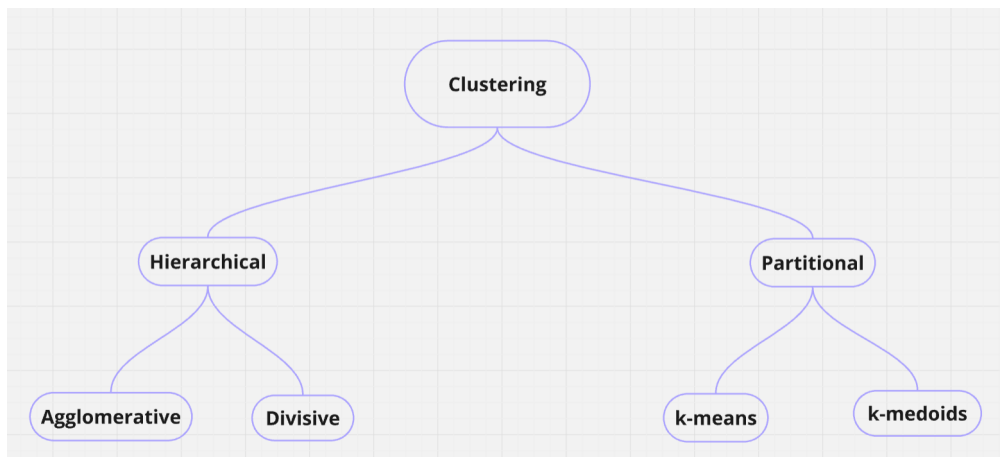# MSMS 201 : CLUSTERING

Ananda Biswas

March 20, 2025

• **Clustering :** Partition data χ into k-groups such that observations in each group are similar and observations in between different groups are far away.

• **Partitional Clustering :** It is simply a division of data objects into non-overlapping subsets such that each object is in only one set.

Two well-known partitional clustering techniques are -

(i) k-means

(ii) k-medoids

In partitional clustering algorithm, our objective is to partition χ in k-clusters such that cost function is minimum.

☞ **k-means clustering :-** defines a prototype in terms of centroid which is usually the mean of a group of points. It is applied when the objects are in continuous n-dimensional space. Centroid may not be an actual data-point.

⊕ **Algorithm for k-means clustering :-**

The $k-$means clustering algorithm is as follows :

(1) select k points as initial centroids

(2) calculate the distance of each data-point from each of the centroids

(3) assign each of the data-points to its closest centroid

(4) relocate the centroids to the average location of the data-points of similar group

And we repeat this procedure until the assignments don't change after the centroid locations were recomputed.

For some type of combinations of proximity measure and types of centroids, k-mean always converges to a solution.

🎯   an example : k-means clustering on a 1D data

🎯   an example : k-means clustering on $iris$ dataset

☞   **k-medoids clustering :-** It defines a prototype in terms of medoids which are the most representative points of a group of points and can be applied to a wide range of data. It requires only a proximity measure for a pair of objects. Medoid must be an actual data-point while centroids mostly never correspond to any actual data-point.

➔   **Algorithm for k-medoids clustering :-**

The $k-$medoids clustering algorithm is as follows :

(1) randomly select k data-points as initial medoids belonging to k clusters

(2) calculate the distance of each data-point from each of the medoids using <u>Manhattan Distance</u> M given by

$$M((X_1, Y_1), (X_2, Y_2)) = |X_1 - X_2| + |Y_1 - Y_2|$$

(3) assign each of the data-points to the cluster from whose medoid the distance M is minimum

(4) calculate the total cost S given by

$$S = \sum_{i=1}^{k} \sum_{j} M(C_i, X_{ij})$$

where C_i is the medoid of i-th cluster and X_{ij} is the jth data-point in ith cluster.

And we repeat this procedure and obtain a new total cost.

Let $G = $ new total cost $-$ old total cost. We stop if $G > 0$.