Module 6: Descriptive and Predictive Modeling

Exercise 2: A complete, non-guided Data Mining pipeline

Preliminary note: this exercise is to be completed in groups of 2 students.

Groups are asked to choose a classification or regression data set from those available in the literature (no more than 5000 examples to avoid lengthy training processes) and to select off-the-shelf models, namely, those learned during the classes and those available in the Scikit-learn Python library. Some public repositories to search and retrieve datasets of varying size and complexity can be found at:

- UCI repository
- https://github.com/caesar0301/awesome-public-datasets

The contribution will be especially valued with regard to:

1) The complexity of the data set in terms of:
    a. Missing values (→ data imputation)
    b. Mixture of categorical and continuous variables
    c. The fusion of different data sources towards the same goal
2) The difficulty of the modeling task:
    a. Clustering (hierarchical, partitional, fuzzy)
    b. Classification (binary, multiclass, multilabel)
    c. Regression (single-step, multi-step)
3) The usage of different models (not only linear), and their justification (NOT just a plain import of the models within the library)
    a. A justified selection of the cross-validation strategy
    b. A proper selection of the hyperparameters tuned for each model
    c. A thorough characterization of the error bias/variance of each model configuration
4) The utilization of a diversity of performance measures:
    a. Graphical plots showing the performance per class/per instance/per model
    b. The evaluation of the statistical significance of the differences between models in terms of performance using e.g. hypothesis contrast test (Wilcoxon) or boxplots
    c. Interpretation of results with respect to the problem at hand
    d. Consequences of modeling errors in the use case producing the data
    e. (When applicable) The representation of the decision regions of the models used in the case of using classifiers
    f. (When applicable) The representation of the learning curve for different levels of complexity of the selected model and / or number of training samples considered
5) The use of concepts not seen during the classes (i.e. non-linear selection of characteristics, selection of samples, balancing with weights), together with an explanation that demonstrates the assimilation of the concepts by the students

Reports can be a DOC document, a PDF document (along with the Python scripts that generate the reported figures and results) or a Jupyter Notebook (with saved checkpoint). Other formats (e.g. link to Google Colab) must be agreed with the professor.

When uploading the report, please indicate name, surname and ID (DNI number) of all members of the team.

Delivery deadline: March 7th, 2021