CM2604 Machine Learning

Clustering Techniques

Week 06 | Prasan Yapa

ROBERT GORDON UNIVERSITY ABERDEEN
TEF Gold
INFORMATICS INSTITUTE OF TECHNOLOGY

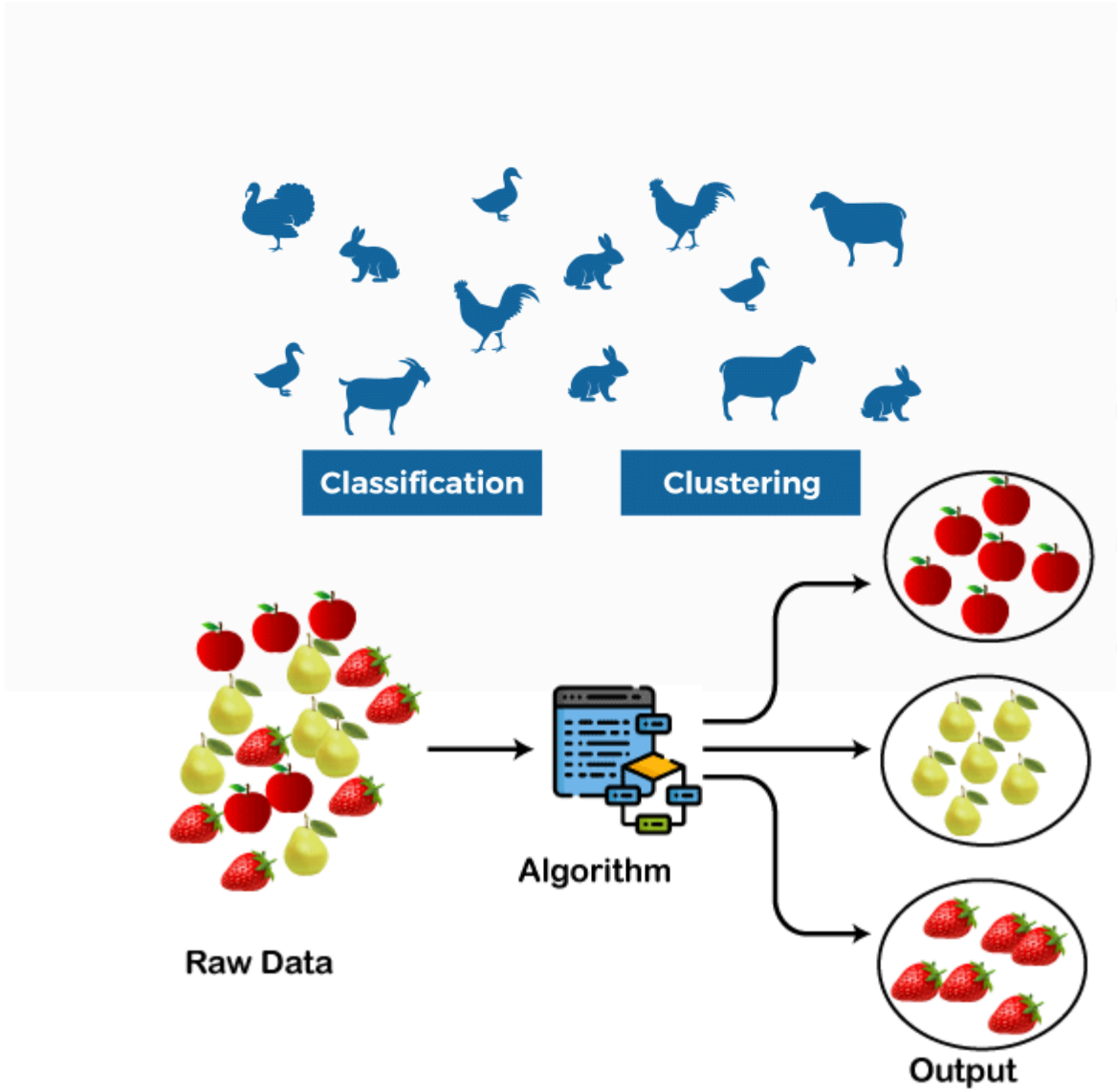
Content

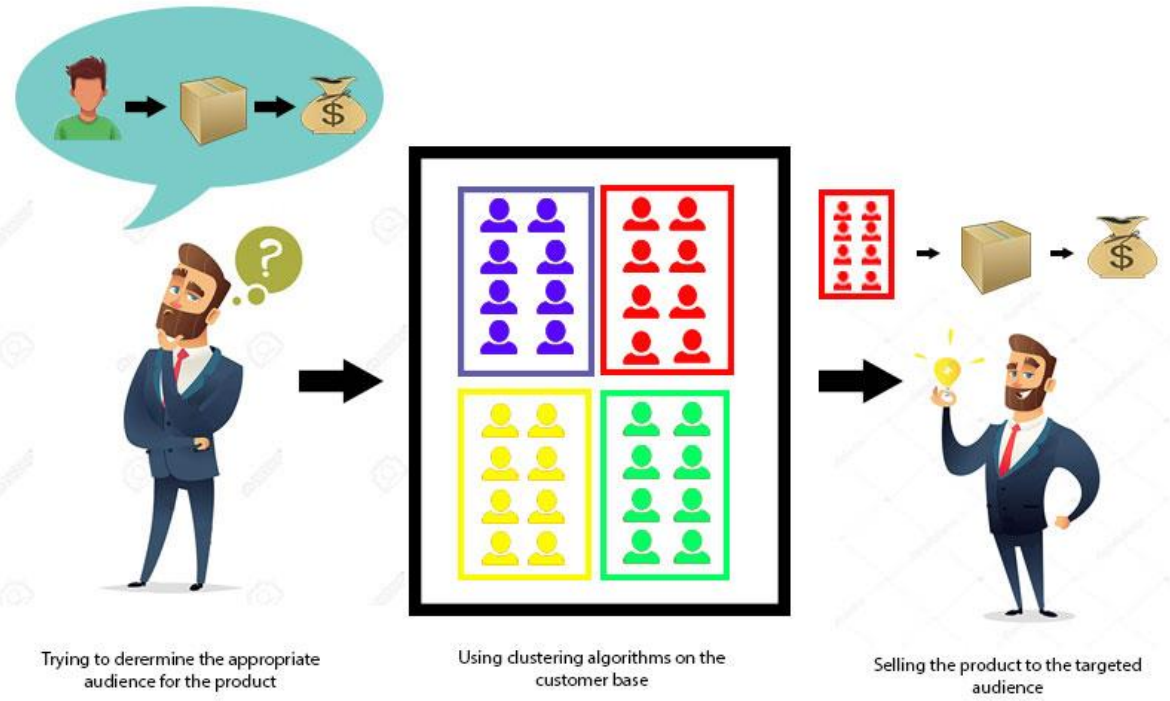
- Clustering in ML
- Classification Vs Clustering
- Use cases
- Notion of a Cluster
- Clustering approaches
- K-means Clustering

Clustering in Machine Learning

- A way of grouping the data points into different clusters, consisting of similar data points.
- Clustering is an unsupervised learning method; hence no supervision is provided to the algorithm.
- The objects with the possible similarities remain in a group that has less or no similarities with another group.
- After applying the clustering technique, each cluster or group is provided with a cluster-ID.

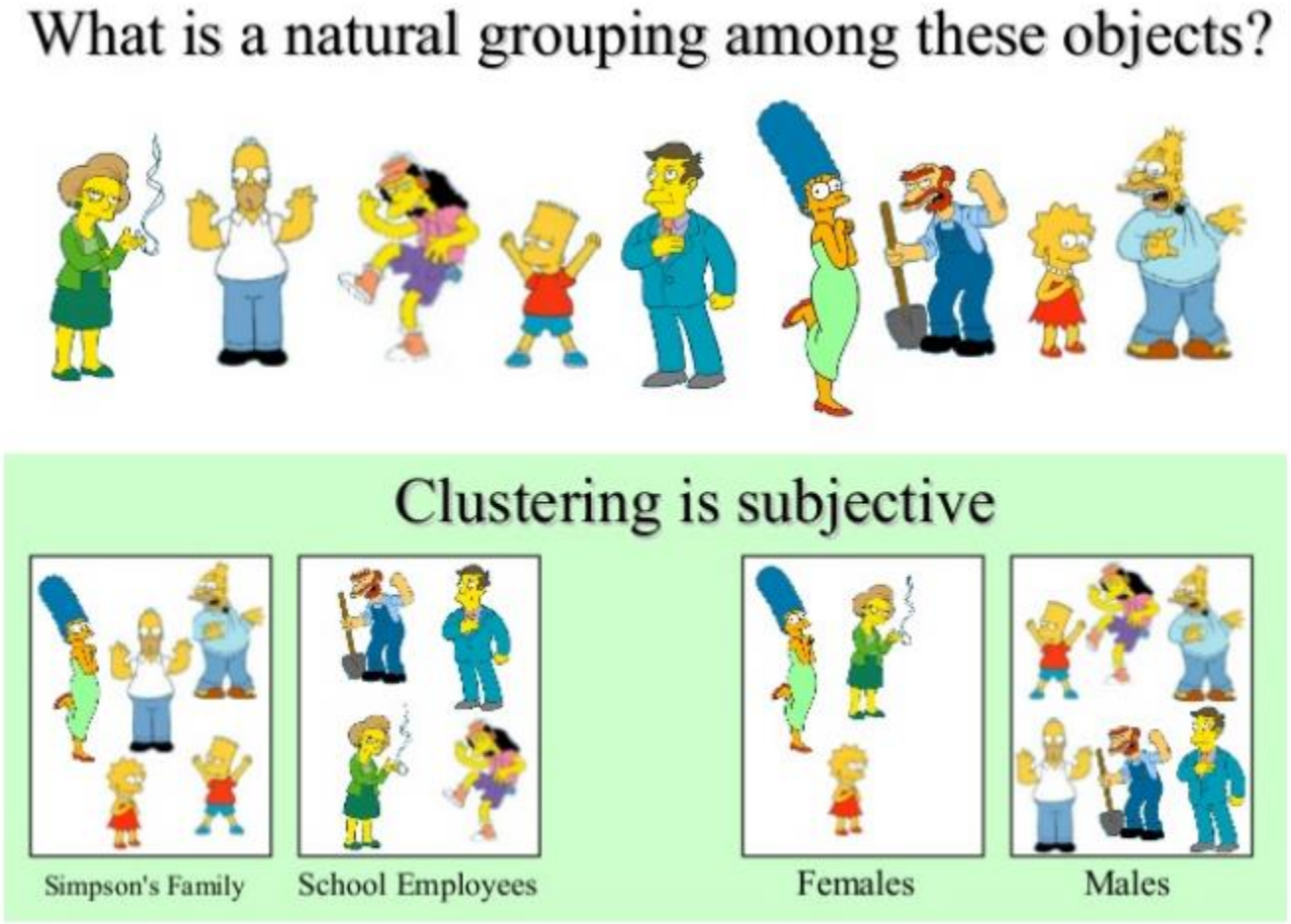
Clustering in Machine Learning



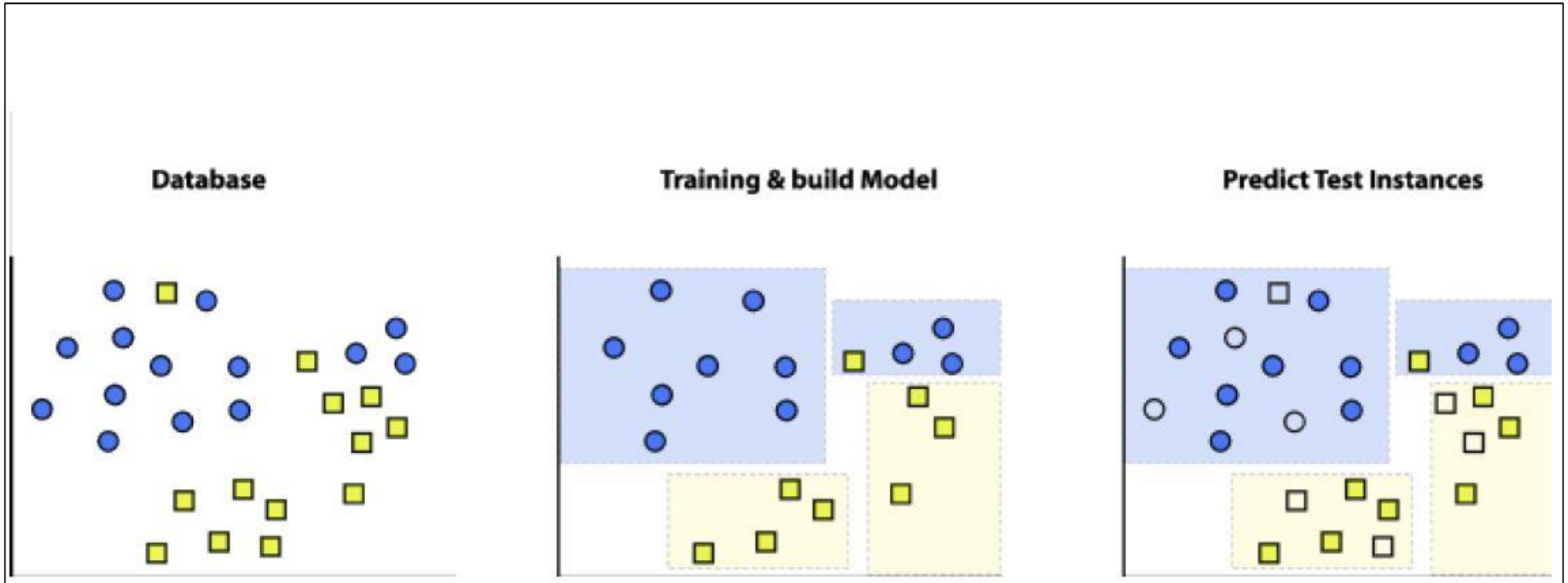


Clustering in Machine Learning

- Market Segmentation
- Statistical data analysis
- Social network analysis
- Image segmentation
- Anomaly detection

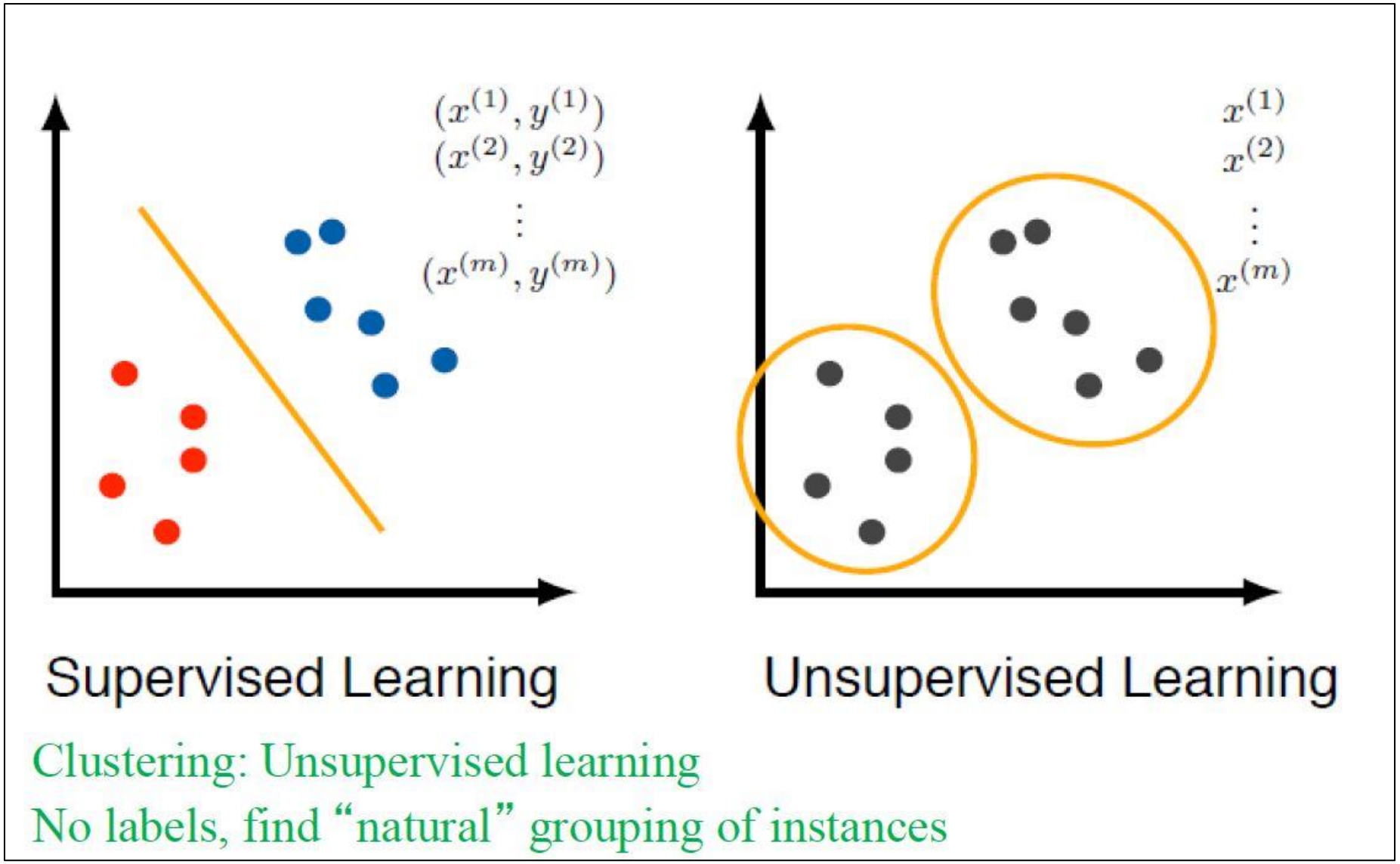


Classification Vs Clustering

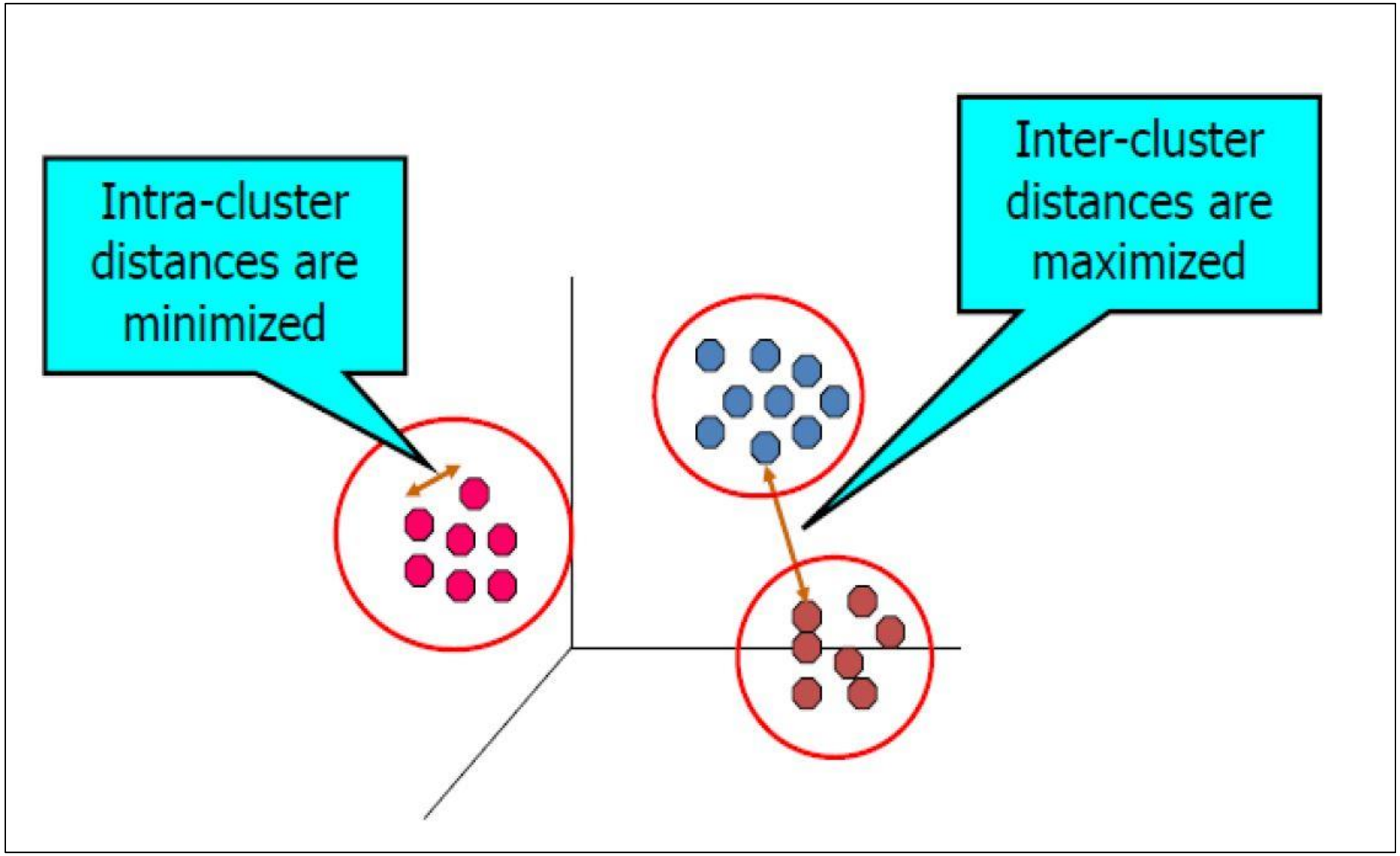


Classification: Supervised learning

Classification Vs Clustering



Intra-cluster Vs Inter-cluster



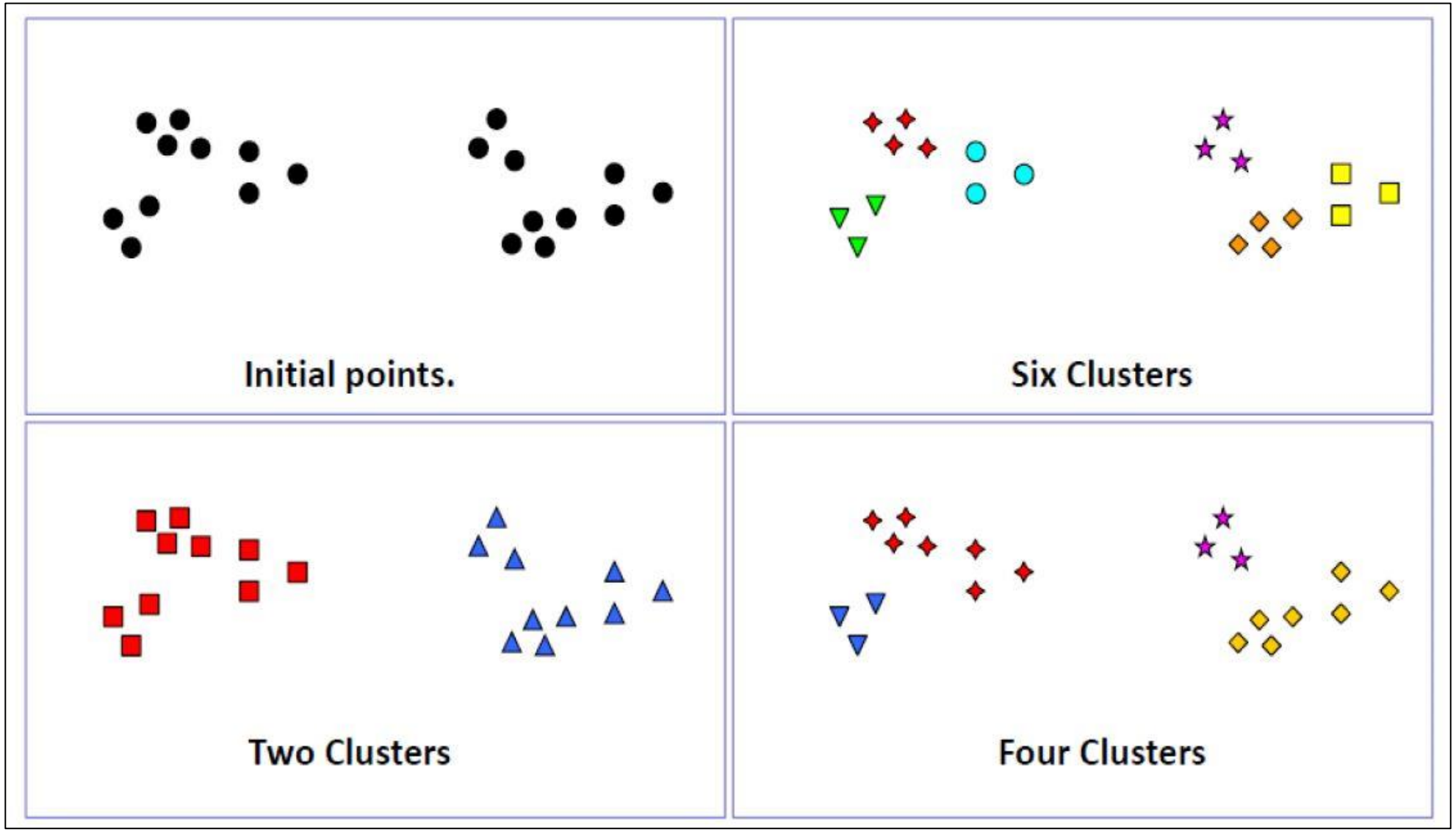
Use cases

- **In Identification of Cancer Cells:** It divides the cancerous and non-cancerous data sets into different groups.
- **In Search Engines:** It does it by grouping similar data objects in one group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.
- **Customer Segmentation:** It is used in market research to segment the customers based on their choice and preferences.
- **In Biology:** It is used in the biology stream to classify different species of plants and animals using the image recognition technique.
- **In Land Use:** The clustering technique is used in identifying the area of similar lands use in the GIS database.

Use cases – Image Recognition

Identify parts of an image that belong to the same object

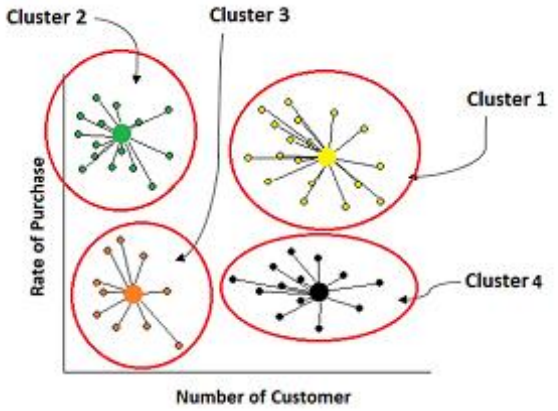
Notion of a Cluster

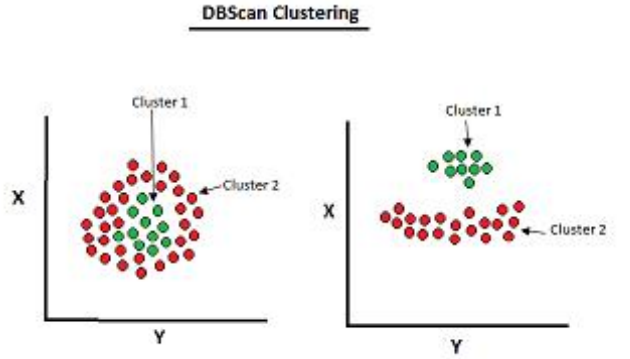


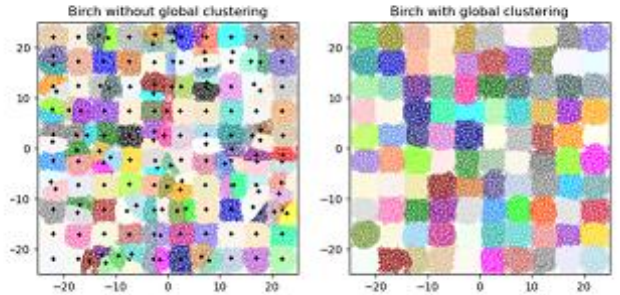
Types of Clustering

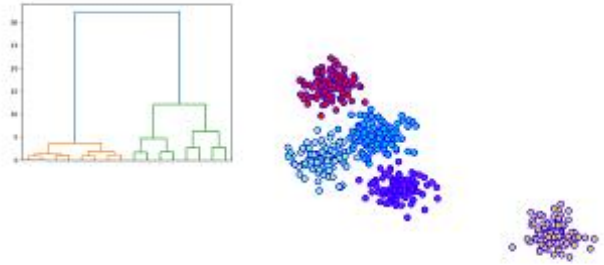
Types of Clustering Methods

- Partitioning Clustering
- Density-Based Clustering
- Distributional Clustering
- Hierarchical Clustering
- Fuzzy Clustering









Partitioning Clustering

- A type of clustering that divides the data into non-hierarchical groups.
- It is also known as the centroid-based method.
- The most common example of partitioning clustering is the **K-Means Clustering** algorithm.
- The dataset is divided into a set of groups, where *K* is used to define the number of pre-defined groups.
- The cluster center is created in such a way that the distance between the data points of one cluster is minimum.

Introduction to K-Means

- Partitioning Clustering Approach
  - a typical clustering analysis approach via iteratively partitioning training data set to learn a partition of the given data space
  - learning a partition on a data set to produce several non-empty clusters (usually, the number of clusters given in advance)
  - in principle, optimal partition achieved via minimising the sum of squared distance (i.e. cost function) to its "representative object" in each cluster

$$E = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} d^2(\mathbf{x}, \mathbf{m}_k)$$

e.g., Euclidean distance

$$d^2(\mathbf{x}, \mathbf{m}_k) = \sum_{n=1}^{N} (x_n - m_{kn})^2$$

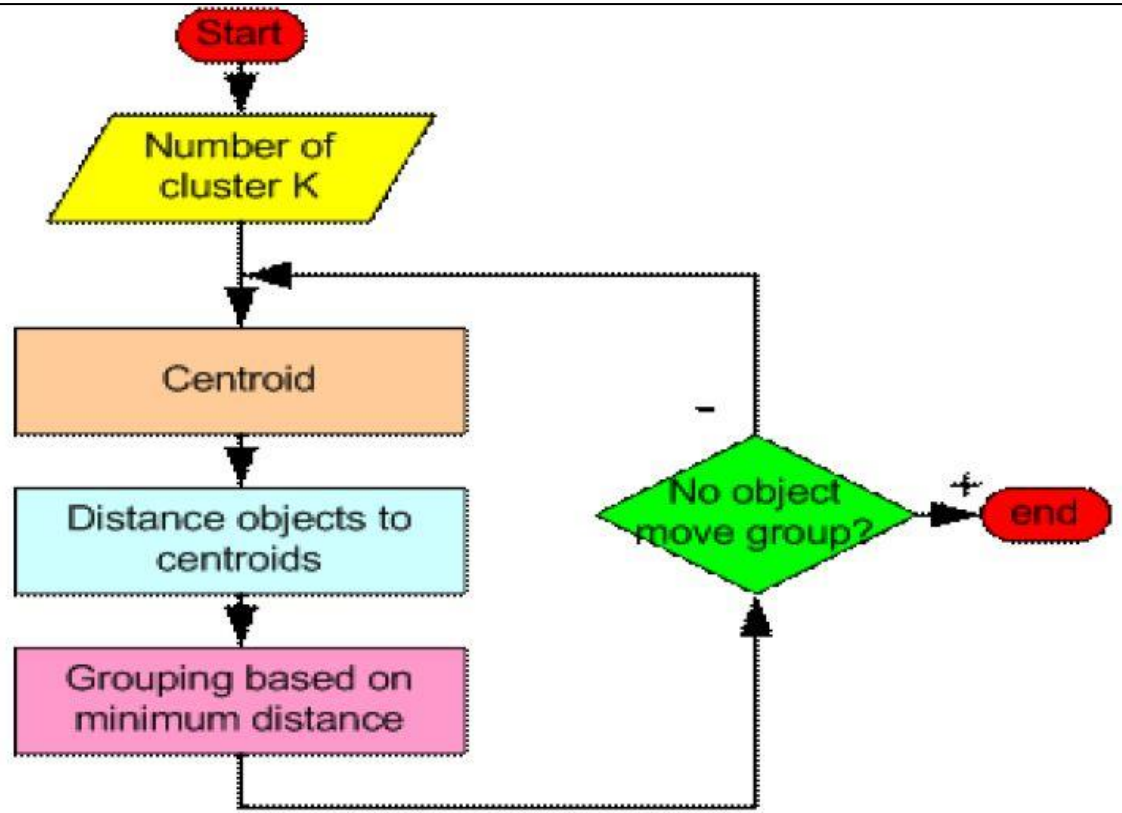
K-Means Algorithm

Given the cluster number *K*, the *K-means* algorithm is carried out in three steps after initialization:

Initialisation: set seed points (randomly)

1) Assign each object to the cluster of the nearest seed point measured with a specific distance metric
2) Compute new seed points as the centroids of the clusters of the current partition (the centroid is the centre, i.e., *mean point*, of the cluster)
3) Go back to Step 1), stop when no more new assignment (i.e., membership in each cluster no longer changes)

How K-Means Clustering algorithm works?



1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Example – Implementation of K-Means (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

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

Step 3:

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

- 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)

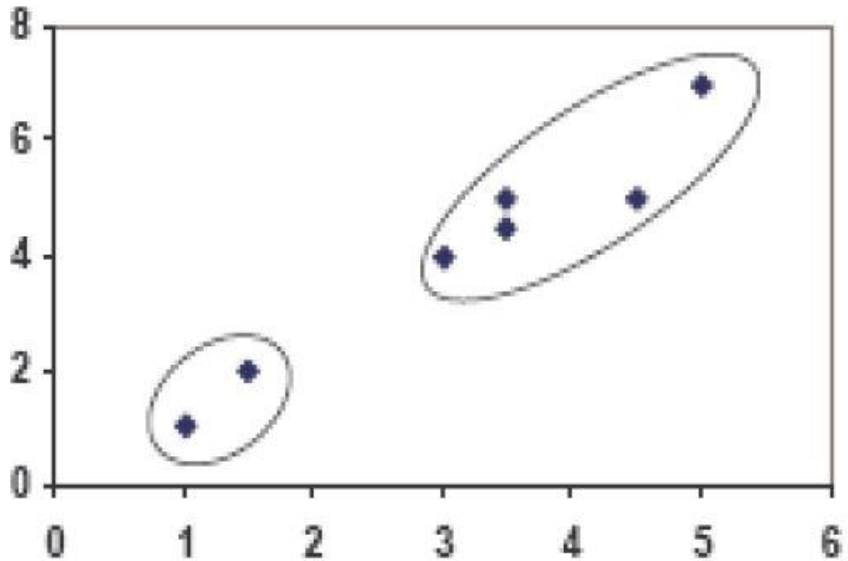
Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

- 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

K-Means Pros & Cons

- Pros
  - Simple, understandable
  - Quick
  - Instances automatically set to clusters
- Cons
  - All instances lead to a single cluster
  - Sensitive to more outliers
  - Cluster must be picked beforehand

Distributional Clustering

- The data is divided based on the probability of how a dataset belongs to a particular distribution.
- The grouping is done by assuming some distributions commonly Gaussian Distribution.
- The example of this type is the **Expectation-Maximization Clustering** algorithm.
- This is based on **Gaussian Mixture Models (GMM).**

Hierarchical Clustering

- In this algorithm, we develop the hierarchy of clusters in the form of a tree.
- This tree-shaped structure is known as the **dendrogram**.
- Sometimes the results of K-means clustering and hierarchical clustering may look similar.
- The hierarchical clustering technique has two approaches: Agglomerative & Divisive.

Questions