K- MEANS CLUSTERING

INTRODUCTION-What is clustering?

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.

In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

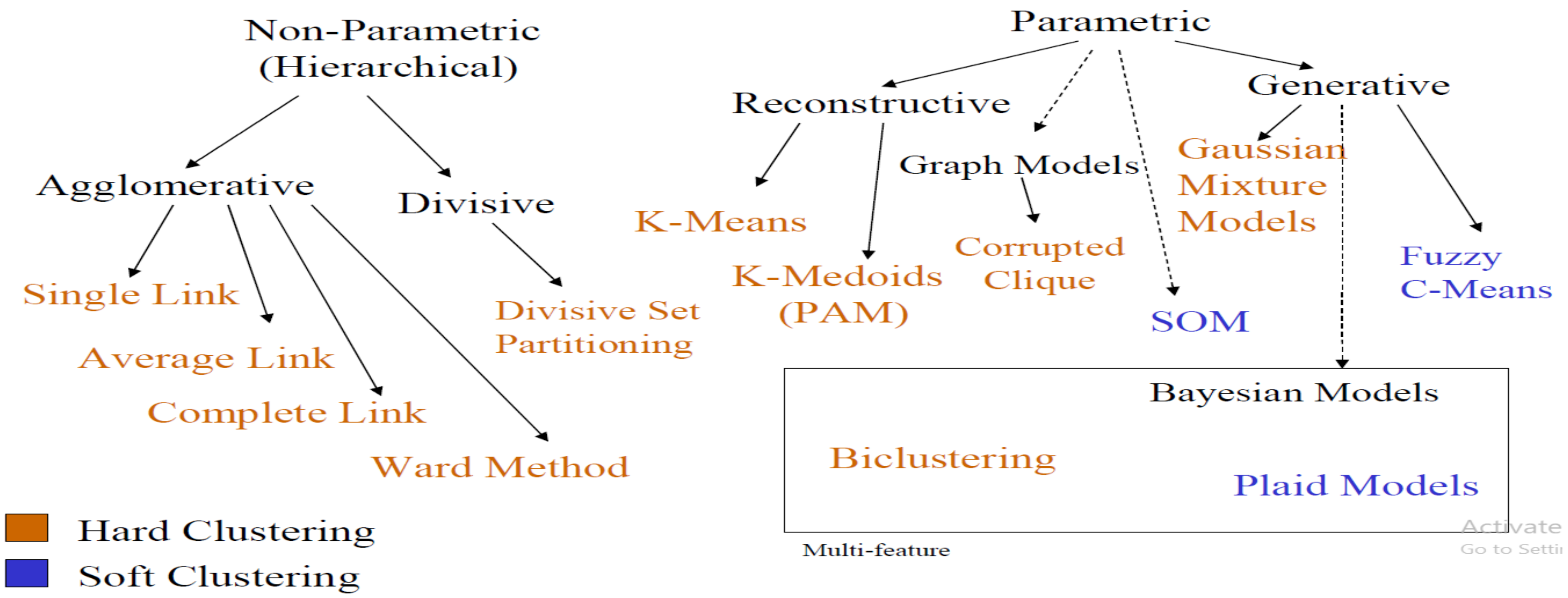
INTRODUCTION-What is clustering?

Let’s understand this with an example. Suppose, you are the head of a rental store and wish to understand preferences of your costumers to scale up your business. Is it possible for you to look at details of each costumer and devise a unique business strategy for each one of them? Not.

But what you can do is to cluster all of your costumers into say 10 groups based on their purchasing habits and use a separate strategy for costumers in each of these 10 groups. And this is what we call clustering.

Types of clusters ?

Clustering Approaches



Non-Parametric (Hierarchical)

Agglomerative

Single Link

Average Link

Complete Link

Ward Method

Divisive

Divisive Set Partitioning

Parametric

Reconstructive

K-Means

K-Medoids (PAM)

Graph Models

Corrupted Clique

SOM

Generative

Gaussian Mixture Models

Fuzzy C-Means

Bayesian Models

Biclustering

Plaid Models

Multi-feature

🟧 Hard Clustering

🟦 Soft Clustering

Types of clusters ?

The various types of clustering are:

1.Connectivity-based Clustering (Hierarchical clustering)

2.Centroids-based Clustering (Partitioning methods)

3.Distribution-based Clustering

4.Density-based Clustering (Model-based methods)

5.Fuzzy Clustering

6.Constraint-based (Supervised Clustering)

What is Hard Clustering & Soft Clustering?

Hard Clustering: In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.

What is Hard Clustering & Soft Clustering?

ML Labs Pvt Ltd

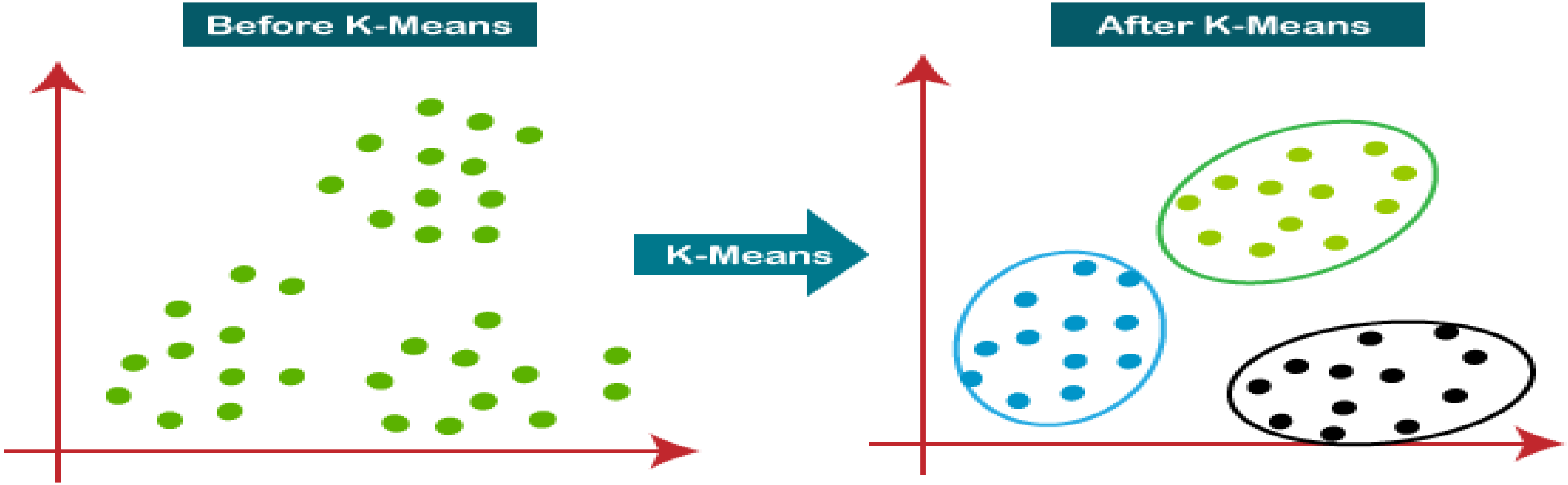
Soft Clustering: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.

Introduction to K-means Clustering:

ML Labs Pvt Ltd

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*.



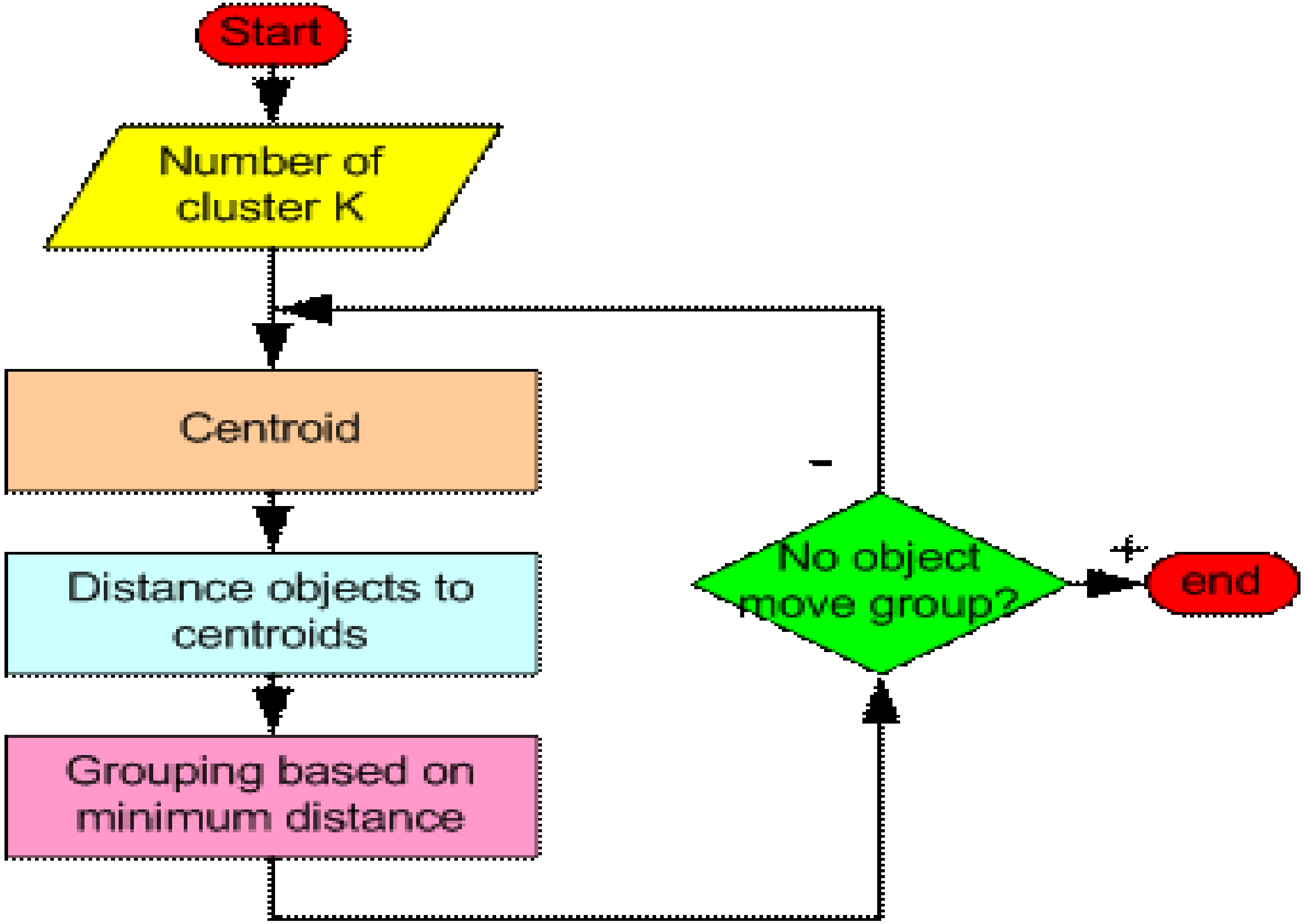
Introduction to K-means Clustering:

The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

- The centroids of the *K* clusters, which can be used to label new data
- Labels for the training data (each data point is assigned to a single cluster)

How the K-Mean Clustering algorithm works?

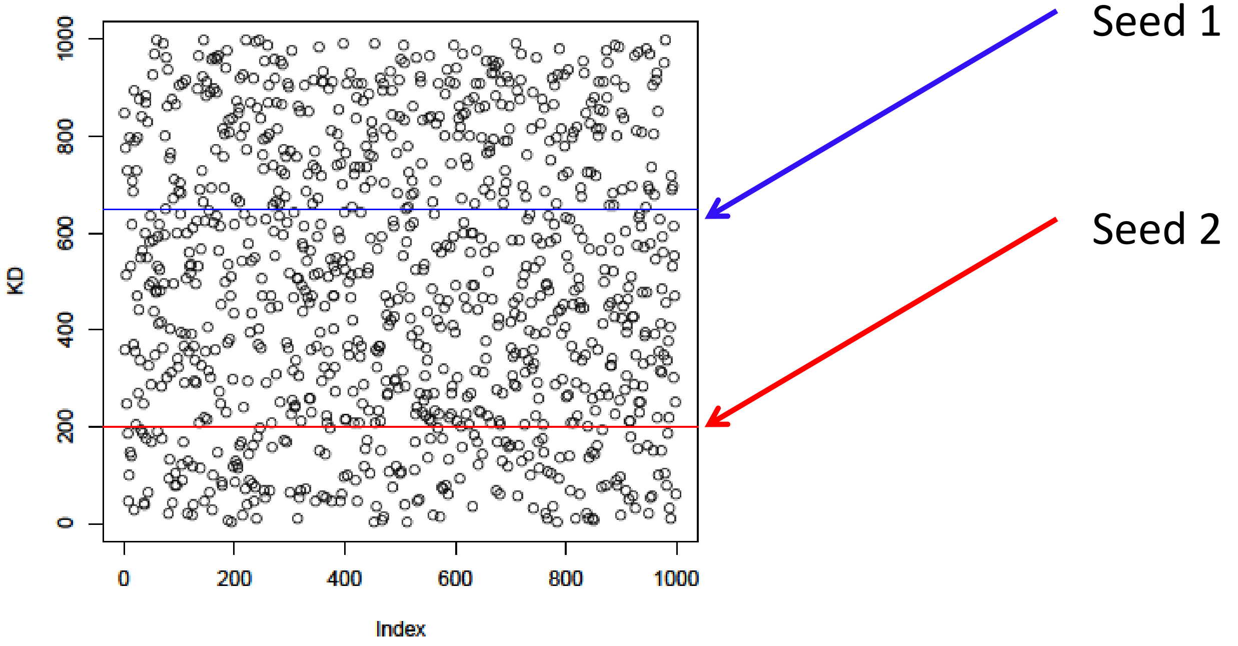
ML Labs Pvt Ltd


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

HOW THE K MEAN CLUSTERING ALGORITHM WORKS?

ML Labs Pvt Ltd

- Step 1: Begin with a decision on the value of k = number of clusters(randomly select initial cluster seeds).



HOW THE K MEAN CLUSTERING ALGORITHM WORKS?

- Step 2: Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following:

    1.Take the first k training sample as single-element clusters

    2. Assign each of the remaining (N-k) training sample to the cluster with the nearest centroid. After each assignment, recomputed the centroid of the gaining cluster.

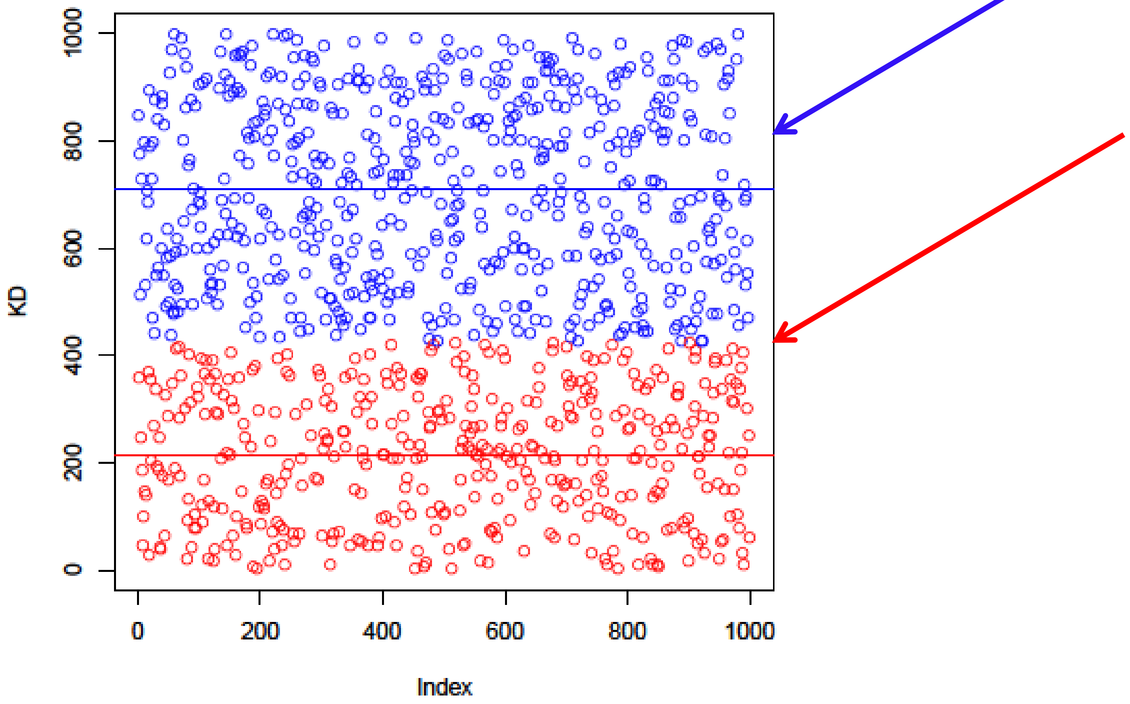
HOW THE K MEAN CLUSTERING ALGORITHM WORKS?

- Step 3: Take each sample in sequence and compute its distance from the centroid of each of the clusters.

If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

HOW THE K MEAN CLUSTERING ALGORITHM WORKS?

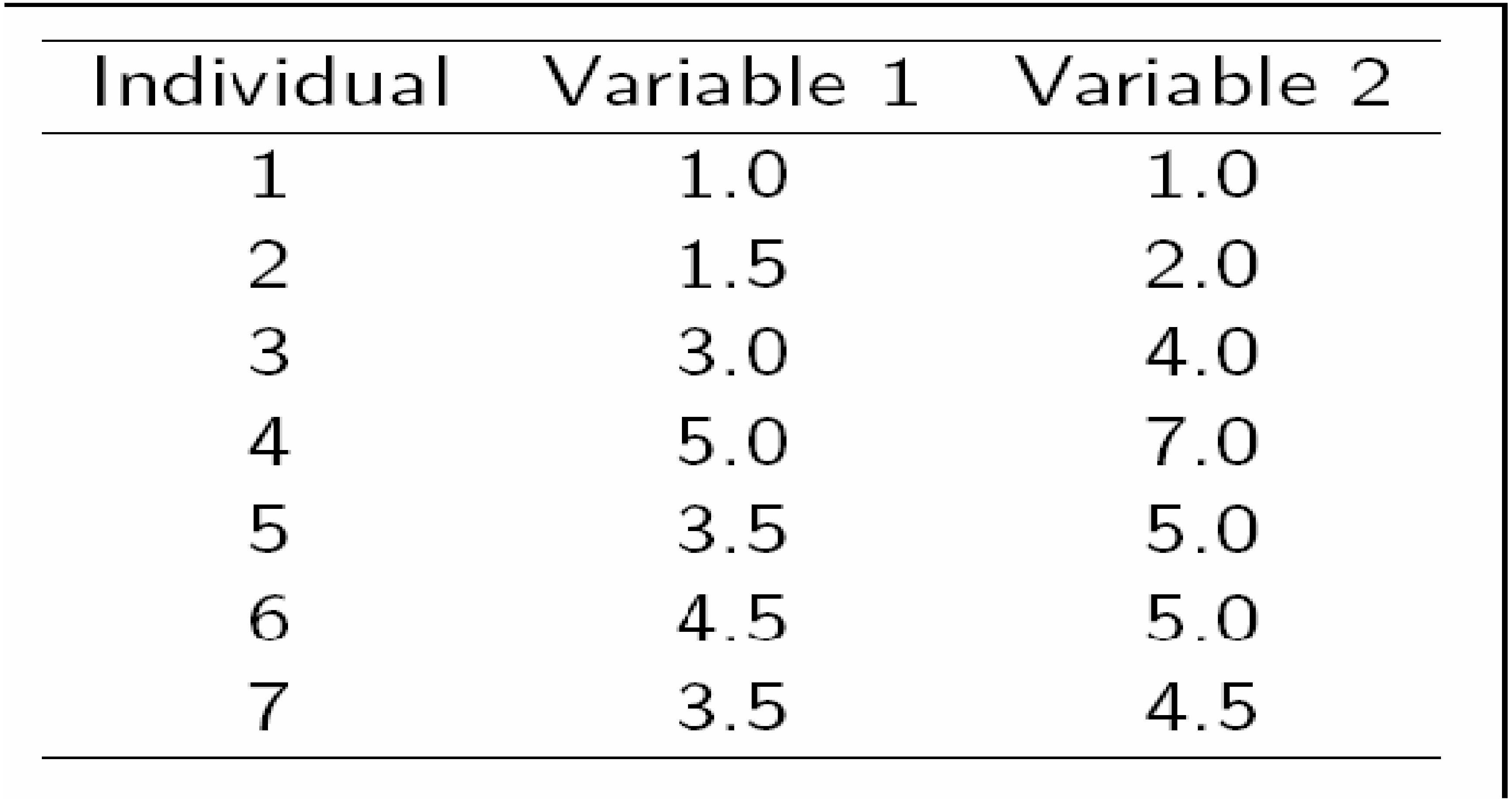
- **Step 4 :Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.**



Seed 1

Seed 2

A Simple example showing the implementation of k-means algorithm (using K=2)

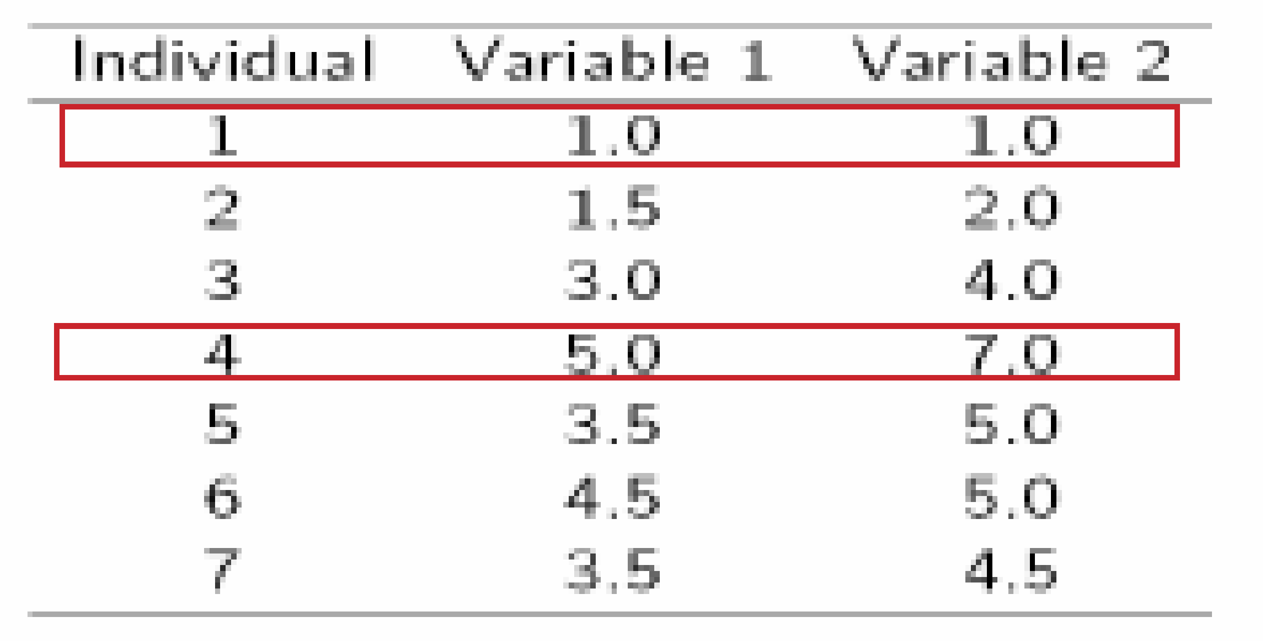


Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

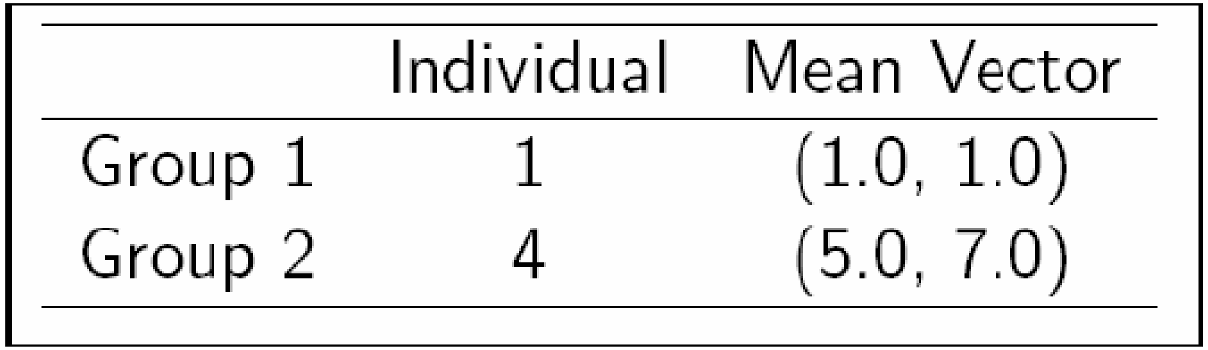
Step 1:

Initialization: Randomly we choose following two centroids (k=2) for two clusters.

In this case the 2 centroid are: m1=(1.0,1.0) and m2=(5.0,7.0).



Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5



	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Step 2:

Thus, we obtain two clusters containing:

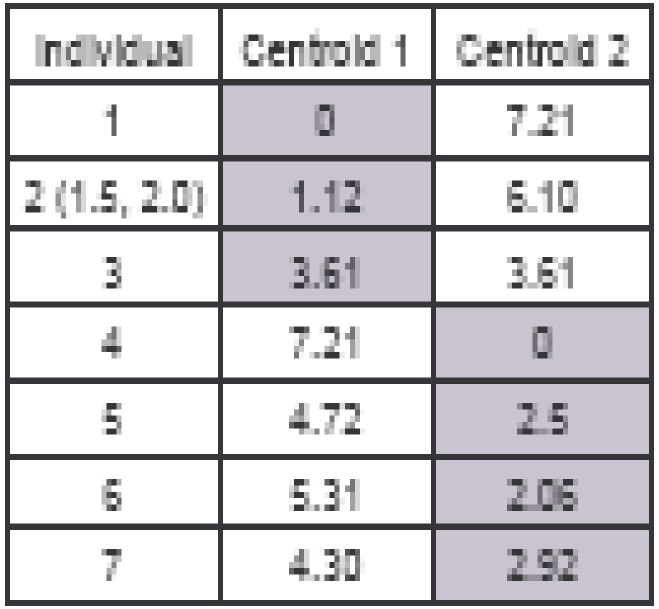
{1,2,3} and {4,5,6,7}.

Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83,2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12,5.38)$$



Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

THE DISTANCE FORMULA

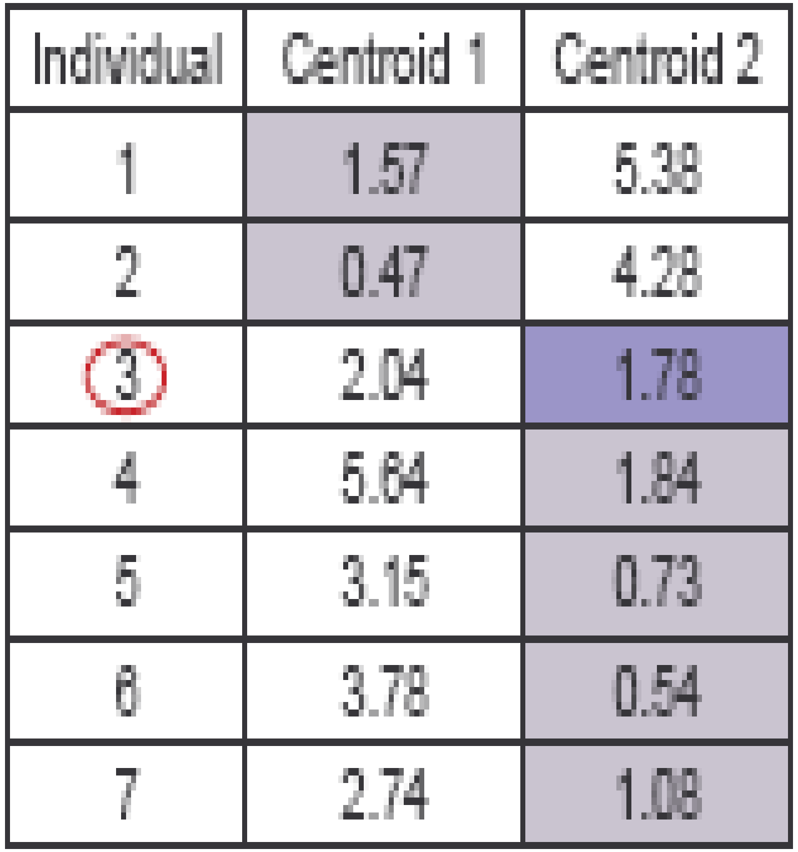
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$d(m_1,2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} = 1.12$$

$$d(m_2,2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

Step 3:

Now using these centroids, we compute the Euclidean distance of each object, as shown in table.



Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

Therefore, the new clusters are:

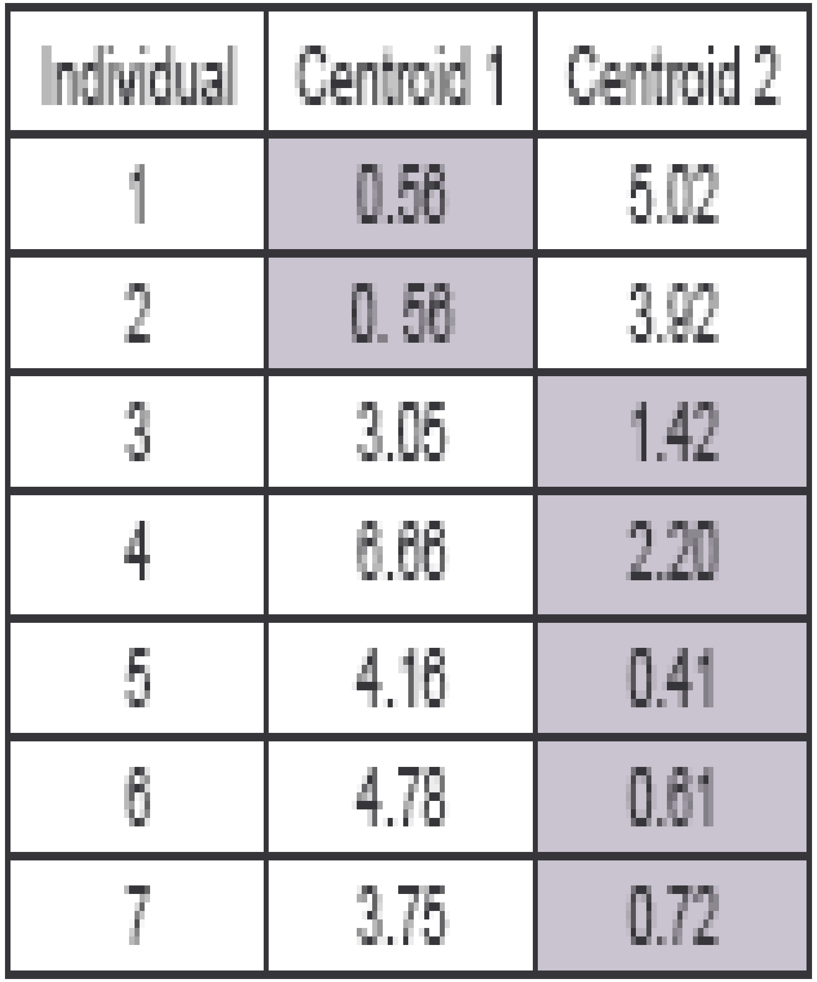
{1,2} and {3,4,5,6,7}

Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

Step 4 :

The clusters obtained are:
{1,2} and {3,4,5,6,7}

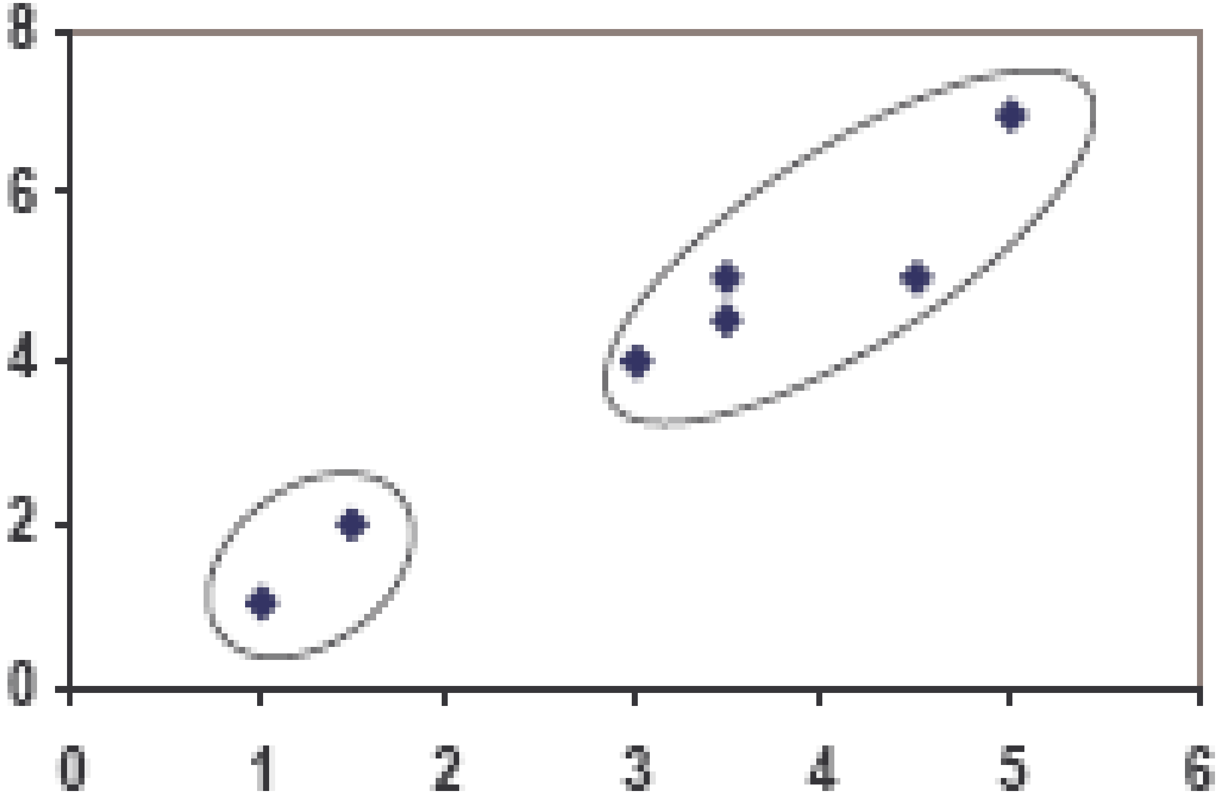
Therefore, there is no change in the cluster.
Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.



Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

Plot:

ML Labs Pvt Ltd



Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

K-Means(Strengths and Weakness):

- Strengths
  - Simple iterative method
  - User provides “K”
- Weaknesses
  - Often too simple → bad results
  - Difficult to guess the correct “K”

K-Means(Bottom Line):

- K-means
  - Easy to use
  - Need to know K
  - May need to scale data
  - Good initial method
- Local optima
  - No guarantee of optimal solution
  - Repeat with different starting values

Business Uses:

The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

Business use cases:

This is a versatile algorithm that can be used for any type of grouping.

Some examples of use cases are:

Behavioural segmentation:

- Segment by purchase history
- Segment by activities on application, website, or platform
- Define personas based on interests
- Create profiles based on activity monitoring

Business use cases:

This is a versatile algorithm that can be used for any type of grouping.

Some examples of use cases are:

Inventory categorization:

- Group inventory by sales activity
- Group inventory by manufacturing metrics

Sorting sensor measurements:

- Detect activity types in motion sensors
- Group images
- Separate audio
- Identify groups in health monitoring

Business use cases:

Detecting bots or anomalies:

- Separate valid activity groups from bots
- Group valid activity to clean up outlier detection

In addition, monitoring if a tracked data point switches between groups over time can be used to detect meaningful changes in the data.

CONCLUSION:

K-means algorithm is useful for undirected knowledge discovery and is relatively simple. K-means has found widespread usage in lot of fields, ranging from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligence, image processing, machine vision, and many others.