STAT 227
Statistical Learning and High Dimensional Data Analysis
Fall, 2024

Instructor: Zehang "Richard" Li
Office Hours: TBD
E-mail: lizehang@ucsc.edu

Course Meeting Times: Tuesday/Thursday 1:30 - 3:05
Location: Crown 104
Website: Through `Canvas`

Course Description: The course presents basic principles of statistical learning, with a focus on connecting statistics and computation in modeling complex, large, and high-dimensional data. We will focus on challenges associated with several learning and inferential problems, with real-world applications in health, social sciences, and engineering. We will explore algorithms, models, computational foundations, and inferential tools associated with the task of analyzing such data. Topics covered in this course include supervised and unsupervised learning, model selection, dimension reduction, matrix factorization, topic models, graphical models, interpretability and causality.

Learning outcomes: By the end of the course, students should be able to:

- Understand a variety of learning algorithms and practical considerations of learning models from data.
- Understand how to apply and assess models, conduct inference, and interpret results in real-world problems.
- Develop deep understanding of an active area of statistical machine learning and data science research through the class project.

Textbook and Course Materials: There is no required textbook for the course. Useful reference include:

- *The Elements of Statistical Learning* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 2001. You can find a downloadable PDF copy from the UCSC library.
- *An Introduction to Statistical Learning with Applications in R*, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013. You can find a downloadable PDF copy at `https://www.statlearning.com/`
- *Probabilistic Machine Learning: an Introduction*, by Kevin P. Murphy, 2022. You can find a downloadable PDF copy at `https://probml.github.io/pml-book/book1.html`

Prerequisites: Single-variable and multi-variable calculus and linear algebra. Previous coursework in statistics and probability (STAT 203 and 205, or equivalent), Bayesian modeling (STAT 206B, or equivalent). Previous programming experience in any programming language.

Assessment and Grading:

- *Homework (30%):* There will be 3 homework sets.
- *Project (70%):* See the project section for details.

Course Expectations: Though attendance is not required, it is strongly recommended. Students may work together on homeworks, but may not copy solutions from other students or from other sources.

Computing: We will be use the R programming language (www.r-project.org) in lectures and homework. But any programming language can be used for the project if needed.

Key Topics:

- Introduction to statistical learning principles.
- High-dimensional regression, model selection and sparsity.
- Linear and non-linear classification.
- Tree-based methods, bagging, and boosting.
- Clustering and latent variable models.
- Learning with fewer labels.
- Undirected graphical models, and structure learning.
- Causality and statistical learning for causal inference.

Project: I can't believe I (can't) break it Each student is expected to complete a course project. The project is based on reading a chosen paper and presenting an in-depth critique of it. The paper should be about an area of statistical machine learning. The paper can be on theory, methodology, or application, but it should contain proposals of some statistical methods for the purpose of analyzing data. The main goal of the project consist of three parts:

1. Understand and summarize the method and the problems this method was proposed to tackle.
2. Understand potential limitations of the method and hypothesize realistic scenarios where the method may fail. Preferably in situations not discussed extensively in the original paper.

3. Design simulation studies and implement and apply the method to your simulated scenarios. Validate or invalidate your hypothesis and explain your findings.

A list of potential papers will be provided, but students are encouraged to choose papers outside of the list as well, upon approval from the instructor. The project is **not** graded based on whether the chosen method can be broken with the simulation. The assessment of the project is based on the demonstrated understanding of the chosen method and the empirical performance of the method.

The project consists of the following milestones:

- Project proposal (5%): a short summary of the background, a short description of why the problem addressed by the paper is interesting/important/challenging, why are you interested in 'breaking' the method, and goals to be accomplished in the project. The proposal should be limited to 1 page.
- Midway presentation (10%): a 5 minute short talk introducing the chosen paper and the preliminary proposals on when it might not work.
- Final presentation (25%) a 20 minute talk by each student, presenting all aspects of the project.
- Final report (30%) a writeup that summarizes the project and findings. The report should be limited to 8 pages in NeurIPS style, including all figures and tables and excluding references.

Support for students with disabilities: UC Santa Cruz is committed to creating an academic environment that supports its diverse student body. If you are a student with a disability who requires accommodations to achieve equal access in this course, please submit your Accommodation Authorization Letter from the Disability Resource Center (DRC) to me privately during my office hours or by appointment, preferably within the first two weeks of the quarter. At this time, I would also like us to discuss ways we can ensure your full participation in the course. I encourage all students who may benefit from learning more about DRC services to contact DRC by phone at 831-459-2089 or by email at drc@ucsc.edu.

Support for students with other difficulties While we sincerely hope that you will be able to pursue your studies peacefully and worry-free, we are aware that in some cases difficulties happen that are beyond your control. You should always feel free and comfortable to bring up any problem with the instructor, but if this is not sufficient, or if you prefer professional help, here are several campus resources that you may want to consider contacting:

- UC Care which is a confidential space to discuss issues of dating violence, sexual assault and stalking.
- Slug Support where you can ask for help on many practical issues, including dealing with a financial crisis, problems with your living situation, computers, books, etc.

- CAPS, which provides counseling and psychological services to students.

Campus advocacy resources & education: Title IX prohibits gender discrimination, including sexual harassment, domestic and dating violence, sexual assault, and stalking. If you have experienced sexual harassment or sexual violence, you can receive confidential support and advocacy at the Campus Advocacy Resources and Education (CARE) Office by calling (831) 502-2273. In addition, Counseling and Psychological Services (CAPS) can provide confidential, counseling support, (831) 459-2628. You can also report gender discrimination directly to the University's Title IX Office, (831) 459-2462. Reports to law enforcement can be made to UCPD, (831) 459-2231 ext. 1. For emergencies call 911. Faculty and Teaching Assistants are required under the UC Policy on Sexual Violence and Sexual Harassment to inform the Title IX Office should they become aware that you or any other student has experienced sexual violence or sexual harassment. If you prefer to speak to someone confidentially, please contact UC Care (see above).

Academic Integrity: Academic integrity is the cornerstone of a university education. Academic dishonesty diminishes the university as an institution and all members of the university community. It tarnishes the value of a UCSC degree. All members of the UCSC community have an explicit responsibility to foster an environment of trust, honesty, fairness, respect, and responsibility. All members of the university community are expected to present as their original work only that which is truly their own. Plagiarism of any kind is unacceptable. All members of the community are expected to report observed instances of cheating, plagiarism, and other forms of academic dishonesty in order to ensure that the integrity of scholarship is valued and preserved at UCSC. Any student found in violation of the UCSC Academic Integrity policy may face both academic sanctions imposed by the instructor of record and disciplinary sanctions imposed by the graduate division. Violations of the Academic Integrity policy can result in dismissal from the university and a permanent notation on a student's transcript. For the full policy and disciplinary procedures on academic dishonesty, students and instructors should refer to the Academic Integrity page at the Division of Undergraduate Education or Graduate Division.