

STAT 4710: Modern Data Mining

Fall 2022

JMHH 260 on Tue/Thu 3:30-5:00pm ~ katsevich-teaching.github.io/stat-4710-fall-2022/

Teaching staff member	Office Hours	Location
Eugene Katsevich (Instructor)	Tue 12:45-2:15pm	ARB 311
Joseph Deutsch (Head TA)	Wed 10:15-11:45am	JMHH F92
Samantha Simon (TA)	Wed 5:15-6:45pm	SHDH 110
James Blume (TA)	Fri 12:00-1:30pm	JMHH F96
James Kuang (TA)	Mon 3:30-5:00pm	Zoom
Alex Chen (TA)	Thu 8:30-10:00am	Zoom
Arnold Jia (TA)	Thu 5:15-6:45pm	Zoom

(JMHH for John M. Huntsman Hall, ARB for Academic Research Building, SHDH for Steinberg-Dietrich Hall)

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 required to have taken two semesters of statistics courses and to be familiar with multiple regression. Students are also required to have programming experience in at least one language (R is preferred, but experience in another language is sufficient).

Course Outline

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: R for Data Mining

- Course introduction, data visualization, data manipulation, data wrangling

Unit 2: Prediction fundamentals

- 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

- Neural networks, optimization, deep learning for image and text processing

Course Textbooks

Our primary textbook (required) is

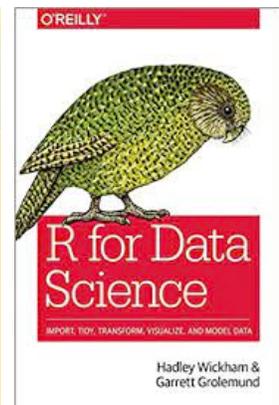
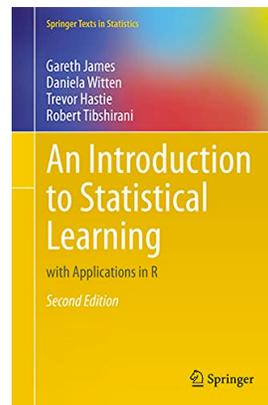
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning*. Second edition. 2021.

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

We will use following textbook for R programming:

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

This textbook is freely available [online](#).



Assignments and assessments

	Homework	Quizzes	Exams
Format	Take home	In class	Take home
Submission	Gradescope	Canvas	Gradescope
Duration	1-2 weeks	30-45 minutes	36 hours
Number of assessments	5 (one per unit)	5 (one per unit)	2 (midterm and final)
Percentage of grade	25%	25%	50%
Involves programming	✓	✗	✓
Course materials allowed	✓	✓	✓
Collaboration allowed*	✓	✗	✗
Internet allowed*	✓	✗	✓

*See course policies below for details.

Course policies

Letter grades

An overall numeric grade will be computed for each student at the end of the semester by weighting the homework, quizzes, and exams according to the above percentages. Final letter grades will then be assigned based on numeric grade thresholds chosen at the discretion of the instructor, but not below those specified in the table below.

Numeric grade	94-100	90-94	87-90	84-87	80-84	77-80	74-77	70-74	67-70	60-67	0-60
Minimum letter grade	A	A-	B+	B	B-	C+	C	C-	D+	D	F

Students can view their minimum letter grade on Canvas based on all graded assignments at any point during the semester. By the grade type change deadline, students will have grades for two homework assignments, three quizzes, and the midterm exam.

Extra credit

Extra credit will be awarded at the discretion of the instructor for participation in class and on Ed Discussion. On Ed Discussion, answering questions will be weighted more heavily than asking questions, with greatest weight given to instructor-endorsed answers. Beyond possible extra credit questions on the homeworks, quizzes, and/or exams, there will be no other extra credit opportunities due to limited time of the teaching staff.

Regrades

Homeworks and exams will be graded through Gradescope, where points will be awarded or deducted based on clear rubrics. Regrade requests for these assignments can also be made through Gradescope. These requests will be considered only in cases when there is a clear discrepancy between the rubric and the grade, and only if submitted within a week of the date the grade was posted. Quizzes are automatically graded; any disputes of the intended answers should be submitted via private message on Ed Discussion.

Late assignment submission

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 rounded up to the nearest whole day. For example, an assignment turned in a minute late is counted as one day late. Late submission is not permitted on the midterm and final exams. At instructor discretion, exceptions to these policies will be provided to students encountering major unforeseen circumstances (e.g. family emergencies) if they obtain a letter from their academic advisor or departmental representative.

Class participation and class recordings

Class participation is strongly encouraged to get the most out of STAT 4710. However, attendance or participation is not mandatory and will not be used as part of the course grade, aside from potential extra credit. Class recordings will be provided on Canvas and all in-class lecture materials will also be posted to the course webpage.

Masking in class and in office hours

Following the [University's guidance](#), "masks are optional (but encouraged) while on campus."

Collaboration

Students are permitted to work together on homework assignments, but must write up and submit solutions individually. In particular, students may not copy each others' solutions. Furthermore, students must disclose all classmates with whom they collaborated on a given homework assignment. Students are prohibited from collaborating on the quizzes or exams.

Allowed materials for assignments and assessments

Students may consult all course materials, including course textbooks, for all assignments and assessments. For programming-based assignments (homeworks and exams), students may also consult the internet (e.g. Stack Overflow) for help with general programming tasks (e.g. how to add a dashed line to a plot). Students may not search the internet for help with specific questions or specific datasets on any homework or exam. In particular, students may not use solutions to problems that may be available online and/or from past iterations of the course.

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 suspected 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.

Communication

The instructor and teaching assistants will hold office hours every week (times listed on the first page). Outside of office hours, students can ask the teaching staff questions on Ed Discussion. Students should only email the instructor in exceptional circumstances.

Quiz and exam makeups

Students unable to take a quiz or exam 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. The quizzes are in-class assessments, but students not feeling well enough to come to class are permitted to take these assessments from home (at the scheduled time). Instructor permission is not needed to do so, but students who are able to take the assessments in class are bound by the [Code of Academic Integrity](#) to do so.

Laptops in class

Students need to bring their laptops to class for programming exercises and quizzes. Outside of these two settings, students are encouraged not to have their laptops out.

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.