# CLUSTERING: GROUPING THINGS TOGETHER

Unit 4

# Introduction

- Cluster analysis or simply clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters.
- The set of clusters resulting from a cluster analysis can be referred to as a clustering.
- Different clustering methods may generate different clusterings on the same data set.
- Clustering is useful in that it can lead to the discovery of previously unknown groups within the data.
- Clustering is known as unsupervised learning because the class label information is not present.
- Common algorithms: K-means, K-mediods, ROCK, DBSCAN etc.

# K-means Clustering

**Algorithm: *k*-means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

**Input:**

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

**Output:** A set of *k* clusters.

**Method:**

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

The *k*-means partitioning algorithm

# DBSCAN

- DBSCAN stands for Density-Based Spatial Clustering Application with Noise.
- Unsupervised machine learning algorithm that makes clusters based upon the density of the data points or how close the data is.
- The points which are outside the dense regions are excluded and treated as noise or outliers. This characteristic of the DBSCAN algorithm makes it a perfect fit for outlier detection and making clusters of arbitrary shape.
- The DBSCAN algorithm takes two input parameters. Radius around each point ( *eps*) and the minimum number of data points that should be around that point within that radius ( *MinPts*).
- In this algorithm, three types of points are considered: **core point**, **border point**, and **outlier**.

# DBSCAN

- If the number of neighbourhood points around **x** is greater or equal to MinPts then **x** is treated as a **core point**. If the neighbourhood points around **x** are less than MinPts but is close to a core point then **x** is treated as a **border point**. If **x** is neither core nor border point then **x** is treated as an **outlier**.
- **Algorithm:**

1. Find all the neighbor points within eps and identify the core points.
2. For each core point if it is not already assigned to a cluster, create a new cluster.
3. Find recursively all its density-connected points and assign them to the same cluster as the core point.
4. Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

# DBSCAN

- A point *a* and *b* are said to be **density connected** if there exists a point *c* which has a sufficient number of points in its neighbors and both points *a* and *b* are within the *eps distance*. This is a chaining process. So, if *b* is a neighbor of *c*, *c* is a neighbor of *d*, and *d* is a neighbor of *e*, which in turn is  neighbor of *a* implying that *b* is a neighbor of *a*.

# ROCK Clustering

- Study yourself