Data Analysis Lab

Degree in Information Technology

2021/2022

Goal

The practical work of Data Analysis Lab subject aims:

- Learn to analyse data sets identifying key variables.
- Apply and adapt math and statistical models for *Machine Learning*.
- Analyse and interpret the results doing a critical analysis.

Description

The practical work will be developed using Python language and it is composed of two different components:

1. The first part includes a statistical analysis (40 %). The student should do an exploratory analysis.
2. In the second part the student must apply *Machine Learning* models, using the statistical analysis performed in part 1 (40 %).

Part I (30 %)

In the first part of the work, the student should:

- Choose a data set. The data sets should be different for each student. The student can use the toy data sets available in Python:
    - Boston House-prices dataset
    - Iris plants dataset
    - Diabetes dataset
    - Physical exercise linnerud dataset.
    - Wine recognition dataset
    - Breast cancer Wisconsin (diagnostic) dataset
    - UCI Datasets
- Make a description of the dataset's characteristics (*e.g.*, domain, size, data types, entities, *etc.*); (5 %)

- Develop a statistical analysis using measures such as average, variance, covariance, or correlations; (10 %)
- Make a graphic and coherent representation of the results; (10 %)
- Critical analysis; (5 %)
- Elaborate the first report (2.5 %) and the corresponding oral presentation (2.5 %).

Part II (70 %)

In the second part, the student should apply a Machine Learning algorithm identifying:

- The data set feature to learn.
- The final goal - prediction or classification (10 %)
- Apply and compare machine learning models (20 %)
- Perform a Cross Validation (20 %)
- A user interface for your predictive or classification application (10 %)
- Elaborate the second report (5 %) and the corresponding oral presentation (5%).

Example: Classify a tumour as benign or malign.

Note that the goal is to explore Machine Learning algorithms and understand how they work. In this context, ***it is forbidden the usage of python library auto-scikit learn***.

Submission

The work must be carried out in group. Each work should be submitted in Moodle using a ZIP archive with the PDF report and all Python scripts. In addition, the student should include the declaration of authorship also available in Moodle.

Each report has a limit of 10 pages. It should include the student identification and a clear description of the Python scripts elaborated to data analysis. The deadlines will be announced in Moodle. The oral presentation is always in the next class. If possible, it will be done presential.

References and Resources

The student should consider the bibliography of this subject as well as the material provided by professor. The development environment is free choice. However, we recommend using Pycharm.

Good work!

Fátima Leal