Project Guidelines

AA1, GCED, UPC

This document contains the guidelines for the practical work (the *project*). Please read with care!

General Information

This project is meant to give you the opportunity to apply the techniques seen during the course to a real-world dataset. The project should cover all aspects of the modelling methodology seen in class from preprocessing to generating a final predictive model together with an assessment of its prediction quality.

The project is to be done in teams of **two** persons; singles are not allowed. Once you have chosen your partner, the next thing will be to select a dataset to work on. You can choose one from the data repositories we are going to provide to you (see below), or propose your own.

In any case, you are expected to hand in a final written report. This document should describe the work carried out, its motivation, the problems encountered and the solutions found together with final results, interpretation and conclusions of your study. More details follow.

To carry out your analysis you should use the language `python`. Remember that there are many useful packages that extend its basic functionality. Certainly you can find inspiration in the notebooks from our weekly laboratory sessions. If you use code or ideas or any kind of resource from elsewhere you should cite it appropriately. Plagiarism will be prosecuted.

Data repositories

The following sites contain a number of very diverse datasets; many correspond to real-world problems. They vary in domain (biology, medicine, economy, etc.), and also in size, type of variables, type of problem (classification or regression), among other things. *Please chose one that interests you!*

- Open ML [https://www.openml.org/search?type=data]
- UCI Repository [http://archive.ics.uci.edu/ml/index.php]
- UCI KDD Archive [http://kdd.ics.uci.edu/summary.data.application.html]
- Statlib [http://lib.stat.cmu.edu/datasets/]
- Delve [http://www.cs.utoronto.ca/~delve/data/datasets.html]
- School of Informatics (U. of Edinburgh) repository [http://www.inf.ed.ac.uk/teaching/courses/irds/miniproject-datasets.html]

Requirements for dataset/problem chosen

1. The dataset of this problem has numerical and categorical variables.
2. The dataset of this problem is not synthetically generated.
3. The dataset of this problem contains more than 10 variables.
4. You need to have enough information about the problem to be able to understand and analyze your results. Just getting random data and feeding it to the machine is not valid.
5. Datasets already pre-processed are not valid. You need a problem which data has any pre-processing work to do.
6. The dataset of this problem contains more than 200 samples.
7. The problem to solve is not one of the simple known problems like the `iris`, `mnist` or `wine`, etc.

How and what to submit

You are expected to submit work on your project on three occasions, only the last one counts toward the project grade:

- **March 20th**: declaration of dataset and team. Some preliminary information on the dataset, such as number of rows/columns and nature of columns should be provided. A 1-page pdf file suffices.
- **April 17th**: first delivery. This is a first approximation to the project and should include preprocessing/cleaning, and some initial modelling. It won't be graded however you will get feedback from your lab instructors within a reasonable amount of time. A pdf report is expected around 7-8 pages.
- **June 5th**: final delivery. Includes final report and code. *Code and report should be submitted **separately*** (namely, a *python notebook* is not a report).

The final report should include:

1. A brief description of the work and its goals, data available, and any additional information that you may have used.
2. Related previous work (if applicable)
3. The data exploration process, including: pre-processing, feature selection/extraction, visualization, clustering, etc.
4. Modeling methods considered, validation protocol and the reasons why the choices were made.
5. Results obtained with each method used (along with best set of parameters), comparison of results.
6. Final model chosen and an estimation of its generalization performance.
7. Scientific and personal conclusions
8. Possible extensions and known limitations.

Note that the report should not describe explanations seen in class; every table or plot should be appropriately described. The style of the report should resemble what you encounter in a scientific publication. Your code should be **reproducible**; that means using "seeds" if your code is stochastic.

Make sure you include a variety of linear and non-linear methods seen during the course.

All deliveries are to be made exclusively through the racó. An appropriate mechanism will be prepared for every delivery. **Important: only one member of the team should upload the material**

For the final delivery, make sure you include in a compressed file the following:

1. The written report (pdf document). It should not exceed **15 pages**; if you need more space, consider placing the secondary information in a **separate appendix file**.
2. Any script or code you have used (`python` notebooks, scripts, or any other code)
3. A flat text file with precise instructions on how to execute and reproduce your results.

Evaluation

Your final project will be evaluated on the basis of the clarity of your report as well as on its technical quality. Conditions for a good score are:

1. The appropriate use of techniques and methods seen in class
2. Care and rigor for obtaining results (resampling protocol, quality metrics, etc.)
3. Quality of obtained results (generalization error, simplicity, interpretability)
4. Quality of written report (conciseness, completeness, clarity, appropriate format of report etc.)

Special attention will be given to insights into the results obtained, gaining knowledge on the data analyzed and obtaining useful conclusions. All experimental decisions should be appropriately justified (on the resampling protocol or preprocessing, for example). Merely applying the methods to the data and showing the table of results is not enough, there has to be an interpretation of the results obtained.