Departamento de Electrónica, Telecomunicações e Informática

EXPLORAÇÂO de DADOS

Homework I: Exploratory Data Analysis

1 Background

Exploratory Data Analysis (EDA) is used to study the data and generate hypotheses for further analysis. On the other hand, you can also use it to prepare the data for machine learning algorithms. Steps in data exploration and preprocessing are

- Identification of variables and data types.
- Statistical Univariate Analysis. Using descriptive parameters and graphical tools (histograms, boxplots)
- Missing value treatment.
- Variable transformations. Transformation to numeric values, scaling, normalization.
- Bivariate Analysis. Statistical parameters (Pearson correlation, ANOVA tests) and graphical tools (heatmaps or scatter plots).

2 Preparation

The Jupyter notebook is an interactive environment for writing and running code. In a notebook all comments and results interpretation can be embedded in as markdown cells, and you may use heading cells to further organize your document. The code cells are the default and it is where python code should be written. The folder *PreProcessingA* contains examples of Jupyter Notebooks whose contents is an illustration of relevant pre-processing steps. You can run the notebooks and answer to the questions that follows the coding blocks.
Pencil-and-Paper Exercises can be found in the Notes. Exercises of section 1 are about the pre-processing concepts.

3 HOME WORK

1. Programming task: Download one data set (possible links in Data Repositories) and construct a Jupyter notebook to perform the exploratory analysis you consider important of deal with the data set. Consider in your analysis the first four items of the section 1. Note that the notebooks available in *PreprocessingA* have illustrative examples applied to a data set.

2. Paper-and-pen task: Answer to two questions of section 1 of the Notes/Exercises. The answer can be handwritten and digitalized. For most of the exercises the answer (with justification) do not fill more than half a page.

The Jupyter Notebook and the Exercises should be download in Elearning.

4 Useful Links

- Jupyter `https://www.dataquest.io/blog/jupyter-notebook-tutorial/`
- Pandas. Data structures and data analysis tools for the Python programming language. `https://pandas.pydata.org/pandas-docs/stable/getting_started/index.html`
- Numpy. Multi-dimensional arrays and matrix. `https://www.tutorialkart.com/numpy-tutorial/`
- The pre-processing module of scikit-learn `https://scikit-learn.org/stable/modules/preprocessing.html`

5 Repositories

UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/index.php`). As possible choices check the following links

- `https://archive.ics.uci.edu/ml/datasets/Adult`
- `https://archive.ics.uci.edu/ml/datasets/HCV+data`
- `https://archive.ics.uci.edu/ml/datasets/South+German+Credit`
- and so on

Other machine learning repositories like Kaggle or OpenML can also be considered.