**UNSUPERVISED LEARNING -FINALISE NEW**

No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning). In some pattern recognition problems, the training data consists of a set of input vectors x without any corresponding target values. The goal in such unsupervised learning problems may be to discover groups of similar examples within the data, where it is called clustering, or to determine how the data is distributed in the space, known as density estimation. To put forward in simpler terms, for a n- sampled space x1 to xn, true class labels are not provided for each sample, hence known as earning without teacher.

**CLUSTERING**

Clustering is an important concept when it comes to unsupervised learning. It mainly deals with finding a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters(groups) if they exist in the data. You can also modify how many clusters your algorithms should identify. It allows you to adjust the granularity of these groups.

Clustering algorithms can be applied in many fields, for instance:

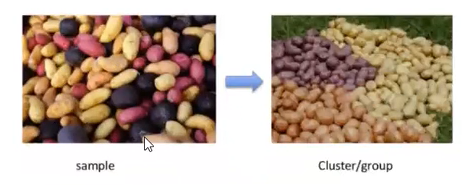
Marketing: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;

Biology: classification of plants and animals given their features;

Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;

Earthquake studies: clustering observed earthquake epicenters to identify dangerous;

World Wide Web: document classification, clustering blog data to cover groups of similar access patterns, Distance-based clustering.



Given a set of points, with a notion of distance between points, grouping the points into some number of clusters, such that internal (within the cluster) distances should be small i.e members of clusters are close/similar to each other.

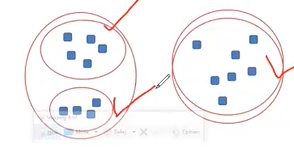
**# Types of clustering**

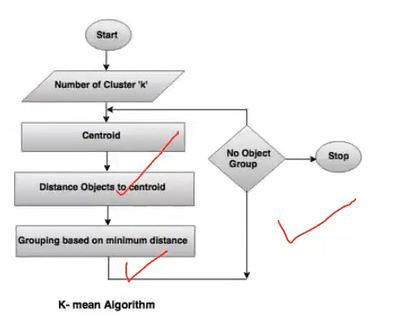
1> KMeans

2> Hierarchical Clustering

3> DESCAN Density Based clustering

**1> KMeans clustering:-** K-means is one of the simplest unsupervised learning algorithms that solves the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centres, one for each cluster. These centroids should be placed in a smart way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated





Let's work with KMeans clustering

**K-means Overview**

Before diving into the dataset, let us briefly discuss how k-means works:

1. The process begins with k centroids initialised at random.

2. These centroids are used to assign points to its nearest cluster.

3. The mean of all points within the cluster is then used to update the position of the centroids.

4. The above steps are repeated until the values of the centroids stabilize

**2> Hierarchical clustering**

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

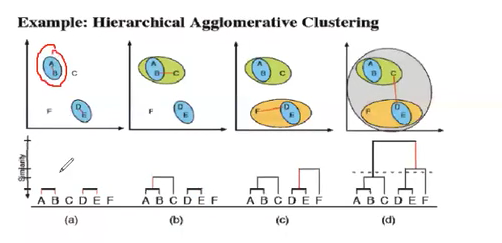
Given a set of N itens to be clustered, and an N distance (or similarity) matrix, the basic process of hierarchical clustering is this:

Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item.

1.let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

2.Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less..

3.Compute distances (similarities) between the new cluster and each of the old clusters. A Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.



The AgglomerativeClustering object performs a hierarchical clustering using a bottom up approach: each observation starts in its own cluster, and clusters are successively merged together

# 3> DBSCAN -Density Based Scan

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex shaped. The central component to the DBSCAN is the concept of care samples which are samples that are in areas of high density

DESCAN uses 2 parameters to find clusters in the data sets cluster radius (epsilon) and minimum number of points needed to form a cluster(n). So, we scan each point in the data and check whether it has n or more number of points around it within the radius epsilon. And if a point meats the criteria, it becomes a core point and the points surrounding it become border points and a cluster is created and so on.

In short DBSCAN creates group where 5 or more datapoints are available,less then 5 data it does not take as group we called them noise and DBSCAN represent those with -1.

