G H Raisoni College of Engineering
SY AI Semester-IV AY 2023-24 Division-A
UCAIP210: Machine Learning Algorithms Practicals
Lab Manual
Practical Teacher : Dr Monika Y. Dangore

Experiment No : 7.1

Aim:

Implement KMeans Clustering (UnSupervised Learning)

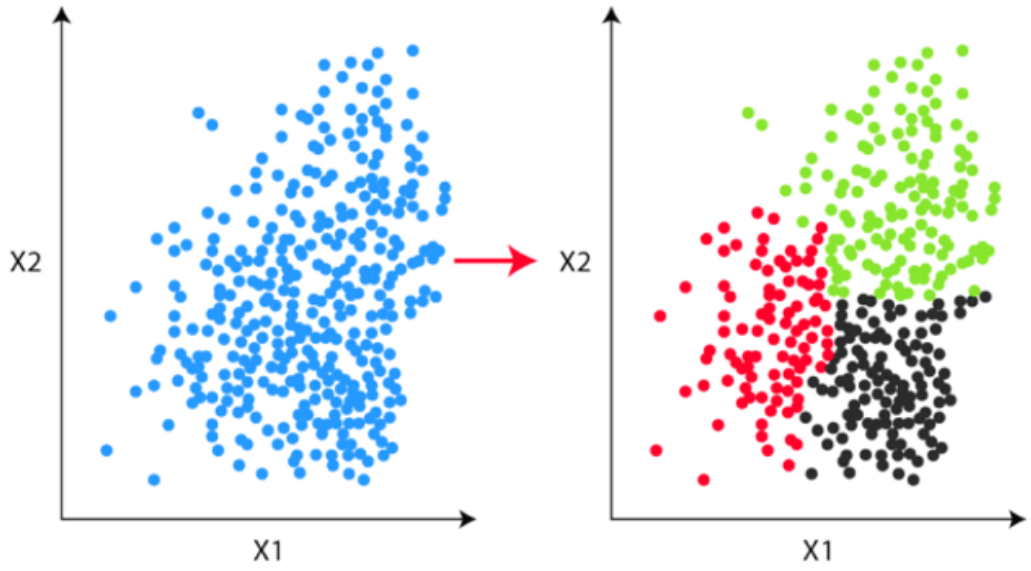
Problem Statement:

Categorize the Countries in to Continetwise Clusters on the basis of their Latitude and Longitude

Introduction to K-Means Clustering:

Clustering is a set of techniques used to partition data into groups, or clusters. **Clusters** are loosely defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters.

Clustering is an unsupervised algorithm where labels are missing meaning the dataset contains only input data points (Xi).



K-means is an unsupervised learning method for clustering data points. The algorithm iteratively divides data points into K clusters by minimizing the variance in each cluster.

First, each data point is randomly assigned to one of the K clusters. Then, we compute the centroid (functionally the center) of each cluster, and reassign each data point to the cluster

with the closest centroid. We repeat this process until the cluster assignments for each data point are no longer changing.

K-means clustering requires us to select K, the number of clusters we want to group the data into. The elbow method lets us graph the inertia (a distance-based metric) and visualize the point at which it starts decreasing linearly. This point is referred to as the "eblow" and is a good estimate for the best value for K based on our data.

The algorithm works as follows:

1. First the algorithm initializes k points, called means or cluster centroids.

3. It categorizes each item to its closest mean, and updates the mean's coordinates, which are the averages of the items categorized in that cluster so far.

4. It repeats the process for a given number of iterations and at the end, clusters are formed.

Experiment No: 7.2

Aim:

Implement Hierarchical Clustering (UnSupervised Learning)

Introduction to Hierarchical Clustering

Hierarchical clustering is a technique in unsupervised machine learning that involves the organisation of data into a hierarchy of nested clusters. Unlike other clustering methods, hierarchical clustering creates a tree-like structure of clusters (dendrogram), which visually represents the relationships between data points.

Hierarchical clustering is based on two main approaches:

(1) Agglomerative Clustering: In this approach, each data item is initially considered as a single cluster. Then, using similarity measures, the two closest clusters are combined and transformed into a larger cluster. This aggregation process continues until all data items are in a single large set.

The working logic of the algorithm can be summarized with the following steps:

a) Initially, each item is considered a set by itself.

b) A similarity or distance matrix is created between all items. This matrix shows the distance or similarity between each pair of items.

c) The closest two clusters (or elements) are found and a new cluster is formed by combining these two clusters. In this step, different linkage methods can be used. The most common attachment methods are:

- **Single Linkage**

  $D(c_1,c_2) = \min D(x_1,x_2)$

  Minimum distance or distance between closest elements in clusters

- **Complete Linkage**

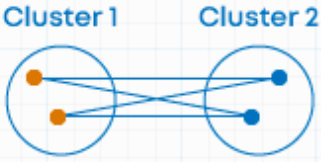
  $D(c_1,c_2) = \max D(x_1,x_2)$

  Maximum distance between elements in clusters

- **Average Linkage**

  $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\Sigma\Sigma D(x_1,x_2)$

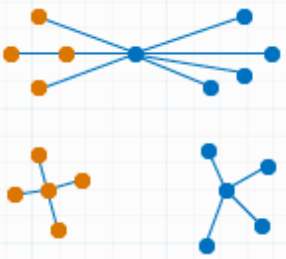
  Average of the distances of all pairs



- **Centroid Method**

  Combining clusters with minimum distance between the centroids of the two clusters

- **Ward's Method**
  - Combining clusters where increase in within cluster variance is to the smallest degree.
  - Objective is to minimize the total within cluster vairance



d) Step 3 is repeated until only one cluster remains in the dataset. This process creates a clustering structure in which items are grouped into a hierarchical tree structure.

The resulting hierarchical clustering structure is often visualized as a dendrogram. A dendrogram is a tree chart showing step-by-step cluster aggregation operations.

(2) Segmentative Clustering: In this approach, the entire dataset is initially considered as a single large cluster. Then, the items within the set are divided into subsets based on their similarity. This division process continues such that each subset groups the most similar items within itself.

Distance matrix is generally used for hierarchical clustering.
The distance matrix is a matrix that contains the distances (measures of difference) between data points. This distance matrix represents the similarities or differences between objects. Usually, the distance between two data points is calculated with metrics such as Euclidean distance, Manhattan distance, Correlation coefficient, Mahalanobis distance.
Here we will use the euclidean distance, the equation of the euclidean distance is as follows:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

where **d** represents the Euclidean distance between two points.
(x_1, y_1) and **(x_2, y_2)** indicate the coordinates of two points.

Program:

Attach the printouts of both the K-Means Clustering and Hierarchical Clustering Programs.

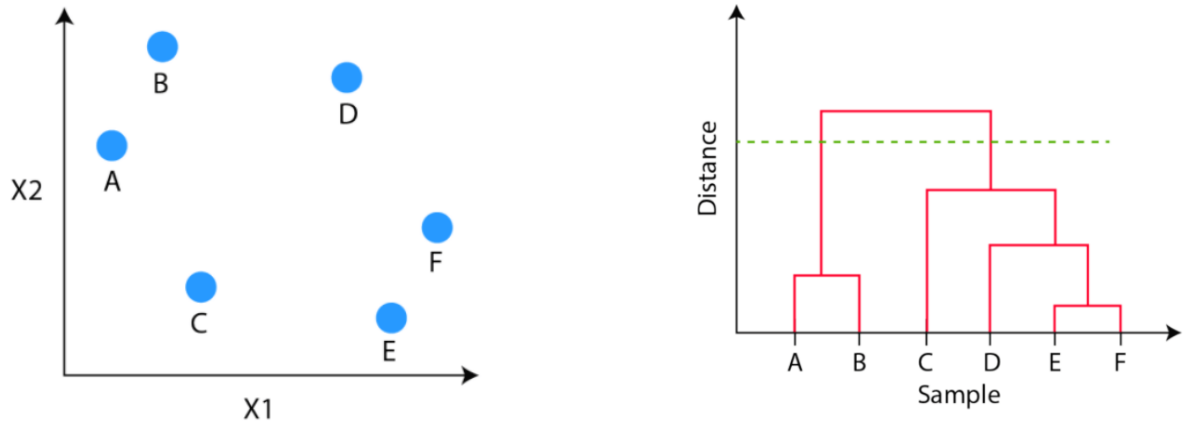
Result:

The concept of K-Means Clustering and Hierarchical Clustering is studied and programs are executed successfully.

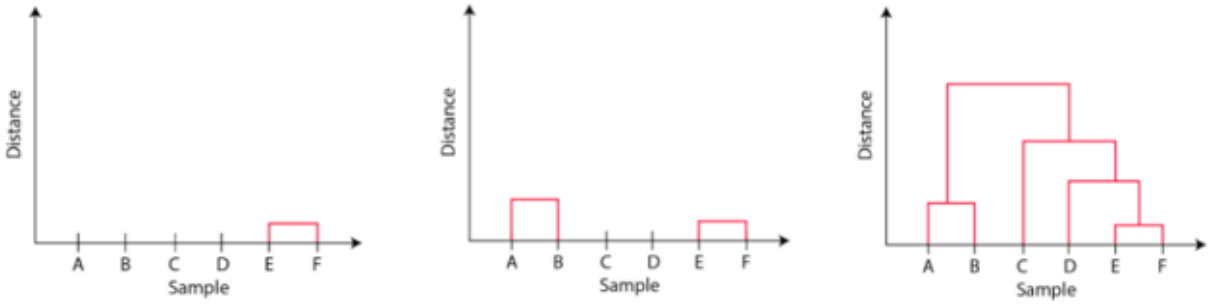
Additional Content (To be skipped from write up)

What is a Dendrogram?

A dendrogram is a tree-like diagram that shows the hierarchical relationship between the observations.



The observations E and F are closest to each other by any other points. So, they are combined into one cluster and also the height of the link that joins them together is the smallest. The next observations that are closest to each other are A and B which are combined together.



This can also be observed in the dendrogram as the height of the block between A and B is slightly bigger than E and F. Similarly, D can be merged into E and F clusters and then C can be combined to that. Finally, A and B combined to C, D, E and F to form a single cluster.

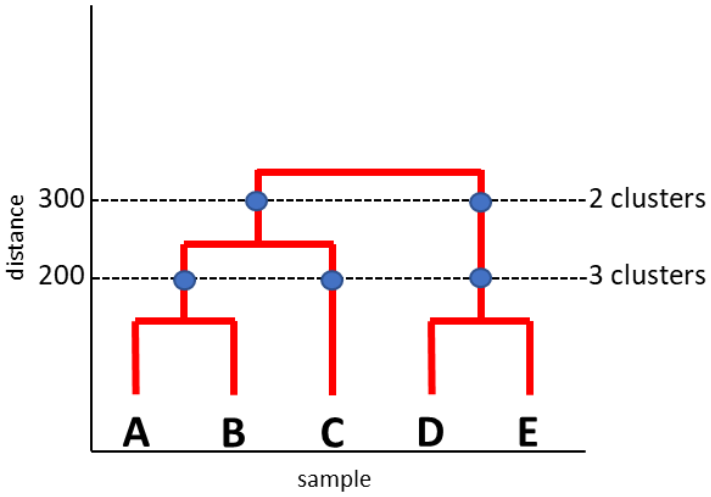
The important point to note while reading the dendrogram is that:

1. The Height of the blocks represents the distance between clusters, and
2. Distance between observations represents dissimilarities.

Now the question is, how do we find the number of clusters using a dendrogram or where should we stop merging the clusters? Observations are allocated to clusters by drawing a horizontal line through the dendrogram.

Generally, we cut the dendrogram in such a way that it cuts the tallest vertical line. In the above example, we have two clusters. One cluster has observations A and B, and a second cluster has C, D, E, and F.

Dendrogram allows us to do a distance cutoff to choose how many clusters we want to obtain. Figure 8 shows the illustration of the cluster number differences between distance cutoff 300 and 200.

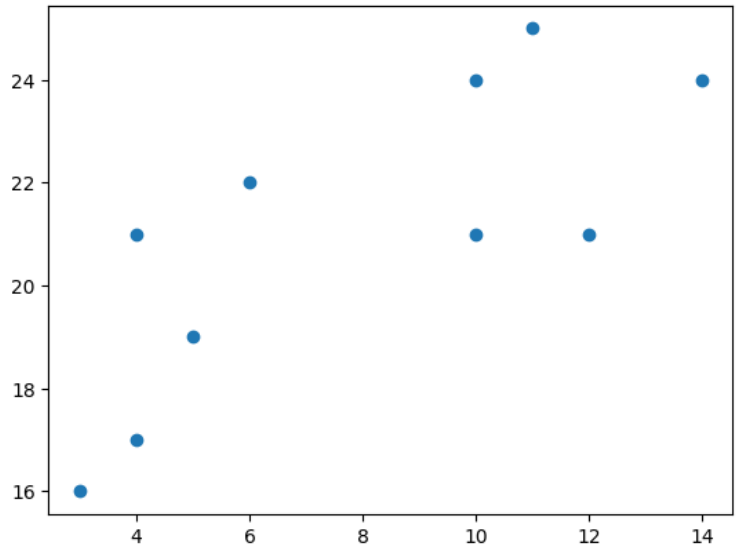


Example Explained:

x = [4, 5, 10, 4, 3, 11, 14 , 6, 10, 12]

y = [21, 19, 24, 17, 16, 25, 24, 22, 21, 21]

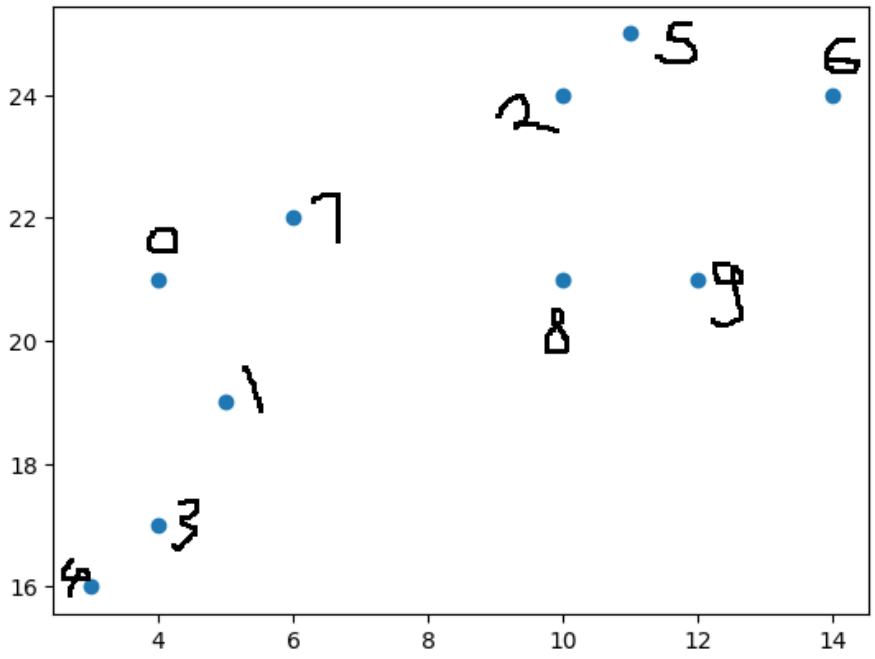
plt.scatter(x, y)

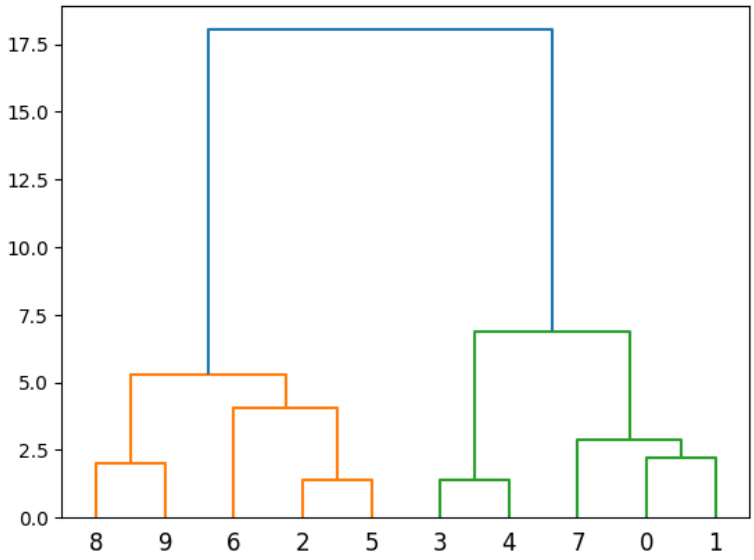


data = list(zip(x, y))

print(data)

[(4, 21), (5, 19), (10, 24), (4, 17), (3, 16), (11, 25), (14, 24), (6, 22), (10, 21), (12, 21)]


24
22
20
18
16
4
6
8
10
12
14


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
8
9
6
2
5
3
4
7
0
1