36-600 – Overview of Statistical Learning and Modeling

Fall 2022

Instructor

	Email	Office Hours
Peter Freeman	`pfreeman@cmu.edu`	WF 10:10 – 11 AM (BH 229G)

Lectures	TR 8:35 – 9:55 AM BH 136A
Useful Texts	*Introduction to Statistical Learning in R* G. James et al. (available on `Canvas` as `Files/ISLR.pdf`) *Modern Data Science wth R* Baumer, Kaplan, & Horton (`mdsr-book.github.io/mdsr2e/`) *R for Data Science* Grolemund & Wickham (`r4ds.had.co.nz/index.html`)
Web Site	`https://canvas.cmu.edu/courses/29946`

This course is intended to expose graduate students outside of the departments of Statistics and of Machine Learning to the essentials of data analyses. The primary goal is to cover *statistical learning*, e.g., the modeling of relationships between some measurements taken in an experiment and some other measurement. However, we will also explore other elements of the "analysis workflow," including data pre-processing and exploratory data analysis.

By the end of this course, you will...

- demonstrate proficiency with statistical terminology, and be able to contextualize methods;
- be able to read datasets into `R` sessions, perform basic exploratory data analyses, and explain the results of these analyses verbally and in writing;
- be able to apply methods of unsupervised learning to data, and to explain the results verbally and in writing;
- be able to compare and contrast regression and classification algorithms and be able to explain what results when you apply them to data; and
- demonstrate your newfound knowledge by, over the course of the semester, completing a series of data analysis projects.

In the end, gaining contextual knowledge of statistical learning and the analysis workflow is the goal, and not learning theory and/or mathematics. The Gauss-Markov theorem won't help you when you need to code a random forest model and interpret its result for your advisor!

This course builds off of the 2019 Statistical Learning Summer Workshop[1] and the course 36-290, a course on statistical learning for sophomore statistics majors. For more about the philosophy behind 36-290 in particular, see the following paper in *The American Statistician*:
`https://www.tandfonline.com/doi/abs/10.1080/00031305.2020.1844293`

[1] `https://github.com/pefreeman/SLSW_2019`

Administrative Remarks

Lectures

Lecture Slides. One or more sets of slides, constructed via `R Markdown`, will be presented in a typical class session. Slide decks in `PDF` format will be available on `Canvas`, under the `Files` hierarchy. A single slide set will contain introductory and contextual detail on a particular topic (e.g., random forest or statistical model assessment). Note that lectures will often take up only a fraction of a class session: the rest of the session will be devoted to the completion of a lab assignment (which will ultimately be due before the next class session begins).

To see the 2021 schedule and slides, go to:

`https://github.com/pefreeman/Statistical-Learning-Overview/tree/main/LECTURES`

Note that the schedule and slides may be updated over the course of this fall's class, thus what is put on `Canvas` will be the most up-to-date material.

Textbook. Students should use the books listed above in two ways: as references that provides more in-depth coverage of the topics introduced in class via the lecture slides; and as a coding resource. In particular, `ISLR` by James et al. includes lab exercises at the end of each chapter; these will help you initially code analyses in `R`.

Attendance. As you are graduate students, I will not require class attendance, but rather expect that you will make every effort to attend every class. There is no need to contact me if you will miss a class.

Laptops. The majority of time in each class session will be devoted to completing lab assignments. Thus you should *always* bring your laptop to class. If you do not have a laptop, or if your laptop breaks, please let me know *immediately* so that you and I can work out an accommodation (such as completing a lab in the BH 140 computer cluster, assuming terminals are available).

Software

R/RStudio/R Markdown. You will work heavily with the `R` programming language via the medium of `R Markdown`/`RStudio`. Details on installing all three (along with important `R` packages) will be given by email prior to the first day of class and reiterated on the first day of class.

Assignments

Labs. Most class sessions will have associated lab assignments. The medium for completing lab assignments will be `R Markdown` files, which you will "knit" to `HTML` format; you will then submit the knitted `HTML` file to `Canvas`. As this is a graduate class, a lab will be graded primarily on whether you made a good-faith effort to complete it. As such, late submissions will be accepted, but will be issued a 20-point penalty. (We will waive the penalty for situations like your lab being two minutes late...but regardless, just turn the labs in!) Note that feedback may be given in class, if we observe common issues; if I share your lab on the screen, I will share it anonymously. You must bring any missing lab score to my attention within one week of the lab being graded and the score being posted to `Canvas`. Feel free to discuss lab assignments with others, but realize that the work you hand in must be your own. Simply copying someone else's work is plagarism; see "Cheating" below. Plus, you won't learn the material!

Analysis Projects. The goal of this class is to learn how to perform, and practice performing, statistical analyses of datasets. Thus your grade will be based not only on lab reports, but also on a series of three data analysis (DA) reports (one revolving around exploratory data analysis, one around the application of linear/logistic regression, and one around the application of machine learning methods), and a final group poster. Each DA report is meant to be a "complete" report showing the details of an analysis. I put complete in quotes because for the earliest report, you will only be able to do, e.g., exploratory data analysis, while for the last report you can do EDA and utilize any and all of the statistical learning methods at your disposal. Note that a DA is *not* a "lab notebook," but rather a summary, a story, something that doesn't mention every dead end or show every plot, but provides enough detail that the reader could independently recreate your analysis if he or she was so inclined. Your DAs will be written in `R Markdown` format and you will submit the knitted `HTML` files to `Canvas`. As with the labs, the primary grading criterion will be whether a good-faith effort is made to create a complete and organized report. All the other material written about labs holds here as well (late submissions, missing scores, collaboration and plagiarism).

As for the group poster: this will involve analyzing a dataset we will provide in the final weeks of class. This poster will be displayed on the Statistics & Data Science departmental website. Your team will turn in at least one poster draft during the last week of class, and as is the case above, the goal here is a good-faith effort, particularly in regards to suggested revisions to your draft(s).

Miscellaneous

Email. All course-related email should be sent to `pfreeman@cmu.edu`. I will respond to it if it is appropriate to do so. Sending email does not shift any responsibility from you to me; you are still responsible for completing your assignments. In particular, do not send complicated questions or requests via email; replies will not be given for email questions or problems requiring lengthy (more than a couple of sentences) responses. These types of communications should be done in person.

Cheating. Cheating or plagarism on labs or the semester project will be dealt with as allowed under CMU policies: `https://www.cmu.edu/policies/student-and-student-life/academic-integrity.html`.

Accommodations. If you require accommodations, please obtain documentation from the Office of Disability Resources (`https://www.cmu.edu/hr/eos/disability/students`). I will make no accommodations without documentation.

Diversity, Equity, and Inclusion. It is my intention that *all* students be well-served by this course, regardless of gender, sexuality, disability, age, socioeconomic status, ethnicity, race, religion, and/or culture. If you observe that you or someone else in the class is experiencing unfair or hostile treatment, please let me know, or contact the Center for Student Diversity and Inclusion directly (`csdi@andrew.cmu.edu`; 412 268-2150).

Mental Health. Many of your syllabi will have verbiage about taking care of yourself.

My take on this is that you have to realize that in the greater scheme of things, your performance in this course is not as important as your physical and mental health. Use your time wisely during the day, and sleep at night. Sleep during the day too, if you need to. (If you sleep in class, just don't snore, OK?) Twenty years from now, you won't remember your grade in this course; in fact, you may have forgotten that you even took a data analysis course. But you will remember if you were generally happy, or if you were stressed beyond belief. Strive for the happy memories; don't take on more courses and more responsibilities than you can reasonably handle. For some of you, this is easier said than done, but do try to scale back if you need to.

If, however, you are struggling and need support, feel free to seek me out. (Would I hold your struggles against you? No, of course not.) Or seek out Counseling and Psychological Services (CaPS; 412-268-2922 or `https://www.cmu.edu/counseling/`), the Re:solve Crisis Network (888-796-8226), and TimelyCare (`https://www.cmu.edu/wellbeing/resources/timely-care.html`). If you or someone you know is in a life-threatening situation, however, forego these resources and call the police immediately (8-2323 on campus, 911 off campus).

Grading

Labs	40%
Projects	45%
Group Poster	15%

Important Dates

Week	Date	What's Happening
12	25 Nov (R)	No Class: Thanksgiving

Final Grade. Grades in this class will be assigned "straight-scale." To reiterate, if you make a good-faith effort to complete all assignments through the semester, your final grade should not be something that you need to worry about!