SDS 323: Statistical Learning and Inference (Spring 2022)
Monday, Wednesday, 3:00 pm - 4:30 pm

Instructor: Nhat Ho (minhnhat@utexas.edu)
Office hours: Monday, Wednesday, 1:30 pm - 2:45 pm

TA:

Goals. Welcome to SDS 323: Statistical Learning and Inference. The goal of this course is to introduce basic concepts and tools in data science, statistics, and machine learning that are widely used in practice to draw inferences about large-scale and real-world data. The course will mainly focus on data analysis and less on mathematical formulations and theorems. You only need to have basic mathematical preparation, such as basic linear algebra or probability at the level of SDS 321 or M 362K, to follow the topics covered in the course. Finally, we will cover many topics (see below) and the hope is that you can use some of the tools studied in this course for real-world applications.

Course materials. There are no required textbooks or course packets to buy. The main point of reference is the lecture notes from the class. There are three online resources relevant to the topics covered in the course:

- [ISL] *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. The book is freely available at http://www-bcf.usc.edu/~gareth/ISL/. This book will cover several topics in the class.
- [DSGI] *Data Science: A Gentle Introduction* by James Scott. The link to this book is at https://jgscott.github.io/STA371H_Spring2018/files/DataScience.pdf.
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The book is freely available at https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf. This optional book is for an in-depth understanding of topics in the class.

Topics. The topics covered in this class are largely based on the two books [ISL] and [DSGI]. Here, "ISL = An Introduction to Statistical Learning" and "DSGI = Data Science: A Gentle Introduction". We will go in order down the following list of topics.

- Warm-up: An introduction to R and a review of basic probability notion (lecture notes)
- Data visualization (DSGI Chapter 1)
- Some introductory concepts of Statistical Learning (ISL Chapters 1 and 2)
- Linear models (ISL Chapter 3)
- Classification (ISL Chapter 4)
- Resampling methods (ISL Chapter 5)
- Linear model selection and regularization (ISL Chapter 6)
- Nonlinear models (ISL Chapter 7)
- Tree-based methods (ISL Chapter 8)

- Support Vector Machines (ISL Chapter 9)
- Unsupervised learning: Clustering, Dimension reduction (ISL Chapter 10)

If time permits, we will also cover the following topics:

- Deep learning, Neural networks (lecture notes)
- Reinforcement learning (lecture notes)

Assignments, exams and grading. There are no in-class and final exams in this course. The final grade in this course is distributed as follows:

- There are 8 homeworks, which count for 60% of the final grade. The homework can be finished either by R or other relevant programming languages, such as Python, Matlab, C/C++.
- The final project counts for 40% of the final grade.

Note that, you are allowed to work in a group of up to four people on the homework and final project.

Attendance and quizzes. Attendance is required for this class. Furthermore, there are quizzes before class and during class. The quizzes are counted as bonus points to your homework and final project.

Late assignments and grace policy. If you are unable to submit your homework in time, please email me or the TA as soon as possible to ask for an extension. For each homework, you are allowed to have a two-day grace period. Each student can only have at most two extensions. After the two-day grace period, late assignments will be penalized 5 points per day late.

Final project details. For the final project, you are encouraged to pick your own data set, pose your own question with this data set and use the tools covered in the course to answer the question. Here are a few good sources of data sets (you do not need to just use data sets from these sources):

- UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets.php`
- CMU Statistical Data Repository: `http://lib.stat.cmu.edu/datasets/`
- Public Data Repositories: `http://statweb.stanford.edu/~sabatti/data.html`

If you cannot find the data set that you like, please email me or the TA and we will provide you some default data sets. However, you will only receive at most 90% of the final project grade if you use these default data sets. Finally, the evaluation of the final project will be based on the technical correctness and intellectual quality of your approach and write-up.

The final project is due on May 8th, 2022. Grace-period is not allowed for the final project. Late projects will be penalized 5 points per day late. You are encouraged to send me or the TA an outline of your question, data set, and technical tools in April so that we can provide you some feedback for your final project.

Final grades. I will use the following minimum thresholds for letter grades:

- A: 92
- A-: 88
- B+: 84
- B: 81
- B-: 78
- C+: 72
- C: 67
- D: 60

Students with Disabilities. Students with disabilities may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 512-471-6259, http://diversity.utexas.edu/disability/.

Diversity and Inclusion. One of my top priorities is that students from all diverse backgrounds and perspectives will be treated equally in this course. Furthermore, I believe that the diversity students bring to this class can be comfortably expressed and be viewed as a resource, strength, and benefit to all students. Please come to me at any time with any concerns.

Harassment Reporting Requirements. Senate Bill 212 (SB 212), which went into effect as of January 1, 2020, is a Texas State Law that requires all employees (both faculty and staff) at a public or private post-secondary institution to promptly report any knowledge of any incidents of sexual assault, sexual harassment, dating violence, or stalking "committed by or against a person who was a student enrolled at or an employee of the institution at the time of the incident". Please note that both the instructor and the TA for this class are classified by SB 212 as mandatory reporters. That means we must share with the Title IX office any information about sexual harassment/assault that is shared with us by a student—whether in-person, via electronic communication, or as part of any class assignment. Note that a report to the Title IX office does not obligate a victim to take any action, but this type of information cannot be kept strictly confidential except when shared with designated "confidential employees." A confidential employee is someone a student can go to and talk about a Title IX matter without triggering any obligation by that employee to have to report the situation so that it will be investigated. A list of confidential employees is available on the Title IX website. The professor and TA for this class are NOT designated confidential employees per SB 212.

Religious Holy Days. By UT Austin policy, you must notify me of your pending absence at least fourteen days prior to the date of observance of a religious holy day. If you must miss an examination, a work assignment, or a project in order to observe a religious holy day, you will be given an opportunity to complete the missed work within a reasonable time after the absence.

Policy on Scholastic Dishonesty. The University of Texas at Austin has no tolerance for acts of scholastic dishonesty. University policies regarding academic honesty and student conduct are outlined in Section 11, Appendix C of the University's General Information Catalog for this academic year. This catalog is the document of final authority for all matters of student conduct. If you are at all unclear about what constitutes scholastic dishonesty in this class or on its assignments, it is your responsibility to ask me for clarification. Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Policies on scholastic dishonesty will be strictly enforced. You should refer to the Student Conduct and Academic Integrity website at http://deanofstudents.utexas.edu/conduct/ to find more detail on official University policies and procedures on scholastic dishonesty as well as further elaboration on what constitutes scholastic dishonesty.