STATS 216
Introduction to Statistical Learning

Stanford University
Quarter: Winter 2021
Time: Mon, Wed
2:30 PM - 3:50 PM

Instructor: Benjamin Seiler
bbseiler@stanford.edu
Zoom Office Hours: Mon 10-12

TA: Theo Misiakiewicz
misiakie@stanford.edu
OH: Thu 3-5

TA: David Fager
dfager@stanford.edu
OH: Tue 12-2

TA: Harrison Li
hli90722@stanford.edu
OH: Wed 10:30-12:30

Summary

This course will provide an overview of supervised learning, with a focus on regression and classification methods. Topics include: linear and polynomial regression, logistic regression and linear discriminant analysis; cross-validation and bootstrap; model selection and regularization methods (ridge and lasso); nonlinear models, splines, and generalized additive models; tree-based methods, random forests, and boosting; support-vector machines; and some unsupervised learning: principal components and clustering (k-means and hierarchical). Computing is done in R, through tutorial sessions and homework assignments. This math-light course is offered via video segments (MOOC style) and in-class problem solving sessions.

Prerequisites

Introductory courses in statistics or probability (e.g. STATS 60/160, 101, or 110), linear algebra (e.g. MATH 51), and computer programming (e.g. CS 105).

Logistics

This course is in the "flipped" format. The lectures are pre-recorded, and students will watch them on their own time. The course material consists of recorded video chunks (typically around 10-12 minutes long), as well as quizzes and review questions. Each week new material will become available, and students are expected to keep up. The lectures follow the course textbook.

In-class sessions focus on hands-on experience, where we will solve problems together while applying topics from lectures directly in R. Our in-class sessions will often require the use of computing. All in-class sessions and office hours will take place remotely using Zoom. All Zoom links will be available on Canvas along with some course notes, this syllabus, and any announcements. The recorded sessions will remain available on Canvas for further review, but we encourage attending all in-class sessions live to best facilitate discussion and interaction, if possible. Office hours will not be recorded.

Goals of the Course

The course covers the entire contents of the textbook with objectives to

- introduce fundamental tools for building predictive models, including some "state-of-the-art" methods in data science
- understand the role of model selection and assessment using cross-validation and randomization
- learn how to use the vast collection of tools in R to implement the methods learned

Computing and R

Students will be required to use R, and the lectures include some instructions in the use of R. R is an excellent open source (free) software for statistical analysis. Please install R during the first week of the course (details below). In class, we will generally be using the RStudio IDE (details below), but other R-friendly environments such as Jupyter Notebooks are acceptable for student use if a strong preference exists.

- R: https://www.r-project.org/
- RStudio is highly recommended for syntax highlighting, package management, document generation, and more: https://www.rstudio.com/ The newest version of RStudio is highly recommended. The way RStudio handles RMarkdown and RNotebook documents is evolving rapidly (e.g., previews of code chunks) so you will be best able to get support with the newest version.
- Latex, which will enable you to create PDFs directly from the RMarkdown feature RStudio with code, output (e.g., graphs), and equations.
    - o Mac users should download macTeX (http://www.tug.org/mactex/downloading.html) from Safari (not Chrome).
    - o Windows users should install MiKTeX (https://miktex.org/download).
    - o Both of these are very large files but unfortunately the lite versions are not sufficient for our needs in this course.

If you have any installation difficulties or questions, please reach out to the instructors.

Textbook

The primary text for this course is ISL:

An Introduction to Statistical Learning, with applications in R, J. Gareth, et. al., ISBN: 9781461471387. Available in pdf through the Stanford libraries or from the book website. Errata and data are also posted on the book website. **Important**: International editions may have missing/swapped exercises, which complicates your learning and may impact your grade (if you submit wrong solutions).

A more advanced treatment of the material beyond the scope of this course is available in the related text ESL:

The Elements of Statistical Learning: Data Mining, Inference, and Prediction, T. Hastie, et. al., (book website). We will not use this book in this course, but it is an excellent resource for those looking to go more in-depth with a particular topic on their own after completing this

course.

Students with Documented Disabilities

Students who may need an academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE). Professional staff will evaluate the request, review appropriate medical documentation, recommend reasonable accommodations, and prepare an Accommodation Letter for faculty. The letter will indicate how long it is to be in effect. Students should contact the OAE as soon as possible since timely notice is needed to coordinate accommodations. Students should also send your accommodation letter to instructors as soon as possible. The OAE is located at 563 Salvatierra Walk (phone: 723-1066, URL: http://oae.stanford.edu).

Assignments and Grading

For homework problems, you can discuss high-level ideas with your peers, but submitted work must be your own. Please indicate any collaborators in the submissions. You cannot collaborate on quizzes or exams. Do NOT share or post the code nor solutions on any forums or the internet.

Homework (45%) We will have biweekly graded homework assignments (approximately four), which will include analysis of datasets, analytical and conceptual problems, and programming assignments. Homework submission is via Gradescope.com. We will manually sync its student list with Canvas. If you don't have access to Gradescope in Week 1, let us know.

Each problem should start on a new page, be properly tagged in Gradescope, and executed correctly and independently from other problems (seed your RNG, if needed). We will only accept one PDF document for each homework. The best way to have all your text, code and code results in one PDF document is to write up your homework as an R markdown file, then either (i) knit it directly to PDF, or (ii) knit it to HTML, then save that HTML file as a PDF. We will not accept screenshots of code. Please reach out to the teaching team if you are having trouble submitting the homework. Homework will typically be due Fridays at 11:59 PM.

Late penalties apply: Each day late (up to three days) is penalized by 10% of the homework value. Homework more than three days late will receive no credit. For sickness, interviews, and other events, up to three late days total are forgiven at the end of the quarter.

"In-class" Midterm (15%): Tentatively planned for the fifth week. Exact logistical details will be announced during the first week of class.

"Take home" Final (30%): Will take place in the final week. Exact logistical details will be announced during the first week of class.

Canvas Quizzes (10%): Quizzes are based on video lectures, slides, and textbook readings. We highly recommend taking these as you watch the videos, instead of saving them until the end of the quarter. Answers can only be submitted once, so please check them carefully before submitting. Quizzes will close 24 hours after the final exam's submission deadline.

Honor Code

Students should be familiar with and act in accordance with the honor code.

Communication and Piazza

Canvas announcements will also be posted to our Piazza page (here). Please post any personal concerns (your grading, your absences, your accommodations, etc.) via *private posts* for the staff. All other concerns that are beneficial to your peers (who may be seeking the same info) should be posted as public messages. If emailing Ben or the TA team, use "[Stats 216]" in the subject line to avoid delays in reply.

Piazza offers a reliable and convenient interface for discussions, learning, and assistance. Its great functionality is only valuable if utilized. So, please format your posts (use Code Blocks, LaTeX, embedded images, other post references, tables, links, etc.). Collaboratively edit the posts to make corrections and clarifications, instead of chaining corrected posts.

Do NOT post homework code or solutions while your peers are still working on them. Give your peers a chance for discovery and learning. We do encourage higher-level discussions of problems/solutions and posts of R code that clarify functionality of tools. The instructor team will try to attend daily, so please be patient. Peer assistance is encouraged, just please avoid posting threads for the sake of posting (the quality does matter). Before asking a question, quick-search prior posts for similar concerns. Anonymous posts are ok, but the teaching staff still sees your identities.

If you have a complaint about the class, please notify the instructor team privately. Keep in mind that the forum answers may not always be correct (typos, question misunderstandings, etc.). Naturally, keep the communication professional, respectful, and cordial.

Please use Piazza for all questions related to lectures, homework and exams. Students may earn up to 3% extra credit by answering other students' questions in a substantial and helpful way.

Video Recordings and Zoom

As mentioned above, in-class Zoom sessions will be recorded for the convenience of students who cannot attend in real time. While we encourage students to keep their video cameras on during class to engender better communication and participation, we understand that some students would prefer not to be recorded. Questions and participation both verbally and in the Zoom chat are allowed and encouraged.