**Clustering in Machine Learning**

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as ***"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 does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

It is an [unsupervised learning](https://www.javatpoint.com/unsupervised-machine-learning) method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.

After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML systems can use this id to simplify the processing of large and complex datasets.

The clustering technique is commonly used for **statistical data analysis.**

Note: Clustering is somewhere similar to the [classification algorithm](https://www.javatpoint.com/classification-algorithm-in-machine-learning), but the difference is the type of dataset that we are using. In classification, we work with the labeled data set, whereas in clustering, we work with the unlabelled dataset.

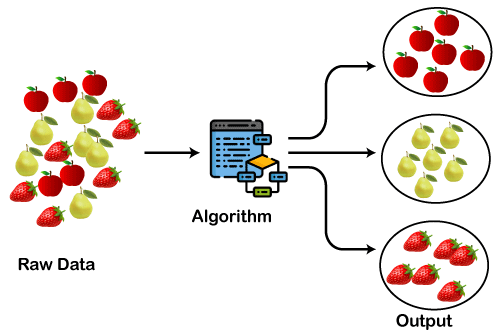
**Example**: Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together. T-shirts are grouped in one section, and trousers are in other sections. Similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily find out the things. The clustering technique also works in the same way. Other examples of clustering are grouping documents according to the topic.

The clustering technique can be widely used in various tasks. Some most common uses of this technique are:

* Market Segmentation
* Statistical data analysis
* Social network analysis
* Image segmentation
* Anomaly detection, etc.

Apart from these general usages, it is used by the **Amazon** in its recommendation system to provide the recommendations as per the past search of products. **Netflix** also uses this technique to recommend the movies and web-series to its users as per the watch history.

The below diagram explains the working of the clustering algorithm. We can see the different fruits are divided into several groups with similar properties.



Types of Clustering Methods

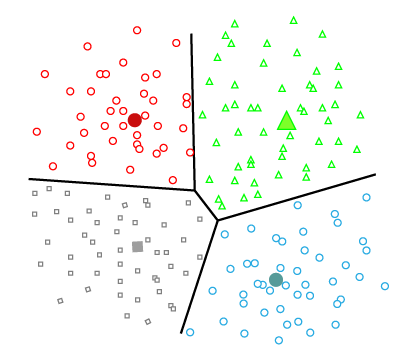
The clustering methods are broadly divided into **Hard clustering** (datapoint belongs to only one group) and **Soft Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

1. **Partitioning Clustering**
2. **Density-Based Clustering**
3. **Distribution Model-Based Clustering**
4. **Hierarchical Clustering**
5. **Fuzzy Clustering**

***Partitioning Clustering***

It is 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**](https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning).

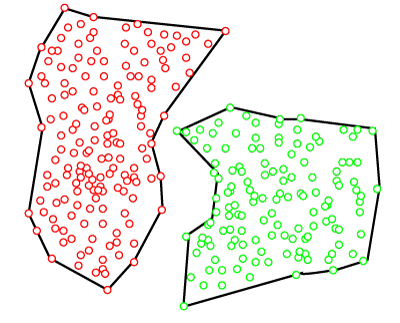
In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.



***Density-Based Clustering***

The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.

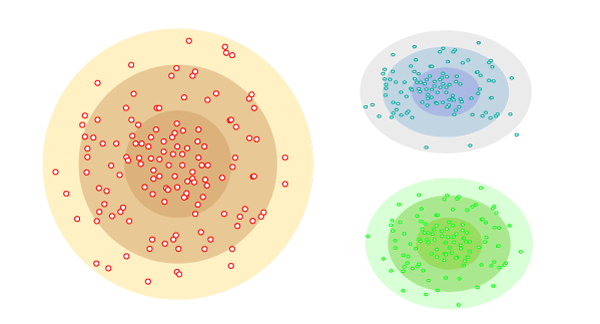
These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.



**Distribution Model-Based Clustering**

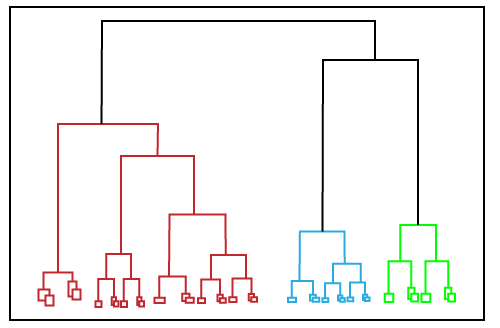
In the distribution model-based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution. The grouping is done by assuming some distributions commonly **Gaussian Distribution**.

The example of this type is the **Expectation-Maximization Clustering algorithm** that uses Gaussian Mixture Models (GMM).



Hierarchical Clustering

Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative Hierarchical algorithm**.



***Fuzzy Clustering***

[Fuzzy](https://www.javatpoint.com/fuzzy-logic) clustering is a type of soft method in which a data object may belong to more than one group or cluster. Each dataset has a set of membership coefficients, which depend on the degree of membership to be in a cluster. **Fuzzy C-means algorithm** is the example of this type of clustering; it is sometimes also known as the Fuzzy k-means algorithm.

***Clustering Algorithms***

The Clustering algorithms can be divided based on their models that are explained above. There are different types of clustering algorithms published, but only a few are commonly used. The clustering algorithm is based on the kind of data that we are using. Such as, some algorithms need to guess the number of clusters in the given dataset, whereas some are required to find the minimum distance between the observations of the dataset.

Here we are discussing mainly popular Clustering algorithms that are widely used in machine learning:

1. **K-Means algorithm:** The k-means algorithm is one of the most popular clustering algorithms. It classifies the dataset by dividing the samples into different clusters of equal variances. The number of clusters must be specified in this algorithm. It is fast with fewer computations required, with the linear complexity of **O(n).**
2. **Mean-shift algorithm:** Mean-shift algorithm tries to find the dense areas in the smooth density of data points. It is an example of a centroid-based model that works on updating the candidates for centroid to be the center of the points within a given region.
3. **DBSCAN Algorithm:** 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.
4. **Expectation-Maximization Clustering using GMM:** This algorithm can be used as an alternative for the k-means algorithm or for those cases where K-means can be failed. In GMM, it is assumed that the data points are Gaussian distributed.
5. **Agglomerative Hierarchical algorithm:** The Agglomerative hierarchical algorithm performs the bottom-up hierarchical clustering. In this, each data point is treated as a single cluster at the outset and then successively merged. The cluster hierarchy can be represented as a tree-structure.
6. **Affinity Propagation:** It is different from other clustering algorithms as it does not require to specify the number of clusters. In this, each data point sends a message between the pair of data points until convergence. It has O(N2T) time complexity, which is the main drawback of this algorithm.

Applications of Clustering

Below are some commonly known applications of clustering technique in Machine Learning:

* **In Identification of Cancer Cells:** The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.
* **In Search Engines:** Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. 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. This can be very useful to find that for what purpose the particular land should be used, that means for which purpose it is more suitable.

# Clustering Evaluation strategies

Clustering is an unsupervised machine learning algorithm. It helps in clustering data points to groups. Validating the clustering algorithm is a bit tricky compared to supervised machine learning algorithms as the clustering process does not contain ground truth labels. If one wants to do clustering with ground truth labels being present, validation methods and metrics of supervised machine learning algorithms can be used. This blog post tries to address evaluation strategies when ground truth labels are not known.

**How Clustering can be evaluated?**

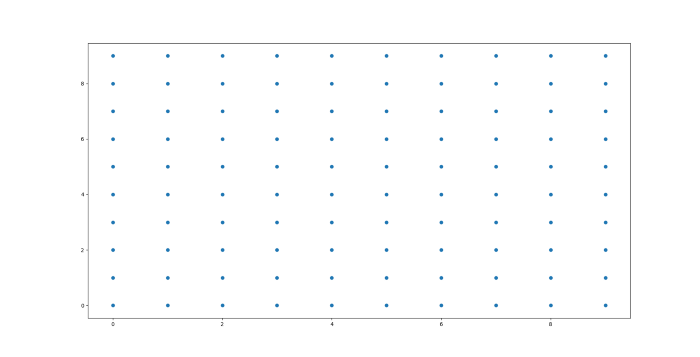
Three important factors by which clustering can be evaluated are

*(a) Clustering tendency (b) Number of clusters,****k****(c) Clustering quality*

# **Clustering tendency**

Before evaluating the clustering performance, making sure that the data set we are working with has clustering tendency and does not contain uniformly distributed points is very important. If the data does not contain clustering tendency, then clusters identified by any state of the art clustering algorithms may be irrelevant. Non-uniform distribution of points in the data set becomes important in clustering.

To solve this, Hopkins test, a statistical test for spatial randomness of a variable, can be used to measure the probability of data points generated by uniform data distribution.



Plot for data from Uniform distribution

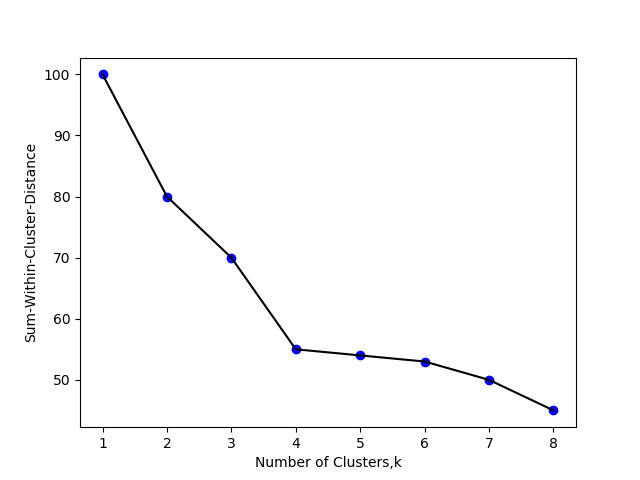
***Null Hypothesis (Ho) :****Data points are generated by uniform distribution (implying no meaningful clusters)****Alternate Hypothesis (Ha):****Data points are generated by random data points (presence of clusters)*  
If H>0.5, null hypothesis can be rejected and it is very much likely that data contains clusters. If H is more close to 0, then the data set doesn’t have clustering tendency.

# **Number of Optimal Clusters,***k*

Some of the clustering algorithms like K-means, require number of clusters, *k*, as clustering parameter. Getting the optimal number of clusters is very significant in the analysis. If *k* is too high, each point will broadly start representing a cluster and if *k* is too low, then data points are incorrectly clustered. Finding the optimal number of clusters leads to granularity in clustering.  
  
There is no definitive answer for finding the right number of clusters as it depends upon (a) Distribution shape (b) scale in the data set (c) clustering resolution required by the user. Although finding the number of clusters is a very subjective problem. There are two major approaches to find optimal number of clusters:  
(1) Domain knowledge  
(2) Data driven approach  
  
**Domain knowledge**— Domain knowledge might give some prior knowledge on finding number of clusters. For example, in the clustering iris data set, if we have the prior knowledge of species (*sertosa, virginica, versicolor*) , then k = 3. Domain knowledge driven *k value*gives more relevant insights.  
  
**Data driven approach**— If the domain knowledge is not available, mathematical methods help in finding out right number of clusters.

*Empirical Method:-*A simple empirical method of finding number of clusters is Square root of N/2 where N is total number of data points, so that each cluster contains square root of 2 \* N  
  
*Elbow method:-*Within-cluster variance is a measure of compactness of the cluster. Lower the value of within cluster variance, higher the compactness of cluster formed.

Sum of within-cluster variance, *W*, is calculated for clustering analyses done with different values of k. *W* is a cumulative measure how good the points are clustered in the analysis. Plotting the *k* values and their corresponding sum of within-cluster variance helps in finding the number of clusters.



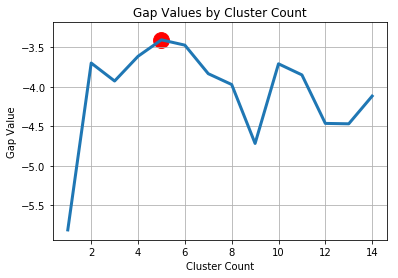
Plot of Sum of within cluster distance vs Number of clusters in Elbow method

Plot shows that number of optimal clusters = 4.  
Initially, Error measure (within-cluster variance) decreases with increase in cluster number. After a particular point, k=4 , Error measure starts flattening. Cluster number corresponding to that particular point, k=4, should be considered as the optimal number of clusters.  
  
*Statistical approach:-*

Gap statistic is a powerful statistical method to find the optimal number of clusters, *k*.

Similar to the Elbow method, the sum of within-cluster (intra-cluster) variance is calculated for different values of k.  
Then Random data points from reference null distribution are generated and Sum of within-cluster variance is calculated for the clustering done for different values of k.

In Simpler words, Sum-of-within-Cluster variance of original data set for different values of k to Sum-of-within-cluster variance of reference data set (null reference data set of uniform distribution) of corresponding values of k is compared to find the ideal *k* value where ‘deviation’ or ‘Gap’ between two is highest. As Gap statistic quantifies this deviation, More the Gap statistic means more the deviation.



Gap statistic Value-Cluster number (k)

Cluster number with maximum Gap statistic value corresponds to optimal number of clusters.

# Clustering quality

Once clustering is done, how well the clustering has performed can be quantified by a number of metrics. Ideal clustering is characterized by minimal intra cluster distance and maximal inter cluster distance.

There are majorly two types of measures to assess the clustering performance.

(i)*Extrinsic Measures* which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

(ii) *Intrinsic Measures*that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.