Final Project Guidelines for STATS 451 Bayesian Data Analysis

Yixin Wang
University of Michigan

October 20, 2022

The final project is an integral part of this course. There is no better way to learn about Bayesian data analysis than by applying it to a research question of your own. The final project is an individual project. You have a lot of freedom in choosing a topic for your final project. The only criterion is that it deeply involves applying Bayesian data analysis to a real-world problem. You choose a dataset, an interesting question about it, and address it with Bayesian modeling.

The project involves three assignments: project proposal, project milestone, and project report. For each, please use 12-point font, and 1-inch margins. **Page limits are without figures; include as many pages of figures as needed.** We offer a LaTex template; you are encouraged to use it but not required.

Project Proposal

The project proposal is an abstract (i.e. a paragraph) that imagines the completed project. We understand that your project will evolve and change over the semester, but writing an abstract early is a good way to plan and think about what it would mean to successfully complete it. (Kaggle is a good place to find interesting datasets.)

We encourage you to refer to computer science conferences (such as *Neural Information Processing Systems* and *International Conference of Machine Learning*) or journals (such as *The Annals of Applied Statistics*, *Journal of the American Statistical Association*, *Journal of Machine Learning Research*) to get a sense of how to write an abstract.

Project Milestone

The project milestone describes the problem you are addressing and discusses some preliminary results. Include what you have completed and what you plan to finish by the end of the semester. As a guideline, the project milestone should (at least) answer the following questions.

- Clarify the purpose of data analysis:
  - What is the problem of interest?
  - What is the final goal you want to achieve?
- To the best of your knowledge, explain the data as thoroughly as possible:
  - How is the data collected?
  - Is the data a simple random sample from some population?

  - What is the potential defect in the data?
  - If you were able to collect the data, what would you have done differently?
  - Are you going to account for the data collecting process in your analysis?
  - Describe simple facts about the data: size, data structure & source, etc.
  - If you are not going to use real data, explain how you are going to generate "realistic" synthetic data that represents your knowledge of the problem.
- Draft a plan for data analysis:
  - What model are you going to use for the data analysis?
  - Have you done any preliminary analysis of the data?
  - What is your timeline for (1) proposing and fitting a preliminary model, (2) examining initial results and adjusting models further if needed, (3) summarizing and writing a final report.
  - What are the (potential) difficulties with your data analysis?

The project milestone is 1–3 pages long.

Project Report

The project report is due at the end of the semester. The report should have a similar format to that of an academic paper. Notebooks or markdowns with annotated code which generates the results from the final project should also be submitted as supplementary materials. You can further include a set of appendices (of any length) to which you can banish any details. (We may not read the appendix when grading your work.)

The project report should (at least) include

- Introduction: Clear description of the problem. Describe the problem you are trying to solve, why it is important or useful, and summarize any important pieces of prior work that you are building upon.
- Dataset: Clear description/visualization of the data. Describe the dataset or datasets you are working with. Show examples from the datasets. If you collected or constructed your own dataset, explain the process you used to collect the images and labels, and why you made the choices you did in the data collection process.
- Method: Clear and thorough description of statistical analysis. Describe the method you are using; this may also contain parts of the implementation of your model, loss function, or other components along with sanity checks to ensure that those components are correctly implemented.

- Experiments: Clear and thorough interpretation of results. Describe the experiments you did, and key results and figures that you obtained. This may interleave explanations of the experiments you run and figures you generate as a result of those experiments.
- Conclusion / Future work: What did you learn in doing this project? What are the shortcomings or failure cases of your work? If you had more time or resources, how would you continue or expand upon the work you have already done?

You should turn in a .pdf file containing your final report, together wih the notebooks/markdowns containing all the code and the generated results (tables, figures etc) that are included in the report. You must run all cells in your notebook to receive credit; we will not rerun your notebook.

The project report is 3–5 pages long, excluding figures or references.

The deadline of the final project report is strict, and late days cannot be applied. All late final projects receive a score of zero. In case of requesting exceptions due to severe medical reasons, a doctor's note and a signature from your graduate (or undergraduate) advisor is needed.

Project Evaluation

We evaluate the project on ambition, significance, originality, technical depth, results, relevance, and writing quality. Two good books about writing are Strunk Jr and White (2007) and Williams (1990).

The following grading criteria for the final project report is based on 100% maximum.

1. Introduction (5 pts)
    - Primary question of investigation should be clearly stated. (2 pts)
    - A clear description of the problem that communicates why the problem is interesting/ important/ worth pursuing. (2 pts)
    - A brief description of the dataset and its source with proper attribution. (1 pt)
2. Analysis (30 pts)
    - At least 4 different models with different likelihoods and priors should be fitted. (10 pts)
    - You need to go through Box-loop for each of those 4 models (15 pts).
    - Each model should be appropriate and well-motivated. For instance, a Poisson likelihood should not be used for a continuous variable. (5 pts)
3. Presentation of results (15 pts)
    - Visualizations should be used judiciously to report findings. (5 pts)

- Figures should be appropriately labeled: x-axis, y-axis, legend, and title. (5 pts)
- Figures and tables should be numbered and referenced properly in the write-up. For example, the writeup should be “Figure 1 shows " not “The above figure shows ". (5 pts)

4. Interpretation of results (20 pts)
    - Output of posterior predictive checks should be properly interpreted not merely stated. (10 pts)
    - All conclusions should be well supported by results. (5 pts)
    - The result should be discussed in the context of the initial question of investigation. (5 pts)
5. Code (10 pts)
    - Code should be organized, clean, and properly commented. (5 pts)
    - Code should not be included in the report. A separate .ipynb file (colab notebook or jupyter notebook) should be submitted. (3 pts)
    - All the code should be executed before submission so that every output is visible. (2 pts)
    - A plagiarized code will result in 0/100 in the project.
6. Writing (20 pts)
    - Report should be clear, well-written, and use academic language. (10 pts)
    - Report is well organized and properly divided into different sections and subsections. (3 pts)
    - Report should not contain substantial grammatical errors. (2 pts)
    - The length of the report should be at least three pages and no more than five pages. (5 pts)

References

Strunk Jr, W. and White, E. B. (2007). *The Elements of Style Illustrated.* Penguin.

Williams, J. (1990). Toward clarity and grace. *Chicago: The University of Chicago.*