STATS 415: Introduction to Data Mining

Winter 2018

Prof. Liza Levina

Logistics

- **Lecture:** TuTh 11:30-1:00, 1202 SEB
- **Contact info:** 459 West Hall, 764-3235, elevina@umich.edu. Please ask all course-related questions on piazza or in class, not by email (unless they are about something personal).
- **Office Hours:** Mon 2:00-3:30pm and Thu 4:00-5:30pm, 459 West Hall.
- **Course Webpage:** Canvas, plus piazza for discussions (can be accessed from Canvas). **You are responsible for obtaining all notes and assignments from Canvas.**
- **GSIs:** April Cho (aprilcho@umich.edu, section 004), Nick Seewald (nseewald@umich.edu, section 002), Weijing Tang (weijtang@umich.edu, section 003).
- **GSI Office hours:** 2165 USB.

Tue 9-10am	Nick
Wed 10-11:30am	April
Thu 9-11am	Nick
Thu 1-4pm	Weijing
Fri 2:30-4pm	April

Text

G. James, D. Witten, T. Hastie, R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R.* Springer. Website: http://www-bcf.usc.edu/ gareth/ISL. A pdf file of the book can be downloaded from the website for free.

Topics covered

This course will provide an introduction to main topics in data mining / statistical learning, including: statistical foundations, data visualization, classification, regression, clustering. Emphasis will be on statistical learning methodology and the models, intuition, and assumptions behind it, as well as applications to real-world problems.

Prerequisites

Math 215 (multivariate calculus) and 214/217 (linear algebra), and one of Stats 401, 406, 412 or 426.

Computing

We will be using the statistical programming language R, which can be downloaded for free from www.r-project.org on any system. A variety of tutorials and manuals for R are available freely online. The GSIs will cover R code for methods we learn in lab sections. Please direct all R questions to the GSIs.

Grading

Homework 20%, midterm 30% (Thu Mar 8, in class), final 30% (Tue Apr 24, 10:30-12:30pm), group project 20 %.

Homework

There will be nearly weekly homework assignments, typically consisting of a data analysis assignment, to be done in R, and/or conceptual and calculation questions that can be done on paper. We strongly encourage R Markdown for writing up the computing part of the homework (or knitr and latex if you prefer). The data reports you submit are expected to be organized and look professional; sloppy writing will result in points taken off, just like incorrect solutions.

The worst homework score will be dropped; therefore **late homework is not accepted.** Your lab GSI will be grading your homework and any questions about homework grading should be directed to her/him first. If you request regrading of your homework or want to raise any issues about points, you must do so **within one week** of the homework being returned.

Collaboration policy

You are allowed and encouraged to *discuss* homework with each other, including in open discussion forums on piazza, but ultimately all questions must be done and written up independently. Identical homeworks will receive no credit and lead to serious consequences. Any form of plagiarism, including from open online sources, in homeworks or project will be taken very seriously and can result in failing the course. If you use any external sources, of any kind, you must cite and credit them properly.

Project

The project will be done in small groups of 3-4 students. The goal of the project is to choose a topic that interests you, find or collect a relevant dataset that can be used to answer some interesting and challenging questions about the topic, formulate these questions and answer them using at least some of the methods covered in class, and write a formal report due at the end of term.

The first step for the project is to form a group and write a one-page project proposal, due after winter break. Feedback on proposals will be provided to make sure the group is on the right track.

Exams

Exams are in class and do not involve a computer. The final is not cumulative. Exams do not test the knowledge of R functions, but may require understanding R output. Sample exams consisting of questions from prior years will be provided.

Piazza and participation bonus

All course-related questions should be asked on piazza. We strongly encourage you to check first to see if someone has already asked your question, and answer other students' questions. The GSIs and the professor will be monitoring piazza, endorsing correct student answers, and answering questions that remain after a discussion. As a bonus, **up to 3 percentage points will be added to your final course grade based on piazza participation** (primarily based on the number of instructor-endorsed answers, but also on asking helpful questions).

Participating in online discussions is a voluntary activity. Thus you will need to register yourself in piazza if you wish to participate. Please follow the link from Canvas for details.