Final project

Advanced Statistical Learning (80619)
Winter 2021
Professor: Aurélie Labbe

Weight: 40% of the final course mark (25% report and 15% presentation)

Due date: April 24th for the reports. Oral presentation during the week of April 26th.

Description of the project

Each team (2 or 3 students) chooses a subject. The list of subjects is provided in the Google sheet below:
https://docs.google.com/spreadsheets/d/1gcu2c2BR7zeThUQ-58IBxO3k4J16ip3XDP76mJrf58s/edit?usp=sharing
You can also propose your own subject, which must be approved by the professor.
Once you choose your subject, your team must be registered in the Google spreadsheet.

The work consists of **preparing a detailed tutorial to perform analyzes related to this subject with R**.
During the course of this project, you are expected to:

- Deepen the subject and do a brief literature review on the subject.
- Do an exhaustive research of the resources available in R to perform analyzes related to this subject. For example: basic R functions, packages on different platforms (CRAN, R-Forge, etc.), code that can be found elsewhere on the web, code developed by you, etc…
- Perform analyses with data (tutorial)

At the end of the term, the students' work will be put on a sharing site and accessible to the other students of the course. This way, all students will have access to a bank of tutorials and code on a variety of topics.

Specific instructions for the team report:

The number of pages is limited to 30 pages **all inclusive** (text, code, outputs, tables, figures). No appendix is accepted. The 30 pages do not include the presentation page and the bibliography.

The report must include:

1. Presentation of the subject.
2. Brief review of literature on this subject. May include web pages.

3. Brief description of the methods.
4. Review of R resources (may include web pages). Include also a summary table with the characteristics of each resource.
5. Examples of analyzes with data and code. These data may be real data available in packages, on specialized sites (e.g. UCI machine learning repository), or artificially generated data. **You cannot reproduce a data analysis available on a web site or in another tutorial.**

Sections 4 and 5 should form the core of your report. The review of R resources (section 4) should be comprehensive. If there are really multiple resources, the examples (section 5) may focus on some of them only. Section 5 should be written in the form of a tutorial, as for example in the R vignettes available in some libraries (see for example https://cran.r-project.org/web/packages/httr/vignettes/quickstart.html).

Everything must be **reproducible**, as far as possible. One should be able to easily reproduce all the results presented in the report. To do so, **we ask you to write your project in an environment that allows automatic integration of the R code and outputs into the text of the report** (Markdown or RSweave for example). In order to ensure reproducibility, remember to:

- Use seeds for random numbers (set.seed function for example)
- Annotate your code
- Indicate at the beginning of the code which packages and data files are required
- Provide data files, as well as the code to import them.

Submission of the project: please submit a single .zip file which contains: a) the pdf document of the report (which includes everything, even the code), b) the data necessary to run your code, c) a .R file containing only the code. This code must be identical to the one given in the report, but you can add the data pre-processing steps if necessary.

Evaluation criteria for the report: they will include the following elements (without fixed weights):

- Quality of the presentation of the subject
- Quality and relevance of the literature review
- Clarity and correctness of the method description
- Relevance, depth and organization of the review of resources
- Relevance, clarity, and usefulness of the data analysis
- complete reproducibility of the results
- general depth of the project
- general presentation of the report

Specific instructions for the individual presentations (15%):

Each student presents his/her work in front of the professor by zoom. Duration: **15 minutes** + questions.

The evaluation criteria are:

- Relevance of the elements presented
- Clarity
- Knowledge of the subject
- Quality of the presentation
- Respect of allocated time
- Quality of the material (PowerPoint or other).

Some common errors in last year's reports

1) Many students used a dataset that was too large or that took too long to compile. You might not need millions of observations to illustrate the use of a method.

2) Instructions related to the installation of specific libraries were not provided (problem with reproducibility)

4) The R code provided was different from the code included in the report

5) In several reports, the general analysis plan was not outlined at the beginning of section 5 (data analysis). Providing the big picture to the reader before going into the details of the analysis is important.

6) Several reports didn't include a reference for the figures and tables included in the report

7) A literature review should go beyond a list of articles listed one by one. We expect this to form a whole that is already digested with a global vision of the problem, issues and challenges, etc.

8) Need to provide the source of data and code (if you need to use some existing code).