Stat 641/543-Statistical Learning project outline and rubrics

Xuewen Lu

Nov. 8, 2022

1 Project Outline

Your term project for this class consists of analyzing **two statistical or machine learning models** you learned from this course with **two data sets** of your choice, **one for classification and one for prediction**, you need to use at least one method for prediction and one method for classification. However, you can use several methods for the same task, e.g., you can compare the performance of several methods, then recommend the best one for your research problem. The procedures for classification and prediction should be clearly explained, including methods for statistical learning, training, test, hyperparameter selection, test error, classification and prediction accuracy based on various evaluation methods, and interpretation of the results.

You have essentially free reign to choose data sets and research questions. The focus of the project is on the proper application of statistical learning methods, not the perceived scientific merit of the research questions. The topic can relate to your thesis, your undergraduate research project, your job (you don't need to share proprietary data), or an independent interest. You can use publicly available data such as kaggle datasets, if such data have already been analyzed by other authors and you want to use it, you should use it in a different and creative way, for example, for answering different research problems or using a different approach for the same problem, you need to cite the references which have used the data you choose. Your data must be real. Your data should have a reasonable sample size n and number of covariates or predictors p. The recommended values are $n \geq 200$ and $p \geq 10$. If you are not sure if your data set is appropriate, you had better consult me first by sending one page project proposal to me. This proposal should formulate question, as concretely as possible, the question you're investigating. You should write a few sentences about why it's an interesting question. You should give some details of the data you'll be studying and the machine learning methods you'll be using, for example, including a data dictionary. Include a short bibliography listing import references.

Note: "the two statistical learning models" means two different methods in analyzing two different types of data, through either supervised learning or unsupervised learning. The method you choose should match the nature of the data, e.g., I don't want to you treat a discrete response as a continuous response, vice vera.

1.1 Purpose of the Project:

To develop your teamwork ability, communication skills and ability to use statistical learning methods in addressing a practical concern in daily life.

1.2 Logistic Requirements:

- You must form a team of 4 students, by yourselves or through my assignment. If by yourselves, chose a team leader and let me know the names and student ID of your team members through the team leader. Team members must work together to finish the project and prepare a written report to be submitted by the team leader by the due date. Under exceptional circumstances, a team could have less than 4 students.

- Prepare and give a 12 to 15-minutes (depending on how many teams we are getting) oral presentation of your project report in class on Dec. 5th and Dec 7th, 2022.

- Name your report "Team # Report.pdf" (your team # will be announced later) and your Rmd file "Team # Report.Rmd", your slides "Team # Slides.pdf" or "Team # Slides.ppt" . You must submit your presentation slides on Dec. 4th, 11:59 pm before the presentation though D2L → Assessment → Dropbox → Project slides, and your Project Report and the associated Rmd file, the data files in the csv format and any graphs used in your report through D2L → Assessments → Dropbox → Project report in pdf and Rmd and data sets on Dec. 11, 11:59 pm.

2 Project Rubrics [Total points=80]

1. (5 points) **A clear and short title and within page limits**, together with your names and student ID. The report must not exceed 25 pages which does not include the Appendix for R codes, additional tables and graphs in the form of paper size=8.5 x 11 inches, font=12pt, 0.5 inch $\leq$ margins $\leq$ 1 inch, line space=18pt.

2. (10 points) **The purpose/motivation of your project**. You must write clearly about the two statistical learning methods and data science issues you want to investigate, the populations, the parameters and variables you want to use in your statistical learning process.

3 (10 points) **Data collection**. Describe the details on how you collected or found your data for your project. Report any issues or difficulties you ran into and how

you handled them.

4. (35 points) **Statistical learning methods and outcomes**. The procedures for classification and prediction should be clearly explained, including methods for statistical learning, training, test, hyperparameter selection, test error, classification and prediction accuracy based on various evaluation methods, and interpretation of the results and outcomes.

5. (3 points) **Conclusion** on the methods and issues you investigated and reflection on anything you learned from doing your project.

6. (2 points) **References**.

7. (10 points) **Appendix**: R codes for the project. This code should be reproducible (i.e., I should be able to run it and obtain the numbers and graphs in your results section) and should not appear anywhere in the main body of the report.

8. (5 points) **Presentation**: You will give a 12 to 15-minutes presentation on your data and your analysis of it using statistical learning methods. These presentations will take place in our regular classroom. Roughly speaking, you should think about allocating time equally in the following manner among your group members:

   - 2 minutes for an introduction into the nature of the data, how it was collected, and the question(s) of interest
   - 5 minutes for describing your methods and how you came up with them, probably accompanied by some exploratory graphs and summary statistics.
   - 5 minutes to describe your results and conclusions.

   You will have 12 to 15 minutes to speak; plan accordingly. If there will be no time for questions after each talk, I will leave questions to the end. Your grade will be based on the following:

   - Reasonable and appropriate choices made in selecting and analyzing data
   - Insightful description of the research questions and conclusions
   - Quality of presentation: interesting, easy to follow, slides and organization are clear, tables and graphs are readable
   - Answering questions if there are any

   Try to avoid those errors shown in the document `BadPresentationBingo.pdf`, which will be made available in D2L.

Please keep the following questions in mind as you prepare your presentation and report:

- What are the main questions I am trying to answer?
- How was the data collected/gathered/sampled?
- Are there any confounding relationships present?
- Are there any interactions present?
- Are your model and methods reasonable? What assumptions is it making?
- What are the limitations of my analysis (assumptions which may not hold, limitations of the data, etc.)?

Appendix: Style Guide

This style guide was written for statistical analysis, except for p-value, it is also suitable for statistical learning.

In the Statistical Analysis section, half of the points will be awarded to "content", meaning the numerical results presented (or, in the model choice part, the logic behind the model selection process), and the other half will be awarded to "communication", meaning the quality of the writing describing those results, interpretations and commentaries.

- **Examples for statistical writing**: Striking the proper balance between the two (finding interesting patterns in the data while refraining from over-interpretation) is perhaps the primary challenge of statistical writing. To help strike this balance, it is often a good idea to explicitly separate specific numeric results from verbal explanations. Here are three ways of doing so:
    - **Separate sentences**: We find that the odds ratio comparing males to females is 2.5, with a 95% confidence interval of (1.9, 3.3). Thus, men are at considerably higher risk of coronary heart disease than women.
    - **Parentheses**: Men are at considerably higher risk of coronary heart disease than women (OR: 2.5, 95% CI 1.9-3.3).
    - **Use of tables/figures**: As shown in Table 2, men are at considerably higher risk of coronary heart disease than women.
- ***p*-values**: It is entirely possible to carry out a thorough analysis and write an excellent report without including a single p-value. No points are explicitly given to the reporting of p-values; whether you include them or not is up to you. However, if you do choose to include them, you must interpret and describe them appropriately. There are more-or-less agreed-upon conventions for describing, in words, the degree of evidence against the null hypothesis that a p-value represents:

	0.10	0.05	0.025	0.01	0.001
Evidence					
against H_0	borderline	moderate	substantial	strong	overwhelming

For example, if the test for whether the regression coefficient for age has a p-value of 0.08, it would be appropriate to say that you have borderline significant evidence that age is associated with your outcome. If $p = 0.04$, there is moderately significant evidence that age is associated with the outcome. If $p = 10^{-8}$, age is clearly associated with the outcome and there probably isn't much point in discussing the p-value other than to assure readers that you have definitive evidence that age is associated with the outcome.

- **Significant digits**: Generally speaking, 2 significant digits is enough when reporting most results - additional digits are often unimportant and distracting. For example, saying that the odds ratio is 2.5 is better than saying that the odds ratio is 2.4762. Unless your sample size is in the millions, it's somewhat absurd to think that you can estimate an odds ratio to the fourth decimal place.

- **Creative use of graphics**: One can make creative use of graphs to illustrate results. These can appear in any of the "Results" subsections. Note that quantity is not equivalent to quality - including pages of automatically produced residual plots, QQ plots, etc., will not benefit you, and indeed will almost certainly lower the "clarity" portion of your grade (that is, clarity vs. insight. Clarity involves concise writing that makes clear points with correct descriptions and accurate interpretations. Insight involves pointing out interesting patterns in the data and creatively putting those observations into words.). All your graphs should have a purpose, and should be discussed in your report. Two important (perhaps debatable) rules to keep in mind are: (1) Anything you could put in a table, you could also put in a figure, and it would probably be better as a figure, and (2) Just about any statistical concept or observation you have about the data, you could probably think of a way to illustrate it with a graph.