STATS 415: Data Mining and Statistical Learning

University of Michigan

Winter 2022

Prof. Jeffrey Regier

- **Course Canvas website:** `https://umich.instructure.com/courses/494514`
- **Lecture:** Tuesday & Thursday, 11:30 am – 1 pm, 1202 School of Education Building (SEB)
- **Primary instructor email:** regier@umich.edu (Please read the section below titled "How to get help" before emailing.)
- Discussion sections are held on Fridays (beginning January 14, 2022) in B760 East Hall
  - **Section 002:** 8:30 am – 10 am, GSI is Kevin Christian Wibisono (kwib@umich.edu)
  - **Section 003:** 1 pm – 2:30 pm, GSI is Brian Manzo (bmanzo@umich.edu)
  - **Section 004:** 2:30 pm – 4 pm, GSI is Easton Huch (ekhuch@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.* Second Edition. Springer. `https://hastie.su.domains/ISLR2/ISLRv2_website.pdf`.

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

Both books can be downloaded from their websites for free.

Topics covered

This course provides an introduction to data mining and statistical learning. It covers statistical foundations of learning and mining methods, dimension reduction, classification, regression, clustering, and neural networks. The course emphasizes models, intuition, and assumptions. For all methods covered, we will learn theory, R code, and applications to real-world problems.

Prerequisites

Multivariate calculus (MATH 215), linear algebra (MATH 214 or MATH 217), and at least one upper-level statistics course (e.g., STATS 401, STATS 412, STATS 425, STATS 426).

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 questions about R code to the GSIs. The homework assignments and the group project must be completed using R Studio and R Markdown.

Grading

Course grades are based on homework (25%), the midterm exam (20%), the final exam (40%), and the group project (15%). Course grade cutoffs are not preset. Instead, the cutoffs will be set so that the distribution of final grades approximately matches that of previous offerings of STAT 415. Historically, the A range in this class is 90–100%, the B range is 80–89%, and the C range is 65–79%. The instructors will not provide preliminary grade estimates.

Homework

There will be weekly homework assignments typically consisting of data analysis, 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 Markdown (rmd) code used to generate results appearing in the pdf. Any rmd code submitted should run without errors.

Homework assignments will also typically involve discussing part of the course textbook through our class Perusall site: `https://app.perusall.com/courses/stats-415-001-wn-2022/`.

Collaboration policy

You are allowed and encouraged to *discuss* homework. Posting on the Canvas discussion board is also encouraged; however, please do not post solutions or parts of solutions. Ultimately, homework must be done and written up independently. Homework submissions with substantial overlap 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 will be administered in class on February 24, 2022. This exam is fully computer-based. You need to use your own laptop to complete the midterm. You are free to bring in any physical material, refer to any file on your computer or hard disk and browse the Internet for other resources. However, you are not allowed to communicate with anyone else by any manner (text, social media, etc.) during the exam. A sample midterm exam will be released on Canvas beforehand.

The final exam will be administered during April 26, 2022, 4 pm – 6 pm. This exam is closed book and does not involve a computer. You are allowed to bring one standard size sheet of paper, writing whatever you want on both sides, and a calculator. The final exam does not test knowledge of R functions, though it may require understanding R output. A sample final exam consisting of questions from prior years will be provided.

If you have SSD-approved accommodations for exams, please submit the documentation at least two weeks in advance. There will be no alternate exam dates unless necessary for SSD accommodations.

Group project

The group project will involve analyzing a provided real dataset and submitting a written report. Form project groups before the midterm, ideally with three members, but two or four is also acceptable. Because the final project is regarded as the final exam on the computational part of this course, any sort of cross-team communication is forbidden.

How to get help

To ask questions about homework, please use our class Canvas discussion board if your questions can be asked without revealing solutions to your classmates. The course instructors will check the Canvas discussion board approximately once a day. Alternatively, you can get answers to your questions about homework during GSI office hours. Please do not use email to ask questions about homework.

To ask questions about the lecture (or lecture slides, or textbook), please visit me during my office hours. For quick questions, I am also generally happy to talk immediately after class. For questions about the textbook, consider asking your questions inline through our class Perusall site. Please do not use email to ask questions about course material.

For questions/concerns about grading, please first contact the GSI who graded your assignment, either by attending this GSI's office hours or by email. If you wish to request regrading of your work, you must do so *within one week* of when the original grade was issued. If you are not satisfied with the answer you get from a GSI and wish to contest the GSI's grading, please email me or visit me during office hours. I will generally defer to the GSIs' judgement on subjective matters (e.g., amount of partial credit).

For questions that are personal (e.g., concerns about keeping up with the class, extended illness), please visit me during my office hours or send me an email to make an appointment.