Project Guidelines

Machine Learning, MDS-MIRI Master

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 modeling 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. In case there is an odd number of people, then we would admit a team of **three**. 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 two deliverables:

1. Written report. This document should describe the work carried out, its motivation, the problems encountered and the solutions found together with final results and conclusions of your study. More details follow. This is a scientific report, not a jupyter notebook as a .pdf. It can include plots in the annex. It should have a bibliography with correct format of citations.
2. Poster or video. The poster should summarize the most important aspects of the project. Alternatively, you can produce a short video (no more than 3-4 minutes) explaining the most important parts of your work. All members of the team should participate in the video more or less equally if you decide to do that. Look for conference poster templates on google. It should be mainly visual - imagine explaining the project to a friend. The report will be more technical.

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.

As a first task, you are expected to submit a 1-page project proposal. It should include the composition of your team, the problem that you are going to work on, the reason why you chose it, together with any references of previous work on this problem if applicable, and a title for your project; please include information on the data such as number of rows, columns, and nature of variables (e.g. if categorical, how many levels).

Finding teammates

I have opened a special "Forum" at the racó so that you can find teammates for the project for those of you who do not know classmates that well.

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 choose 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]

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

The final report should include:
We need to explain our reasons for different pre-processing decisions.
Explain validation protocol and reasons for choosing it.
Results need to be clearly explained for each method and from which partition they come from.

1. A brief description of the work and its goals, data available, and any additional information that you may have used. *Include an introduction.*
2. Related previous work (if applicable) *Use Google Scholar to look for relevant examples from this data or something similar. Be brief.*
3. The data exploration process, including: pre-processing, feature selection/extraction, visualization, clustering, etc. *Explain pre-processing choices, like imputation, transformations, etc.*

It is recommended to try at least three methods from different families.

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. *Summarize results in table to save space.*
6. Final model chosen and an estimation of its generalization performance. *Choose final model and explain why.*
7. Scientific and personal conclusions *Explain which parts of project were difficult or interesting*
8. Possible extensions and known limitations.

It is important not to get distracted with extra things and and leave out one of these parts of the statement. This is the grading criteria.

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**. *strong penalization for extra pages!!*
2. Any script or code you have used (python notebooks, scripts, or any other code) *Github link is okay.*
3. A flat text file with precise instructions on how to execute and reproduce your results.

Code needs to be provided, but the teachers should not need to use it for the evaluation. If they need to use it, it indicates a bad problem with the report.

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.

The poster or video

Additionally, you are asked to either produce a poster or a video. These are meant to be a summarization of your work where you briefly explain the most important points from your work (what you have done, how and why). This will serve as the basis of 15% of your final course grade that corresponds to a *reasoning* skill. Summarize important parts of project. A video could be 3-4 minutes talking about some slides.

Dates Look for conference poster templates on google. It should be mainly visual - imagine explaining the project to a friend. The report will be more technical.

- 1-page project proposal: April 8th
- Project submission: June 10th.
- Video/Poster: June 15th.