Introduction to Clustering

(Hierarchical & Partition Clustering)

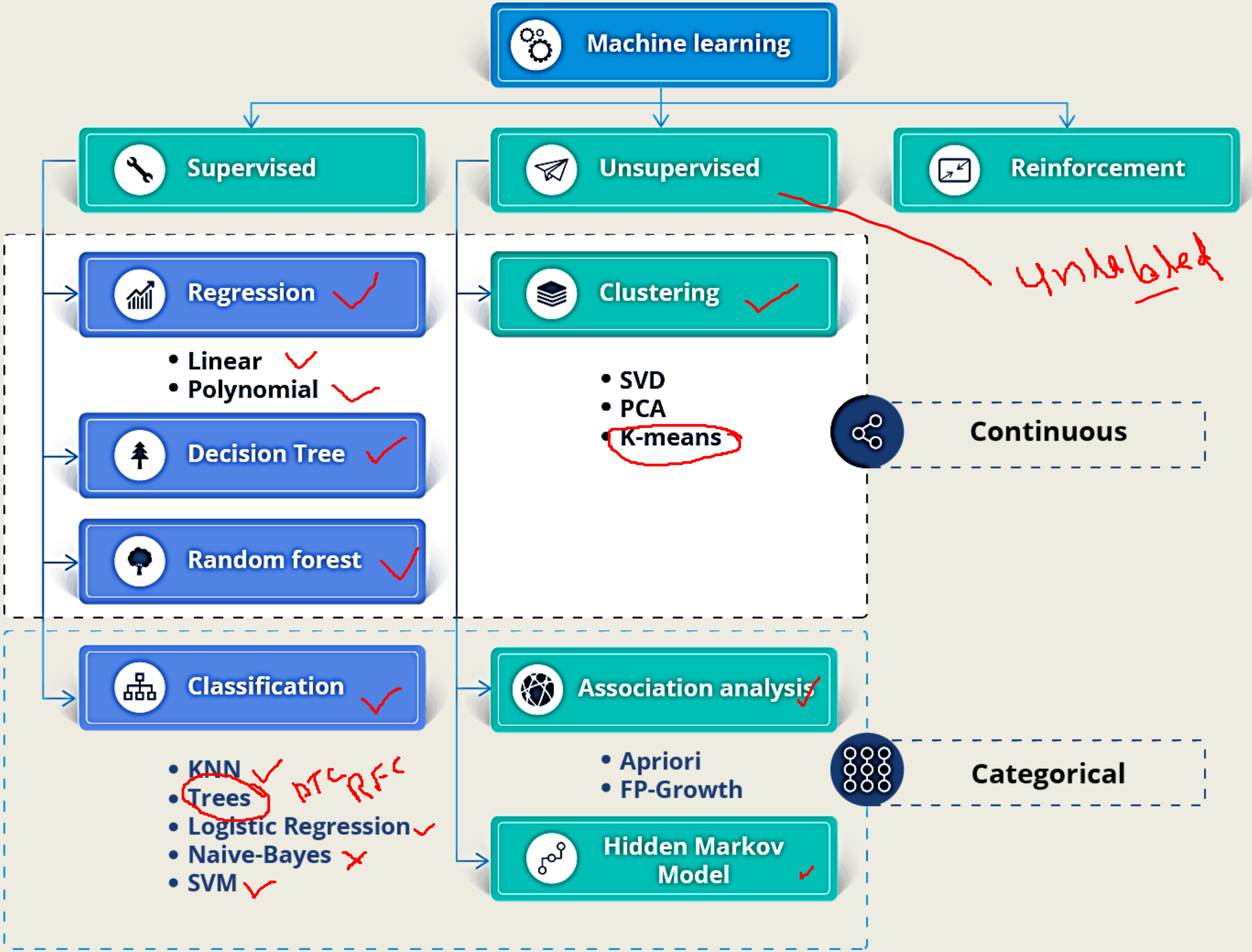
Dr. Virendra Singh Kushwah

Assistant Professor Grade-II

School of Computing Science and Engineering

Virendra.Kushwah@vitbhopal.ac.in

7415869616



Machine learning

Supervised

Unsupervised unlabeled

Reinforcement

Regression ✓

- Linear ✓
- Polynomial ✓

Decision Tree ✓

Random forest ✓

Clustering ✓

- SVD
- PCA
- K-means

Continuous

Classification ✓

- KNN ✓
- Trees DTC RFC
- Logistic Regression ✓
- Naive-Bayes ✗
- SVM ✓

Association analysis ✓

- Apriori
- FP-Growth

Hidden Markov Model ✓

Categorical

- A cluster refers to a small group of objects. Clustering is grouping those objects into clusters. In order to learn clustering, it is important to understand the scenarios that lead to cluster different objects.

What is Clustering?

M C_1

F C_2

- Clustering is dividing data points into homogeneous classes or clusters:

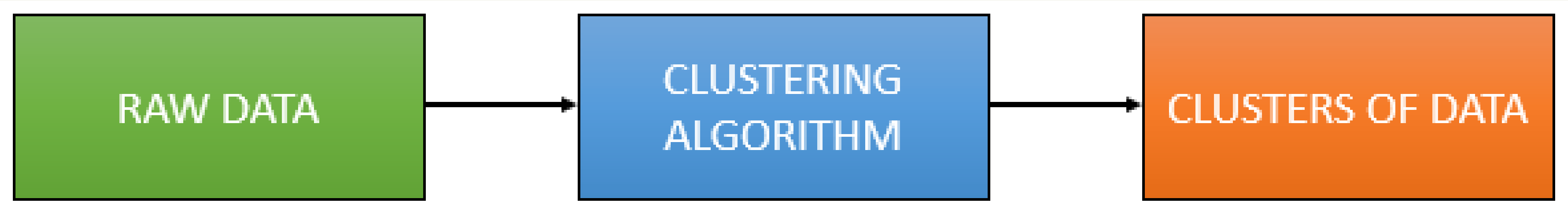
- ✓ Points in the same group are as similar as possible
- ✓ Points in different group are as dissimilar as possible

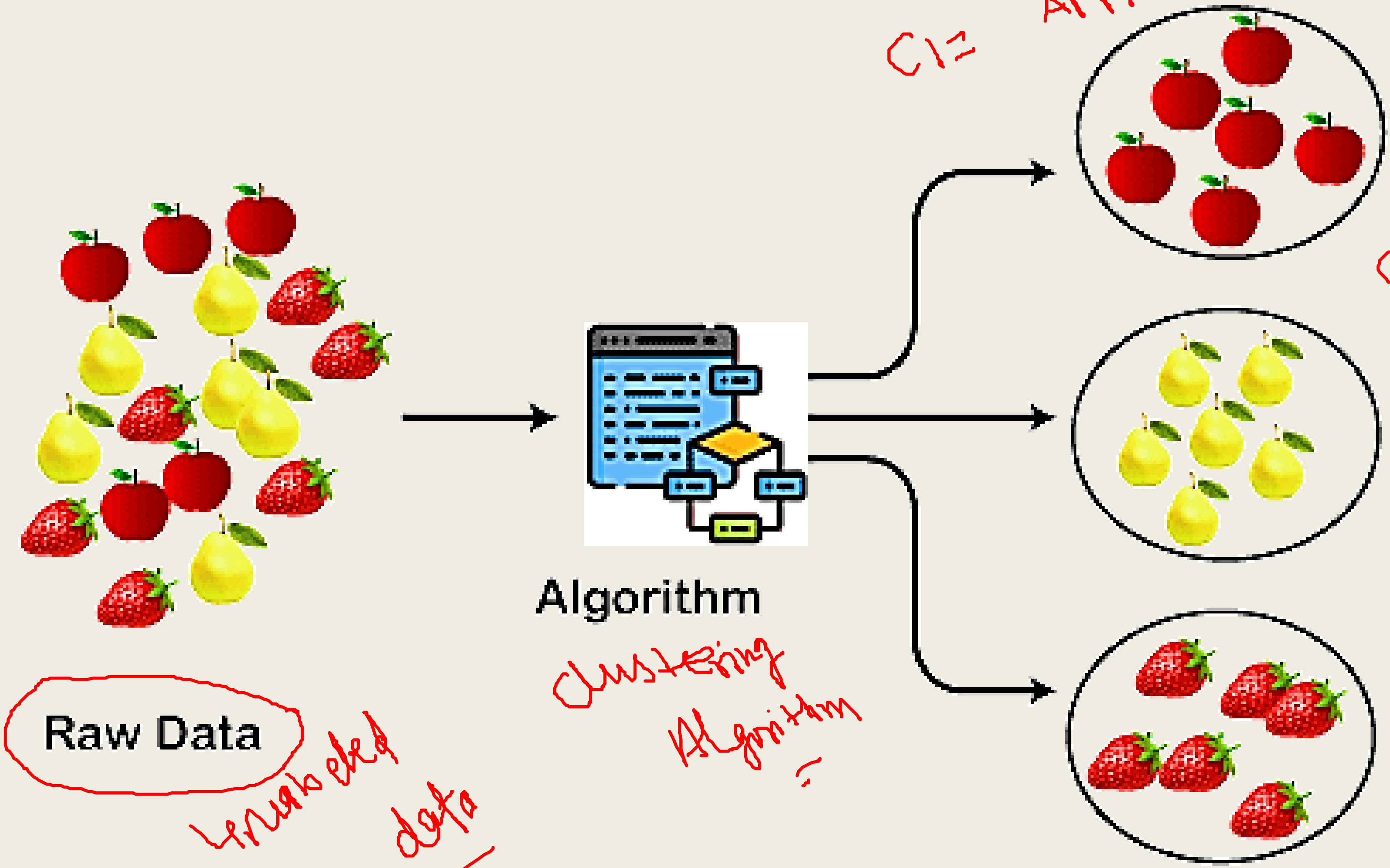
- When a collection of objects is given, we put objects into group based on similarity. Based on Features

Clustering Algorithms

- A Clustering Algorithm tries to analyze natural groups of data on the basis of some similarity. It locates the centroid of the group of data points. To carry out effective clustering, the algorithm evaluates the distance between each point from the centroid of the cluster.

- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. ↳ unsupervised learning





Raw Data

unlabeled data

Algorithm

Clustering Algorithm

Output

C1= Apple

C2= pear

C3= Strawberry

Application of Clustering

- Listed here are few more applications, which would add to what you have learnt.

1. Clustering helps marketers improve their customer base and work on the target areas. It helps group people (according to different criteria’s such as willingness, purchasing power etc.) based on their similarity in many ways related to the product under consideration.
2. Clustering helps in identification of groups of houses on the basis of their value, type and geographical locations.
3. Clustering is used to study earth-quake. Based on the areas hit by an earthquake in a region, clustering can help analyze the next probable location where earthquake can occur.

- Clustering is a type of unsupervised learning mechanism. It basically analyzes the points and clusters them based on similarities and dissimilarities.

Chair ○ R ○ B
○ G

- Applications of Clustering in different fields

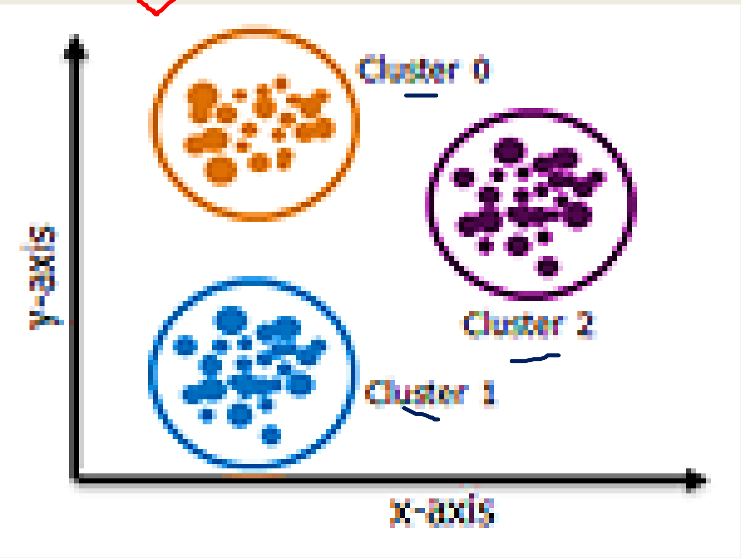
- ✓ 1. Marketing : It can be used to characterize & discover customer segments for marketing purposes.
- ✓ 2. Biology : It can be used for classification among different species of plants and animals.
- ✓ 3. Libraries : It is used in clustering different books on the basis of topics and information.
- ✓ 4. Insurance : It is used to acknowledge the customers, their policies and identifying the frauds.
- ✓ 5. City Planning : It is used to make groups of houses and to study their values based on their geographical locations and other factors present.
- ✓ 6. Earthquake studies : By learning the earthquake affected areas we can determine the dangerous zones.

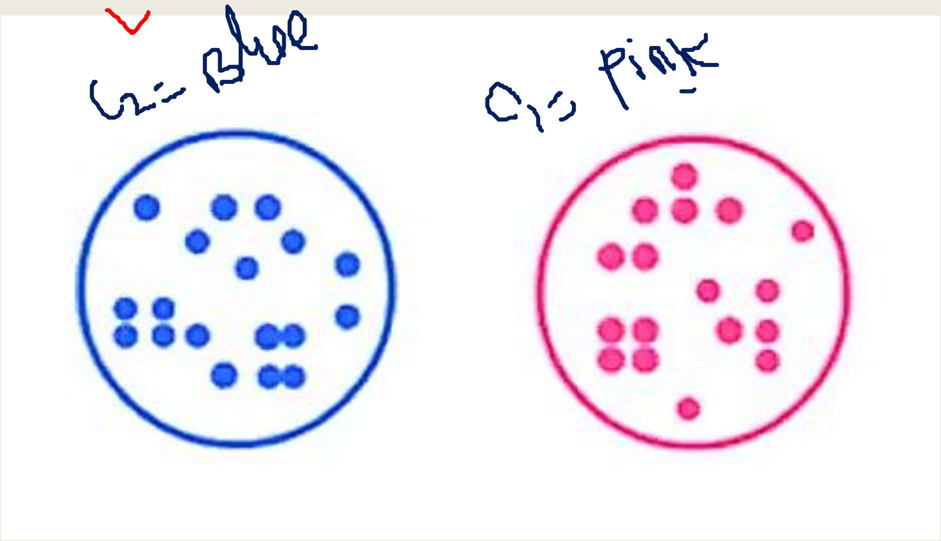
Comparision

Regression	Classification	Clustering
Supervised Learning	Supervised Learning	Unsupervised Learning
Output is a continuous quantity	Output is a categorical quantity	Assigns data points into clusters
Main aim is to forecast or predict	Main aim is to compute the category of the data	Main aim is to group similar items clusters
Eg: Predict stock market price	Eg: Classify emails as spam or non-spam	Eg: Find all transactions which are fraudulent in nature
Algorithm: Linear Regression	Algorithm: Logistic Regression	Algorithm: K-means

Now, once you have understood what is clustering, let's look at different ways to achieve these clusters.

1 • Exclusive Clustering: In exclusive clustering, an item belongs exclusively to one cluster, not several. In the image, you can see that data belonging to cluster 0 does not belong to cluster 1 or cluster 2. k-means clustering is a type of exclusive clustering.

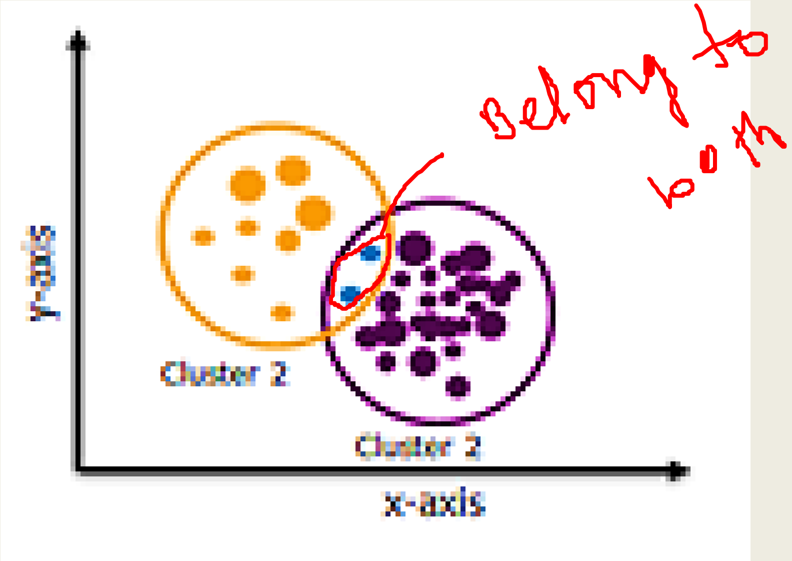




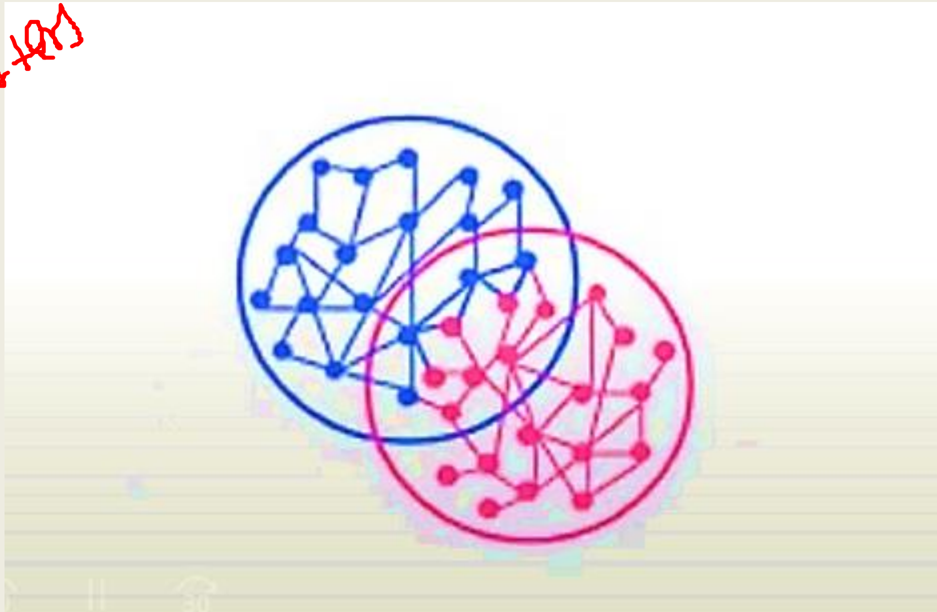
Any DP → C_1

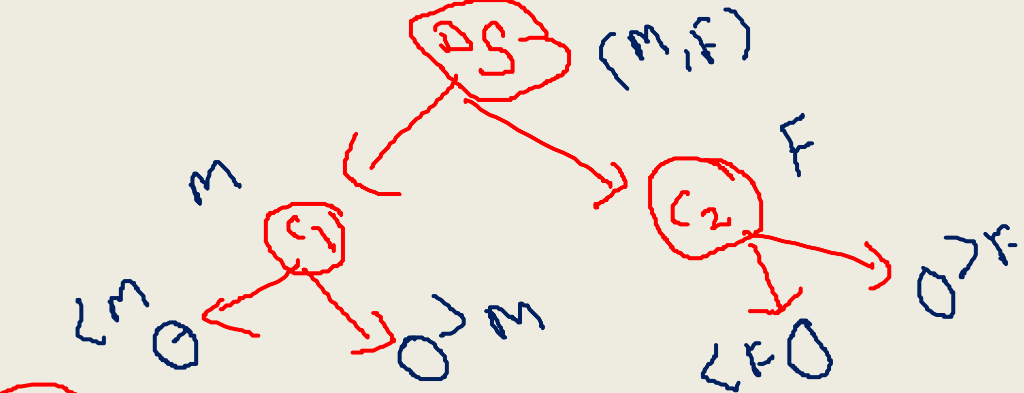
→ C_2 →

2 • Overlapping Clustering: Here, an item can belong to multiple clusters with different degree of association among each cluster. Fuzzy C-means algorithm is based on overlapping clustering.

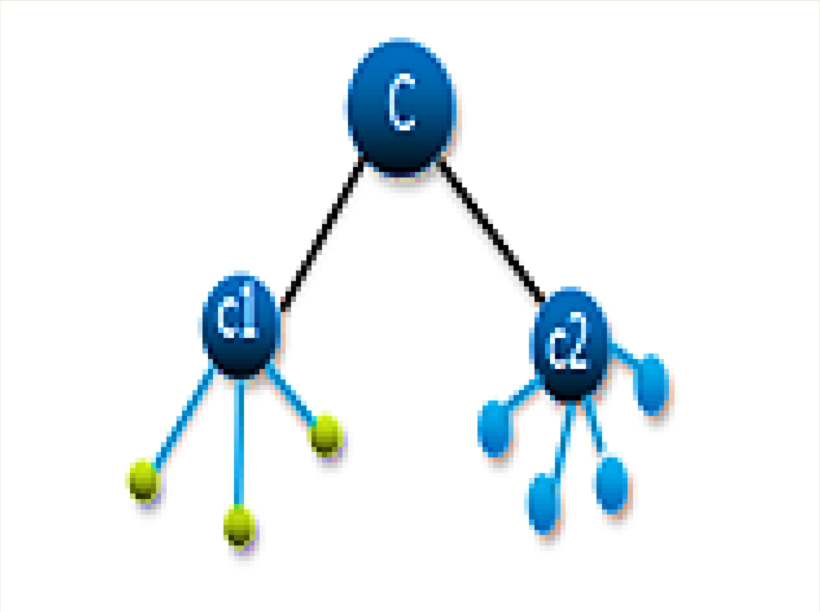


Belong to both clusters





3 • Hierarchical Clustering: In hierarchical clustering, the clusters are not formed in a single step rather it follows series of partitions to come up with final clusters. It looks like a tree as visible in the image.



point

While implementing any algorithm, computational speed and efficiency becomes a very important parameter for end results.

Types of Clustering

- Clustering is a type of unsupervised learning wherein data points are grouped into different sets based on their degree of similarity. There are primarily two categories of clustering:

- Hierarchical clustering ✓
- Partitioning clustering ✓

- Hierarchical clustering is further subdivided into:

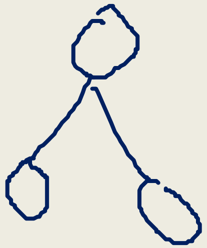
1 - Agglomerative clustering (B-U)
2 - Divisive clustering (T-D)

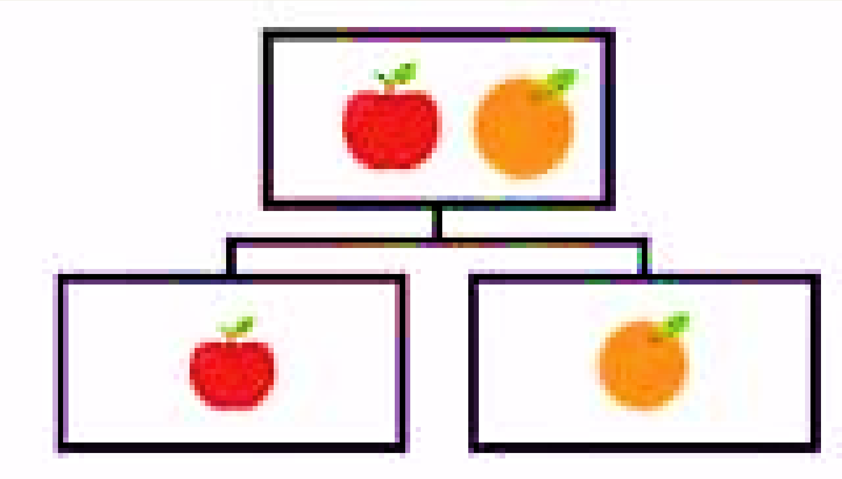
- Partitioning clustering is further subdivided into:

2 - K-Means clustering (Exclusive)
1 - Fuzzy C-Means clustering (Overlapping)

Hierarchical Clustering

- Hierarchical clustering uses a tree-like structure, like so:





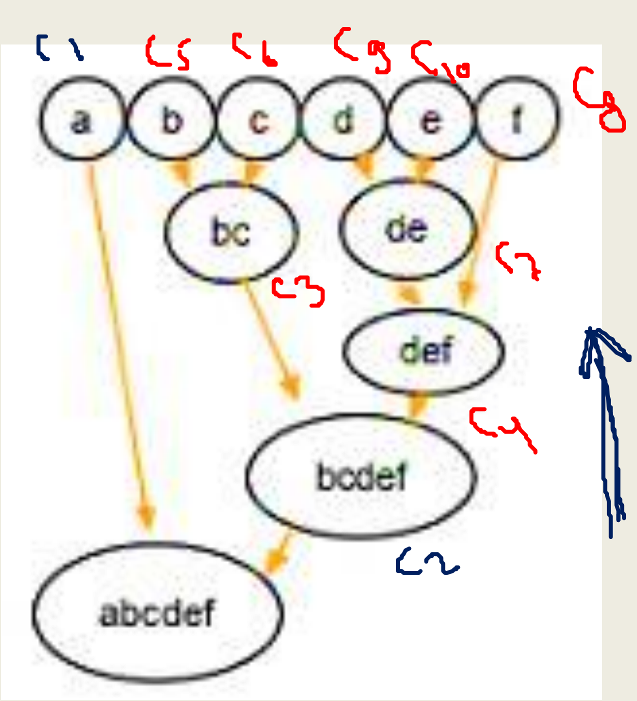
C_1

Apple

C_2

orange

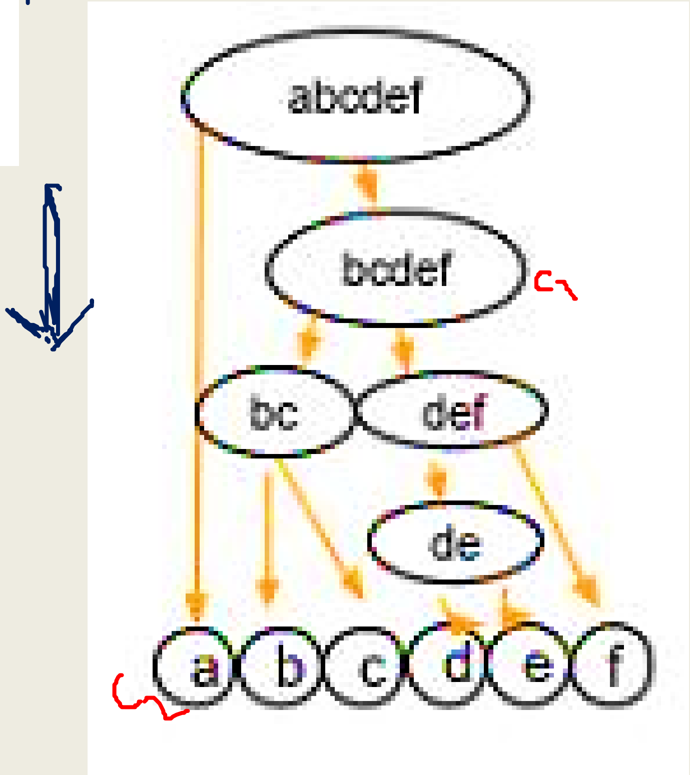
- In agglomerative clustering, there is a bottom-up approach. We begin with each element as a separate cluster and merge them into successively more massive clusters.



clustering will be done by B-U

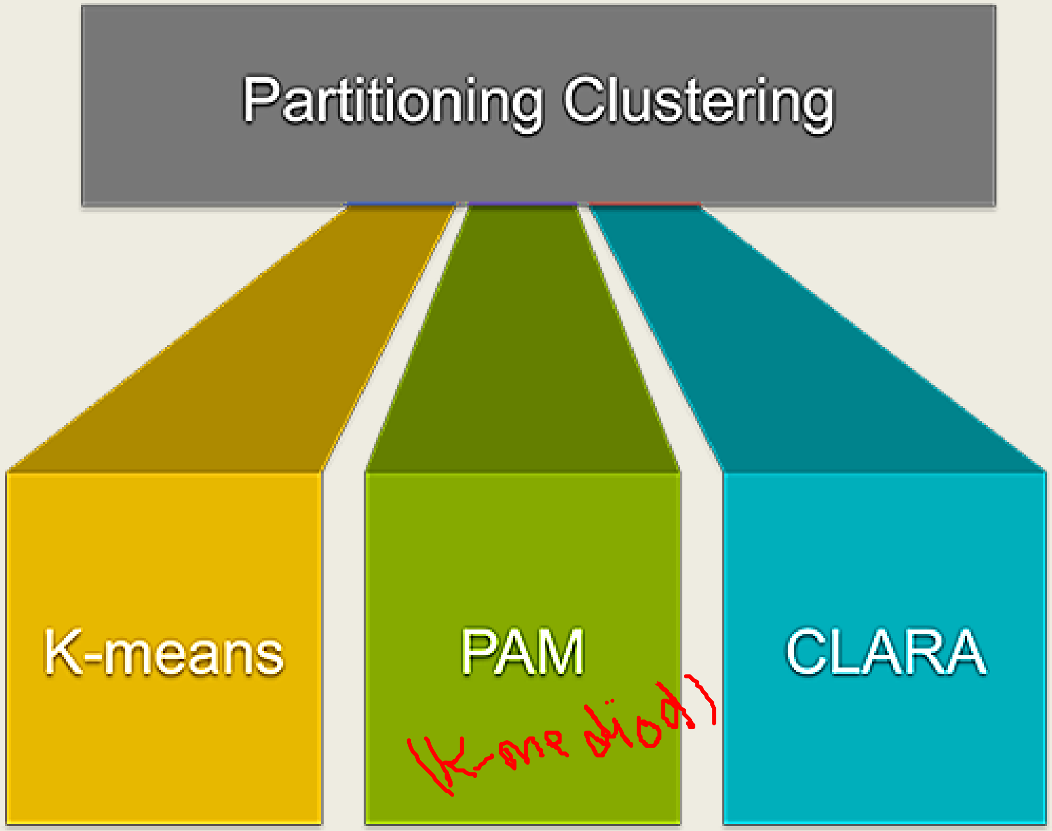
- Divisive clustering is a top-down approach. We begin with the whole set and proceed to divide it into successively smaller clusters.

T-D



What is Partitioning in Clustering?

- The most popular class of clustering algorithms that we have is the iterative relocation algorithms. These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained. There are many algorithms that come under partitioning method some of the popular ones are
- K-Means, ✓
- PAM- partition around medoids (k-Medoid), median
- CLARA algorithm (Clustering Large Applications) etc.



Partitioning Clustering

K-means

PAM

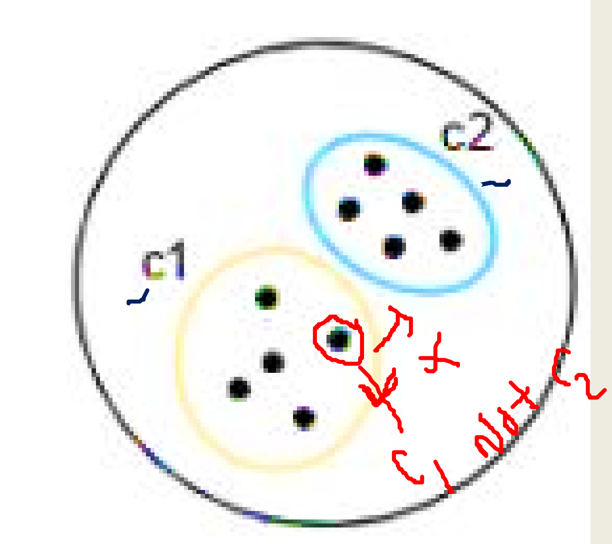
(K-medoid)

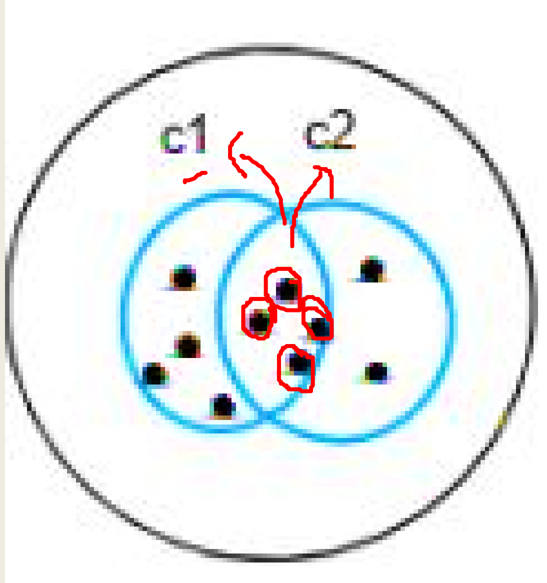
CLARA

we have "K-mode"

Partitioning Clustering

- Partitioning clustering is split into two subtypes - K-Means clustering and Fuzzy C-Means.
- In k-means clustering, the objects are divided into several clusters mentioned by the number 'K.' So if we say K = 2, the objects are divided into two clusters, c1 and c2, as shown:
- Here, the features or characteristics are compared, and all objects having similar characteristics are clustered together.
- Fuzzy c-means is very similar to k-means in the sense that it clusters objects that have similar characteristics together. In k-means clustering, a single object cannot belong to two different clusters. But in c-means, objects can belong to more than one cluster, as shown.





k-means Clustering

- k-means clustering is one of the simplest algorithms which uses unsupervised learning method to solve known clustering issues. k-means clustering require following two inputs.

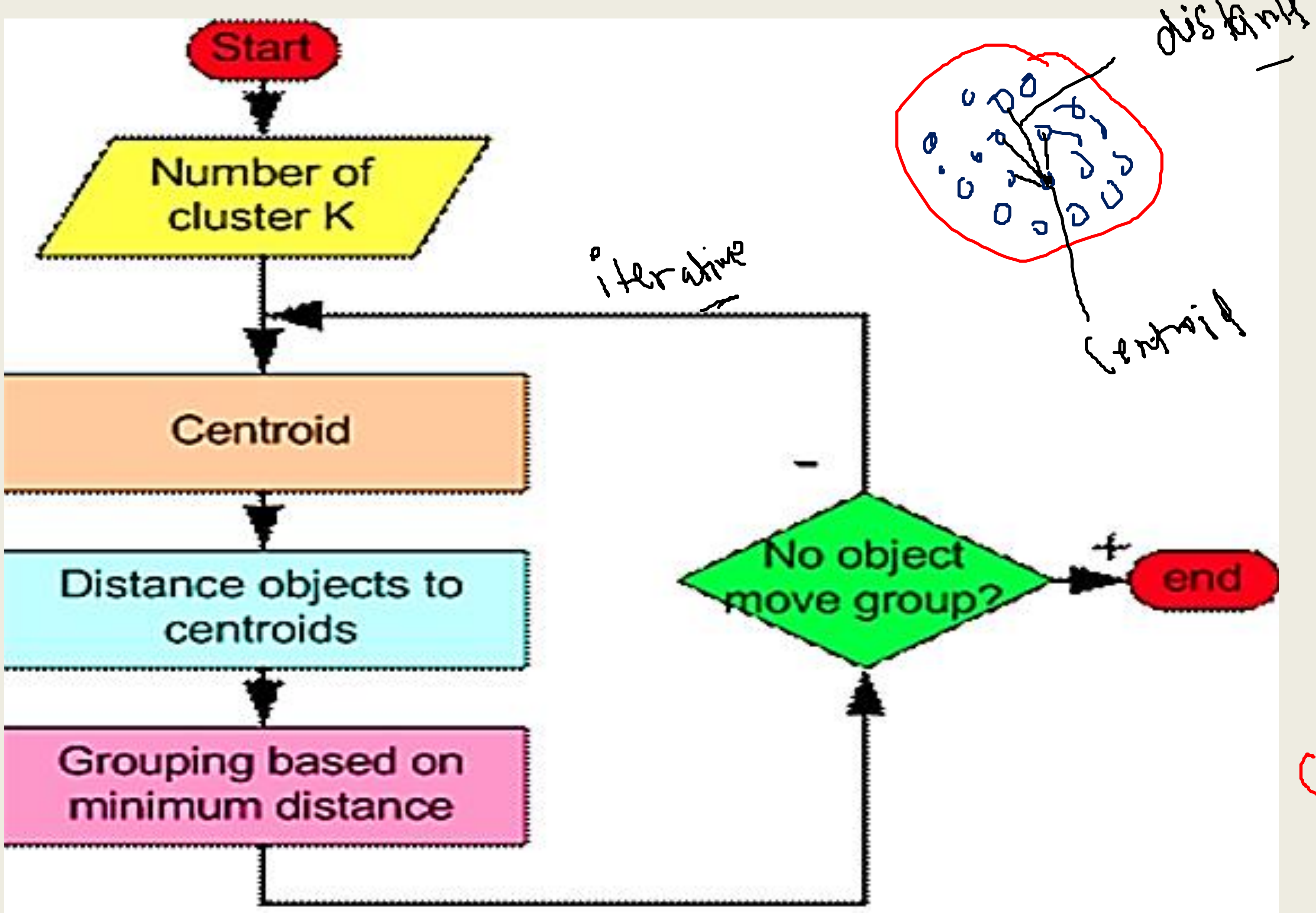
- ✓ k = number of clusters
- ✓ Training set(m) = {x1, x2, x3,……….., xm}

→ Data points

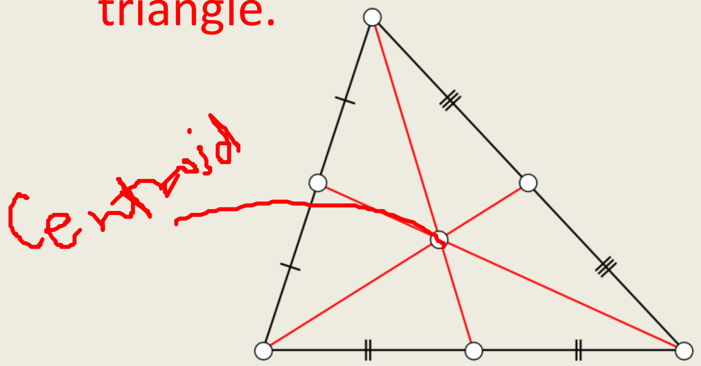
K-Means Algorithm (A centroid based Technique)

- It is one of the most commonly used algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups.
- It classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e., high intra-class similarity), whereas objects from different clusters are as dissimilar as possible (i.e., low inter-class similarity).
- In k-means clustering, each cluster is represented by its center (i.e, centroid) which corresponds to the mean of points assigned to the cluster.
- The basic idea behind k-means clustering consists of defining clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized.

Flowchart of k-means algorithm



The **centroid** is the centre point of the object. The point in which the three medians of the triangle intersect is known as the **centroid** of a triangle. It is also defined as the point of intersection of all the three medians. The median is a line that joins the midpoint of a side and the opposite vertex of the triangle.



Steps involved in K-Means Clustering

- The first step when using k-means clustering is to indicate the number of clusters (k) that will be generated in the final solution.
- The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids.
- Next, each of the remaining objects is assigned to it's closest centroid, where closest is defined using the Euclidean distance between the object and the cluster mean. This step is called "cluster assignment step".

data Point Centroid

- After the assignment step, the algorithm computes the new mean value of each cluster. The term cluster "centroid update" is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means.
- The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until convergence is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration.