Clustering: Cluster Analysis (“data segmentation”) is an exploratory method for identifying homogeneous groups (“clusters”) of records

•  Similar records should belong to the same cluster

•  Dissimilar records should belong to different clusters

Two types of clustering algorithms:

Hierarchical methods: Begin with n clusters; sequentially merge similar clusters until 1 cluster is left.

Non Hierarchical methods: Pre-specified number of clusters, assign records to each of the clusters.

Hierarchical Clustering Algorithm

•  Start with n clusters

•  Step 1: two closest records

•  At every step, pair of records/clusters with smallest distance are merged –  two records are merged, –  or single record added to an existing cluster, –  or two existing clusters are combined

Non Hierarchical Clustering Algorithm

1.  For a user-specified value of K, partition dataset into K initial clusters.

2.  For each record, assign it to cluster with closest centroid

3.  Re-calculate centroids for the “losing” and “receiving” clusters. Can be done

•  after reassignment of each record, or

•  after one complete pass through all records

4.  Repeat Steps 2-3 until no more reassignment is necessary

Each cluster is represented by the center of the cluster

Given K, the number of clusters, the K-Means clustering algorithm is outlined as follows

Select K points as initial centroids

* Repeat

Form K clusters by assigning each point to its closest centroid

Re-compute the centroids (i.e., mean point) of each cluster

* Until convergence criterion is satisfied

Applications of clustering:

Clustering is broadly used in many real time applications like pattern recognition, data analysis,image processing and market research

* Clustering is used in outlier detection applications such as credit card fraud detection.
* It helps in targeted marketing

Advantages:

* Relatively simple to implement.
* Scales to large datasets
* Robust towards outlier detection
* Clusters are non hierarchical and they do not overlap in k-means
* With a large number of variables computationally k-means is faster than hierarchical

Disadvantages:

* Requires number of clusters in advance in K-means
* Hierarchical clustering won’t work well in presence of noise and outliers
* Problem with handling categorical attributes
* Hierarchical clustering is non scalable