Course Project Theory of Statistics I, 625.725.81

Overview

You will be picking a paper from a set of four topic areas (eight total papers to choose from). This paper will be your entry point to your project. You'll then also be picking from a set of data sources (six candidates). Your project will involve applying (and understanding) the ideas in your selected paper to the dataset that you have chosen. We encourage posting of your progress and your ideas on Blackboard. Teams are welcome to form but are not required.

This approach will enable the entire class to learn connections to the eight papers. All of the papers will be covered in class and so you will not miss out by picking one as your focus area. We encourage each of you to become 'near-experts' in your chosen paper and to share your growing knowledge and expertise with others as we go through the term.

Topic Areas

1. **Model Selection/Hypothesis Testing**

   - Articles
     - Sharma, Pratyush, Marko Sarstedt, Galit Shmueli, Kevin H. Kim, and Kai Oliver Thiele. "PLS-based model selection: The role of alternative explanations in information systems research." *Journal of the Association for Information Systems*20, no. 4 (2019): 4.
     - Jung, Yoonsuh. "Multiple predicting K-fold cross-validation for model selection." *Journal of Nonparametric Statistics* 30, no. 1 (2018): 197-215.

2. **Categorical Selection/Logistical Regression**

   - Articles
     - Norton, Edward C., Bryan E. Dowd, and Matthew L. Maciejewski. "Marginal effects—quantifying the effect of changes in risk factors in logistic regression models." *Jama*321, no. 13 (2019): 1304-1305.
     - Sur, Pragya, and Emmanuel J. Candès. "A modern maximum-likelihood theory for high-dimensional logistic regression." *Proceedings of the National Academy of Sciences* 116, no. 29 (2019): 14516-14525.

3. **Bayesian Regression/Testing**

   - Articles
     - Gelman, Andrew, Ben Goodrich, Jonah Gabry, and Aki Vehtari. "R-squared for Bayesian regression models." *The American Statistician* (2019).
     - Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho. "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)." *Bayesian Analysis* 15, no. 3 (2020): 965-1056.

4. **Sequential Regression/High-dimensional Regression**

    - Articles
        - Duarte, Kévin, Jean-Marie Monnez, and Eliane Albuisson. "Sequential linear regression with online standardized data." *Plos one* 13, no. 1 (2018): e0191186.
        - Bertsimas, Dimitris, and Bart Van Parys. "Sparse high-dimensional regression: Exact scalable algorithms and phase transitions." *The Annals of Statistics* 48, no. 1 (2020): 300-323.

Data Sources

1) Census, 2) Sports, 3) Cars, 4) Housing, 5) heart-disease and NHANES

Learning Objectives

- Apply the concepts and tools in the course.
- Gain experience in successfully matching model to data.
- Demonstrate a high level of programming competence.
- Explain the mathematical basis for the modeling methodologies used.
- Produce a detailed written report describing your process and findings.

Instructions/Steps

First step: Project Topic (Module 3) – Submit your preferences from the candidate topic list. Please choose 3 and rank them. Instructors will review. Your assignment will favor your first choice, but it may be second or third choice to enable a reasonable distribution of students on each topic

Second step: Pseudo-code and Outline (Module 7) – This is mainly a planning document. We do not enforce an arbitrary definition of what it means to have a complete pseudo-code or outline. Use this as an opportunity to plan the project and to consider what needs to be done in the remaining part of the course.

Third step: Bibliography (Module 10) – This does not need to be a final bibliography, but it will help you in bringing all your materials together. The set of papers should be ones that you've read and thought about and so will be citing in your final project report.

Fourth step: Full Methods (Module 11) – This may be further modified in your final project. The narrative deliverable will help you focus on documenting what you've done. Also discuss how the project work connects with the coursework.

Fifth step: Final Report (Module 14) – Prepare and submit a final written report (see Guidelines for Final Report below).

Deliverable and Submission Information

Week Due	Deliverable	Submission Instructions	Percentage of Project Grade	Grading Elements/Criteria
Module 3	Rank-order your preferences from the set of four candidate topic areas	Post ideas in discussion board	5%	● Present your ideas ● Explain why you find them interesting and challenging
Module 7	Pseudo-code and Outline: to help with your planning and document your progress to this date Use this to help you determine what parts will be the most challenging to complete on-time.	Post to designated discussion forum	5%	● Quality of your thought process and your ability to stage what remains ahead
Module 10	Bibliography This should be papers that you've spent time on and learned from reading -- they should be ones that you will include in the final writeup.	Post to designated discussion forum	10%	● Clear and present connection between your papers ● Your rationale for choosing these papers and not others
Modules 11	Methods Describe your analysis methods, 3-4 pages. This is a check-in point. You should be about 50% complete on the project.	Upload to assignment link in module 11	20%	● Key deliverable (note higher point count) that should document how your project work connects with class work and with the paper chosen from the topic list.

Module 14	● Final written report ● Associated code This should be an 8-12 page written report with the code and data associated separately Please consider a github repo for hosting your code (along with data if that is possible). This will allow you (and others) to grow the code ideas after class ends.	Upload to assignment link in module 14	60%	● Comprehensive analyses ● Suitability of modeling approaches ● Successful execution in your codebase ● Textural and graphical output and interpretation

Guidelines for Final Report

Title page: containing project name
Executive Summary (1 page): High level description of your project and findings
Project Description: Present the dataset used and the associated analyses performed. Do so in such a way that your analyses can be replicated. Highlight the connection to your entry paper from the topic choices
Methods: Describe how your results were determined and the mathematical and statistical methods that you used.
Results: Provide textural and graphical results of your analyses along with interpretation of those results.
Conclusions: Discuss the results in terms of suitability of the chosen analysis methods for the data, difficulties encountered, and suggestions for alternative types of analyses.
Appendices: Provide the datasets and the code that you used (commented to facilitate understanding what you did. As always, explicitly use randomization seeds where applicable to permit replicability). This may be, alternatively, posted into a Github Repo.

Bibliography: Cite relevant bibliography and sources of data.

Plagiarism

Plagiarism is defined as taking the words, ideas or thoughts of another and representing them as one's own. If you use the ideas of another, provide a complete citation in the source work; if you use the words of another, present the words in the correct quotation notation (indentation or enclosed in quotation marks, as appropriate) and include a complete citation to the source.