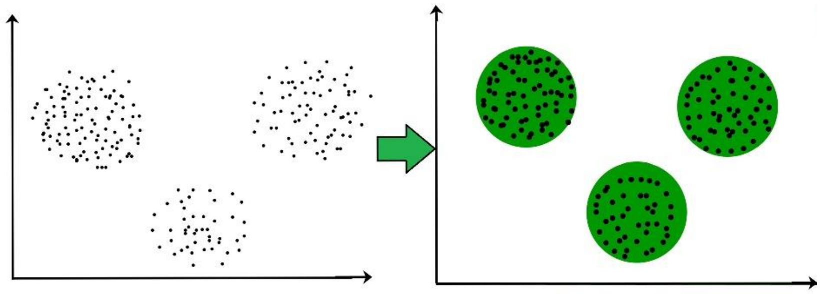
Clustering in Machine Learning

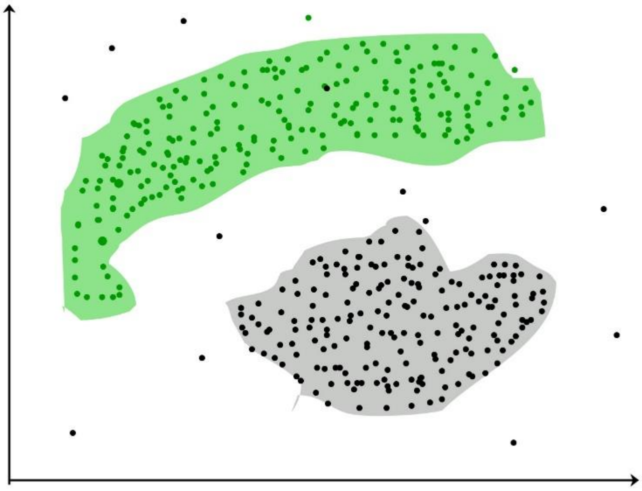
Introduction to Clustering

It is basically a type of unsupervised learning method . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.
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 and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



It is not necessary for clusters to be a spherical. Such as :



Why Clustering ?

Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for a good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions which constitute the similarity of points and each assumption make different and equally valid clusters.

Clustering Methods :

- **Density-Based Methods :** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and 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 forms 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 operation 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.

Measuring Clustering Performance

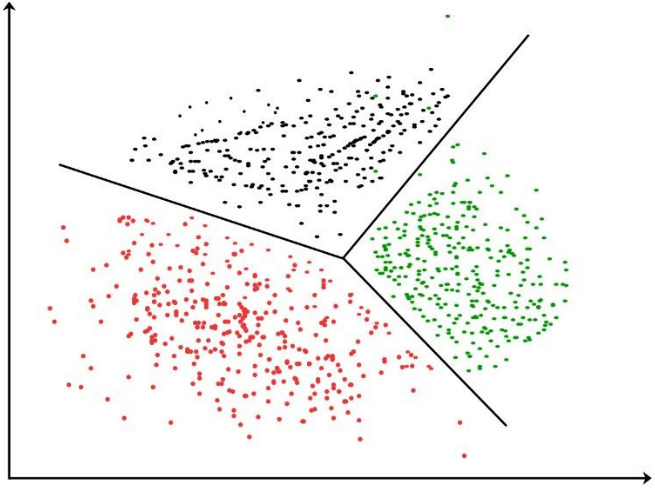
One of the most important consideration regarding ML model is assessing its performance or you can say model’s quality. In case of supervised learning algorithms, assessing the quality of our model is easy because we already have labels for every example.

On the other hand, in case of unsupervised learning algorithms we are not that much blessed because we deal with unlabelled data. But still we have some metrics that give the practitioner an insight about the happening of change in clusters depending on algorithm.

Before we deep dive into such metrics, we must understand that these metrics only evaluates the comparative performance of models against each other rather than measuring the validity of the model’s prediction. Followings are some of the metrics that we can deploy on clustering algorithms to measure the quality of model.

Clustering Algorithms :

K-means clustering algorithm – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster .



Mean-Shift Algorithm

It is another powerful clustering algorithm used in unsupervised learning. Unlike K-means clustering, it does not make any assumptions hence it is a non-parametric algorithm.

Hierarchical Clustering

It is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics.

Applications of Clustering

We can find clustering useful in the following areas −

Data summarization and compression − Clustering is widely used in the areas where we require data summarization, compression and reduction as well. The examples are image processing and vector quantization.

Collaborative systems and customer segmentation − Since clustering can be used to find similar products or same kind of users, it can be used in the area of collaborative systems and customer segmentation.

Serve as a key intermediate step for other data mining tasks − Cluster analysis can generate a compact summary of data for classification, testing, hypothesis generation; hence, it serves as a key intermediate step for other data mining tasks also.

Trend detection in dynamic data − Clustering can also be used for trend detection in dynamic data by making various clusters of similar trends.

Social network analysis − Clustering can be used in social network analysis. The examples are generating sequences in images, videos or audios.

Biological data analysis − Clustering can also be used to make clusters of images, videos hence it can successfully be used in biological data analysis**.**