Practical NO:-1.

Title:- Implementing k-means clustering.

Objective:-

(1) unsupervised Data segmentation - Group similar data points into clusters for pattern recognition.

(2) Optimizing cluster selection - Identify the optimal number of clusters for better accuracy.

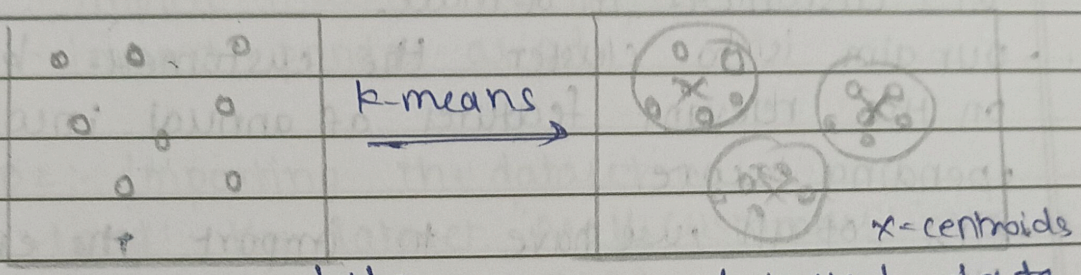
Theory:-

(1) k-means clustering.

- k-means clustering is an unsupervised learning algorithm that is used to solve the clustering problems in ML or datascience.
- K-means clustering is an unsupervised learning algorithm which groups the unlabelled dataset into diff clusters.
- Here K-defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be 2 clusters, and for K=3 there will be 3 clusters & so on.
- It allows us to cluster the datapoints into diff groups and a convienient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.
- It is a centroid-based algorithm, where each cluster is associated with a centroid.

Step 6:- If any reassignment, which means reassign each datapoint to the new closest centroid of each other, then go to step 4, or else go to finish.

step 7:- The model is ready.



unlabelled data labelled data.

Elbow method.

- It is the simplest and most commonly used iterative type of supervised learning algorithm
- unlike supervised learning, we don't have labelled data in k-means.
- some other unsupervised learning algorithm are PCA (principle component analysis), k-mediod etc.
- In k-means, we randomly initialize the k-number of cluster centroids in the data and iterates these centroids until no change happens to the position of the centroid.
- Now we will use Euclidean distance or Manhantt distance as the metric to calculate the distance of the point from the nearest cluster centroid of the clusters thus formed.
- We will continue this for a given number of iterations until the position of the centroid doesn't change.

Advantages.

(1) simple & Efficient - Early to implement & computationally faster than hierarchical clustering
(2) scalability - Works well with large datasets.
(3) Adaptability - Can be used in various domains for the clustering tasks.

Disadvantages.

(1) Sensitive to initial centroids - The final cluster depend on initial random centroid selection.
(2) Affected by outliers - outliers can skew the centroid & leads to poor clustering.

Conclusion:-

The practical demonstrated that k-means clustering effectively groups data points into clusters based on similarity while it is efficient & scalable, choosing the right no. of clusters is crucial for meaningful insights.

R. Lambe