MAT0043: Data Analysis Project

Project Due: **Friday, December 17 (6pm)**
Review Due: **Tuesday, December 21 (midnight)**

As part of the final examination for this course, you are required to analyse a data set of your own choosing. The data set may already exist, or you may collect your own data using a survey or by conducting an experiment. You can choose the data based on your interests or based on work in other courses or research projects.

There are many avenues that you may pursue for this project and I encourage you to be creative even if you do not think you will necessarily get "great" results. Here are some ideas:

1. *Comparison of algorithms:* Throughout the course, we discuss various algorithms and their properties. How do various algorithms perform on the same set of data? Apply (and compare) these techniques to a real regression or classification problem.
2. *Designing new algorithms:* Often times, algorithms do not work like expected and may need to be adapted or modified to better fit the assumptions inherent in the data. What work needs to be done to adapt a model to an interesting dataset that you have found? For example, how might you solve multiple related classification tasks?
3. *Missing information:* Various real world regression/classification problems involve missing components in the input vectors. How can you deal with such missing information? Do you expect your method to degrade rapidly if more information is missing?
4. *Imbalanced classification:* In classification problems, an imbalanced data set occurs when there is an unequal representation of classes (e.g., more 0 than 1's). In certain areas such as fraud detection, medical diagnosis and risk management, severe imbalance is relatively common and a concerning problem. Often times, the interest is in correctly classifying the minority class. Which approaches are available to deal with imbalanced data?
5. *Bayesian intake:* You can approach most of the topics discussed throughout the course from a Bayesian perspective (e.g, Bayesian linear/logistic regression, Bayesian variable selection, Bayesian trees, etc.), and `R` packages may be available to implement these techniques. In terms of performance, how do the frequentist and Bayesian approaches compare on an interesting regression/classification data set? In terms of conclusions you can draw about your data, which additional insights (if any) do you gain with a Bayesian approach?
6. *New topics:* The project can be an *applied* survey of a branch of machine learning that we do not explore (or not in detail). For example, hidden Markov models, Bayesian networks, reinforcement learning algorithms.

Below are a list of data repositories that might be of interest to browse. You are not limited to these resources, and in fact you are encouraged to venture beyond them. But you might find something interesting there:

- TidyTuesday
- Global Health Data Exchange
- WHO Data Collections
- Demographic and Health Surveys Data
- World Bank Health Data
- WHO/UNICEF Joint Monitoring Program on Water Supply, Sanitation, and Hygiene
- UN Refugee Data Finder
- Kaggle datasets
- OpenIntro datasets
- Awesome public datasets
- Youth Risk Behavior Surveillance System (YRBSS)
- Harvard Dataverse
- United Nations Data
- United Nations Statistics Division
- Italian Statistics (Istat) Data
- Italian Population Housing Census Data
- ECB Statistical Data Warehouse

Goals

The goal of this project is for you to demonstrate proficiency in the techniques we have covered in class (and beyond, if you like) and apply them to a novel data set in a meaningful way.

The goal is not to do an exhaustive data analysis, i.e., you do not have to apply every single statistical procedure you have learnt, but rather let me know that you are proficient at asking meaningful questions and answering them with results of a data analysis, that you are proficient in using R (or another programming language), and that you are proficient at interpreting and presenting the results. Focus on methods that help you begin answering your research questions. Also, critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data, and appropriateness of the statistical analysis should be discussed here.

The project is very open ended. You should create some kind of compelling visualisation(s) of the data in R. There is no limit on what tools or packages you may use. Pay attention to your presentation: neatness, coherency, and clarity will count.

Groups

You can collaborate with other students on a common project (**teams of max three people**). Each group will submit a (unique) report. The expectations for the project scope will increase depending on the number of students in each group, and for groups of two or three people, I will also expect a short paragraph to explain the role of each group member along with the final report. Do not include this note in your final report (which has to be anonymous), send it directly to me via email upon submission of your group project.

Project report

Your goal is to submit a cohesive project report that conveyes that you have mastered your applied problem and the techniques that help you answer your research question.

The following is a rough outline of what you are expected to include in your final report:

- An introduction, describing the problem you are solving, the motivation for it, and a brief summary of related work and background material on your project (if applicable). Being clear and explicit about this makes it easier for your reader
- A data section providing information about your data set, with relevant descriptive statistics and exploratory data analysis. Include the citation for your data, and (if available) link to the source
- Methodological work and results. If you are using a niche or cutting-edge algorithm not covered in class, you may want to explain your algorithm in a few paragraphs. You should not worry about getting *great* results. The idea and your understanding of the statistical machine learning issues involved are much more important than getting *great* results
- Conclusions, summarising your report and reiterating key findings. You may also want to note limitations in your study and include ideas for possible future research
- References (if applicable)

Format & length

Your write up should be **at most 5 pages** (including figures and tables) and submitted as a pdf. This is not very long, so you will need to be concise. Every sentence should add something to your paper. Your font size must be greater than or equal to 11pt.

Tone

Write as if you were explaining your results to whoever would be interested in your research question, whether this is other scholars in your field or peers sharing your interest in the topic. Keep in mind this audience may or may not have a statistical background. You must be statistically accurate and use correct statistical terminology, but must also explain your conclusions in a way that a potential reader will little statistical literacy &/or knowledge of your application can understand.

Code

Do not include code within (or at the end of) your report, so that your document is neat and easy to read. Instead, send your file(s) (and data, if appropriate) directly to me together with your report. Code should be clearly commented and annotated, and I should be able to run it with minimal intervention. *Exception*: If you want to highlight something specific about a piece of code, you are welcomed to show that portion.

Submission

The report is due on **Friday, December 17 (6pm)** via email (silvia.montagna@unito.it). **Author names or any other personally identifying information should not appear anywhere in the document (for anonymised peer review purposes).** Late submissions will not be accepted.

Peer review

The peer review process will mimic the process of how scientific papers and grant proposals are evaluated. Your project will be anonymised and assigned to at least one peer reviewer. Each student

will also receive the report of another group, which you have to review. The review process is double-blind - both authors and reviewers will remain anonymous to each other.

As a peer reviewer, you are mainly expected to comment on the technical quality of the solutions. You are also expected to comment on if all scientific questions have been answered, if relevant figures/tables have been included, and if the language of the report is satisfactory. Potential criteria that you can follow for the review are:

- Content: What is the quality of research and/or policy question and relevancy of data to those questions?
- Correctness: Are statistical procedures carried out and explained correctly?
- Writing and presentation: What is the quality of the statistical presentation, writing, and explanations?
- Creativity and critical thought: Is the project carefully thought out? Are the limitations carefully considered? Does it appear time and effort went into the planning and implementation of the project?

The review report can be submitted as a simple text file (.txt) or as a pdf. There is no strict page limit for the review report, but a guideline is 1/2-1 page in plain text.

You will get the review on your report from your reviewer(s), and comment and grade from me. While the review your receive for your own project will not affect your grade, a small part of your project grade will be based on completion of the review assigned to you.

You will receive the project to review on **Monday, December 20** and your review is due by the end of the day on **Tuesday, December 21.**

Grading

The final report will be judged based off of the clarity of the report, the relevance of the project to topics taught in MAT0043, the novelty of the problem, and the technical quality and significance of the work. The overall writing quality, including grammar, spelling, and organisation will also be evaluated.