# Software Development for Data Analysis

# Cluster Analysis

## Preliminaries

- *Clustering* or *unsupervised* classification algorithms are used for determining natural grouping in data or for providing a convenient classification of data.
- Unlike *supervised* classification, such as *Discriminant Analysis*, the unsupervised classification does not employ any *a priori* information regarding grouping. From this perspective, Cluster Analysis is an exploratory method.
- The data concerning the analysis are values regarding the relationship among observations and the studied variables, values considered as distances or dissimilarities.

# Cluster Analysis

## Preliminaries

- The determined groups or classes of data (observations or variables) consist of instances located at closer distances or having lesser dissimilarities.
- Cluster Analysis concerns rather the observations than the variables, without excluding the latter approach from being employed.
- It might be used for :
  - identifying fundamental features of data;
  - obtaining some advantageous representation of data for further analysis;
  - storing and retrieving data more efficiently (faster access to groups of data with similar characteristics).

# Cluster Analysis

## Classification Algorithms

Clustering algorithms belong to one of the following categories:

- Hierarchical algorithms:
  - Bottom-up, agglomerative - start with the points as individual clusters and, at each step, merge the closest pair of clusters (K-nearest neighbors or K-nn)
  - Top-down, divisive - start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain.

# Cluster Analysis

## Classification Algorithms

- Partitional algorithms:
  - Usually start with a random (partial) partitioning
  - Refine it iteratively
    - K-Means clustering
    - Model based clustering

# Cluster Analysis

## Classification Algorithms

Hierarchical vs. Partitional Clustering

- A distinction among different types of clustering is whether the set of clusters is nested or un-nested.
- A partitional clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
- A hierarchical clustering is a set of nested clusters that are organized as a tree.

# Cluster Analysis

## Classification Algorithms

Why Hierarchical Clustering?

1. It does not assume a particular value of $k$, as needed by K-Means clustering.
2. The generated tree may correspond to a meaningful taxonomy.
3. Only a distance or "proximity" matrix is needed to compute the hierarchical clustering.

# Cluster Analysis

## Basic Agglomerative Clustering

Basic agglomerative hierarchical clustering algorithm:

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.

# Cluster Analysis

## Defining Proximity Between Clusters

- The key operation of basic agglomerative clustering is the computation of the proximity between two clusters.
- The definition of cluster proximity differentiates the various agglomerative hierarchical methods or techniques.
- MIN (**single link**), MAX (**complete link**), **group average**, **centroid** and **median** are graph-based proximities.
- **Ward**'s method is a prototype-based proximity

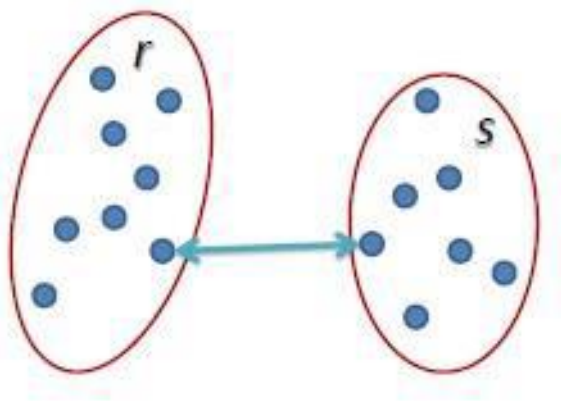
# Cluster Analysis

## Defining Proximity Between Clusters

MIN (Single Link) Proximity

- Defines cluster proximity as the shortest distance between two points, $x$ and $y$, that are in different clusters, $A$ and $B$:

$$d(A, B) = min\ \{d(x - y)\}\ for\ x \in A,\ y \in B$$



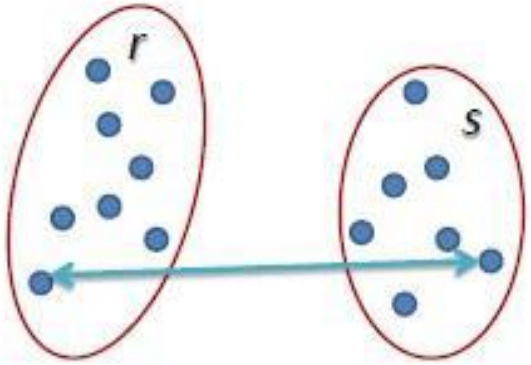
# Cluster Analysis

## Defining Proximity Between Clusters

MAX (Complete Link) Proximity

- Defines cluster proximity as the furthest distance between two points, $x$ and $y$, that are in different clusters, $A$ and $B$:

$$d(A, B) = max\ \{d(x - y)\}\ for\ x \in A,\ y \in B$$



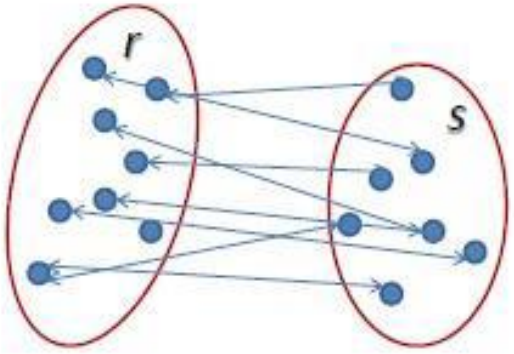
# Cluster Analysis

## Defining Proximity Between Clusters

Group Average Proximity

- Defines cluster proximity as the average distance between two points, $x$ and $y$, that are in different clusters, $A$ and $B$ (the number of points in cluster $A$ is $n_A$ and in cluster $B$ is $n_B$):

$$d(A, B) = \Sigma\{d(x - y)/ n_A n_B\} \text{ for } x \in A, y \in B$$



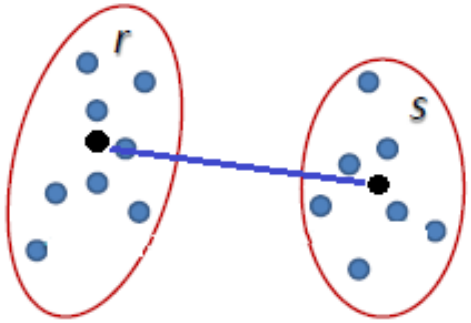
# Cluster Analysis

## Defining Proximity Between Clusters

Centroid method

- Defines the proximity of the cluster as the average distance between the centers of gravity $g_A$ and $g_B$ of two different clusters, A and B (the number of points in cluster $A$ is $n_A$ and in cluster $B$ is $n_B$):

$$d(A, B) = \Sigma \{ d(g_A - g_B) / n_A n_B \} \text{ for } g_A \in A,\ g_B \in B$$



# Cluster Analysis

## Discussion of Proximity Methods

- Single link is “chain-like” and good at handling non-elliptical shapes, but is sensitive to outliers.
- Complete link is less susceptible to noise and outliers, but can break large clusters and favors globular shapes.
- Group average and the centroid method are intermediate approaches between the single and complete link.

# Cluster Analysis

## Ward's Method

- Assumes that a cluster is represented by its centroid, and measures the proximity between two clusters in terms of the increase in sum of the squared error (SSE) that results from merging the two clusters:

  $d(A, B) = SSE_{A \cup B} - SSE_A - SSE_B$

  where $A$ and $B$ are clusters.
- Note that for hierarchical clustering, the SSE starts at 0.

# Cluster Analysis

## Distances used for computing proximity

- *Euclidian distance:*

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$$

- *Standardized Euclidian distance:*

$$d(x, y) = \sqrt{(\frac{x_1 - y_1}{\sigma_1})^2 + \ldots + (\frac{x_n - y_n}{\sigma_n})^2}$$

- *Manhattan (city-block or taxicab) distance:*

$$d(x, y) = |x_1 - y_1| + \ldots + |x_n - y_n|$$

# Cluster Analysis

## Distances used for computing proximity

- *Mahalanobis distance:*

  $d(x, y) = \sqrt{(x - y)^T C^{-1} (x - y)}$ ,

  where $C^{-1}$ is the inverse of variance-covariance matrix of the data set.

  – If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance.

  – If the covariance matrix is diagonal, then the resulting distance measure is called a standardized Euclidean distance.

# Cluster Analysis

## Distances used for computing proximity

- *Chebyshev distance:*

  $d(x, y) = max\{|x_1 - y_1|, \dots, |x_n - y_n|\}$

- *Correlation, as distance between variables:*

  $d(X, Y) = corr(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{(X - \bar{X})(Y - \bar{Y})}{\sigma_X \sigma_Y}$ ,

  where $cov(X, Y)$ is the covariance between the variables $X$ and $Y$, while $\sigma_X$ and $\sigma_Y$ are the corresponding standard deviations.

# Cluster Analysis

## Hierarchical algorithms

Having given a set of objects $\Omega = \{w_1, w_2, \ldots w_n\}$, observations / instances or variables a **hierarchy** ***H***, is an ensemble ordered subsets of $\Omega$, aggregated on certain levels, and with the following properties:

1. $\Omega \in H$, meaning that the aggregated subset at the highest level contains all the observations / instances;
2. For any $w_i$, $i$=1,$n$ ($n$ is the number of objects), there is $\{w_i\} \in H$ which forms base, terminal subsets;
3. For any $h$, $h' \in H$ there is the inference: $h \cap h' \neq \emptyset \Rightarrow h \subset h'$ or $h' \subset h$.

# Cluster Analysis

## Hierarchical algorithms

- The graphical output of hierarchical algorithms is the dendrogram.
- On one of the axes of the dendrogram are represented the distances and, on the other one the set of objects.
- The dendrogram emphasizes the aggregation distances of an hierarchy.

# Cluster Analysis

## Choosing the number of clusters

- Depending of the *nature of the problem*, the number of groups or classes is decided by the domain specialist.
- *Variation of the aggregation distance*. It is chosen the partition which corresponds to the maximum difference in terms of distance between the hierarchical clusters.