Advanced Machine Learning (MDS)
Guidelines for the projects, Fall 2025-26

Lluís Belanche

October 21, 2025

Abstract

This document provides guidelines for the correct development of the practical work of the course, consisting of **three independent projects**. Each project corresponds to one of the three major parts of the course syllabus. Students must apply the different concepts and models lectured during each part to solve real problems, providing feasible solutions intended for the final user. For each project, 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

1.1 Structure and rationale

The course practical work is organized into three independent projects, each aligned with one of the three main parts of the syllabus:

- **Project 1**: Focuses on methods from Part I (Bayesian thinking for ML, GLMs, Bayesian generative methods, Perceptrons, Delta rule and its variants, theoretical foundations, ...and applications thereof)
- **Project 2**: TBA
- **Project 3**: TBA

Each project is independent and will be evaluated separately.

1.2 Project Selection Options

For *each* of the three projects, there are three possibilities:

1. Choose a practical problem from one of the provided data repositories or problems (see Section 4) and develop a solution (a classification or regression model);
2. Bring your own problem (theoretical, practical or both);
3. Choose a paper (a scientific publication) that motivates you and try to reproduce its results and evetually go one step further; in this case, you are responsible for getting the necessary data (if applicable, same or similar as in the paper)

All work must be original!

Important considerations:

- You may use the same dataset across multiple projects, but you **must** apply different methodologies corresponding to each part of the course;
- Alternatively, you may choose different problems for each project;

- Each project must focus on the techniques covered in its corresponding course part;
- **Make sure you have the required computational power for each project**;
- **Make sure you can get a sufficient amount of data, depending on the task**.

1.3 General Guidelines

- You can choose to explore any problem that motivates you. In every case, you are expected to write a complete report for each project 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 for each project is (strictly) limited to 10 pages**, not including Appendixes but inclusing References; a template will be provided;
- **Theoretical rigor is essential in all projects**: Regardless of whether your work is empirical, analytical, or mixed, you must demonstrate deep understanding of the methods you use, provide rigorous justification for all choices, use precise mathematical notation where appropriate, clearly state assumptions and their implications, and analyze why methods work or fail. Theoretical rigor applies to experimental design, method selection, result interpretation, and all aspects of your work;
- The choice of a project related to LLMs is not recommended given the high complexity of these models;
- It is expected that you make a proposal for each project for preliminary evaluation. Specific deadlines are provided in Section 6. 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, which techniques from the corresponding course part you plan to apply, a couple of fundamental references, a preliminary title, and a list of team members; additionally, data sources and computational resources you expect to have. Among all, what you *expect to learn* from doing the project;
- The computer language used for the modeling part can be R, python or Julia. Remember that there are many useful packages which you can use to extend the basic capabilities. Pre-processing or any other non-modeling tasks may be done in any other languages of your choice, such as Perl;
- 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;
- If you use ChatGPT (or another similar tool) in the document, indicate it every time it is used. We want to evaluate your work, not "someone" else's! You can find information on how to cite it in Citing Generative AI.

2 Deliverables and Delivery Mechanism

For each of the three projects, the deliverable should include 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"; 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 10 pages. If more space is really needed, place information of secondary importance in a **separate appendix file**
2. One or more files containing all the necessary code (no githubs)

3. One or more files containing all the necessary datasets, or links to them
4. Additional files with the rest of the code in other languages that you may have used (e.g., for pre-processing or plotting)
5. A flat text file with precise instructions on every step needed to reproduce your final results

3 Evaluation

Each project will be evaluated independently. The grade for each project will be partly based on the clarity of your report, so please make sure each 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 for the corresponding course part, the stated results should be accurate, and technical results should be correct and complete, whether they are your own work or not. **Theoretical rigor, deep understanding of methods, and rigorous justification of all choices are essential components of technical quality.**

For each project, these are the conditions for a high score (in this order!):

1. The proper use of techniques and methods from the corresponding course part
2. Theoretical rigor and deep understanding: precise mathematical formulations, clear statement of assumptions, rigorous justification of choices, and critical analysis of method properties
3. The care and rigor for obtaining the results (resampling protocol, quality metrics, statistical significance, proper experimental design)
4. The quality of the obtained results (generalization error, simplicity, interpretability)
5. The quality of the written report (conciseness, completeness, clarity)

Final grade: The three projects will contribute equally to the practical work component of the course (i.e., each project represents approximately (33+1/3)% of the practical grade). **The rubric used by the teacher will be delivered with sufficient notice.**

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. Note however that in many occasions the dataset has been preprocessed (often multiple times) and may not correspond to the original one.

It is not a bad idea to use synthetic problems (they have been generated by a program), because their characteristics are completely known. Their study is interesting as a prior benchmark for a number of reasons, including meaningful (as well as significant) comparisons of different learning algorithms.
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. Given the heterogeneous nature of these sites, the prior submission of a proposal for each project is of the utmost importance.

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.

5 Pre-processing

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

6.1 Project 1

- **October 21, 2025**. Project 1 start
- **November 17, 2025**. Project 1 delivery

6.2 Project 2

- **November 18, 2025**. Project 2 start
- **December 16, 2025**. Project 2 delivery

6.3 Project 3

- **December 17, 2025**. Project 3 start
- **January 16, 2026**. Project 3 delivery

Remember: The projects are to be developed in **groups of 2/3 people**. Only one member of each team should submit information (**always via the "Racó"** at https://raco.fib.upc.edu). The same team composition must be maintained across all three projects unless exceptional circumstances arise (to be discussed with the instructor).