**UNSUPERVISED LEARNING**

Unsupervised Learning

Unsupervised Learning is a type of MachineLearning algorithm used to draw interferences from datasets consisting of input data without responses.

What is Clustering?

* Clustering is the process of dividing the datasets into groups, consisting of similar data points
* It means groping of objects based on information found in the data, describing the objects or then relationship

Why clustering use?

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. Sometimes, partitioning is the goal.

Where it is used?

* Retail store
* Banking
* Insurance Company
* Amazon

Types of Clustering:-

* Exclusive clustering
* Overlapping clustering
* Hierarchical clustering

Exclusive Clustering

Exclusive clustering is as name suggests and stipulates that each data object can only exist in one cluster.

Overlapping Cluster

Overlapping clustering methods allow data points to belong to more than one cluster. Among overlapping clustering algorithm, partitioning methods are more popular mainly because of their simplifying and effective on large datasets.

Hierarchical Clustering

It is an algorithm that groups similar objects into groups called clusters.

k-Means Clustering

The process by which objects are classified into a predefined number of groups. So that they are as much dissimilar as possible from one group to another group, but as much as similar as possible within each group.

k-Means Clustering Steps

1. First we need to decide the number of clusters to be made(Guessing)
2. Then we provide centroids of all the clusters(Guessing)
3. The Algorithm calculates Euclidian distance of the points from each centroid and assigns the point to the closed cluster
4. Next the centroids are calculated again, when we have our new cluster
5. The distance of the points from center of clusters are calculated again and points are assigned to the cluster
6. And then again the new centroid for the cluster is calculated
7. These steps are repeated until we have a repetition in centroids or new centroids are very close to the previous ones

How to decide the number of clusters?

1. Compute clustering algorithm for different values of k
2. For each k, calculates the total within cluster sum of square(WCSS)
3. Plot the curve of WSS according to the number of clusters k

Pros and Cons: k-Means Clustering

* Simple, Understandable
* Items automatically assigned to clusters
* Must define number of clusters
* All items forced into clusters
* Unable to handle noisy data and outliers

Fuzzy c-Means Clustering

Fuzzy c-means is an extension of k-Means, the popular simple clustering technique.

Fuzzy clustering (also referred to as soft clustering) is a form of clustering in which each data point can belong to more than one cluster.

Pros and Cons: c-Means Clustering

* Allows a data, point to be in multiple dusters
* A more natural representation of the behavior of genes
* Genes usually are involved in multiple function
* Need to define, the number of clusters
* Need to determine membership cut-off value
* Clusters are sensitive to initial assignment of centroids
* Fuzzy c-means is not a deterministic algorithm

Hierarchical Clustering

Hierarchical clustering is an alternative approach which builds a hierarchy from the bottom-up, and doesn’t require us to specify the number of clusters beforehand.

Pros and Cons: Hierarchical clustering

* No assumption of a particular number of clusters
* May corresponds to meaningful taxonomies
* Once a decision is made to combine two clusters it can’t be undone
* Too slow for large datasets