# Cluster Analysis

## (Unsupervised Learning Part I)

**Unsupervised learning vs supervised learning**

In supervised learning, we assumed that there are existing labeled classes in the population separated by distinct features. In unsupervised learning, there are no classes a priori. We find patterns and structures within the data. Examples of unsupervised learning we have learned are principal component analysis and factor analysis.

Both supervised learning are unsupervised learning are essential in machine learning and artificial intelligence.

The most basic technique in unsupervised learning is data clustering, in which we use data characteristics themselves to partition the data into subgroups, commonly called clusters. A common name for the grouping process is cluster analysis. also called data segmentation. The rule of partition can be used to assign future observations to individual clusters.

**Data form for unsupervised learning**

For unsupervised learning, all information is in a dataset of the form

| | variable $X_1$ | variable $X_2$ | $\cdots$ | variable $X_k$ | $\cdots$ | variable $X_p$ |
|---|---|---|---|---|---|---|
| observation 1 | $x_{11}$ | $x_{12}$ | $\cdots$ | $x_{1k}$ | $\cdots$ | $x_{1p}$ |
| observation 2 | $x_{21}$ | $x_{22}$ | $\cdots$ | $x_{2k}$ | $\cdots$ | $x_{2p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| observation $j$ | $x_{j1}$ | $x_{j2}$ | $\cdots$ | $x_{jk}$ | $\cdots$ | $x_{jp}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| observation $n$ | $x_{n1}$ | $x_{n2}$ | $\cdots$ | $x_{nk}$ | $\cdots$ | $x_{np}$ |

A common task in unsupervised learning is to seek structures in data themselves according to certain similarity measures. The objects to be grouped can be either observation items or variables.

**Finding cluster structure among observations**

Often grouping on observation items is desired, then each of the $n$ observed item vectors is to be assigned to one and only one of the $g$ clusters or groups. The observations in the same cluster should share similar properties in their feature values in $X_1, \cdots, X_p$, according to certain similarity measure, defined by the researchers.

After creating groups, the grouped or clustered data are of the form

| $X_1$ | $X_2$ | $\cdots$ | $X_p$ | Assigned cluster |
|---|---|---|---|---|
| $x_{1,11}$ | $x_{1,12}$ | $\cdots$ | $x_{1,1p}$ | 1 |
| $x_{1,21}$ | $x_{1,22}$ | $\cdots$ | $x_{1,2p}$ | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{1,n_11}$ | $x_{1,n_12}$ | $\cdots$ | $x_{1,n_1p}$ | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{g,11}$ | $x_{g,12}$ | $\cdots$ | $x_{g,1p}$ | $g$ |
| $x_{g,21}$ | $x_{g,22}$ | $\cdots$ | $x_{g,2p}$ | $g$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{g,n_g1}$ | $x_{g,n_g2}$ | $\cdots$ | $x_{g,n_gp}$ | $g$ |



with $n_1 + \cdots + n_g = n$.

The data used for unsupervised learning are called "unlabeled data", in contrast to the data used for supervised learning, which already has class labels, thus called "labeled data".

**Finding cluster structure among variables**

If the desired grouping is on variables rather than on observations, the $p$ variables $X_1, \cdots, X_p$ are to be partitioned into $K$ variable groups with $p_1, \cdots, p_K$ variables in each group respectively, $p_1 + \cdots + p_K = p$.

The partitioned data and variable then will be of the following form:

| $X_1^{(1)}$ | $X_2^{(1)}$ | $\cdots$ | $X_{p_1}^{(1)}$ | $X_1^{(2)}$ | $X_2^{(2)}$ | $\cdots$ | $X_{p_2}^{(2)}$ | $\cdots$ | $\cdots$ | $X_1^{(K)}$ | $\cdots$ | $X_{p_K}^{(K)}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_{1,11}$ | $x_{1,12}$ | $\cdots$ | $x_{1,1p_1}$ | $x_{2,11}$ | $x_{2,12}$ | $\cdots$ | $x_{2,1p_2}$ | $\cdots$ | $\cdots$ | $x_{K,11}$ | $\cdots$ | $x_{K,1p_K}$ |
| $x_{1,21}$ | $x_{1,22}$ | $\cdots$ | $x_{1,2p_1}$ | $x_{2,21}$ | $x_{2,22}$ | $\cdots$ | $x_{2,2p_2}$ | $\cdots$ | $\cdots$ | $x_{K,21}$ | $\cdots$ | $x_{K,2p_K}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{1,n1}$ | $x_{1,n2}$ | $\cdots$ | $x_{1,np_1}$ | $x_{2,n1}$ | $x_{2,n2}$ | $\cdots$ | $x_{2,np_2}$ | $\cdots$ | $\cdots$ | $x_{K,n1}$ | $\cdots$ | $x_{K,np_K}$ |

The data are used to forming groupings of the variables,

$$(X_1, X_2, \cdots, X_p) \longrightarrow \begin{cases} (X_1^{(1)}, \cdots, X_{p_1}^{(1)}) & \leftarrow cluster\ 1 \\ \vdots & \\ (X_1^{(K)}, \cdots, X_{p_K}^{(K)}) & \leftarrow cluster\ K \end{cases} \qquad (p_1 + \cdots + p_K = p)$$

Most of our examples are on clustering of observations. The clustering methods discussed can easily apply to clustering of variables correspondingly.

**Objectives and sujectivity of unsupervised learning**

In unsupervised learning, the objectives are to find patterns in data for further analysis and applications.

For example, for meteorologists, it is desirable to use observed meteorological data to group similar weather patterns, then categorize them and analyze the patterns to aid forecasting.

The researcher's objective usually is beyond a one-level partitioning of the objects. Hierarchical structures in the subgroups is also of interest. These are two overlapping albeit different objectives. Hierarchical clustering can be obtained by grouping the clusters at each level of the hierarchy. The result will be a tree-like representation of the clusters.

Cluster analysis has inherent subjectivity. It is often an exploratory process used as a part of a more comprehensive data analysis.

## 1 Similarity measures

Cluster analysis seeks natural or reasonable groupings of objects. After clustering, every objects belongs to one and only one group, called a cluster. Items in the same cluster are similar to each other, and items in different clusters are dissimilar, by some similarity measure. The number of clusters are often not known, contrast to classification where the number and the labels of the groups are given.

Therefore a key notion essential to cluster analysis is the definition of similarity. In order to conduct cluster analysis, we need to measure the degree of similarity (or dissimilarity) between every pair of items. In practice, similarity measures are either come with the data or developed from data.

Choice of similarity measures

There are many similarity or dissimilarity measures, all with certain degrees of subjectivity.

In developing similarity measures, several important aspects should be considered.

- The nature of the variables
- The scales of measurement
- Subject matter knowledge

**Proximity Matrix**

Pairwise similarity of objects can be represented in a proximity matrix.

For $n$ observations each with measurements of $p$ variables or features, the proximity matrix $D$ for the $n$ observations is of dimensions $n \times n$, where the $(i, j)$th entry $d_{ij}$ is a similarity (or dissimilarity) coefficient between observation pair $i$ and $j$. The similarity coefficient $d_{ij}$ quantifies the degree of closeness of measured feature values $(x_{i1}, \cdots, x_{ip})$ for observation $i$ and $(x_{j1}, \cdots, x_{jp})$ for observation $j$.

Examples of common distance and similarity coefficients for pairs of items

- Euclidean distance $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})'(\boldsymbol{x}-\boldsymbol{y})}$
- Statistical or Mahalanobis distance $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$, $\boldsymbol{S}$ the sample covariance matrix
- Minkowski metric: $d(\boldsymbol{x}, \boldsymbol{y}) = (\sum_{i=1}^{p} |x_i - y_i|^m)^{1/m}$. Important cases in application: $m = 1, 2, \infty$.
- Canberra metric: $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{|x_i| + |y_i|}$
  Other names: normalized Manhattan distance, city block or taxicab distance.
- Czekanowski coefficient $d(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p}(x_i + y_i)}$

**Distance metric basics**

$d(\cdot, \cdot)$ is a distance function (or a metric) if

- $d(x, y) = d(y, x)$ for any $x, y$. (Symmetry)
- $d(x, x) = 0$ for all $x$
- $d(x, y) > 0$ if $x \neq y$.
- $d(x, y) \leq d(x, w) + d(w, y)$ for all $x, y, w$. (Triangle inequality)

**Diistance and similarity coefficients for pairs of sets**

Examples $A, B \subset \mathbb{R}^d$ are (measurable) sets.

- Jaccard similarity of $(A, B)$ = size of intersection/union = $\frac{|A \cap B|}{|A \cup B|}$
- Simpson overlap coefficient of $(A, B) = \frac{|A \cap B|}{\min\{|A|, |B|\}}$
- Sørensen-Dice coefficient of $(A, B) = \frac{2|A \cap B|}{|A| + |B|}$



# 2 Hierarchical Clustering Methods

Hierarchical cluster methods proceed by a series of successive mergers, or in the opposite direction, by a series of divisions.

- The bottom-up approach

  Agglomerative hierarchical methods start with each individual object in a cluster of its own, then continuously merge similar clusters. Eventually there is only one cluster consisting of all the objects.

- The top-down approach

  Divisive hierarchical methods begin with a single cluster of all the objects, then split this cluster into more clusters, and then iteratively divides clusters into sub-clusters. At last there is only one object in each cluster.

The results of the clustering process can be illustrated by a two-dimensional diagram called **dendrogram**, which continuously branches from the top, with the final branches at the bottom leading to the objects in the resulting clusters. Often the vertical axis ("height") indicates the distance between the merging clusters.

Different definitions of distance between clusters give rise to different clustering results, thus different dendrograms.

**Linkage methods**

Linkage methods in cluster analysis refer to methods of clustering based on the definitions of distances between groups of objects.

Suppose $D = [d_{ij}]_{n \times n}$ is a symmetric distance (or similarity) matrix of $N$ objects.
For any two groups $U$ and $V$, denote by $d(U, V) = d_{U,V}$ the distance between the two groups.
There are different definitions of the distance between two groups. Three linkage methods are discussed below.

- **Single linkage** (minimum distance or nearest neighbor)

  The single linkage method defines the distance between clusters as the distance between the **nearest pair** of members, one from each cluster. The clustering process goes as the following.

  – Start with each object being a cluster of its own.
  – Find two objects (as two clusters), say $U, V$, with the shortest distance among all pairs of objects.
  – Form a cluster of the two objects, denoted as (UV).
  – Define the distance between the cluster $(UV)$ and any other object $W$ as

  $$d_{(UV),W} = \min\{d_{U,W}, d_{V,W}\}$$

  – Find two objects, or an object and a cluster, with the shortest distance among all pairs of objects and the cluster $(UV)$, then merge the pair.
  – Iteratete process.

- **Complete linkage** (maximum distance)

  The complete linkage method defines the distance between clusters as the distance between the **furthest pair** of members, one from each cluster. The clustering process are based on similar procedure as the single linkage, group objects together by the minimum distance, but updates the distance using maximum:

  $$d_{(UV),W} = \max\{d_{U,V}, d_{V,W}\}$$

- **Average linkage** (average distance)

  The average linkage method defines the distance between clusters as the **average distance of all pairs**, one from each cluster.

$$d_{(UV),W} = \frac{\sum_i \sum_k d_{i,k}}{N_{UV} N_W}$$

  where $i$ and $k$ denote member in $(UV)$ and $W$ respectively, and $N_{(UV)}, N_W$ are the numbers of objects in the subgroups $(UV)$ and $W$ respectively.

Remarks

- Given a specific linkage method, hierarchical clustering produces a sequence of clusters, which is often displayed in a dendrogram — an upside down tree structure. At the terminal leaves, all points are in their own cluster; at the root, all points are in one cluster.
- The vertical distance between two connected nodes in a dendrogram is proportional to the dissimilarity between two clusters at the two connected nodes.
- Single linkage merges two clusters when one pair of items are close, even though other pairs may be far apart. Thus single linkage clusters could be more spread out. On the other hand, complete linkage could produce more compact cluster. Average linkage strikes a balance, however sensitive to monotone changes of the similarity measure.
- Given pairwise dissimilarities, hierarchical clustering produces a consistent result (without initialization issues).
- The choice of the measure of dissimilarity matters and has effect on the cluster structure.

## 3 Nonhierarchical clustering — K-mean cluster method

Originated by MacQueen (1967), K-means has become a popular algorithm of clustering (also Lloyd's algorithm). The algorithm assigns each item to the cluster having the nearest mean or centroid (center of mass). For a given $k$, the objective is to form clusters $C_1, \cdots, C_K$ that **minimize the within cluster sum of squares**

$$\sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

where $\boldsymbol{m}_i$ is the mean or centroid of objects in cluster $C_i$, the norm $\|\cdots\|$ is usually the Euclidean distance in $\mathbb{R}^p$.

In its simple version, the original MacQueen K-means process consists of three steps.

Specify $K$, the number of clusters desired.

1. Initialize $K$ points as cluster centers, or initialize membership.

   Partition the items into K initial clusters with identified centers.

2. Proceed through the list of items:
   - Assignment: Assign an item to the cluster whose centroid is nearest.
   - Update: Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.



3. Repeat Step 2 until no more reassignments take place.

Variations in algorithm

There are many variations and improvements on the original MacQueen iteration, such as Hartigan-Wong and Lloyd algorithms. A simplified version can stated as the following.

1*. Randomly assign each observation to one of the $K$ clusters.

2*. Iterate until the cluster assignments stop changing:

   - Compute the cluster centroids..
   - Assign each observation to the cluster whose centroid is closest.

Newer algorithms improved the convergence speed. The difference in the update steps may result in different final cluster assignments.

Remarks on K-means clustering method

- With each iteration step of the K-means algorithm, the within-cluster variations (or centered sums of squares) decrease and the algorithm converges.
- Initialization effects
  - Typically multiple runs from random initial cluster assignments are conducted.
  - The model with the lowest within cluster sum of squares will be selected.
  - Final clustering assignment depends on the chosen initial cluster centroids.
- Cluster centroids can be served as a prototype of the cluster.
- Geometrically, K-means method is closely related to the **Voronoi tessellation** partition of the data space with respect to cluster centers.
- The algorithm converges to a local optimum. There is no guarantee of the achievement of global optimum.
- The number of clusters K is fixed for the algorithm, and the choice of $K$ depends on the goal.
- Choice of K

  Data-based methods for estimating $K$ view the with-cluster sum of squares or similar measures as functions of $K$. Optimal $K$ may be obtained by cross validation or other criteria.
- K-means can be computationally intensive (NP hard).
- K-means minimize within-cluster distance while maximizing between-cluster distance.

## 4 Clustering methods based on statistical models

The clustering methods discussed above are popular ones. There are clustering methods based on other measures, such as by loss of information during merging.

Statistical models in cluster analysis treat the membership as probabilities. The most common model for clustering is the "Mixture Models". Mixture models and the EM algorithm used for estimating mixture model parameters are important in applications on their own right. They are discussed in a separate section.