STAT 471: Modern Data Mining

Fall 2021

Course and Instructor Information

Course information

Classroom:	Jon M. Huntsman Hall, Room 360
Class time:	Tue/Thu 10:15-11:45am
Canvas:	https://canvas.upenn.edu/courses/1597404
Github:	https://github.com/katsevich-teaching/stat-471-fall-2021
Piazza:	https://piazza.com/upenn/fall2021/stat471
Gradescope:	https://www.gradescope.com/courses/285259

Instructor: Eugene Katsevich

Office:	311 Academic Research Building
Email:	ekatsevi@wharton.upenn.edu

Teaching staff and office hours

Name	Office Hours	Location
Eugene Katsevich (Instructor)	Tue 1:00-3:00pm	ARB 311
Shuxiao Chen (Head TA)	TBD	TBD
Emily Guo (TA)	TBD	TBD
James Blume (TA)	TBD	TBD
Amanda Xu (TA)	TBD	TBD
Gantavya Pahwa (TA)	TBD	TBD
Tanya Thangthanakul (TA)	TBD	TBD

Course Description

With the advent of the internet age, data are being collected at unprecedented scale in almost all realms of life, including business, science, politics, and healthcare. Data mining—the automated extraction of actionable insights from data—has revolutionized each of these realms in the 21st century. The objective of the course is to teach students the core data mining skills of exploratory data analysis, selecting an appropriate statistical methodology, applying the methodology to the data, and interpreting the results. The course will cover a variety of data mining methods including linear and logistic regression, penalized regression, tree-based methods, and deep learning. Students will learn the conceptual basis of these methods as well as how to apply them to real data using the programming language R.

Prerequisites

Students are expected to have taken two semesters of statistics courses and to be familiar with multiple regression. Prior programming experience is helpful but not required.

Course Outline

(tentative and subject to change)

The course is structured into five units. The content of each unit will be presented over the course of four lectures, with an additional lecture devoted to a unit review and quiz.

Unit 1: Introduction to data mining

- Predictive modeling, exploratory data analysis, data wrangling, linear regression

Unit 2: Tuning predictive models

- Model complexity, bias-variance trade-off, cross-validation, classification

Unit 3: Regression-based methods

- Logistic regression, regression in high dimensions, ridge regression, lasso regression

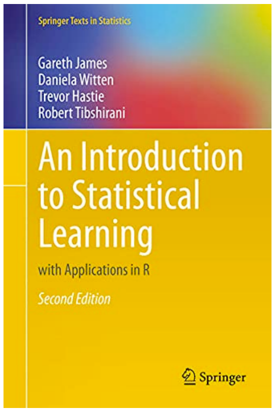
Unit 4: Tree-based methods

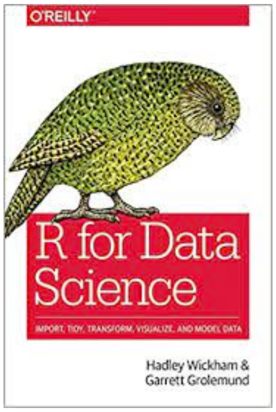
- Growing decision trees, tree pruning, bagging and random forests, boosting

Unit 5: Deep learning

- Single and multi-layer neural networks, optimization and computation, deep learning for image and text processing

Course Textbooks





Our primary textbook (required) is

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning.* Second edition. 2021.

This textbook[1] is available for purchase at the Penn Bookstore and freely available online.

We will use following textbook for R programming:

- Hadley Wickham and Garrett Grolemund. *R for Data Science.* 2016.

This textbook is freely available online.

[1] Students wishing to purchase a hard copy are advised to purchase the second edition. Students who already have a hard copy of the first edition need not purchase a hard copy of the second edition (assigned readings will be given with reference to both editions).

Course Logistics

- Course materials (lecture notes, homework, exams) will be distributed via Github. This private repository is accessible to enrolled students; see Canvas for instructions to join.
- Students will submit and receive feedback on homework and exams through Gradescope.
- The instructor and teaching assistant will hold office hours every week (times listed on the first page). Outside of office hours, students can ask questions about the course content on Piazza (rather than emailing the teaching staff). Students are encouraged to answer each others' questions, for which the instructor may award extra credit. Students should email the instructor with administrative questions.
- Students will use R, RStudio, R Markdown, Git, and Github to complete assignments and exams. Instructions to set up these tools are available on the course Github page and **a computing tutorial will be offered on Wednesday, September 1 (5:15-6:45pm, location TBA) to help students get set up.**

Assignments and Exams

Assessment is based on homework and quizzes for each unit, as well as a midterm exam and a final project.

Homework (25%)

There will be five homework assignments, one at the end of each unit. These homework assignments will involve conceptual questions as well as R programming questions. Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source and from consulting solutions to problems that may be available online and/or from past iterations of the course.

Quizzes (25%)

There will be five 25-minute, individual-work, open-book quizzes, one at the end of each unit. Quizzes will take place during the last 25 minutes of class on the fifth lecture of each unit, with the first hour of this class devoted to a unit review. Quizzes will be administered through Canvas, so students are asked to bring their laptops to class. Students unable to do so will be provided paper copies of the quiz.

Midterm exam (25%)

The midterm exam will take place on Monday, October 25 from 7-9pm. This individual work, open book exam will have a similar format to the homework, involving conceptual questions as well as R programming questions. **A midterm review session will be offered on Friday, October 22 (5:15-6:45pm, location TBA).**

Final project (25%)

In the final project, students will apply the methods they learned in class to tackle data mining problems of personal interest to them. Working individually or in teams of two, students will identify an analysis goal and find a dataset relevant to this goal. **The final project report is due on Sunday, December 19 by 11:59pm.**

Course Grades

An overall numeric grade will be computed for each student at the end of the semester by weighting the homework, quizzes, midterm, and final according to the above percentages. Letter grades will then be assigned based on numeric grade thresholds chosen at the discretion of the instructor.

The class mean and standard deviation will be posted for each assignment and assessment. Furthermore, a class distribution of interim numeric course grades based on the first three homeworks and quizzes as well as the midterm exam will be posted in advance of the grade type change deadline of October 29.

Extra credit will be awarded at the discretion of the instructor for participation in class and on Piazza. On Piazza, answering questions will be weighted more heavily than asking questions, with greatest weight given to instructor-endorsed answers.

Course Policies

Late homework

To offset the effect of relatively common difficult circumstances (computer crash, job interview, PDF compilation problem), **each student will get three "free" late days for homework submission over the course of the semester. No late penalty will be assessed for these three late days, with no need to request or justify this accommodation.** After a student has used his or her late days, each additional late day will come with a penalty of 10 points (out of 100). No homework will be accepted more than three days after the deadline. Lateness will be determined by the Gradescope timestamp and measured in whole days. Exceptions to this policy will be provided to students encountering major unforeseen circumstances (e.g. family emergencies) if they obtain a letter from their academic advisor or a departmental representative.

Quiz and midterm exam makeups

Students unable to take a quiz or the midterm at the scheduled times should notify the instructor as soon as possible, and makeups will be offered at the discretion of the instructor. A foreseen conflict (e.g. another class has an exam scheduled at the same time) must be corroborated with evidence of the conflict and an unforeseen conflict (e.g. family emergencies) must be corroborated with a letter from an academic advisors or departmental representative.

Regrades

All assignments will be graded through Gradescope, where points will be awarded or deducted based on clear rubrics. Regrade requests, which can also be submitted through Gradescope, will be considered only in cases when there is a clear discrepancy between the rubric and the grade. **A regrade request must be submitted within a week of the date the grade was posted.**

Academic integrity

In accordance with Penn's Code of Academic Integrity, students must comply with the course collaboration policies described in this syllabus and in the assignment instructions. **All academic integrity violations will be reported to the Office of Student Conduct and all**

assignments where violations occurred will receive grades of zero. If you have any questions about collaboration policies, please do not hesitate to contact the instructor.

Accessibility for students with disabilities

The instructor is committed to creating a learning experience that is as accessible as possible. Students with disabilities should reach out to the Office of Student Disabilities Services (SDS) by calling 215-573-9235 (services are confidential) and email the instructor. The instructor will then work with the student and SDS to provide reasonable accommodations.