STAT 471: Modern Data Mining

Spring 2021

Instructor: Eugene Katsevich
Email: ekatsevi@wharton.upenn.edu
Office: upenn.zoom.us/my/ekatsevi

Section 1: Tue/Thu 12:00-1:30pm
Section 2: Tue/Thu 3:00-4:30pm
Website: canvas.upenn.edu

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.

Course Outline

(tentative and subject to change)

Introduction

- Data mining in the 21st century
- Statistical learning problem setup
- Prediction, association, and causation
- Regression versus classification
- R and the data mining workflow

Unit 1: Linear modeling

- Single and multiple linear regression
- Single and multiple logistic regression
- Beyond linearity

Unit 2: Model building

- K-nearest neighbors
- Model flexibility and the bias-variance trade-off
- Model evaluation
- Model selection via cross-validation

Unit 3: Penalized regression

- Linear and logistic LASSO regression
- Ridge and elastic net regression
- Dimensionality considerations

Unit 4: Tree-based methods

- Classification and regression trees
- Bagging and random forests
- Boosting

Unit 5: Advanced topics

- Multi-class logistic regression
- Text, image, and speech processing
- Deep learning

Wrap-up

- What is the “best” machine learning method?
- Computational considerations
- Beyond prediction

R Programming

- R Markdown
- Tidy paradigm (tidyverse)
- Data import (readr)
- Data manipulation (dplyr, tidyr)
- Data visualization (ggplot2)

Prerequisites

Two semesters of statistics courses, and familiarity with multiple regression. Prior programming experience is helpful but not required.

Course Format and Logistics

Format

This course will be conducted in the virtual "flipped classroom" format, blending 30-minute asynchronous lectures with 50-minute synchronous Zoom-based discussions and activities. Asynchronous lecture videos can be watched either in advance or during the first 30 minutes of class time and synchronous class sessions will begin 30 minutes into class time; see the table above.

	Watch asynchronous lecture	Attend synchronous class
Section 1	Tue/Thu 12:00-12:30pm or in advance	Tue/Thu 12:30-1:20pm
Section 2	Tue/Thu 3:00-3:30pm or in advance	Tue/Thu 3:30-4:20pm

Organization

The course is structured into five units: linear modeling, model building, penalized regression, tree-based methods, and advanced topics. Each unit introduces a different class of statistical learning models and/or strategies for choosing among them. The asynchronous lecture material will introduce the statistical methodologies relevant to the corresponding unit, and the synchronous class sessions will involve discussions of the material and hands-on application of these methodologies to datasets using R.

Logistics

All course materials (including asynchronous lecture videos, assignments, and exams) will be posted on Canvas. All synchronous sessions (including class and office hours) will be conducted via Zoom. The Zoom rooms for these sessions are listed on Canvas. Each synchronous class session will be recorded and posted to Canvas.

R Tutorials

Teaching assistants will offer two R tutorials during the first week of class. **There will be a basic R tutorial on Wednesday, January 20th from 4:30-5:30pm and an advanced R tutorial on Friday, January 22nd from 4:30-5:30pm.** These tutorials are optional but students —especially those without prior experience with R—are strongly encouraged to attend.

Communication

The instructor and teaching assistants will hold office hours every week (times listed in the table below). Outside of office hours, students can ask questions about course material and assignments on Piazza (piazza.com/upenn/spring2021/stat471). The TAs will monitor Piazza, but students are also encouraged to answer each others' questions (for which the instructor may award extra credit). Students should email the instructor with administrative questions.

	Email	Office Hours
Eugene Katsevich (Instructor)	ekatsevi@wharton.upenn.edu	TBD
Shuxiao Chen (Head TA)	shuxiaoc@wharton.upenn.edu	TBD
Rachel Levin (TA)	raclevin@wharton.upenn.edu	TBD
Amy Liu (TA)	liuamy@sas.upenn.edu	TBD
Tanya Thangthanakul (TA)	tthang@wharton.upenn.edu	TBD
Aaron Yu (TA)	aaronyu@sas.upenn.edu	TBD

Assignments and Exams

Students are expected to participate in class by responding to polls. At the end of each of the first four units, there will be a homework assignment and a quiz. There will also be a midterm exam and a final project. These assignments and assessments are detailed below.

Student participation (5%)

The instructor will solicit student input using the software Poll Everywhere both before and during the synchronous class sessions. Each student's submissions over the course of the semester will comprise 5% of his or her grade. Additionally, students who actively participate in class or answer other students' questions on Piazza will receive up to 5% extra credit.

Homework (25%)

There will be four homework assignments, one at the end of each of the first four units. These homework assignments will involve conceptual questions as well as R programming questions. Students can work in teams of up to three people. Submission can be done through Gradescope or Canvas; one per team. Students can use Piazza to find team members.

Quizzes (20%)

There will be four 20-minute quizzes, one at the end of each of the first four units. Quizzes will be individual work, open book, and multiple choice. Each quiz will take place at the end of a synchronous class session (i.e. 1:00-1:20pm for Section 1 and 4:00-4:20pm for Section 2).

Midterm exam (25%)

The midterm exam will take place on Monday, March 22 from 6-8pm. This individual work, open book exam will involve conceptual questions as well as R programming questions.

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 in teams of up to three, students will identify an analysis goal and either collect or find a relevant dataset. **The final project report (maximum 15 pages) is due on Sunday, May 2 by 11:59pm.** Selected teams will be invited to showcase their projects at the Data Science Live event on Friday, April 30 (see e.g. DSL from Fall 2019).

Accommodations will be made for students with disabilities or COVID-related constraints; please see the Course Policies section below for more details.

Course Schedule

(tentative and subject to change)

Day	Asynchronous	Synchronous	Day	Asynchronous	Synchronous
Wed 1/20	Basic R tutorial 4:30-5:30pm		**Tue 3/16**	Unit 3 Lecture 4	R Lab
Thu 1/21	Introduction	R Lab	**Thu 3/18**	Units 1-3 Review	Q&A, Quiz 3
Fri 1/22	Advanced R tutorial 4:30-5:30pm		**Sun 3/21**	**Homework 3 due by 11:59pm**	
Tue 1/26	Introduction	R Lab	**Mon 3/22**	**Midterm exam 6-8pm**	
Thu 1/28	Unit 1 Lecture 1	R Lab	**Tue 3/23**	Unit 4 Lecture 1	R Lab
Tue 2/2	Unit 1 Lecture 2	R Lab	**Thu 3/25**	Unit 4 Lecture 2	R Lab
Thu 2/4	Unit 1 Lecture 3	R Lab	**Tue 3/30**	No Class	
Tue 2/9	Unit 1 Lecture 4	R Lab	**Thu 4/1**	Unit 4 Lecture 3	R Lab
Thu 2/11	Unit 1 Lecture 5	Q&A, Quiz 1	**Tue 4/6**	Unit 4 Lecture 4	R Lab
Sun 2/14	**Homework 1 due by 11:59pm**		**Thu 4/8**	Unit 4 Lecture 5	Q&A, Quiz 4
Tue 2/16	Unit 2 Lecture 1	R Lab	**Sun 4/11**	**Homework 4 due by 11:59pm**	
Thu 2/18	Unit 2 Lecture 2	R Lab	**Tue 4/13**	Unit 5 Lecture 1	R Lab
Tue 2/23	Unit 2 Lecture 3	R Lab	**Thu 4/15**	Unit 5 Lecture 2	R Lab
Thu 2/25	Unit 2 Lecture 4	Q&A, Quiz 2	**Tue 4/20**	Unit 5 Lecture 3	R Lab
Sun 2/28	**Homework 2 due by 11:59pm**		**Thu 4/22**	Unit 5 Lecture 4	R Lab
Tue 3/2	Unit 3 Lecture 1	R Lab	**Tue 4/27**	Wrap-up	Discussion
Thu 3/4	Unit 3 Lecture 2	R Lab	**Thu 4/29**	Wrap-up	Discussion
Tue 3/9	Unit 3 Lecture 3	R Lab	**Fri 4/30**	Data Science Live	
Thu 3/11	Spring Break		**Sun 5/2**	**Final project due by 11:59pm**	

Course Materials

Textbooks

Our required textbook is known as ISLR and is freely available online:

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*, First Edition, 2013, Springer New York.

The following book is recommended for learning R and is freely available online:

- Garret Grolemund and Hadley Wickham. *R for Data Science*, 2016, O'Reilly.

The following optional book is an advanced reference and is freely available online:

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Second Edition, 2008, Springer.

The following optional book (available on Canvas) is a reference for general statistical methods:

- Fred Ramsey and Daniel Schafer. *The Statistical Sleuth,* Third Edition, 2013, Brooks/Cole.

Laptops
Laptops are required for hands-on data analysis, an essential part of the course.

Software
Students must download the free statistical computing language R and the integrated development environment RStudio.

Course Policies

Late work
Assignments may be turned in up to three days late, but with a 15% penalty per day. Lateness will be determined by the Gradescope timestamp; for example 12:01am is considered late.

Zoom
Students should treat the Zoom classroom as though it were a normal classroom. Therefore, **students are strongly encouraged to keep their videos on unless their network bandwidth prevents them from doing so.** As in a normal classroom, students are encouraged to speak up if they have questions. Students should feel free to do so in whatever way they feel comfortable: by unmuting themselves and speaking at any time, by using the "raise hand" Zoom feature and waiting to be called on, or by typing their question in the Zoom chat.

Regrades
Grading of all non-multiple-choice assignments, i.e. homework, midterm, and final project, will be through Gradescope. In this system, 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. Grades will not necessarily stay the same or increase as a result of a regrade request.

Make-ups
Quiz or midterm make-ups will be offered only if external circumstances prevent students from taking these assessments at the appointed times and students notify the instructor in advance.

COVID-related accommodations
Accommodations will be made for students with COVID-related constraints, such as geographical location or poor internet connectivity. Students unable to attend synchronous lectures due to their geographical location will be provided with alternate means for class participation. Students unable to take the midterm at the scheduled time will be given an alternate time slot appropriate to their time zone. Late penalties on assignments will be waived for students who encounter technical difficulties with submission, provided they take a screenshot of the issue and email the instructor as soon as possible.

Accessibility
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 notify the instructor by February 1. The instructor will then work with the student and SDS to provide reasonable accommodations.