Advanced Machine Learning (MDS)
Project Guidelines, Fall 2022

Lluís Belanche

October 17, 2022

Abstract

This is a brief guide for the correct development of the practical work of the course (the **project**). The students must apply the different concepts and models lectured during the course to solve a real problem, providing a feasible solution intended for the final user. The students must write a complete **report** describing the work carried out, the problems encountered and the solutions envisaged, as well as the final results and conclusions of the study.

Please read this document carefully!

1 General Information

There are two basic possibilities to develop a term project.

- Choose a practical problem from one of the provided data repositories or specific problems (see below) and develop a solution (a classification or regression model);
- Bring your own problem (theoretical, practical or both); in this case you are responsible for getting the necessary data, if needed

In both cases –but specially in the latter– you will need to announce your choice and get the "go ahead" answer from the teacher.

The most important **guidelines** are:

1. You can choose to explore any problem that motivates you. In every case, you are expected to write a complete report describing the work carried out, its motivation, the problems encountered and the solutions envisaged, and the final results and conclusions of the study. **The final main text is (strictly) limited to 20 pages**;
2. It is expected that you make a proposal for the project for preliminary evaluation. Proposals should be submitted through the "Racó" no later than **October 25, 2022**. Submit your proposal (preferably) as a **1-page pdf**; it is enough that one member of the team submits this through the "Racó". Your project proposal should specify: which problem you want to tackle, why you choose this problem, a couple of fundamental references, a preliminary title, and a list of team members;
3. The computer language used for the modeling part can be R (`http://cran.r-project.org/`) or python (`https://www.python.org`) or both. Remember that there are many useful packages (especially for R) which you can use to extend its basic capabilities. Pre-processing or any other non-modeling tasks may be done in any other languages of your choice;
4. If needed, additional information on the methods or on the problems may be obtained. Some of the web repositories (see Section 4) contain previous usage of the data; information can also be gathered from textbooks, other courses, domain experts, the web ... and maybe from the teachers. Please acknowledge or cite properly everything you use.

2 Deliverables and delivery mechanism

The final report should include:

1. A description of the work and its goals, the available data, and any additional information that you have gathered and used
2. A description of related previous work (if applicable)
3. The data exploration process (pre-processing, feature selection/extraction, clustering, visualization, etc)
4. The resampling protocol; the modelling methods considered, reasoning the choice
5. The results obtained with each chosen method (along with the best set of parameters) and a comparison of these results
6. The final model chosen and an estimation of its generalization error
7. Scientific and personal conclusions
8. Possible extensions and known limitations
9. References

You will be required to submit the full code and a brief text file with instructions on how to execute your code (make sure that your results are *reproducible*, for example, by using "seeds" in random processes, etc.). Nothing needs to be delivered on paper. The report should *not* include technical explanations seen in class; please do not include tables or plots without explanation.

All deliveries are to be made exclusively through the "Racó". An appropriate mechanism will be prepared for every delivery. For the final delivery, please be sure to include the following (please compress everything into a single file):

1. A document (written report). This document has to open with a (maybe hyperlinked) standard pdf reader and should not exceed 20 pages. If more space is really needed, place information of secondary importance in a **separate appendix file**
2. One or more containing all the necessary code
3. Additional files with the rest of the code in other languages that you may have used (e.g., for pre-processing or plotting)
4. A flat text file with precise instructions on every step needed to reproduce your final results

3 Evaluation

The grade will be partly based on the clarity of your report, so please make sure your final report is well organized and clearly written. There should be an introductory part explaining the basics of your work, and a conclusions section, basically stating what you know compared to what you knew before the work started; also any gaps, possible extensions or limitations in your development should be noted and explained. Your work will also be evaluated based on technical quality. This means that the techniques you use should be reasonable, the stated results should be accurate, and technical results should be correct and complete, whether they are your own work or not.

In summary, these are the conditions for a high score (in this order!):

1. The proper use of techniques and methods presented in class
2. The care and rigor for obtaining the results (resampling protocol, quality metrics, statistical significance)
3. The quality of the obtained results (generalization error, simplicity, interpretability)
4. The quality of the written report (conciseness, completeness, clarity)

4 Data repositories

You can browse the following data repositories, among others:

- 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
- The Statlib database
  http://lib.stat.cmu.edu/datasets/
- The Delve project
  http://www.cs.utoronto.ca/~delve/data/datasets.html
- The School of Informatics (University of Edinburgh) repository:
  http://www.inf.ed.ac.uk/teaching/courses/irds/miniproject-datasets.html
- Luis Torgo's compilation of datasets (regression only)
  http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html

Additionally, you may want to have a look at "competition" web sites, like `kaggle`, which hosts a growing number of datasets: https://www.kaggle.com/datasets.
Given the heterogeneous nature of this site, the prior submission of a proposal is of the utmost importance. Most of the problems are real-world tasks. Their origins are very diverse, not only regarding the area of work (biology, geophysics, medicine, etc) but because they show different data characteristics. For example, there are great differences in the number of variables and examples, number of classes, intrinsic difficulty, lost values, various errors, mixed nominal and/or continuous variables, etc. Other problems are synthetic (they have been generated by a program), and their characteristics are completely known. However, their study is interesting for a number of reasons, including meaningful (as well as significant) comparisons of different learning algorithms.

Some problems are easier in some aspects and more difficult in others. Therefore, the selection of the particular problem does not have a lot of importance for the grade. In particular, it is not at all advisable that you start to test problems to see how they "behave". It is recommended that you make the decision by the interest that it raises in you.

The main "warning" we give is to be aware of the large computational needs that some (or many) of the chosen methods may have. If you choose a problem with a number of rows in the order of 10^5 or more, you will surely get into serious demands of CPU/GPU and RAM ... Unless you do something to reduce these needs, it is wise to keep this number in the order of $10^2 - 10^5$.

5 Pre-processing (prior to analysis)

Each problem requires a different approach in what concerns data cleaning and preparation, and the selection of the particular information you are going to use can vary; this pre-process is very important because it can have a deep impact on future performance; it can easily take you a significant part of the time. It is then strongly advised that you analyze well the data before doing anything, in order to gauge the best way to pre-process it. In particular, you shall pay attention to the following aspects (not necessarily in this order):

1. treatment of lost values (missing values)
2. treatment of anomalous values (outliers)
3. treatment of incoherent or incorrect values
4. elimination of irrelevant variables
5. (possible) elimination of redundant variables

6. coding of non-continuous or non-ordered variables (nominal or binary)
7. extraction of new variables that can be useful
8. normalization of the variables (e.g. standardization)
9. transformation of the variables (e.g. correction of skewness and/or kurtosis)

6 Delivery dates

- **December 22, 2022**. First round.
- **January 16, 2022**. Second round (will incur a penalty of 0.5 points)

Remember: The project is to be developed in **groups of 2 people**. Only one member of each team should submit information (**always via the "Racó"** at https://raco.fib.upc.edu). For your convenience, a "Forum" is open starting today, to facilitate the finding of mates.