DATA 5322 Statistical Machine Learning II
Spring Quarter 2025
Practical Homework Guidelines

Overview

Practical homework will be assigned with each of the 4 covered sections and will explore the full data science methodology of collecting, processing, and analyzing data as well as communicating results. The goal of these assignments is to give you practical experience solving real-world data science problems. They will demonstrate the use of computational and coding skills and will emphasize visualization and communication skills.

Assignments are built for students to showcase mastery of course content, but will be based on real data and may require additional resources, data cleaning, or other tools that are not directly presented in class. As such, it is expected that these assignments will take longer to complete. Each practical homework will have two deliverables: (1) your code and (2) communication of your work. The communication deliverables should explain the concept/theory you learned in your own words, the methodology you took to investigate the data, and your findings. The format will vary and will include presentations, posters, formal and informal written communication, but the same content should be covered.

Submission and Grading

The main goal in this course is your learning and development, which is most impactful through trial and error and in response to feedback. The assessment structure is intended to allow you to learn from mistakes and make corrections. Each practical assignment, covering one chapter (two weeks) of course material, comes with two opportunities for submission. The first submission is due at the start of the new material. You will receive a score and feedback and will have one week to improve your work based on these comments. **All points can be earned back in the revision with no penalty.** That is, if you receive a 65% on your homework, you have one week to revise and resubmit in response to this feedback and replace your grade with a 100%. However, first submissions must be on time and should reflect full effort to complete the assignment properly. Points will be explicitly assigned for effort and completion of the first submission, i.e. half-done or missing first submissions cannot earn back full points on the revision. For any revised submissions, points will also be assigned for how completely the feedback was addressed, i.e. if you only revise only one section/aspect of your work when many comments were given, you will lose points.

Expected Contents

Good data communication is accurate, clear, and concise. Try your best to convey your understanding and process fully while including only the most important details and results. Language, especially writing, should be clear and use correct spelling and grammar– expectations will be high, so please utilize resources such as the Writing Center or English Language Learning Center for support. Plots, figures and tables should be thoughtfully designed (i.e. not simply a screenshot of your code output!), must be high-resolution, and should be captioned and labeled.

Each communication deliverable must, at minimum, include the following sections. See each assignment description for additional content that should be included.

1. Title: A descriptive title for your work
2. Introduction: This section should set the goals of work. Introduce the application topics you are studying, the particular question(s) you wish to investigate, and a description of the data

set to be investigated (e.g. where it is from, what general variables are available). Clearly state the models/methods you will be using.

3. Theoretical Background: This section should demonstrate your full understanding of the theory we learned in class. It should summarize each model we covered in the section, including how it works, how to appropriately tune/apply it, how to assess its performance, how to interpret its results, and considerations/limitations and appropriate uses of the method. It does not need to be too detailed or rigorous, but should talk about each model/method from that section.

4. Methodology: A short description of data processing/cleaning, including selection of variables, and a detailed description of how you implemented and tested the models. You should be sure to discuss hyperparameter tuning (what values you tried, how you selected the best), error metrics, comparisons, cross-validation, etc.

5. Results: Presentation of your computational results (nice plots!), including any metrics that should be used to justify your model is well-implemented. Comparisons between models are highly encouraged– make sure to directly show this in plots or tables for quick comparisons.

6. Discussion: This section should contain all the interpretations and takeaways from your computational results. Be sure to compare and contrast models on metrics like errors and run time. Interpret your model results in the context of the application– share key findings and their relevance. Your models may not be very good– that is ok! Make sure to discuss where, how, and why your models might be performing poorly, limitations of your work, and improvements or exentsions you could do, if relevant.

7. Conclusions: Discussion of the broader impacts of your findings and how your work could be used in the real world.

8. Bibliography/References: Numbered citations (i.e. not just URLs) to any sources you used, including figures and diagrams taken from online, blog posts or repositories you referenced code from, software packages you made use of, etc.