Unsupervised Learning

Unsupervised learning refers to a set of statistical techniques for exploring and discovering knowledge from a multivariate data, without building a predictive models.

It makes it possible to visualize the relationship between variables, as well as, to identify groups of similar individuals (or observation)

The most popular unsupervised learning methods, include:

- **Principal component methods**, which consist of summarizing and visualizing the most important information contained in a multivariate data set.

- **Cluster analysis** for identifying groups of observations with similar profile according to a specific criteria. These techniques include hierarchical clustering and k-means clustering.

Principal components methods

The type of principal component methods to use depends on variable types contained in the data set. These are three known methods:

1. **Principal Component Analysis (PCA**), which is the most popular multivariate analysis method. The goal of PCA is to summarize the information contained in a continuous (i.e quantitative) multivariable data by reducing the dimensionality of the data without loosing important information.

2. Correspondence Analysis (CA), which is an extension of the principal component analysis for analyzing large contingency table formed by two qualitative variables (or categorical data).

3. Multiple Correspondence Analysis (MCA), which is an adaptation of CA to a data table containing more than two categorical variables.

Required R packages

- FactoMineR for computing principal components methods
- Factoextra for visualizing the output of FactoMineR

Cluster Analysis

Cluster analysis is used to identify groups of similar objects in a multivariate sets collected from fields such as marketing, bio-medical and geo-spatial.

Type of clustering. There are different types of clustering methods, including:
Partitioning clustering: Subdivides the data into a set of k groups.
Hierarchical clustering: Identify groups in the data without subdividing it.

Distance measures. The classification of observation into groups requires some methods for computing the distance Or the (dis) similarity between each pair of observations. The result of this computation is known as a dissimilarity or distance matrix. There are different methods for measuring distances, including:

- Euclidian distance
- Correlation based-distance

Data standardization

Before cluster analysis, it's recommended to scale (or normalize) the data, to make the variable comparable. This is particularly recommended when variables are measured in different scales (e.g. : kilograms, kilometers, centimeters,..); otherwise, the dissimilarity measures obtained will be severely affected.

R functions for scaling the data: scale(), applies scaling on the column of the data (variables).

Partitioning clustering

Partitioning algorithms are clustering techniques that subdivide the data sets into a set of k groups where k is the number of groups pre-specified by the analyst.

There are different types of partitioning clustering methods. The most popular is the **K-means clustering,** in which, each cluster is represented by the center or means of the data points belonging to the cluster. The K-means method is sensitive to outliers**.**

An alternative to k-means clustering is the K-medoids clustering or PAM (Partitioning Around Medoids), which is less sensitive to outliers compared to k-means.