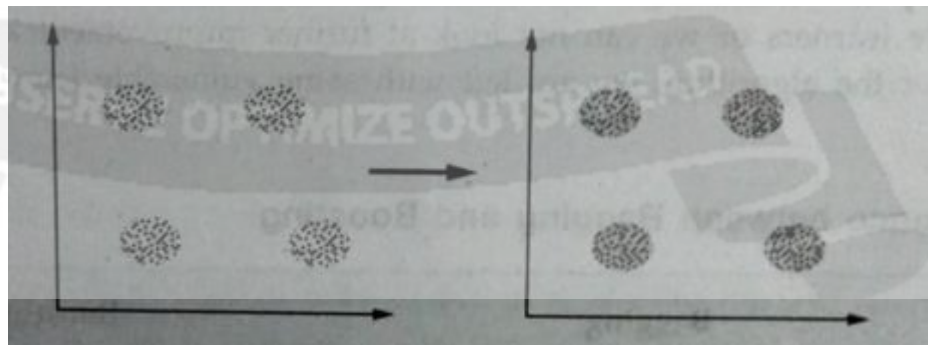## 4.3 CLUSTERING

- Given a set of objects, place them in groups such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups,

- Cluster analysis can be a powerful data-mining tool for any organization that

needs to identity discrete groups of customers, sales transactions, or other types of

behaviors and things. For example, insurance providers use cluster analysis to detect fraudulent claims and banks used it for credit scoring.

- Cluster analysis uses mathematical models to discover groups of similar customers

based on the smallest variations among customers within each group.

- Cluster is a group of objects that belong to the same class. In another words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.

- Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the sub class shares a common trait. It helps a user understand the natural grouping or structure in a data set.

- Various types of clustering methods are partitioning methods, hierarchical clustering, fuzzy clustering, density based clustering and model based clustering.

- Cluster analysis is process of grouping a set of data objects into clusters.

Desirable properties of a clustering algorithm are as follows:

1. Scalability (in terms of both time and space)
2. Ability to deal with different data types.

3. Minimal requirements for domain knowledge to determine input parameters.
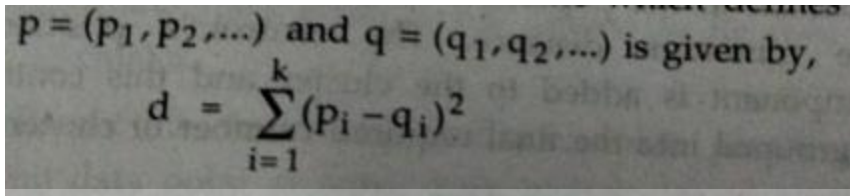4. Interpretability and usability.

Clustering of data is method by which large sets of data are grouped into clusters of smaller sets of similar data. Clusters can be considered the most important supervised learning problems

• A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to the other clusters.

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance : two or more objects belong to the same duster they are "close" acccording to a given distance (in this case geometrical distance) This is called distance based clustering.
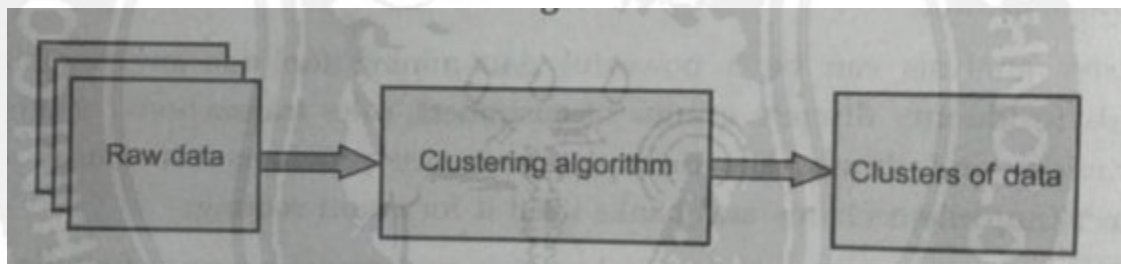
- Clustering, means grouping, of data or dividing a large data set into smaller data sets of scene similarity.
- A clustering, algorithm attempts to find natural groups components or data based on some similarity. Also, the clustering, algorithm finds the centroid of a group of data sets
- To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.
- **Cluster centroid:** The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the cluster. Each cluster has a well defined centroid.
- **Distance**: The distance between two points is taken as a common metric to as see the similarity among the components of population. The commonly used distance measure is the euclidean metric which defines the distance between two points

$$p = (p_1, p_2, \dots) \text{ and } q = (q_1, q_2, \dots) \text{ is given by,}$$
$$d = \sum_{i=1}^{k} (p_i - q_i)^2$$

- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply criterion, in sucha way that the result of the clustering will suit their needs.

- Clustering analysis helps construct meaningful partitioning of a large set of objects: Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing etc.

- Clustering algorithms may be classified as listed below:

  1. Exclusive clustering
  2. Overlapping clustering
  3. Hierarchical clustering
  4. Probabilisitic clustering

- A good clustering method will produce high quality clusters high intra- class similarity and low inter class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by it's ability to discover some or all of the hidden patterns.

- Clustering techniques types. The major clustering techniques are,

  a) Partitioning methods
  b) Hierarchical methods
  c) Density-based methods.