Introduction to Machine Learning / Statistical Analysis and Learning

Professor Varun Rai
University of Texas at Austin, Spring 2024

17 Jan - 24 Apr, Wednesdays 9am-12pm at LBJ School, SRH 3.122
PA 397C (59815) / EER 396 (26399) / INF (27864)

Faculty
Varun Rai
Professor, LBJ School of Public Affairs
Professor, Mechanical Engineering
Sid Richardson Hall, Unit 3, Room: SRH 3.232
rai@austin.utexas.edu; 512-471-5057

Office hours: Mon 5:00 pm - 6:00 pm,
Wed 1:30 pm – 2:30 pm and by appointment

COURSE DESCRIPTION

Large datasets are increasingly becoming available across many sectors such as healthcare, energy, and online markets. This course focuses on methods that allow "learning" from such datasets to uncover underlying relationships and patterns in the data, with a focus on predictive performance of various models that can be built to represent the underlying function generating the data. The course starts with a review of basic statistical concepts and linear regression. But the course will focus mostly on introducing students to regression, classification, and clustering techniques beyond linear regression, such as tree-based approaches, support vector machines, and unsupervised learning (e.g., hierarchical clustering). This course is intended for first- and second-year Masters students. Ph.D. students with an interest machine learning models may also find this course useful.

In covering the material from the assigned textbook and complementary selected readings (e.g. journal articles), this course will emphasize both on formulaic and conceptual understanding of the discussed methods. As necessary, the instructor will draw on material from outside the textbook for driving conceptual clarity and showcasing the application of the methods learned to a broad range of practical problems.

PREREQUISITES

Basic grasp of statistics and linear regression would be helpful. However, all relevant concepts will be reviewed in class. Problem sets will include applied problems, including some from the textbook, that will require programming in R. All coding for this course will be in R. In the beginning, the instructor will point students to preparatory resources in R to provide the necessary background and toolsets in R that will be necessary in solving the problem sets.

CLASS FORMAT, ATTENDANCE, READINGS, AND ELECTRONIC DEVICES

The class meets in person, via a once-a-week lecture-style format. Consistent with central UT policy, I will not make individual accommodations for students to attend the seminar virtually, as that would severely disrupt the educational experience for everyone. To safeguard the classroom space as a place where students can try out ideas and speak freely, I also do not allow recordings of the seminar: audio OR video. Anyone violating this policy is

subject to disciplinary consequences by the LBJ School.

Attendance is mandatory, and crucial to student success in the course. It is in class discussions that students apply the concepts they encounter in their weekly readings to ongoing policy discussions. As per UT Austin policy, students must notify the instructor of any pending absence at least fourteen days prior to the date of observance of a religious holy day. If students must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, they will be given an opportunity to complete the missed work within a reasonable time after the absence. Students are each allowed one "freebie" absence. Any further unexcused absences reduce a student's participation grade by 10%. Late submission of assignments will incur a 25% grade penalty per late day.

Students are expected to complete the required readings each week ***prior*** to the class meeting for the unit and to contribute to the class discussion. **Required readings will include sections from the textbook and**, occasionally, a selection of 2-4 papers that will be posted on Canvas at least one week prior to the relevant unit.

Students may use their laptops or tablets in class, but only for the purpose of referring to their notes and readings directly related to this course. Any other use of electronic devices is not allowed.

ASSIGNMENTS AND GRADING

1. *CLASS PARTICIPATION (10%):* Participation grades depend on both quantity and quality of participation, including the effort and goodwill students put into our class meetings. Students should not only read, but also formulate questions, complement understanding through relevant materials outside the course, engage with the ideas raised by their peers, pushing each other toward a deeper understanding of the concepts we encounter. After the first five weeks, I will give feedback on each student's participation, so they can make any necessary adjustments. I expect full and active participation by all students during all class sessions. As outlined above, students are each allowed one "freebie" absence. Any further unexcused absences reduce a student's participation grade by 10%.

2. *WEEKLY READING REACTION PAPERS (10%):* Students are expected to complete the required readings each week *prior* to the class meeting for the unit and to contribute to the class discussions. **By every Tuesday 11:59 pm** (except for the first and last weeks), students will submit a ~500 words (two pages max) reaction paper based on the readings for the week. The goal is to be analytical and questioning, not descriptive. The reaction papers should therefore not summarize the readings—instead, they should address which points of the readings stood out and why. This could be because students were introduced to new fundamental insight(s) they weren't aware of before, because they seem in tension with other points in the readings, or because they are unclear or raised questions. As the course progresses, the reaction papers should try to build cumulatively, referring to or building upon readings and concepts discussed in prior weeks. Canvas will display the grade for each paper as "complete" or "incomplete." If a submitted paper does not meet the standards I expect, I will offer comments on why this is so. If a student receives no comments, they should assume that the paper was satisfactory.

3. PROBLEM SETS (20%): There will be four problem sets (PS) over the course of the semester. PS will be announced via email and posted on Canvas. The due date for each assignment will be noted on the assignment. Submit your assignments (including code) via Canvas. Late submission of assignments will incur a 25% grade penalty per late day.

4. IN-CLASS TESTS (40%): Two 90 minutes tests on March 6 and April 17.

5. GROUP PROJECT (20%): Each student will work in a group of 3 to 4 students on a term project to develop a predictive modeling analysis using real-world dataset(s). A project proposal will be due by 5pm on Friday, 23rd February. Each team will deliver a short (~15-20 min) presentation in class on 24th April and submit a 15-20 page project report due by 5pm on Monday, April 29. All students in the same project group will, by default, get the same grade for the project. However, if a group feels that this may be unfair for their project (for example, because of greatly different workload/contribution), the group must let me know by the last class day (24th April).

GRADING

Final grades will be determined by the following formula:

1. Participation in class discussions	10%
2. Weekly reading reaction papers	10%
3. Assignments	20%
4. In-class test (20% each)	40%
5. Group project	20%

REQUIRED READINGS

An Introduction to Statistical Learning, 2ed (2021), by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. A free pdf of the text and all associated datasets are available at: **https://www.statlearning.com/**. The course outline below provides details and schedule of the specific material to be covered from the book.

Relevant material from the textbook for each unit may be augmented with 2-4 additional readings, typically highly accessible and relevant journal articles selected by the instructor. These readings will be posted on Canvas at least one week prior to the unit when these readings will be discussed.

COURSE POLICIES AND DISCLOSURES

HONOR CODE

The University of Texas at Austin strives to create a dynamic and engaging community of teaching and learning where students feel intellectually challenged; build knowledge and skills; and develop critical thinking, creativity, and intellectual curiosity. As a part of this community, it is important to engage in assignments, exams, and other work for your classes with openness, integrity, and a willingness to make mistakes and learn from them. The UT Austin honor code champions these principles:

I pledge, as a member of the University of Texas community, to do my work honestly, respectfully, and through the intentional pursuit of learning and scholarship.

The honor code affirmation includes three additional principles that elaborate on the core theme:

- I pledge to be honest about what I create and to acknowledge what I use that belongs to others.

- I pledge to value the process of learning in addition to the outcome, while celebrating and learning from mistakes.
- This code encompasses all of the academic and scholarly endeavors of the university community.

The honor code is more than a set of rules, it reflects the values that are foundational to your academic community. By affirming and embracing the honor code, you are both upholding the integrity of your work and contributing to a campus culture of trust and respect.

ACADEMIC INTEGRITY EXPECTATIONS

Students who violate University rules on academic misconduct are subject to the student conduct process. A student found responsible for academic misconduct may be assigned both a status sanction and a grade impact for the course. The grade impact could range from a zero on the assignment in question up to a failing grade in the course. A status sanction can range from a written warning, probation, deferred suspension and/or dismissal from the University. To learn more about academic integrity standards, tips for avoiding a potential academic misconduct violation, and the overall conduct process, please visit the Student Conduct and Academic Integrity website at: http://deanofstudents.utexas.edu/conduct.

ARTIFICIAL INTELLIGENCE

The creation of artificial intelligence tools for widespread use is an exciting innovation. These tools have both appropriate and inappropriate uses in classwork. The use of artificial intelligence tools (such as ChatGPT) in this class shall be **permitted on a limited basis**. You will be informed as to the assignments for which AI may be utilized. You are also welcome to seek my prior-approval to use AI writing tools on any assignment. In either instance, AI writing tools should be used with caution and proper citation, as the use of AI should be properly attributed. Using AI writing tools without my permission or authorization, or failing to properly cite AI even where permitted, shall constitute a violation of UT Austin's Institutional Rules on academic integrity. If you are considering the use of AI writing tools but are unsure if you are allowed or the extent to which they may be utilized appropriately, please ask.

ACCOMMODATIONS FOR DISABILITIES

Students with disabilities may request appropriate academic accommodations from Disability and Access (D&A), 512-471-6259, https://diversity.utexas.edu/disability/. If D&A certifies your needs, I will work with you to make appropriate arrangements.

ACCOMMODATIONS FOR RELIGIOUS HOLIDAYS

Please refer UT Austin academic policies and procedures regarding absence from class for the observance of a religious holy day: https://catalog.utexas.edu/general-information/academic-policies-and-procedures/attendance/

MENTAL HEALTH ACCOMMODATIONS

I urge students who are struggling for any reason and who believe that it might impact their performance in the course to reach out to me if they feel comfortable. This will allow me to provide any resources or accommodations that I can. If immediate mental health assistance is needed, call the Counseling and Mental Health Center (CMHC) at 512-471-3515. Outside CMHC business hours (8a.m.-5p.m., Monday-Friday), contact the CMHC 24/7 Crisis Line at 512-471-2255.

IMPORTANT SAFETY INFORMATION

CARRYING OF HANDGUNS ON CAMPUS

Students in this class should be aware of the following university policies related to Texas' Open Carry Law:

- Students in this class who hold a license to carry are asked to review the university policy regarding campus carry.
- Individuals who hold a license to carry are eligible to carry a concealed handgun on campus, including in most outdoor areas, buildings and spaces that are accessible to the public, and in classrooms.
- It is the responsibility of concealed-carry license holders to carry their handguns on or about their person at all times while on campus. Open carry is NOT permitted, meaning that a license holder may not carry a partially or wholly visible handgun on campus premises or on any university driveway, street, sidewalk or walkway, parking lot, parking garage, or other parking area.
- Per my right, I prohibit carrying of handguns in my personal office. Note that this information will also be conveyed to all students verbally during the first week of class. This written notice is intended to reinforce the verbal notification, and is not a "legally effective" means of notification in its own right.

TITLE IX DISCLOSURE

Beginning January 1, 2020, Texas Education Code, Section 51.252 (formerly known as Senate Bill 212) requires all employees of Texas universities, including faculty, to report to the Title IX Office any information regarding incidents of sexual harassment, sexual assault, dating violence, or stalking that is disclosed to them. Texas law requires that all employees who witness or receive information about incidents of this type (including, but not limited to, written forms, applications, one-on-one conversations, class assignments, class discussions, or third-party reports) must report it to the Title IX Coordinator. Before talking with me, or with any faculty or staff member about a Title IX-related incident, please remember that I will be required to report this information.

Although graduate teaching and research assistants are not subject to Texas Education Code, Section 51.252, they are mandatory reporters under federal Title IX regulations and are required to report a wide range of behaviors we refer to as sexual misconduct, including the types of misconduct covered under Texas Education Code, Section 51.252. Title IX of the Education Amendments of 1972 is a federal civil rights law that prohibits discrimination on the basis of sex – including pregnancy and parental status – in educational programs and activities. The Title IX Office has developed supportive ways and compiled campus resources to support all impacted by a Title IX matter.

If you would like to speak with a case manager, who can provide support, resources, or academic accommodations, in the Title IX Office, please email: supportandresources@austin.utexas.edu. Case managers can also provide support, resources, and accommodations for pregnant, nursing, and parenting students.

For more information about reporting options and resources, please visit: https://titleix.utexas.edu, contact the Title IX Office via email at: titleix@austin.utexas.edu, or call 512-471-0419.
campus safety

The following are recommendations regarding emergency evacuation from the Office of Emergency Management, 512-232-2114:

- Students should sign up for Campus Emergency Text Alerts at the page linked above.
- Occupants of buildings on The University of Texas at Austin campus must 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.
- For more information, please visit the Office of Emergency Management.

UNIVERSITY RESOURCES

For a list of university resources that may be helpful to you as you engage with and navigate your courses and the university, see the University Resources Students Canvas page.

COURSE OUTLINE

Course Overview and Introduction
Readings Week 1. Jan 17: No readings

Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised. Broadly speaking, supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy. With unsupervised statistical learning, there are inputs but no supervising output; nevertheless, we can learn relationships and structure from such data.

Statistical Learning
Readings Week 2. Jan 24: ISL Chapter 1 and Chapter 2 + Papers posted on Canvas

Chapter 2 introduces the basic terminology and concepts behind statistical learning. This chapter also presents the K-nearest neighbor classifier, a very simple method that works surprisingly well on many problems.[1]

Linear Regression
Readings Week 3. Jan 31: Chapter 3 + Papers posted on Canvas

Chapter 3 reviews linear regression, the fundamental starting point for all regression methods.

Classification and Linear Discriminant Analysis
Readings Week 4. Feb 7: Sections 4.1- 4.3 + Papers posted on Canvas
Week 5. Feb 14: Sections 4.4 - 4.5 + Papers posted on Canvas

In Chapter 4 we discuss two of the most important classical classification methods, logistic regression and linear discriminant analysis.

Resampling Methods
Readings Week 6. Feb 21: Chapter 5 + Papers posted on Canvas

A central problem in all statistical learning situations involves choosing the best method for a given application. Hence, in Chapter 5 we introduce cross-validation and the bootstrap, which can be used to estimate the accuracy of a number of different methods in order to choose the best one.

Linear Model Selection and Regularization
Readings Week 7. Feb 28: Sections 6.1- 6.2 + Papers posted on Canvas
Week 8. Mar 6: Sections 6.3 - 6.4 + Papers posted on Canvas

Much of the recent research in statistical learning has concentrated on non-linear methods. However, linear methods often have advantages over their non-linear competitors in terms of interpretability and sometimes also

[1] Descriptions taken from *ISL*.

accuracy. Hence, in Chapter 6 we consider a host of linear methods, both classical and more modern, which offer potential improvements over standard linear regression. These include stepwise selection, ridge regression, principal components regression, partial least squares, and the lasso.

March 6: In-class Exam #1

Tree-Based Methods

Readings Week 9. Mar 20: Chapter 8 + Papers posted on Canvas

In Chapter 8, we investigate tree-based methods, including bagging, boosting, and random forests.

Support Vector Machines

Readings Week 10. Mar 27: Sections 9.1 - 9.2 + Papers posted on Canvas
Week 11. Apr 3: Sections 9.3 + Sections 6.3 - 6.4 + Papers posted on Canvas

Support vector machines, a set of approaches for performing both linear and non-linear classification, are discussed in Chapter 9.

Unsupervised Learning

Readings Week 12. Apr 10: Chapter 12 + Papers posted on Canvas

In Chapter 12, we consider a setting in which we have input variables but no output variable. In particular, we present principal components analysis, K-means clustering, and hierarchical clustering.

Deep Learning

Readings Week 13. Apr 17: Chapter 10 + Papers posted on Canvas

In this unit we will discuss the basics of neural networks and deep learning, and then go into specializations for specific problems, such as convolutional neural networks (CNNs) for image classification, and recurrent neural networks (RNNs) for time series and other sequences.

Apr 17: In-class Exam #2

Apr 24: Project presentations & Course Wrap-Up