M.R.R. Project 2022

A. Charantonis, M. Mougeot, S. Oger, J.Park

November 1, 2022

Abstract

In this project you will have to apply the notions introduced in the course of M.R.R on a dataset. It will provide a basis for your final evaluation. As such, you are considered to have read through this document pertaining to the particulars of this evaluation.

1 Introduction

Regularized regression methods are, in a statistician's toolbox, one of the most useful tools. In this project you will be called to:

1. Describe the dataset you will work on.
2. Detail the model selection methodology you will apply, which can include (among others):
   - using a baseline model;
   - performing variable selection through stepwise and/or penalized regression methods;
   - using one method that is the extension of what was seen in the course. Select among the KNN, K-means + Regression, group-Lasso, or Elastic-Net methods.
3. Justify the choice of one of these models over the others.

In this document you will find a link to your associated data-set, a more detailed explanation of our expectations throughout the five project sessions, as well as information on your evaluation.

You will need to treat this project as a full research project on the given data-set and be able to show maturity both in *what* you present us and in *how* you present it.

2 Getting the Data

2.1 Binomials

As mentioned during the practicals you will be evaluated in Binomials. Each Binomial will have a data-set attributed to them, which will be found starting the 7th of November at:

https://tinyurl.com/MRRPROJECT2022.

If you are reading this, it means that binomials can no longer be changed and you have either already filled the form as you were asked during the last practical, or you will soon be randomly attributed one. You cannot change your binomial.

The data-sets concern either a classification or a regression problem. They may be associated to a scientific publication or detailed appendices which can give you extra information on the subject mater. If that is the case you are expected to have read the associated publication. If that is not the case, you should perform a small bibliographical search on the topic before delving into the dataset.

Plots	Labeled	Good Quality	Grade
Yes	Yes	Yes	Positive
Yes	Yes	No	Negative
Yes	No	No	NEGATIVE
Yes	No	Yes	NEGATIVE
No	Yes	Yes	**NEGATIVE**
No	Yes	No	**NEGATIVE**
No	No	No	Negative
No	No	Yes	**NEGATIVE**

Table 1: An example of a well labeled and cross-referenced table, which presents the effect that having plots, labeling them correctly and making sure they are of good quality, have on our grading of the A4 page you need to provide on your data-set description. The capitalization and boldness of letters correspond to the intensity on the grading of the page.

2.2 Statistical Analysis and Description

Before the **third** project session, we will expect a ***printed*** A4 page containing all the information you would provide to someone who knows nothing about your project in order to summarize to them:

- The **names** of the authors of this document, their binomial's number and the name of the dataset explored.
- The general nature of the problem.
- The explanatory variables available.
- The target variable and its links to the explanatory variables.

Deposit this ***printed*** A4 page before the third session in M.Charantonis's mailbox in the teacher's lounge at the first floor. Any delay will **penalize** your final grade.

You should try to make this document accessible, readable and include some plots or tables that illustrate your remarks. Do not forget to clearly label such plots or tables and cross-reference them in the main body of the text, such as the one seen in Table 1, which shows the effect that missing labels and poor quality figures have on the grading of your A4 page. In it we can notice that commenting on non-existent plots is worse than not commenting on them. This is inferred by interpreting the capitalization and boldness of the letters as an indicator of the intensity of the effect on grading. You can use both sides of the A4 paper.

3 Methodology

Simply applying functions without a clear objective is a lost cause. Before the **fifth project session**, we will be requiring to give us a **printed** A4 page containing your projected methodology to select a regression or classification model for the problem. The impact and delivery instructions are the same as in Section 2.2.

As such we expect a well thought-out document elaborating the different steps you have taken and/or plan to take in order to select the optimal model for the task. As mentioned in Section 1, we expect a justification in the ordering of the different phases of your project.

Should you wish to use an extra-curricular method to regress or classify your data, if that method is not in the list of methods provided in Section 1, then you need the oral approval of one of the course's teachers.

4 Evaluation

The evaluation of this project will be in two parts:

- The two A4 pages detailing, respectively, the data-set and the methodology on and with which you will be working.

- The 8 minutes presentation (4 minutes per student) and 4 minutes of questions each binomial will give at the sixth project session, the 12th of December, 2022.

This year the project can be done either in R or in python, at your discretion. Should you use python, keep in mind that the objective is to produce a detailed statistical analysis of the problem studied, not just an application of functions. This will be reflected in the questions we are going to be asking you throughout the practical and during the presentation. As such we expect you to fully understand the mathematical underpinning of any function you use.

You will also be required to deposit your R Markdown or .ipython notebook file of the project in a depository whose address we will give during one of the project sessions. You should **strictly** follow this naming convention for this .Rmd file: BINOMIAL_N_FAMILYNAME1_FAMILYNAME2.Rmd or BINOMIAL_N_FAMILYNAME1_FAMILYNAME2.ipynb . Do not forget to indicate your names and binomial number in your RMarkdown or ipynb file too.

4.1 The A4 Pages

Both sides of the A4 pages can be used, should you consider that important. You will be evaluated on the pertinence, conciseness, ease to read and clarity of the pages. The A4 pages must:

- Contain your binomial's family names.
- Be ***printed and delivered*** before the start of the third and fifth project sessions, respectively, for the data-set report and the methodology report.

4.2 The Oral Presentation

For your oral presentation you need to provide a .pdf file of maximum 12 slides, which will need to be deposited ***2 days*** beforehand in the corresponding repository, the address to which will be communicated during the fifth session. **Any** other file format, or **any** delay in providing the slides in a .pdf format will ***be heavily penalized***. Should you miss the deadline you will furhtermore have to give your presentation without any supporting slides. You need to ***strictly*** follow this naming convention: `BINOMIAL_`N`_FAMILYNAME1_FAMILYNAME2.pdf`, where N is the binomial number number found in the link in section 2.1.

We will have with us the corresponding printed A4 paper reports of your binomial. Each member of the binomials should:

- Present for 4 minutes during the oral presentation.
- Be able to answer questions both on the project and on the theory of the methods used.

The time limits will be scrupulously followed. In that time you will need to present the project, the database, your methodology and conclude with your final model selection, justifying your choices along the way. You can and should refer to any associated research article provided alongside the data-set or bibliographical research you did on the project. The 4 minutes are to be consecutive for each participant.