PHP 2650 Syllabus

Instructor:
Alice Paul
Email: alice_paul@brown.edu

Office Hours: TBD

Teaching Assistant:
TBA

Lectures: Tuesdays and Thursdays 10:30-11:50

Course Goals

This course will introduce modern statistical learning tools with a focus on tools developed for big data. It covers three interconnected components: statistical machine learning methods, the underlying algorithms, and computational tools. This course will focus on the principal techniques to analyze data from start to finish: managing large data, exploring patterns, framing statistical problems, building efficient computational algorithms, and writing reports. Topics will include data management, feature engineering, clustering, convex optimization algorithms, tree/ensemble methods, and predictive modeling.
After you complete this course, you should be able to:

- Manage and explore big data sources.
- Apply and evaluate results from statistical learning tools including clustering, decision trees, and neural networks.
- Understand the underlying assumptions and mathematical foundations of these methods.
- Compare the computational complexity of different statistical learning approaches.
- Write a data analysis reports utilizing the above concepts and synthesizing the results.

Course Logistics

Prerequisites: This course is limited to biostatistics graduate students. Students who have taken 2510/2511 may request permission to enroll.

Course Website (Canvas): The canvas site will contain all the information for this course including this syllabus, office hours, weekly course info, and posted grades. Make sure that you are enrolled on Canvas and that you check the sites regularly.

Reading: The course will cover material found in the textbooks below, which are all recommended and not required. Additional resources, such as video lectures, journal articles, and notes, may be provided as needed.

G James, D Witten, T Hastie, and R Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. Free online. `http://www-bcf.usc.edu/~gareth/ISL/`.

T Hastie, R Tibshirani, and J Friedman (2009). *The Elements of Statistical Learning.* Free online. `http://statweb.stanford.edu/~tibs/ElemStatLearn/`.

If your Brown undergraduate financial aid package includes the Book/Course Material Support Pilot Program (BCMS), concerns or questions about the cost of books and course materials for this or any other Brown course (including RISD courses via cross-registration) can be addressed to bcms@brown.edu. For all other concerns related to non-tuition course-related expenses, whether or not your Brown undergraduate financial aid package includes BCMS, please visit the Academic Emergency Fund in E-GAP (within the umbrella of "E-Gap Funds" in UFunds) to determine options for financing these costs, while ensuring your privacy

Course Assessment

Your course grade will be based on:

Homework Assignments	50%
Final Project	25%
Take-Home Final Exam	25%
Total	100%

Homework Assignments

There will be five homework assignments each containing a few theory questions and one open-ended, data analysis prompt. Students are expected to write-up a response to the latter in a report format.

Final Project

For the final project, students will choose an extension of a topic seen in class. They will write an online post explaining the concept and detailing an application. More details to come. There will be no extensions given for the final project.

Take-Home Final Exam

The final exam will be a 48-hour take-home exam with a mix of theory and short applications. There will be no extensions given for the final exam.

Late Assignments

Every student has a bank of ten late days that they can use towards the homework assignments where each day buys you an additional 24 hours. You do not need to email me to use these days as they will be counted automatically, but you should think about saving them for when you really need them. After using up your late days, there is a 10% deduction in score for each additional late day. Late days may not be applied to the final exam or project.

Credit Hours and Time Expectations

Over 14 weeks, students are expected to spend a total of 40 hours in class, 20 hours on class preparation, 75 hours completing homework assignments, 15 hours on the final exam, and 30 hours on the final project, for an overall total of 180 hours.

Grading Scheme

All homework problems will be graded on the following scale

- 10: Your solution is correct, justifies your approach, applies methods appropriately, and is well-written with useful tables and figures.
- 8: Your solution has minor areas with room for improvement. For example, you may have a small error or need more detail in your response.

- 6: Your solution needs moderate changes. For example, you may not provide justification for your approach or show a minor misunderstanding of the methods used.
- 4: Your solution has major errors or areas with room for improvement. The response may be poorly written or show a major misunderstanding of the theory.
- 0: No response or not enough effort to judge.

You may resubmit up to one homework assignment per week to improve your score. You may reference the provided course solution and any other course material, but you should use it to inform your own solution rather than copying from it. Additionally, you must work individually on the re-submission. Included with your revised homework assignment you should also include a short (2 paragraph) reflection talking about what you changed, why you changed it, and what you learned.

Resubmissions will also be graded out of 10 points with 8 points for the resubmitted solution (graded on a 0/8 scale) and 2 points for your reflection (graded on a 0/1/2 scale). Your updated homework will grade will then be

$$\max\left(\text{ original grade }, \frac{(\text{new grade}) + (\text{original grade})}{2}\right).$$

The reason for using this grading scheme is to give you opportunities to improve and to encourage mastery of the material.

Schedule

The following schedule is tentative. Assignments are shown in bold.

Week 1 (1/26) Intro to the Course
Week 2 (1/31, 2/2) Big Data in R
Week 3 (2/7, 2/9) Big-O Analysis
Week 4 (2/14, 2/16) Dimension Reduction
Assignment 1 (2/17) Week 5 (2/23) Dimension Reduction Continued
Week 6 (2/28, 3/2) Clustering
Assignment 2 (3/3) Week 7 (3/7, 3/9) Clustering Extensions
Week 8 (3/14, 3/16) Optimization
Assignment 3 (3/17)
Week 9 (3/21, 3/23) Regression Trees
Week 10 (3/28, 3/30) Spring Break
Week 11 (4/4, 4/6) Bagging and Boosting
Assignment 4 (4/7)
Week 12 (4/11, 4/13) Neural Networks
Week 13 (4/18, 4/20) Deep Learning
Week 14 (4/25, 4/27) Deep Learning Extensions
Assignment 5 (4/28)
Week 15 (5/2, 5/4) Extra Topics
Final Project (5/11)
Take-Home Final Exam (5/15-5/17)

Additional Information

CEPH Competency Statements For Masters students only:

- Demonstrate a foundation in statistical theory and methods for standard designs and analyses encountered with biomedical data.
- Identify and implement statistical techniques and models for analysis of data.
- Acquire knowledge and skills in research methodologies to collaborate with substantive investigators.
- Recognize key research designs and be able to assist in developing plans for their implementation.
- Apply programming skills to analyze data and develop simulation studies.
- Attain proficiency in management, documentation of study data for use in practical statistical analysis.
- Formulate a public health question in statistical terms.
- Develop proficiency in making oral, written and poster presentations of work to statistical and non-statistical colleagues.
- Evaluate research reports and proposals on the basis of their scientific integrity, validity, and the strength of the quantitative analysis.
- Review and evaluate the use of biostatistical methods in public health or biomedical field of study.
- Demonstrate proficiency in the language of the public health or biomedical field of studies.

For PhD students only:

- Identify and implement advanced statistical models for the purposes of estimation, comparison, prediction, and adjustment in non-standard settings.
- Develop and plan research studies with complex sampling schemes.
- Apply programming skills to analyze data and develop simulation studies.
- Acquire knowledge and skills in research methodologies to collaborate with substantive investigators.
- Formulate a public health question in statistical terms.
- Choose a study design to address the public health question.
- Identify important scientific problems in an area of biomedical or public health research outside of biostatistics/statistics that require the development of innovative biostatistical methodology for their solution.
- Review and evaluate the use of biostatistical methods in public health or biomedical field of study.
- Develop proficiency in making oral, written and poster presentations of work to statistical and non-statistical colleagues.
- Communicate effectively with public health experts, relying upon a basic understanding of human health and disease.
- Prepare written summaries of quantitative analyses for journal publication, presentations at scientific meetings, grant applications, and review by regulatory agencies.

Academic Integrity You are expected to maintain the highest level of academic integrity in the course. Any violation will be penalized according to the Brown academic integrity code: `https://www.brown.edu/academics/college/degree/policies/academic-code`. In particular, sharing materials outside of this class will be treated as an academic code violation.

You are expected to communicate and interact with your classmates. You are welcome and encouraged to discuss projects and exercises with others as long as these interactions are consistent with your educational goals. For example, if you read someone else's code in order to learn a new technique and apply it to your own implementation, that is a good way to learn. But if you copy code you don't understand instead of learning to develop your own, you are violating the academic integrity code and undermining your own education.

Accessibility and Accommodations

Brown University is committed to full inclusion of all students. Please inform me early in the term if you may require accommodations or modification of any of course procedures. You may speak with me after class, during office hours, or by appointment. If you need accommodations around online learning or in classroom accommodations, please be sure to reach out to Student Accessibility Services (SAS) for their assistance (seas@brown.edu, 401-863-9588). Students in need of short-term academic advice or support can contact one of the academic deans in the College.

It is my intent that students from all diverse backgrounds and perspectives will be well served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. It is my intent to present materials and activities that are respectful of diversity: gender, sexuality, disability, age, socioeconomic status, ethnicity, race, and culture. I expect students to conduct themselves in a way that respects the experiences and identities of all members of the class. Please let me know ways to improve the effectiveness of the course for you personally or for other students or student groups. In addition, if any of the class meetings conflict with your religious events, please let me know so that I can make arrangements for you.