UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

Course project on Advanced Engineering Data Analysis

MUEI, MUEA, and MASE
Prof.: Daniel Fernández

Any data matrix contains information about the generating phenomenon. The practice will consist of choosing a real problem and applying multivariate techniques to reveal the hidden information contained in the data set, as a prior step to modeling. The problem can be chosen from any repository.
Some examples of repositories are:

- UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/index.php
- Kaggle: https://www.kaggle.com/datasets
- IFCS Data Repository: https://ifcs.boku.ac.at/repository/challenge2/
- Google public datasets: https://cloud.google.com/bigquery/public-data/
- Eurostat database: https://ec.europa.eu/eurostat/data/database

You should get a dataset with a set of variables and 500 observations as a minimum. Namely, approximately every observation will have an identification variable ID (it can be the name of a person, a brand, a code, etc.), and 10 additional variables **at least**: a binary categorical variable, two polyatomic categorical variable (with two or more categories) and seven numerical variables. It will be appreciated the difficulty and originality of the dataset. If the chosen dataset does not meet the requirements, it will be punished in the mark.

It is still possible if you are interested in working with your own data set. In that case, you must ask for the **approval of the professor**.

The student group must perform a multivariate approach of the data matrix (i.e., implementation, visualization, and interpretation of the results, which must be contextualized).

The student must write a complete report upon the solution envisaged.

Suggested steps for conducting the practice

1. **Selection**: The student group will select a problem and its associate data set. You must read the corresponding documentation trying to understand what the goals of the problem are.
2. **Pre-process of data**: The student group will perform a preliminary summary of the data (exploratory data analysis), which will allow the detection of errors, outliers, and missing values and, thus, take the appropriate measures of correction. According to the problem and data, it may be necessary to perform a selection of variables (feature selection) and /or a derivation of new explanatory variables (feature extraction) if the problem requires it.
3. **Preparing data for analysis**: The student group will choose the type of protocol for the validation, i.e., split up the data set in a training and a test sample to assess the quality of the final model. Depending on the data size, it won't make sense to have a separate test data file.
4. **Analysis**: The student group will perform a multivariate analysis (descriptive analysis) of the training data set, leaving the test data as supplementary. As a group, you must discuss and choose which multivariate techniques are the more appropriate for your problem. Among other possible results, you must at least show a visualization of the information, detection of the hidden latent factors, the synthesis of the complexity by clustering, and interpretation of the results, which must be contextualized

Note: Regarding section 2, the visual/graphical part of exploratory data analysis is not studied in the MVA course. This is something that you already have the competencies and must research by yourself exploring ways of doing it in books or on the website.

The structure of the report should include:

1. A comprehensive description of the problem and its associated data set.
2. The pre-process of data
3. The protocol of validation
4. The visualization performed
5. The interpretation of the latent concepts.
6. The clustering performed.

7. The interpretation of the obtained clusters.
8. Discussion about the differences of the test sample concerning the training one.
9. Scientific and personal conclusions.

Final remarks:

- The weight of the project in the final grade is 40%.
- The student group will be formed by 3-4 people.
- Oral presentation: March 28th.
- The final report must be delivered in PDF or HTML format and updated in Atenea before the oral presentation.
- All members of the group must present and answer questions during the oral presentation.
- The analysis must be performed using R, RStudio, and RMarkDown.
- There is no extension limit but use common sense.