ML-MIRI Project Guidelines, Spring 2019

Lluís Belanche, Marta Arias

March 22, 2019

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. A **poster** will also be asked, in order to complement the presentation. Both deliveries will be graded.

Please read this document carefully!

1 General Information

All students enrolled in ML-MIRI are required to complete a term project. There are two basic possibilities:

1. 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);
2. Bring your own problem (theoretical, practical or both); in this case you are responsible for getting the necessary data, if needed

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 main text is (strictly) limited to 12 pages**. An accompanying poster will also be asked, summarizing the main information of the project (problem definition, main findings, ...) in an attractive format. All posters will be uploaded to the teacher's docency webpage and made public for everyone to have a look.

It is expected that you make a proposal for the project for preliminary evaluation. Proposals should be submitted through the "Racó" no later than **April 9th, 2019**. Submit your proposal 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.

The computer language used for the modeling part must be R (http://cran.r-project.org/). Remember that there are many useful packages 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.

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 yes, from the teachers. Please acknowledge or cite properly everything you use.

2 Deliverables and delivery mechanism

The final report should include:

1. A brief description of the work and its goals, the available data, and any additional information that you have gathered and used

2. A brief description of related previous work (if applicable)

3. The data exploration process (pre-processing, feature selection/extraction, clustering, visualization, etc)

4. The validation protocol, and the modeling methods considered, reasoning the choice

5. The results obtained with each 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

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 (compress everything to a single file):

1. A document (written report). This document has to open with a standard pdf reader and should not exceed 12 pages. If more space is needed, place information of secondary importance in a **separate appendix file**

2. One or more text files (.r or .R) containing all the necessary R code

3. Additional files with the rest of the code in other languages (perl, python, awk, etc) 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

3.1 Final report

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)

3.2 Posters

As you probably know, there is a "generic competence" (or *soft skill*) associated to this course: REASONING, which is worth a 15% of the final grade. This grade is obtained from a poster. More information and help on how to create good posters will be given at due time.

4 Data repositories

Browse the following data repositories:

- Open ML
  https://www.openml.org/search?type=data
- UCI Repository
  http://www.ics.uci.edu/~mlearn/MLRepository.html
- 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

... and *choose one* of the problems. Many of the problems therein 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.

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 Key dates

- **April 2, 2019** (preferred date). Project proposals (1-page pdf file) to be submitted through the "Racó". Maximum date: April 9, 2019
- **June 22, 2019**. Final report and code to be submitted through the "Racó".
- **June 27, 2019**. Poster deliveries

Remember: The project is to be developed in **groups of 2 people**; groups of 3 people will be allowed subject to an explicit ack and based on a reasonable cause. Only one member of each team should submit information through the "Racó". The address is https://raco.fib.upc.edu. For your convenience, a "Forum" is open starting today, to facilitate the finding of mates.