ECO 395M: Data Mining and Statistical Learning

Course details

- Instructor: James Scott (james.scott@mccombs.utexas.edu)
- Course website: http://www.github.com/jgscott/ECO395M/
- Teaching assistant: Rui Zou (ruizuo11@utexas.edu)

Office hours: MW 1-2 PM in CBA 6.478

Overview

This is a master's level course on data mining/machine learning for students in the master's program in Economics at UT-Austin. 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. You'll see a lot of data sets, but not a lot of proofs. (Obviously the mathematical foundations of machine learning are important---and very interesting. They're just not what this course is about.)

The course homepage will have a detailed unit-by-unit description of what we're doing in class, together with topic-by-topic reading lists.

The prerequisites are the Statistics and Probability course and the Econometrics course in the Econ master's program.

Lectures

Lectures will take place in person in BRB 1.118.

Software

We'll rely on the following software:

- R and RStudio, a free, platform-independent graphical front-end for R. Make sure you have both R and RStudio installed, along with the RMarkdown package.
- Python, specifically in the form of Jupyter notebooks. You can either install Python and Jupyter on your machine
- Other software: please install Git and create a GitHub account, if you don't already have one. You will use GitHub for version control and to submit your assignments.

Readings

The course readings are all free and available online. Our main reference is *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani. The book is freely available here. I'll refer to it as "ISL" in the course outline. We will pretty much move beginning to end through this book, in order.

I'll also occasionally draw on two other references:

- *Elements*: selections from *Elements of Statistical Learning*, by Hastie, Tibshirani, and Friedman. A standard reference on data mining from a more statistical perspective. Referrered to as "Elements" in the course outline. This is a more advanced treatment of much of the material in ISL, and I will highlight selections from this book as optional, supplemental reading.
- *DSGI*: selections from *Data Science in R: A Gentle Introduction*, by James Scott.

Outline of topics

The core topics we will cover in this course are as follows. The corresponding readings are in parentheses. See the course homepage for more details on each topic.

- The data scientist's toolbox: R; Markdown and RMarkdown; version control with Git and Github (lecture notes).
- Statistical learning: some introductory concepts (ISL Ch 1-2)
- Linear regression (ISL Ch 3).
- 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 (readings TBA).

- Networks: basic concepts and visualization.
- Recommender systems.
- Working with text data.
- Causal inference.

Assignments 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 the exercises here, along with their due dates.
- 40% final project, which will be like a bigger, more complex version of your homework problems.

Homework

The homework assignments consist mainly of analyzing some data and writing a report on what you've found. Here's the submission protocol:

- Prepare your report as a single RMarkdown file.

- Knit the RMarkdown file to a Markdown (.md) output. Consult the RMarkdown documentation if you need help specifying this output format. Do not knit to HTML.
- E-mail the links for both the output (.md file) and the raw RMarkdown file (.Rmd) to our TA with the subject line: "ECO 395M Homework N: Your name(s) here". Obviously, you should substitute the correct number for N (e.g. 1, 2) as well as your own name/group members' names.

Do not send an attachment. Do not knit to an HTML file.

Late assignments and grace policy

Sometimes we have bad days, bad weeks, and bad semesters. In an effort to accommodate any unexpected, unfortunate personal crisis, I have built a grace 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, I encourage you to e-mail me and our TA as soon as possible to notify us that you are using our grace policy.

All other late assignments will be penalized 10 points per day or partial day that they are late.

Final project

The assignment for the final project is simple: pose an interesting question; collect a relevant data set; and use the data, in conjuction 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 dictacted 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.

Projects are due at 5:00 PM, US Central Time, on Monday, April 24. Because of the quick turn-around required to grade final projects, I unfortunately cannot extend the grace policy to encompass the project. But remember, you have all semester to get this sorted.

Groups are allowed

You are welcome to work on the assignments and the final project in groups of up to 3 people. (Groups aren't required; you can work on your own if you wish.) If you are working in a group, put all of your names in alphabetical order at the top of each assignment, and submit a single set of files for all of you.

If you'd like to work in a group but are having trouble coordinating with other class members, please let our TA or me know and we'll do our best to place you in a group.

Grade cutoffs

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.

Miscellaneous policies and notices

If you need help, please just ask :-)

Your success in this class is important to me. We will all need accommodations becausewe all learn differently. If there are aspects of this course that prevent you from learning or exclude you, please let me know as soon as possible. Together we'll develop strategies to meet both your needs and the requirements of the course. I also encourage you to reach out to the student resources available through UT. I am happy to connect you with a person or Center if you would like. This includes any of the following:

- Services for Students with Disabilities. This class respects and welcomes students of all backgrounds, identities, and abilities. If there are circumstances that make our learning environment and activities difficult, if you have medical information that you need to share with me, or if you need specific arrangements in case the building needs to be evacuated, please let me know. I am committed to creating an effective learning environment for all students, but I can only do so if you discuss your needs with me as early as possible. I promise to maintain the confidentiality of these discussions. If appropriate, also contact (Services for Students with Disabilities)[https://diversity.utexas.edu/disability/], 512-471-6259 (voice) or 1-866-329- 3986 (video phone).
- Counseling and Mental Health Center. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress. Yet all of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful. If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. (http://www.cmhc.utexas.edu/individualcounseling.html)

Names and pronouns

Class rosters are provided to me with each student's legal name, but I will gladly address you with whatever name and pronouns you feel most comfortable with. Please let me your preferences!

Academic Integrity

Each student in the course is expected to abide by the University of Texas Honor Code: "As a student of The University of Texas at Austin, I shall abide by the core values of the University and uphold academic integrity." Plagiarism is taken very seriously at UT. Therefore, if you use words or ideas that are not your own (or that you have used in previous class), you must cite your sources. Otherwise you will be guilty of plagiarism and subject to academic disciplinary action, including failure of the course. You are responsible for understanding UT's Academic Honesty and the University Honor Code which can be found at the following web address: http://deanofstudents.utexas.edu/sjs/acint_student.php

Q-drops

If you want to drop a class after the 12th class day, you'll need to execute a Q drop before the Q-drop deadline, which typically occurs near the middle of the semester. Under Texas law, you are only allowed six Q drops while you are in college at any public Texas institution. For more information, see: http://www.utexas.edu/ugs/csacc/academic/adddrop/qdrop

Important Safety Information:

If you have concerns about the safety or behavior of fellow students, TAs or Professors, call BCAL (the Behavior Concerns Advice Line): 512-232-5050. Your call can be anonymous. Trust your instincts and share your concerns.

The following recommendations regarding emergency evacuation from the Office of Campus Safety and Security, 512-471-5767, http://www.utexas.edu/safety/.

- Occupants of buildings on The University of Texas at Austin campus are required to evacuate buildings when a fire alarm is activated. Alarm activation or announcement requires exiting and assembling outside.
- Familiarize yourself with all exit doors of each classroom and building you may occupy. Remember that the nearest exit door may not be the one you used when entering the building.
- Students requiring assistance in evacuation shall inform their instructor in writing during the first week of class.
- In the event of an evacuation, follow the instruction of faculty or class instructors. Do not re-enter a building unless given instructions by the following: Austin Fire Department, The University of Texas at Austin Police Department, or Fire Prevention Services office.
- Link to information regarding emergency evacuation routes and emergency procedures can be found at: www.utexas.edu/emergency

Further recommendations regarding emergency evacuation from the Office of Campus Safety and Security, 512-471-5767, http://www.utexas.edu/safety/

Title IX and SB 212 Reporting

Title IX is a federal law that protects against sex and gender based discrimination, sexual harassment, sexual assault, sexual misconduct, dating/domestic violence and stalking at federally funded educational institutions. UT Austin is committed to fostering a learning and working environment free from discrimination in all its forms. When sexual misconduct occurs in our community, the university can:

1. Intervene to prevent harmful behavior from continuing or escalating.
2. Provide support and remedies to students and employees who have experienced harm or have become involved in a Title IX investigation.
3. Investigate and discipline violations of the university's relevant policies.

Faculty members and certain staff members are considered "Responsible Employees" or "Mandatory Reporters," which means that they are required to report violations of Title IX to the Title IX Coordinator. Both I and the TA for this course are a Responsible Employees and must report any Title IX related incidents that are disclosed to us in writing, discussion, or one-on-one. If you want to speak with someone for support or remedies without making an official report to the university, or without invoking an employee's legal obligation to report anything, please email advocate@austin.utexas.edu. For more information about reporting options and resources, visit (titleix.utexas.edu) or contact the Title IX Office at titleix@austin.utexas.edu.

Moreover, 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." 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.