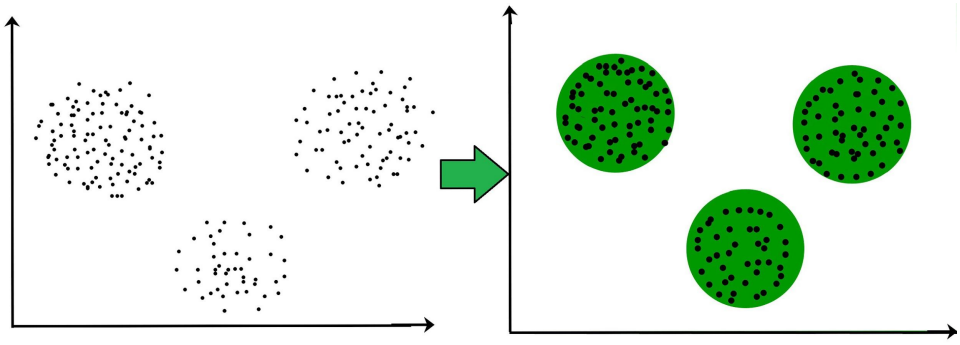
Clustering

- It is a type of unsupervised learning algorithm
- It is a way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.



- It is not necessary for clusters to be spherical.
- The clustering technique can be widely used in various tasks.
    - Market Segmentation
    - Statistical data analysis
    - Social network analysis
- Also used by Amazon for recommendation systems
- Eg: K - means algorithm, DBSCAN

Clustering Distances measures

- The classification of observations into groups requires some methods for computing the distance or the (dis)similarity between each pair of observations.
- There are many methods to calculate this distance information

1. Euclidean Distance

It can be simply explained as the ordinary distance between two points.
Mathematically it computes the root of squared differences between the coordinates between two objects.

1. *Euclidean distance*:

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2. Manhattan Distance

This determines the absolute difference among the pair of the coordinates.

2. *Manhattan distance*:

$$d_{man}(x, y) = \sum_{i=1}^{n} |(x_i - y_i)|$$

Other dissimilarity measures exist such as correlation-based distances, which is widely used for gene expression data analyses.

1. Pearson correlation distance
2. Eisen correlation distance
3. Spearman correlation distance
4. Kendall correlation distance

The clustering methods are broadly divided into two main subgroups

- **Hard Clustering**

In hard clustering, each data point either belongs to a cluster completely or not.

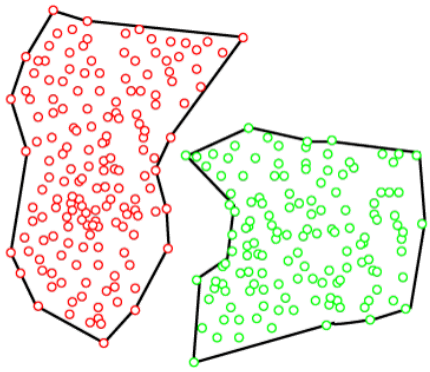
- **Soft Clustering**

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.

Clustering Methods

Density-Based Methods:

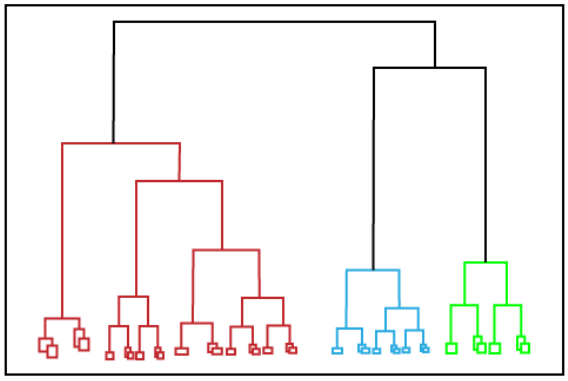
- These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space.
- These methods have good accuracy and the ability to merge two clusters.
- Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), etc.



Hierarchical Based Methods:

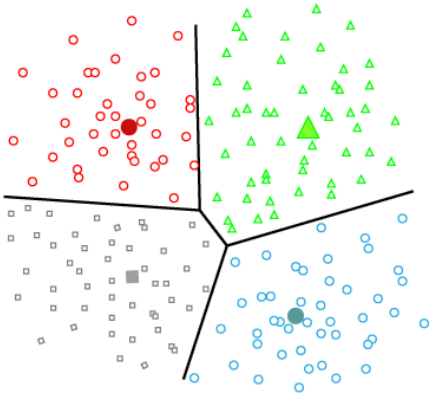
- The clusters formed in this method form a tree-type structure based on the hierarchy.
- New clusters are formed using the previously formed one.
- It is divided into two category

    - Agglomerative (bottom-up approach)
    - Divisive (top-down approach)
- Examples CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies), etc.



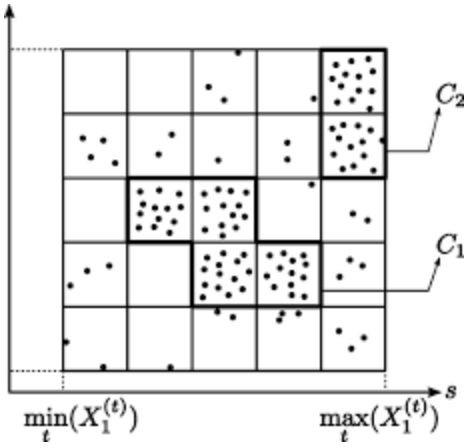
Partitioning Methods:

- These methods partition the objects into k clusters and each partition forms one cluster.
- This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter
- example K-means, CLARANS (Clustering Large Applications based upon Randomized Search), etc.



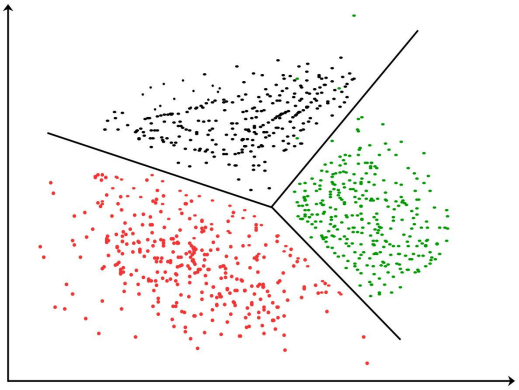
Grid-based Methods:

- In this method, the data space is formulated into a finite number of cells that form a grid-like structure.
- All the clustering operations done on these grids are fast and independent of the number of data objects
- example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest), etc.



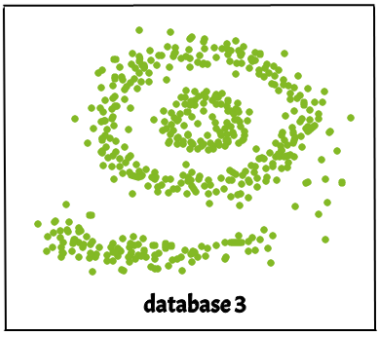
K- means clustering

- It is the simplest unsupervised learning algorithm that solves clustering problems.
- K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.
- It is fast with fewer computations required, with the linear complexity of O(n).



DBSCAN

- It stands for Density-Based Spatial Clustering of Applications with Noise.
- It is an example of a density-based model similar to the mean-shift, but with some remarkable advantages.
- In this algorithm, the areas of high density are separated by the areas of low density.
- Because of this, the clusters can be found in any arbitrary shape.



Applications of Clustering in different fields

- **Marketing**: It can be used to characterize & discover customer segments for marketing purposes.
- **Biology**: It can be used for classification among different species of plants and animals.
- **Libraries**: It is used in clustering different books on the basis of topics and information.
- **Insurance**: It is used to acknowledge the customers, their policies and identifying the frauds.

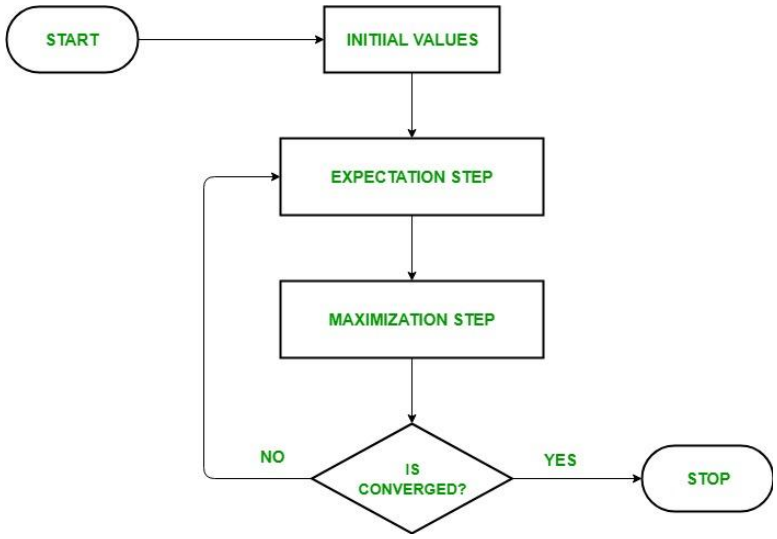
Expectation-Maximization clustering

- It is an unsupervised clustering Algorithm
- It is an iterative method to find (local) maximum likelihood in statistical models, where the model depends on unobserved latent variables.
- This algorithm is actually at the base of many unsupervised clustering algorithms in the field of machine learning.

Algorithm

Given a set of incomplete data, consider a set of starting parameters.

- **Expectation step** (E – step): Using the observed available data of the dataset, estimate (guess) the values of the missing data.
- **Maximization step** (M – step): Complete data generated after the expectation (E) step is used in order to update the parameters.
- Repeat step 2 and step 3 until convergence.



The essence of Expectation-Maximization algorithm is to use the available observed data of the dataset to estimate the missing data and then using that data to update the values of the parameters.