# Clustering

Clustering or cluster analysis is a machine learning technique, which groups the unlabeled 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 unlabeled 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 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 system 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, 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 unlabeled dataset.**

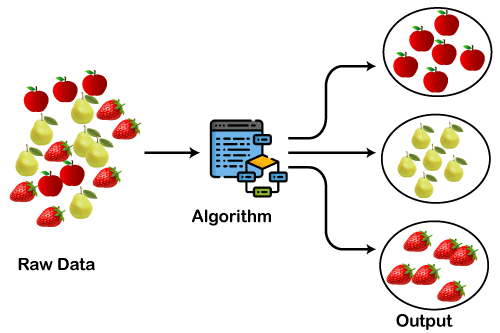
**Example**: Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together. Such as the t-shirts are grouped in one section, and trousers are at other sections, similarly, at vegetable 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)**K-means Clustering**– Using this algorithm, we classify a given data set through a certain number of predetermined clusters or “k” clusters.

2) **Hierarchical Clustering** – follows two approaches Divisive and Agglomerative.

Agglomerative considers each observation as a single cluster then grouping similar data points until fused into a single cluster and Divisive works just opposite to it.

3) **Fuzzy C means Clustering** – The working of the FCM Algorithm is almost similar to the k-means clustering algorithm, the major difference is that in FCM a data point can be put into more than one cluster.

4) **Density-Based Spatial Clustering** – Useful in the application areas where we require non-linear cluster structures, purely based on density.

# K-Means Clustering

K-Means Clustering is an Unsupervised Learning algorithm, used to group the unlabeled dataset into different clusters/subsets.

* Now you must be wondering what does ‘k’ and ‘means’ in the k-means Clustering means??

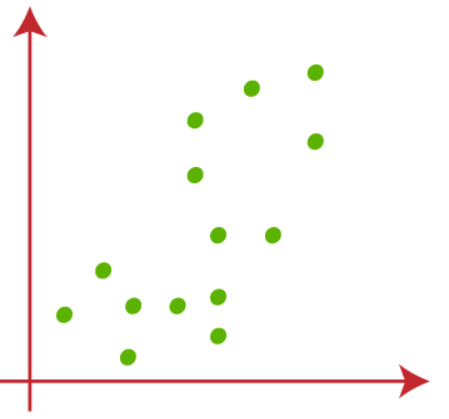
Putting a rest to all your guess here ‘k’ defines the number of pre-defined clusters that need to be created in the process of clustering say if k=2, there will be two clusters, and for k=3, there will be three clusters, and so on. As it is a centroid-based algorithm, ‘means’ in k-means clustering is related to the centroid of data points where each cluster is associated with a centroid. The concept of a centroid based algorithm will be explained in the working explanation of k-means.

Mainly the k-means clustering algorithm performs two tasks:

* Determines the most optimal value for K center points or centroids by a repetitive process.
* Assigns each data point to its closest k-center. Cluster is created with data points which are near to the particular k-center.

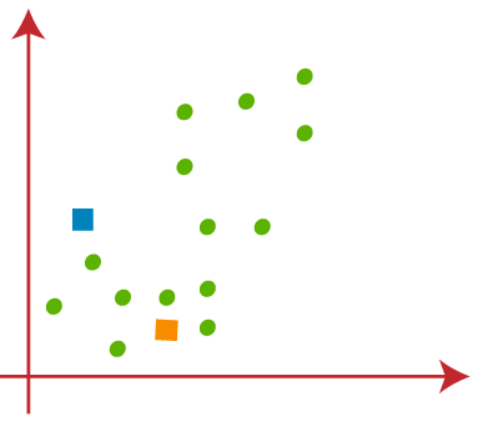
# How does k-means Clustering work?

Suppose, we have two variables X1 and X2, scatter plot below-

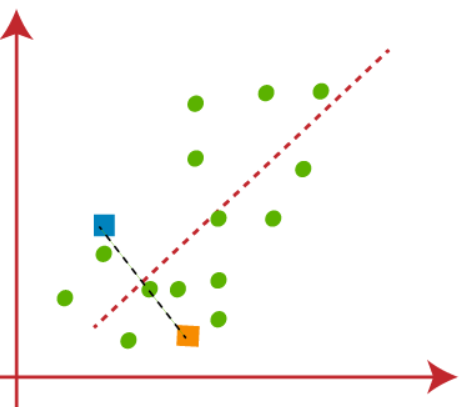


(1) Let’s take the value of k that is the number of pre-defined clusters to be 2(k=2), so here we will be grouping our data into 2 clusters.

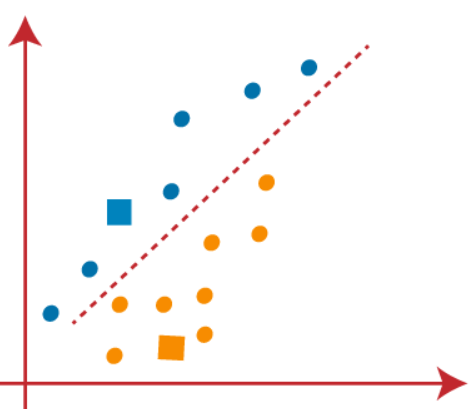
Random k points need to be chosen to form the clusters. No restrictions on the selection of random k points can be from inside the data and outside as well. So, here we are considering 2 points as k points (which are not part of our dataset) shown in the figure below-



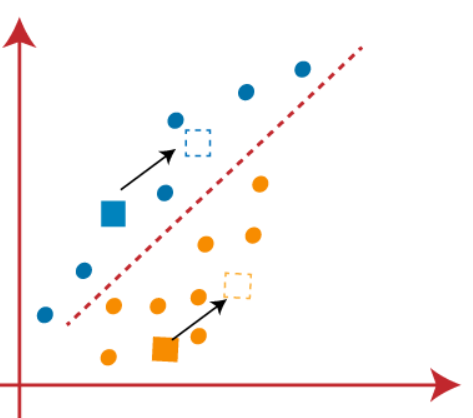
(2) The next step is to assign each data point of the dataset in the scatterplot to its closest k-point, this will be done by calculating Euclidean distance between each point with k point and draw a median between both the centroids, shown in the figure below-



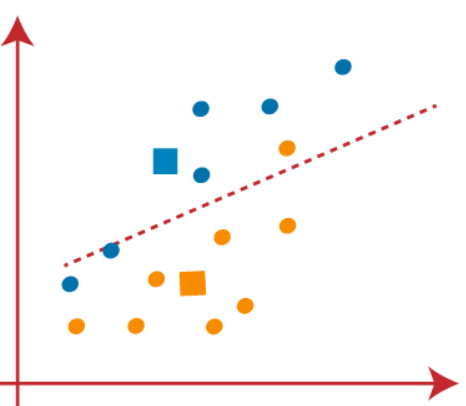
We can clearly observe that the point to the left of the red line are near to K1 or the blue centroid and the points to the right of the red line are near to K2 or the orange centroid.



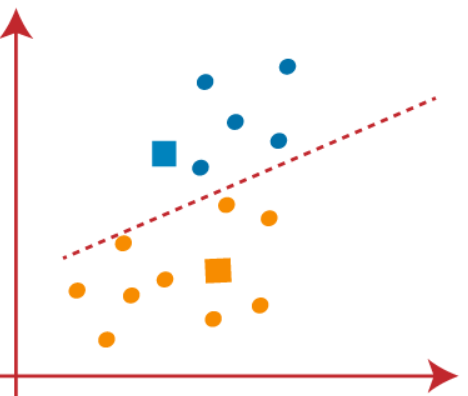
(3) As we need to find the closest point, so we will repeat the process by choosing a new centroid. To choose the new centroids, we will compute the center of gravity of these centroids and will find new centroids as below-



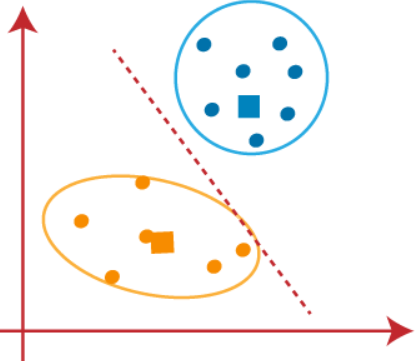
(4) Now, we need to reassign each data point to a new centroid. For this, we have to repeat the same process of finding a median line. The median will be like below-



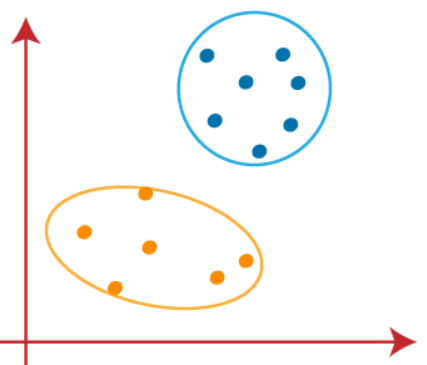
In the above image, we can see, one orange point is on the left side of the line, and two blue points are right to the line. So, these three points will be assigned to new centroids



We will keep finding new centroids until there are no dissimilar points on both sides of the line



We can now remove the assumed centroids, and the two final clusters will be as shown in the below image



So, far we have seen how the k-means algorithm works and the various steps involved to reach the final destination of differentiating clusters.