Statistical Learning

Bachelor in Data Science and Engineering

First Homework: Unsupervised Learning

Deadline: November 17, 2024 at 23:59

General Objectives

- Apply unsupervised-learning tools to an open data set.
- To do that, think about some goals or questions you want to answer.
- Then, download the variables (input) you need to attain previous goals
- The bigger and more diverse the input, the better the grades
- Apply all the unsupervised tools you can (remember each tool may have different versions)
- Get convenient insights and conclusions from the output of the tools

Open Data

Some interesting links to get data (but you can use any other):

- http://datos.madrid.es/
- http://datos.gob.es/
- http://open-data.europa.eu/es/data/
- http://data.gov/
- http://quandl.com/
- http://datacatalog.worldbank.org/
- https://research.stlouisfed.org/fred2/
- https://archive.ics.uci.edu/ml/index.html
- http://www.statsci.org/datasets.html
- http://lib.stat.cmu.edu/DASL
- http://www.umass.edu/statdata/statdata/
- http://www.philender.com/courses/multivariate/data.html
- http://biostatistics.iop.kcl.ac.uk/publications/everitt/
- http://www.oecd.org/statistics/

Open Data

Alternatively, you can use R-libraries to connect open data:

- World Health Organization `https://www.who.int/gho/en/` using the rgho R-library
- Air Quality `https://openaq.org/` using the ropenaq R-library
- World Bank `https://data.worldbank.org` using the WDI or wbstats R-library
- Organization for Economic Cooperation and Development `https://data.oecd.org` using the OECD R-library
- Financial and economic data using the quantmod R-library or the Quandl
- Weather data using the NOAA R-library
- Etc.

Evaluation

1. Data preprocessing (outliers, NA, feature engineering, etc.): 2 points
2. Visualization tools to get insights before the tools: 1 points
3. Principal Component Analysis: 2 points
4. Factor Analysis: 2 points
5. Clustering tools: 3 points

Important: you can take inspiration from other notebooks but you need to cite always the source.

Files to be evaluated

The evaluation process will be given by the following four files:

- **Document** (.pdf). A document with a maximum length of 5 pages that is a report of the work done. The report should contain a **summary** of the whole process, with some graphs, results, R outputs, and references. The conclusions shown in the paper must be reasoned from the perspective of the course.

- **Data**. The data file must be present.

- **Notebook** (.Rmd). The file must be fully executable. Errors will not be corrected. If not executed, the assignment will not be evaluated.

- **Notebook** (.html). The notebook will also be delivered in html format with comments of what the student is doing to verify with the document.

Upload the 4 files in a single zip folder to Aula Global. The **absence** of any file will result in the assignment **not being evaluated**.

Instructions

- The assignment has to be done in pairs.
- The assignment will be uploaded by both members of the pair.
- The evaluation will be the same for both members.
- **No assigment will be accepted after the deadline**.