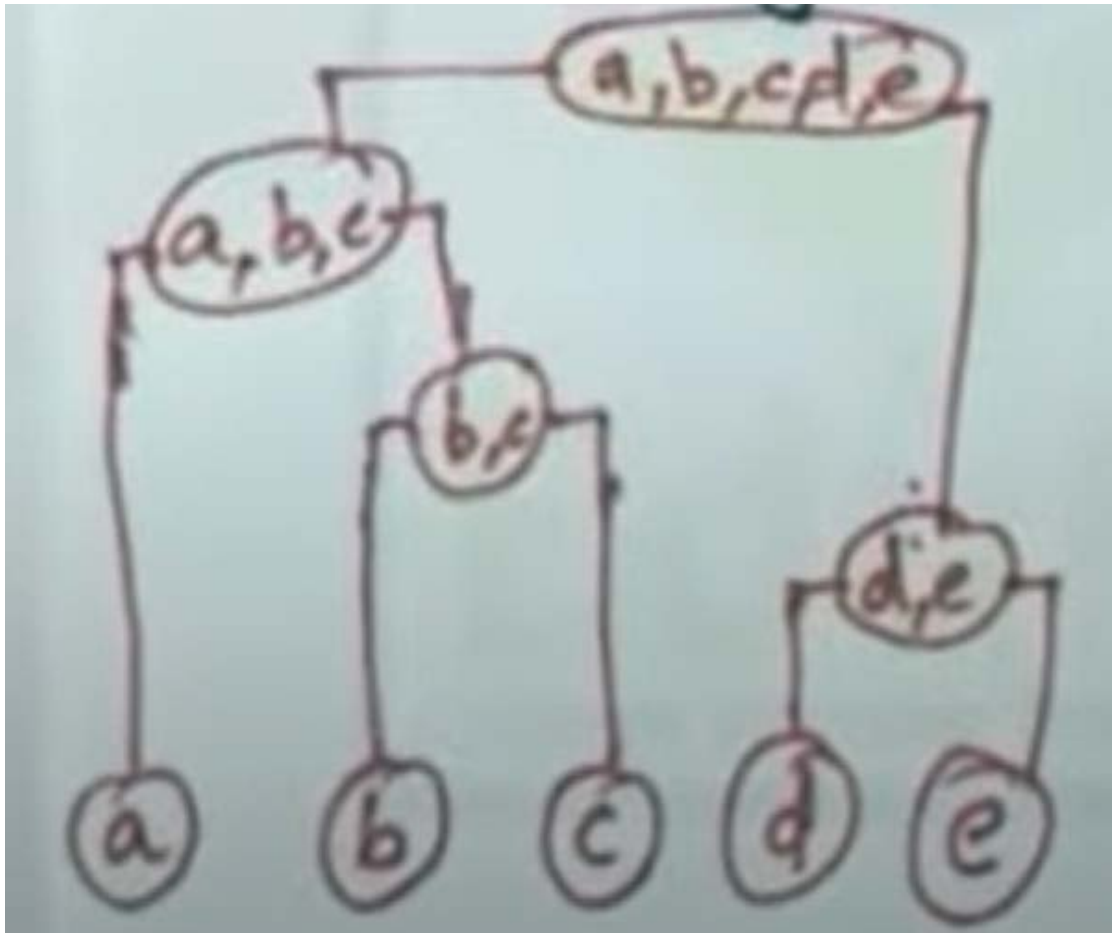
Dendrogram

Clustering Approaches

- **Hierarchical method**
  i. Agglomerative clustering
  ii. Divisive clustering

Clustering Approaches

- **Hierarchical method - Agglomerative clustering**
  - One of the most common types of hierarchical clustering used to group similar objects in clusters
  - Agglomerative clustering is also known as AGNES (Agglomerative Nesting)
  - In agglomerative clustering, each data point act as an individual cluster and at each step, data objects are grouped in a bottom-up method

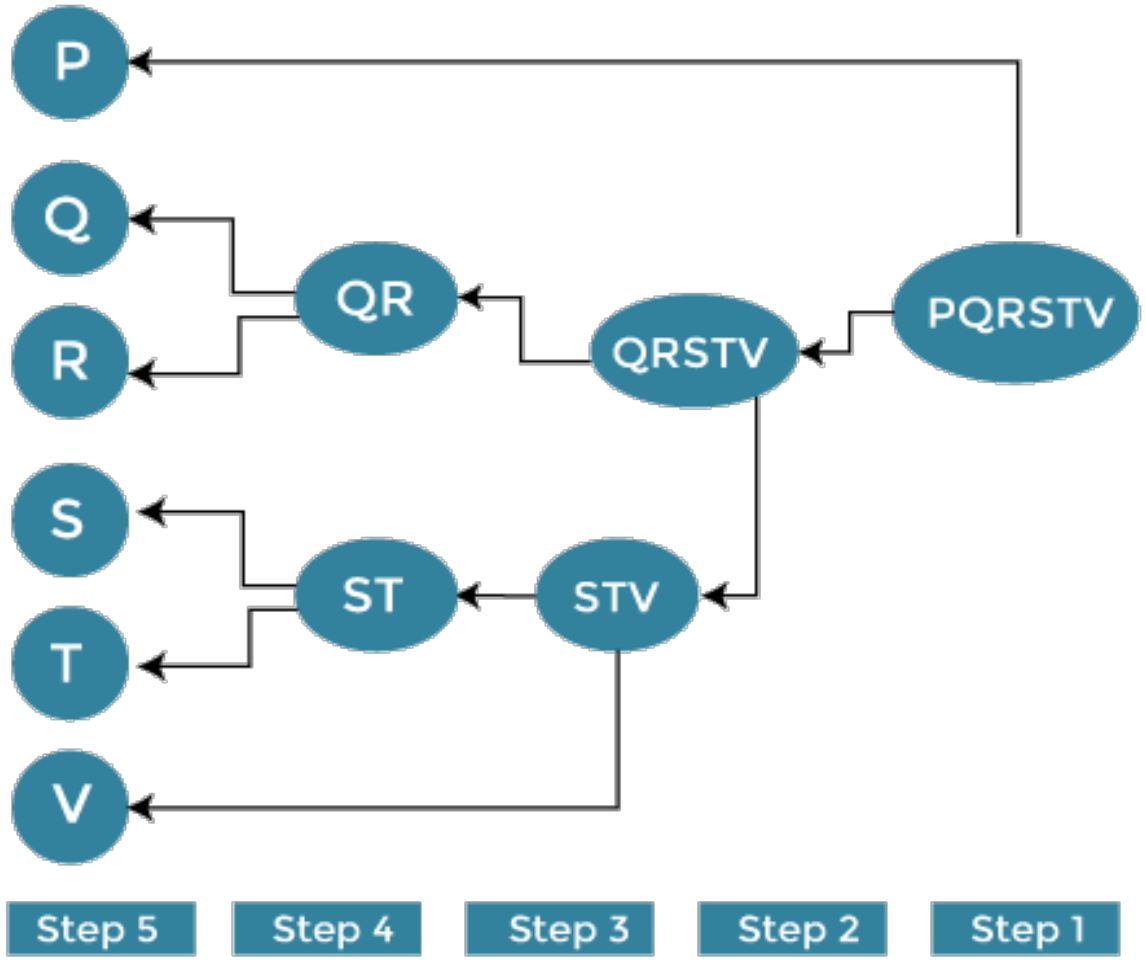


Clustering Approaches

- **Hierarchical method - Agglomerative clustering**
  - Algorithm
    - Determine the similarity between individuals and all other clusters. (Find proximity matrix)
    - Consider each data point as an individual cluster
    - Combine similar clusters
    - Recalculate the proximity matrix for each cluster
    - Repeat step 3 and step 4 until you get a single cluster

Clustering Approaches

- **Hierarchical method - Divisive clustering**
  - Is exactly the opposite of Agglomerative Hierarchical clustering
  - In Divisive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the cluster
  - The separated data points are treated as an individual cluster
  - Finally, we are left with N clusters



Clustering Approaches

- **Hierarchical method - Identify the distance between the clusters**
  - **Single Linkage:** The distance between the two clusters is represented by the shortest distance between points in those two clusters

$$d(A,B) = \max\{d(x,y) : x \epsilon A,\ y \epsilon b\}$$

  - **Complete Linkage:** The distance between the two clusters is represented by the maximum distance between points in those two clusters

$$d(A,B) = \min\{d(x,y) : x \epsilon A,\ y \epsilon b\}$$

  - **Average Linkage:** The distance between the two clusters is represented by calculating the average distance between points in those two clusters

$$d(A,B) = \frac{1}{|A|\,|B|} \sum_{x \epsilon A, y \epsilon b} d(x,y)$$

Clustering Approaches

- **Hierarchical method - Identify the distance between the clusters**
  - ➢ non-numerical data, distance between texts or words can be found by using

    Leventein distance
    - Consider the example,

      Suppose we have two words - **kitten** , **sitting**

      Distance between these two words can be determined by considering the factors such as number of insertion, number of deletion and number of substitution

      i.e., k is substituted with s

      e is substituted with i

      insertion of g at the end.

      therefore, the distance is 3

Clustering Approaches

- **Density-based method**
  - refers to one of the most popular unsupervised learning methodologies used in model building and machine learning algorithms
  - The general idea behind density-based clustering is growing a given cluster as long as the density in the neighborhood exceeds some threshold
  - The algorithm or technique comes under density-based clustering is DBSCAN
  - DBSCAN stands for **Density Based Spatial Clustering of Applications with Noise** and useful
    - To find the clusters of arbitrary shape
    - To handle noise

Clustering Approaches

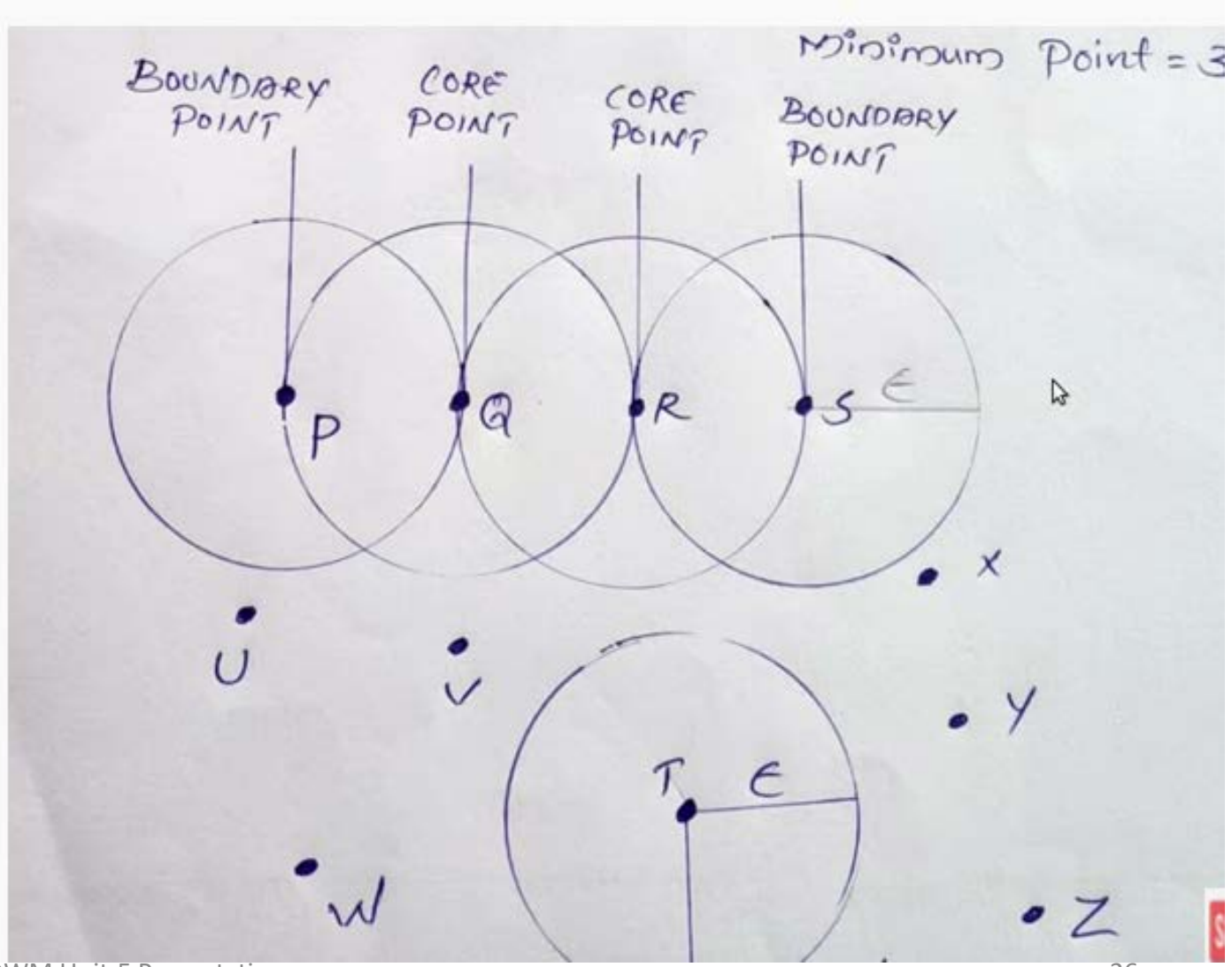
- **Density-based method – DBSCAN parameters**
  - Two parameters required for DBSCAN algorithm
    i. **Epsilon (Σ):** Represents the radius of the region
    ii. **MinPts:** Minimum number of points that should come inside radius (threshold)
  - Based on these two parameters, points are classified as **Core point**, **Boundary/ Border point**, or **Outlier/Noise**
    - **Core Point:** The point that satisfies minimum point condition
    - **Boundary point:** The point does not have sufficient minPts and is coming under core point
    - **Noise:** The point does not have sufficient minPts and does not become a neighbor of core point

Clustering Approaches

- **Density-based method – DBSCAN parameters**
  - Based on these two parameters, points are classified as **Core point**, **Boundary/ Border point**, or **Outlier/Noise**
    - **Direct Density Reachable (DDR):** A point ‘p’ is Direct Density Reachable from a point ‘q’, if ‘q’ is the core point and ‘p’ is the neighbor of q
    - **Density Reachable:** Two points are density reachable if there is a chain of DDR point that link these two points

Clustering Approaches

- **Density-based method – DBSCAN**



Self study

- **Outlier detection**