## ****What are Clusters?****

**The word cluster is derived from an old English word, ‘clyster, ‘ meaning a bunch. A cluster is a group of similar things or people positioned or occurring closely together. Usually, all points in a cluster depict similar characteristics; therefore, machine learning could be used to identify traits and segregate these clusters. This makes the basis of many applications of machine learning that solve data problems across industries.**

**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.**

**Why Clustering?   
Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for 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 that constitute the similarity of points and each assumption make different and equally valid clusters.**

## Examples of clustering

**Instances that benefit from data cluster analysis:**

* **Optimizing city planning**
* **Customizing**[**training sets**](https://h2o.ai/wiki/training-sets)**for professional athletes**
* **Detecting spam threats and criminal activity**
* **Identifying misinformation**
* **Analyzing documents**
* **Personalizing advertisements to customers**
* **Tracking online business traffic**

**Clustering Methods :**

### Connectivity-based Clustering (Hierarchical Clustering)

**Hierarchical clustering, also known as connectivity-based clustering, is based on the principle that every object is connected to its neighbors depending on their proximity distance (degree of relationship). The clusters are represented in extensive hierarchical structures separated by a maximum distance required to connect the cluster parts. The clusters are represented as Dendrograms, where X-axis represents the objects that do not merge while Y-axis is the distance at which clusters merge**

**1.1 Divisive Approach**

**This approach of hierarchical clustering follows a top-down approach where we consider that all the data points belong to one large cluster and try to divide the data into smaller groups based on a termination logic or a point beyond which there will be no further division of data points. This termination logic can be based on the minimum sum of squares of error inside a cluster, or for categorical data, the metric can be the GINI coefficient inside a cluster.It must be taken into account that this algorithm is highly “rigid” when splitting the clusters – meaning, once a clustering is done inside a loop, there is no way that the task can be undone.**

**1.2 Agglomerative Approach**

**Agglomerative is quite the contrary to Divisive, where all the “N” data points are considered to be a single member of “N” clusters that the data is comprised into. We iteratively combine these numerous “N” clusters to a fewer number of clusters, let’s say “k” clusters, and hence assign the data points to each of these clusters accordingly. This approach is a bottom-up one, and also uses a termination logic in combining the clusters. This logic can be a number-based criterion (no more clusters beyond this point) or a distance criterion (clusters should not be too far apart to be merged) or a variance criterion (increase in the variance of the cluster being merged should not exceed a threshold, Ward Method)**

### Centroid-based or Partition Clustering

**Centroid-based clustering is the easiest of all the clustering types in data mining. It works on the closeness of the data points to the chosen central value. The datasets are divided into a given number of clusters, and a vector of values references every cluster. The input data variable is compared to the vector value and enters the cluster with minimal difference.The K-Means algorithm lies in this category.**

**Density-Based Methods: Density-based clustering method *considers density ahead of distance*. Data is clustered by regions of high concentrations of data objects bounded by areas of low concentrations of data objects. The clusters formed are grouped as a maximal set of connected data points.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.**

* **DBSCAN: Density-based Spatial Clustering of Applications with Noise   
  These data points are clustered by using the basic concept that the data point lies within the given constraint from the cluster center. Various distance methods and techniques are used for the calculation of the outliers.**

### Distribution-Based Clustering

**Until now, the clustering techniques as we know them are based on either proximity (similarity/distance) or composition (density). There is a family of clustering algorithms that take a totally different metric into consideration – *probability*.**

**Distribution-based clustering creates and groups data points based on their likely hood of belonging to the same probability distribution (Gaussian, Binomial, etc.) in the data.**

**It is a *probability-based distribution* that uses statistical distributions to cluster the data objects. The cluster includes data objects that have a higher probability to be in it. Each cluster has a central point, the higher the distance of the data point from the central point, the lesser will be its probability to get included in the cluster.**

**A major drawback of density and boundary-based approaches is in specifying the clusters apriori to some of the algorithms and mostly the definition of the shape of the clusters for most of the algorithms. There is at least one tuning or hyper-parameter which needs to be selected and not only that is trivial but also any inconsistency in that would lead to unwanted results.**

**Distribution-based clustering has a vivid advantage over the proximity and centroid-based clustering methods in terms of flexibility, correctness, and shape of the clusters formed. The major problem however is that these clustering methods work well only with synthetic or simulated data or with data where most of the data points most certainly belong to a predefined distribution, if not, the results will overfit.**

### 5. Fuzzy Clustering

**Fuzzy clustering generalizes the *partition-based clustering method* by allowing a data object to be a part of more than one cluster. The process uses a weighted centroid based on the spatial probabilities.The steps include initialization, iteration, and termination, generating clusters optimally analyzed as probabilistic distributions instead of a hard assignment of labels.The algorithm works by assigning membership values to all the data points linked to each cluster center. It is computed from a distance between the cluster center and the data point. If the membership value of the object is closer to the cluster center, it has a high probability of being in the specific cluster.At the end iteration, associated values of membership and cluster centers are reorganized. Fuzzy clustering handles the situations where data points are somewhat in between the cluster centers or ambiguous. This is done by choosing the probability rather than distance.**

### 6. Constraint-based (Supervised Clustering)

**The clustering process, in general, is based on the approach that the data can be divided into an optimal number of “unknown” groups. The underlying stages of all the clustering algorithms are to find those hidden patterns and similarities without intervention or predefined conditions. However, in certain business scenarios, we might be required to partition the data based on certain constraints. Here is where a supervised version of clustering machine learning techniques comes into play.**

**A constraint is defined as the desired properties of the clustering results or a user’s expectation of the clusters so formed – this can be in terms of a fixed number of clusters, the cluster size, or important dimensions (variables) that are required for the clustering process.**

* **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.**

**Clustering Algorithms :**[**K-means clustering algorithm**](https://www.geeksforgeeks.org/k-means-clustering-introduction/)**– It is the simplest unsupervised learning algorithm that solves clustering problem.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.**

**The k-means**[**clustering**](https://www.javatpoint.com/clustering-in-machine-learning)**algorithm mainly performs two tasks:**

* **Determines the best value for K center points or centroids by an iterative process.**
* **Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.**

****

**How does the K-Means Algorithm Work?**

**The working of the K-Means algorithm is explained in the below steps:**

**Step-1: Select the number K to decide the number of clusters.**

**Step-2: Select random K points or centroids. (It can be other from the input dataset).**

**Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.**

**Step-4: Calculate the variance and place a new centroid of each cluster.**

**Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.**

**Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.**

**Step-7: The model is ready.**

**Applications of K-Means Clustering**

**K-Means clustering is used in a variety of examples or business cases in real life, like:**

* **Academic performance**
* **Diagnostic systems**
* **Search engines**
* **Wireless sensor networks**

**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.**

**City Planning: It is used to make groups of houses and to study their values based on their geographical locations and other factors present.**

**Earthquake studies: By learning the earthquake-affected areas we can determine the dangerous zones.**