SDS 323: Statistical Learning and Inference (Spring 2020)
M W, 3:00 – 4:15 p.m., PAR 301

Instructor: James Scott (james.scott@mccombs.utexas.edu)
Office: GDC 7.516
Office hours: M W 1:30 – 2:45 PM (GDC 7.516)

TA: Rimli Sengupta (rimlisengupta@gmail.com)
Office hours: TBA

Background, goals, prereqs. Hi! Welcome to SDS 323. This course is a practical introduction to statistical learning, defined broadly as using data to draw conclusions about real-world systems, using a mix of tools from statistics and machine learning. This is **not** a very mathematically rigorous course. Our focus is on data analysis, not theorems; you'll see a lot of data sets and very few proofs. But you do need some basic mathematical preparation to understand the data-analysis tools we'll be covering. To that end, I do assume that you understand probability at the level of SDS 321 or M 362K. I also assume a working knowledge of differential calculus and basic high-school level descriptive statistics (what is a mean, standard deviation, box-and-whisker plot, etc). The course is intended as an overview, rather than an in-depth treatment of any particular topic. We will move fast and cover a lot, but we will focus on practical applications rather than theory.

Topics covered. We will go in order down the following list of topics. The corresponding readings are in parentheses. See the course homepage for more details on each topic. DSGI = "Data Science: A Gentle Introduction"; ISL = "Introduction to Statistical Learning" (see Materials section below).

- The data scientist's toolbox: R; Markdown and RMarkdown; version control with Git and Github (lecture notes).
- Warm-up: a bit of R and a bit of probability to start the semester (lecture notes)
- Basic data visualization (DSGI chapter 1).
- Fitting equations (DSGI chapter 2).
- Statistical learning: some introductory concepts (ISL Ch 1-2)
- Linear models (ISL Ch 3, DSGI Ch 6).
- Classification (ISL Ch 4).
- Resampling methods (ISL Ch 5).
- Regularization and feature selection in linear models (ISL Ch 6).
- Nonlinear models (ISL Ch 7).
- Trees and ensembles (ISL Ch 8).
- Latent feature models and principal component analysis (ISL Ch 10).
- Clustering: k-means and hierarchical clustering (ISL Ch 10).

If there's time, we will try to cover some or all of the following supplemental topics.

- Networks: basic concepts and visualization.
- Association-rule mining.
- Working with text data.
- Causal inference.

Course materials. There are no required textbooks or course packets to buy. Your main point of reference will be the lecture notes from class. As a reference on many topics we will also refer to two free online resources:

- *Data Science: A Gentle Introduction*, by James Scott. This is posted as a PDF file on the course web page. Referred to as "DSGI" in the course outline.
- *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani. The book is freely available at http://www-bcf.usc.edu/~gareth/ISL/. I'll refer to it as "ISL" in the course outline. We will cover about 75% of this book.

Communication. I will use Canvas pretty much only to send e-mails. I will post all assignments, announcements, etc., on the class GitHub site: https://github.com/jgscott/SDS323

Software. The use of R, RMarkdown, and GitHub are mandatory for this course. More information on these tools is available on the class website. For those without much programming experience, R is a great first language for data-science applications. Certainly many of you will be more comfortable in Python. If that's you, consider this an opportunity to learn a second platform—one that is not as powerful an all-purpose language as Python, but that is better for doing the kind of interactive data exploration, plotting, and iterative model fitting at the heart of good data science.

Assignments, exams and grading. There are no in-class exams and no final exam. Your grade for this course will come from:

- 60% homework. I will assign four homework assignments throughout the semester, which count for 15% of your final grade each. I will post assignments on the course home page, along with their due dates. Data-analysis assignments must be completed in RMarkdown and turned in via GitHub.
- 40% final project (see below), due on Wednesday, May 6, 2020.

You are allowed to work on the homework and projects in groups of up to 4 people. If you work in a group, please turn in one copy with all your names on it. If you want to work with a group but don't know many people in the class, that's OK! Please reach out to me and I will do my best to make introductions.

Plus/minus grades will be used for the final class grade for C grades and above. I use the following minimum thresholds for letter grades:

- A: 94.0
- A-: 90.0
- B+: 87.0
- B: 84.0
- B-: 80.0
- C+: 77.0
- C: 70.0
- D: 60.0

I do not round grades. Attendance is not an explicit component of your class grade.

Late assignments and grace policy. Sometimes we have bad days, bad weeks, and bad semesters. In an effort to accommodate any unexpected personal crises, I have built a grace-period policy into the course: that is, a one-time, three-day grace period for **one homework assignment**. You do not have to utilize this policy, but if you find yourself struggling with unexpected personal events, just e-mail me and our TA as soon as possible to notify us that you are using your one-time free grace period. If you subsequently turn in your assignment within 72 hours after the initial due date/time, there will be no penalty. All other late assignments will be penalized 10 points per day (or partial day) late, including if you turn in your "grace-period" after the three-day window.

Final project details. The assignment for the final project is simple: pose an interesting question; collect a relevant data set; and use the data, in conjunction with the tools we have learned in class, to answer the question you have posed. Make sure to address any shortcomings in the answer provided by your data and analysis. You will be evaluated both on the technical correctness (50%) and the overall intellectual quality (50%) of your approach and write-up.

This assignment is purposely open-ended, allowing you considerable freedom to follow a path dictated by your own intellectual curiosity. Strive to write something that a statistically literate person of wide-ranging interests (for example, a future employer) would find engaging and impressive.

The deliverable dates for the project are as follows:

- 5 PM on Wednesday, April 15, 2020: 2-page (max) project prospectus outlining the question, proposed methods, and data sources you hope to pursue for your project. The prospectus is ungraded, but it is an opportunity for you to get feedback on your idea and approach. If you don't turn in a project prospectus on time, then you will not receive feedback from me on your idea.
- 5 PM on Wednesday, May 6, 2020 (our last class day of the spring semester): the final project is due. Because of the quick turn-around required to grade final projects, I unfortunately cannot extend the grace policy to encompass the project. Late projects will be penalized 10 points per day or partial day, and if you turn in a late project, you may receive a temporary "Incomplete" grade in the course. But remember, you have all semester to get this sorted. If you do not turn in a final project, you will receive a grade of F for the course.

In case getting a data set proves too difficult, I will provide a "default" data set and project. If you use this data set, I will impose a 93% ceiling (i.e. an A-) on your grade. The last 7% is an incentive to be more creative and go with your own project.

Important Notifications

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. It is my intent that students from all diverse backgrounds and perspectives be well served by this course, that all students' learning needs be addressed, and 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 a class, 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.

Campus Safety. Please note the following key recommendations regarding emergency evacuation, provided by the Office of Campus Safety and Security, 512-471-5767, More info at: https://preparedness.utexas.edu/.

- Occupants of buildings on The University of Texas at Austin campus are required to evacuate buildings and assemble outside when a fire alarm is activated.

- Familiarize yourself with all exit doors of each classroom and building you may occupy.
- If you need evacuation assistance, inform the instructor in writing asap.
- In the event of an evacuation, follow the instruction of faculty or class instructors.
- Do not re-enter a building unless given instructions by Austin or UT police or fire authorities.
- Behavior Concerns Advice Line (BCAL): 512-232-5050 or on-line.
- In case of emergency, further information will be available at: http://www.utexas.edu/emergency.