STATS 415: Data Mining and Statistical Learning

University of Michigan, Ann Arbor

Fall 2019

Prof. Jeffrey Regier

- **Lecture:** Tuesday & Thursday, 2:30pm–4pm, 260 Weiser Hall
- **Contact info:** 454 West Hall, 734-647-1670, regier@umich.edu. Please ask all course-related questions on Piazza or in class, not by email, unless they are about something personal.
- **Course Canvas website:** `https://umich.instructure.com/courses/327682`
- **Course Piazza website:** `https://piazza.com/umich/fall2019/stats415005fa2019`
- **Lab section 006:** Tuesday 4pm–5:30pm. B760 East Hall. GSI is Brook Luers (luers@umich.edu)
- **Lab section 007:** Thursday 5:30pm–7pm. B760 East Hall. GSI is Yangyi Lu (yylu@umich.edu)
- **Office hours:** see Canvas.

Textbooks

Required: G. James, D. Witten, T. Hastie, R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R.* Springer. `http://www-bcf.usc.edu/~gareth/ISL`.

Optional: T. Hastie, R. Tibshirani, J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer. `http://web.stanford.edu/~hastie/ElemStatLearn`.

Both books can be downloaded from their websites for free.

Topics covered

This course will provide an introduction to data mining / statistical learning. It covers statistical foundations of learning and mining methods, dimension reduction, classification, regression, and clustering. Emphasis will be on the models, intuition, and assumptions. For all methods covered, we will learn theory, R code, and applications to real-world problems.

Prerequisites

Multivariate calculus, linear algebra, and an upper-level statistics course (please see the LSA Course Guide for details). STATS 250 alone will not suffice.

Computing

We will be using the statistical programming language R, which can be downloaded for free from www.r-project.org. Many R tutorials are available freely online. The labs will cover R code for methods we learn in lecture. Please direct all R questions to the GSIs. Homework must be done in R, but for the final project you may use any language.

Grading

Quizzes 10%, homework 20%, midterm 20%, final 30%, group project 20%. There are no preset grade cutoffs in this class. The cutoffs will be set so that the distribution of final grades approximately matches that of previous STAT 415 classes, and so that students with very close totals will receive the same final grade.

Quizzes

Brief quizzes will be held weekly during labs, beginning the week that the first homework assignment is due. They will check your understanding of 1) the topics on the most recently due homework, and 2) the assigned reading. The lowest two quiz grades for each student will be dropped; no make-up quizzes will be given.

Homework

There will be nearly weekly homework, typically consisting of a data analysis assignment, to be done in R, and conceptual and/or derivation questions. The lowest homework score will be dropped; therefore *late homework is not accepted.* Homework will be submitted electronically through Canvas as a pdf, along with any R code used to generate results appearing in the pdf. Any R code submitted should run without errors.

Collaboration policy

You are allowed and encouraged to *discuss* homework. Posting on Piazza is also encouraged; however, please do not post solutions or parts of solutions on Piazza. Ultimately, homework must be done and written up independently. Homeworks with substantial overlaps will receive no credit and be referred to the office for academic integrity. Any form of plagiarism, including from open online sources, will be referred to lsajudicial@umich.edu, and can result in failing the course. If you use external sources, you must cite and credit them.

Exams

The midterm exam is in class on Tuesday October 22. The final exam is on Friday December 13 from 4pm–6pm. Both exams are closed book, and do not involve a computer. You can bring one standard size sheet of paper, writing whatever you want on both sides, and a calculator. Exams do not test knowledge of R functions, but may require understanding R output. If you have SSD-approved accommodations for exams, please submit the documentation at least two weeks in advance. There will be no alternate test dates unless necessary for SSD accommodations.

Regrading requests

Questions about homework grading should be directed to the GSIs first; questions about graded exams should be submitted to the professor in writing. If you wish to request regrading of your work, you must do so *within one week* of when the original grade was issued.

Project

The project will be done in groups of 2 or 3 students from the same lab section. For the project, you will select one or more datasets from the UCI Machine Learning Repository, formulate several challenging questions about the data, and answer them using methods covered in class. You will present your findings in an oral presentation during the last week of classes and in a written report due on the last day of class. The first step is to form a group and write a one-page project proposal.

Piazza and participation bonus

All course-related questions (outside of class) should be asked on Piazza. Please check first to see if someone has already asked your question, and answer other students' questions. We will be checking Piazza approximately once a day, endorsing correct answers, and answering questions that remain. As a bonus, *up to 2 percentage points* will be added to your final course total based on class participation, primarily from Piazza reports.