

CLUSTER ANALYSIS: BASIC CONCEPTS & ALGORITHMS

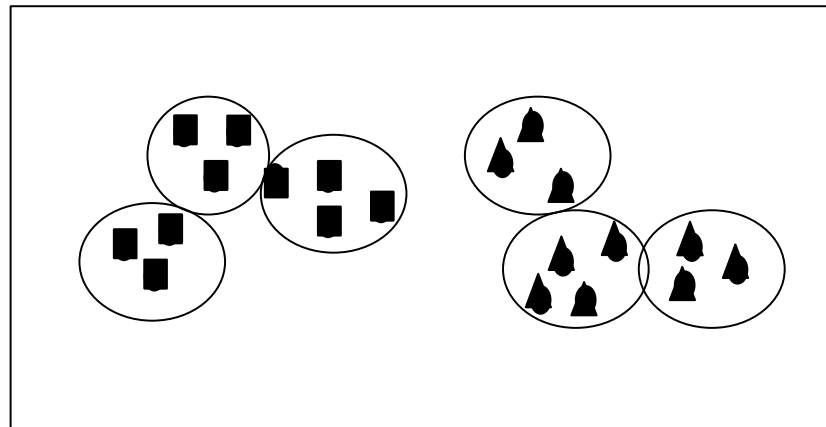
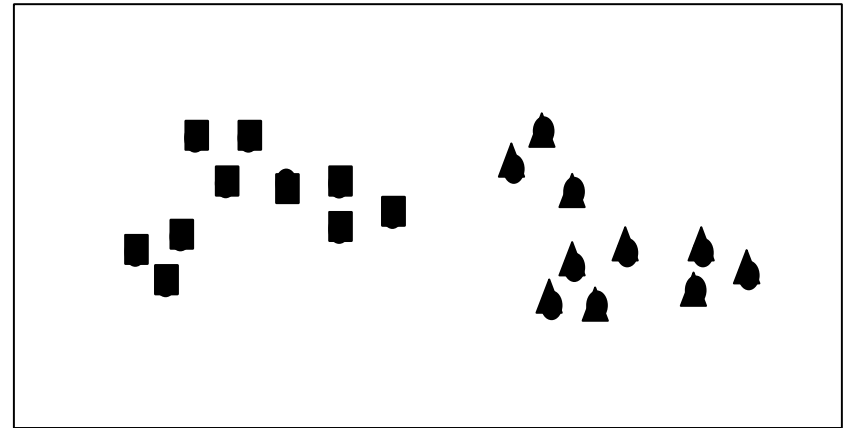
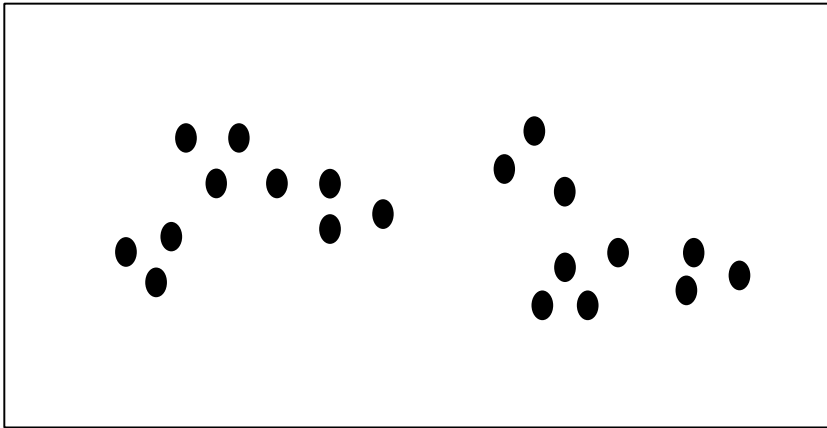
Understanding

- Cluster analysis divides data into groups (clusters) that are meaningful, useful or both.
- Examples:
 - Biology: large amount of genetic information
 - Information Retrieval: WWW billion of web pages, group into a small number of clusters.
 - Business: to segment customers into a small number of groups.

Cluster Prototypes

- Summarization: such as regression or PCA with $O(m^2)$
- Compression: such as vector quantization, where highly similar, loss information, reduction.
- Efficiently finding nearest neighbors: to compute pairwise distance between all points.

Different ways of clustering



Different Types of Clustering

- Hierarchical vs Partitional
 - hierarchical is a set of nested clusters that are organized as a tree.
 - partitional is simply a division of the set of data objects onto non-overlapping subset.
- Exclusive vs Overlapping vs Fuzzy
 - exclusive, an object is only in a cluster
 - overlapping, an object can be in some clusters
 - fuzzy, an object is related to all clusters in $[0,1]$

Different Types of Clustering

- Complete vs Partial
 - complete is assigns every object to a cluster.
 - partial is just part of object to a cluster.

Different Types of Clusters

- Well separated, sufficiently close (similar) compare to other groups.
- Prototype-based, each object is closer to the prototype that defines the cluster than to the prototype to any other cluster.
- Graph-based, nodes are objects and link represent connections among objects.
- Density-based, region is surrounded by a region of low density.
- Shared Property,

K-means

- prototype-based, paritional clustering
- based on centroids

1: select K points as initial centroids

2: **repeat**

3: Form K clusters by assigning each point to its closet centroid.

4: Recompute the centroid of each cluster

5: **until** Centroids do not change

Agglomerative Hierarchical Clustering

- Agglomerative: start with the points as individual clusters and at each step, merge the closest pair of clusters. This requires defining a notion of cluster proximity.
- Divisive : start with one, all inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain.

Agglomerative Hierarchical Clustering

- 1: compute the proximity matrix, if necessary
- 2: **repeat**
- 3: merge the closest two clusters
- 4: update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
- 5: **until** only one cluster remains

DBSCAN

- 1: label all points as core, border or noise points
- 2: Eliminate noise points
- 3: Put an edge between all core points that are within EPS of each other
- 4: Make each group of connected core points into a separate cluster
- 5: Assign each border point to one of the clusters its associated core points.