Assignment No. 4

Aim: Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method.

Dataset link : https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

Objective:

- The Basic Concepts of K-means algorithm.
- Implementation logic of K-means algorithm.

Theory:

Introduction:

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labeled, outcomes.

A cluster refers to a collection of data points aggregated together because of certain similarities. You'll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.

Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

The _means' in the K-means refers to averaging of the data; that is, finding the centroid. To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

The centroids have stabilized—there is no change in their values because the clustering has been successful.

The defined number of iterations has been achieved.

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into *k* clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low.

Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity.

The k-means algorithm proceed as follows

First, it randomly selects k of the objects, each of which initially represents a cluster mean or center.

For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.

It then computes the new mean for each cluster. This process iterates until the criterion function converges.

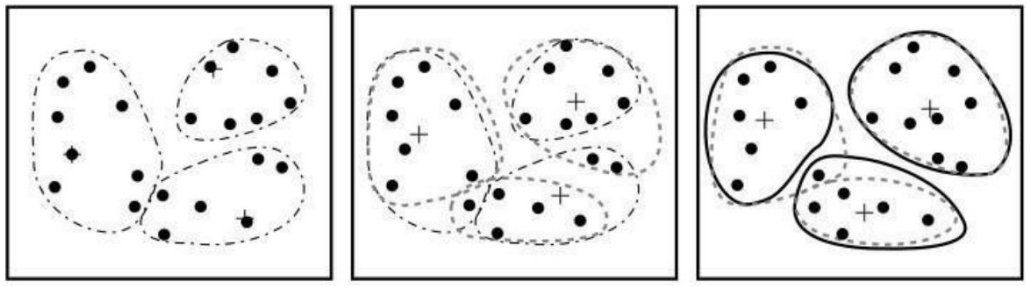
Typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2,$$

where *E* is the sum of the square error for all objects in the data set; p is the point in space representing a given object; mi is the mean of cluster *Ci* (both p and mi are multidimensional).

In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible.

The k-means procedure is summarized in following figure.



(a) (b) (c)

Above figure represent clustering of set of objects based on clustering methods. In figure, mean of each cluster marked by a

K-means Algorithm

The k-means algorithm for partitioning, where is cluster center is represented by the mean value of the object in the cluster. .

Algorithm:

1. Arbitrary choose k objects from D as the initial cluster centers
2. Repeat
3. (re)Assign each object to the cluster to which the object is most similar, Based on the mean value of the objects in the cluster;

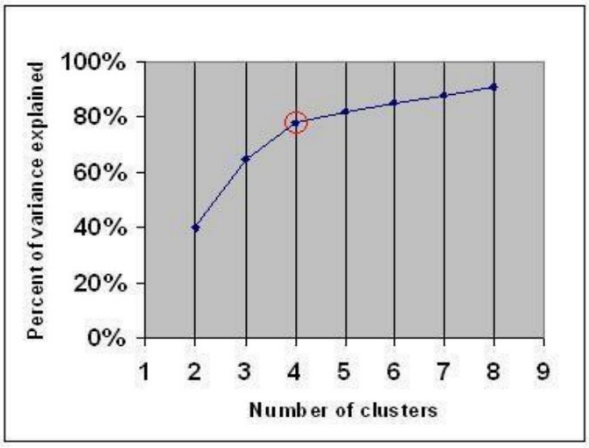
4. Update the cluster means, i.e., calculate the mean value of the objects for each cluster,

5. Until no change

Elbow method:

In cluster analysis, the **elbow method** is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. The

same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.



Advantages

- Easy to implement.
- An instance can change cluster (move to another cluster) when the centroids are re-computed.
- If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.
- K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Disadvantages

- Difficult to predict number of clusters.
- Initial seed have a strong impact on the final results.
- The order of the data has impact on the final results.
- Sensitive to scale: rescaling your datasets (normalization or standardization) will completely change results. While this itself is not bad, not realizing that you have to spend extra attention to scaling your data might be bad.
- Difficult to predict K-Value.
- With global cluster, it didn't work well.
- Different initial partitions can result in different final clusters.
  It does not work well with clusters (in the original data) of Different size and Different density.

Conclusion: We have studied the k-means clustering algorithm and also implemented successfully.