STATISTICAL INFERENCE. FINAL PROJECT GUIDELINES

The final project should be an applied data science project that involves the data analysis methods seen in class (or extensions thereof). Each project should consists of two parts where you use different methods from each part of the class. For Part 1, typical examples would consider a problem that requires applying regression methods where there is a significant amount of parameters, e.g. to predict or explain some given outcome(s). For Part 2, typical examples would either apply unsupervised learning to explore the relationship between multiple outcomes or multiple covariates, or supervised learning methods where one applies robust regression or hierarchical/random effects models. The data used for part 2 can potentially be the same as used for part 1, but this is not necessary. It may depend on the goal of the project.

Key facts:

- Final report due: **December 21, 2021**
- The projects are in groups of 2 people. If there's an odd number of students, then one of the groups will have 3 people.
- The max length of the report is 10 pages (15 pages if there's a group with 3 people). References and appendix do not count towards the page limit. Any R/Python/Stan code should go into the appendix.
- You choose the project's theme and what dataset(s) to analyze. You're encouraged to discuss with the lecturers the suitability of the project.
- In case you have a project in mind that is not an applied data science project you must first get approval from the lecturers.

Structure of the report

The report should be structured as a research paper with the following sections.

(1) Abstract
The abstract is optional, depending on your available space. It should consist of 1 paragraph consisting of the motivation for your paper and a high-level explanation of the methodology you used/results obtained.

(2) Introduction. [$\approx$ 0.5 page]
Explain the problem and why it is important. Discuss your motivation for pursuing this problem. Give some background if necessary. Clearly state what the input and output is. Be very explicit: "The input to our algorithm is an: image, amplitude, patient age,rainfall measurements, grayscale video, etc.. We then use a: lasso regression , clustering, hidden markov model, etc. to output a predicted: age, stock price, cancer type, music genre, etc." Being explicit about this makes it easier for readers.

(3) Related work. [≈ 0.25 page]
You should find existing papers, group them into categories based on their approaches, and discuss their strengths and weaknesses, as well as how they are similar to and differ from your work. In your opinion, which approaches were clever/good? What is the stateof-the-art? Do most people perform the task by hand? You should aim to have at least 5 references in the related work. Include previous attempts by others at your problem, previous technical methods, or previous learning algorithms.

(4) Dataset [≈ 1 page]
Describe your dataset: how many training/validation/test examples do you have? Is there any preprocessing you did? What about normalization or data augmentation? What is the resolution of your images? How is your time-series data discretized? Include a citation on where you obtained your dataset from. Depending on available space, show some examples from your dataset. You should also talk about the variables you used. Try to include examples of your data in the report (e.g. include an image, show a waveform, etc.).

(5) Regression (first part of the course). [≈ 4 pages]
   (a) Methods. Discuss your choice of data analysis methods. The description should be accurate enough to ensure that others can reproduce your results. Make sure to include relevant mathematical notation. For each method, give a short description (≈ 1 paragraph) of how it works. Again, we are looking for your understanding of how the methods work.
   (b) Results. Main findings, adequately supported by tables and figures.

(6) Unsupervised / supervised learning (second part of the course). [≈ 4 pages]
   (a) Methods. Discuss your choice of data analysis methods. The description should be accurate enough to ensure that others can reproduce your results. Make sure to include relevant mathematical notation. For each method, give a short description (≈ 1 paragraph) of how it works. Again, we are looking for your understanding of how the methods work.
   (b) Results. Main findings, adequately supported by tables and figures.

(7) Discussion. [≈ 0.25 page]
Summarize your report and reiterate key points. Which methods/algorithms were the highest performing? Why do you think that some algorithms worked better than others? For future work, if you had more time, more team members, or more computational resources, what would you explore?

(8) Appendices.
Include the references and code.

Please note that Sections (5) and (6) can be interchanged. For example, for some problems you would first do an exploratory data analysis using clustering, PCA, ICA etc in order to determine which variables you will use in the second part. Also, for some projects it may be more elegant to not introduce all data at once, but discuss the extra data for part (6) in this section.

Marking criteria

- Originality of the project (5 marks)
- Clarity of the exposition, efficacy of figures/tables in conveying the results (15 marks)
- Data analysis chosen and technical soundness (25 marks)
- Critical assessment (10 marks)