Data Analysis Lab

Degree in Information Technology

2024/2025

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 (60 %).

Part I (40 %)

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

- Choose a data set. The data sets should be different for each group. The student can:
  - UCI Datasets
  - Kaggle Datasets
  - Other repositories
- 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, correlations, etc.; (5 %)
- Create new features to enrich the analysis (2.5 %)
- Develop graphical analysis (10 %)

- Make a graphical dashboard with a coherent representation of the results;. For the Graphical User Interface (GUI), you can use the technologies at your convenience. (10 %)
- Critical analysis in the report; (2.5 %)
- Do the normalisation and standardization of your data and analyse the new data distributions. (2.5 %)
- Elaborate the first report and the corresponding oral presentation (2.5 %).

Part II (60 %)

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

- The data set feature to learn (5 %) .
- Describe the final goal - prediction or classification (5 %)
- Apply and compare machine learning models (10 %)
  - Linear Regression
  - Logistic Regression
  - Ridge and lasso regression (including alpha analysis)
  - Naïve Bayes
  - SVM including the analysis of multiple kernels
  - K-NN explaining the optimal number of neighbours
  - Decision trees. Draw a legible tree and analyse the gini
  - Ensemble methods. In the case of Random Forest it is required an image of the tree
  - Neural Networks: single layer and multi layer
- Analysis of the fit time per machine learning model (5 %)
- Apply the Principal Component Analysis and compare accuracy and time execution of the previous machine learning models. Use the dimensionality reduction based on SVD. (5 %)
- Identify a clustering analysis to execute in your dataset. Employ clustering and hierarchical clustering drawing dendograms. It should include the analysis of the optimal number of clusters. (5 %)
- Perform a Cross Validation (5 %)
- Analysis of the RMSE or Precision, recall or F-measure using proper graphs
- Join to the dashboard the predictive or classification application using the best algorithm identified in the previous steps (5 %)
- Elaborate the second report and the corresponding oral presentation (5 %)

Challenges: (10 %)

- Employ recommendation algorithms to your dataset when possible. **If it is not possible you should justify why.**

- Using the ***auto-scikit learn*** check if the best algorithm provides the same results than your previous analysis.

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.

The report 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