MATH 40024/50024: Computational Statistics

Fall 2019 (Aug 22, 2019 - Dec 8, 2019)

Lecture times: Mondays/Wednesdays/Fridays 11:00AM-11:50AM

Location: MSB 158

Instructor: Tsung-Heng Tsai | ttsai1@kent.edu | Office Hours: Mondays/Wednesdays/Fridays 12PM-1PM or by appointment

Textbooks: There is no required textbook but course notes will be provided throughout the course. Useful references are:

1. *Foundations of Applied Statistics*, J. Storey, 2019
2. *R for Data Science*, H. Wickham and G. Grolemund, 2017

Plagiarism: Be familiar with the university's academic integrity policy on cheating and plagiarism. (https://www.kent.edu/policyreg/administrative-policy-regarding-student-cheating-and-plagiarism)

Syllabus: here.

Course Objectives

Computation plays an essential role in modern statistics and machine learning. This course aims to develop a broad working knowledge of modern computational statistics. The topics include tools for exploratory data analysis, simulation, computational techniques for optimization, numerical integration, statistical modeling, statistical inference, and statistical prediction.

The course will use the programming language R. In many cases the course will rely on the existing implementations of statistical methods, but some programming effort will also be required.

After successful completion of the course, the students will understand the underlying principles of modern computational methods used in statistics, and be able to (1) use computational techniques to organize, explore, summarize, and analyze data, (2) use computation as a tool to investigate the properties of statistical methods, and (3) appropriately apply and/or develop computational methods to solve statistical problems.

Schedule

(Subject to change.)

Week 1 (Fri Aug 23): Introduction

- Course expectations
- Reproducible research
- R, Rstudio, R Markdown

Week 2 (Mon Aug 26 - Fri Aug 30): Fundamentals of R

- Basic data structures
- Indexing and iteration
- Function
- Readable and efficient R code

Week 3 (Mon Sept 02 - Fri Sept 06): Data tidying

No class Monday: Labor Day.

- Introduction to tidyverse
- Tidy data
- Data tidying with `tidyr`

Week 4 (Mon Sept 09 - Fri Sept 13): Data transformation

- Pipes `%>%`
- Data transformation with `dplyr`
- Split-apply-combine
- Relational data

Week 5 (Mon Sept 16 - Fri Sept 20): Data Visualization

- Visualization with `ggplot2`
- Layered grammar of graphics
- Principles and practice

Week 6 (Mon Sept 23 - Fri Sept 27) Data wrangling

- Strings, factors, date-times
- Exploratory data analysis
- More on data visualization and transformation

Week 7 (Mon Sept 30 - Fri Oct 04): Random variable and simulation

- Random number generator
- Simulation

Week 8 (Mon Oct 07 - Fri Oct 11): Midterm Exam

No class Friday: Fall Break.

- Review for midterm exam
- **Midterm exam** (in class) on **Wednesday Oct 9**

Week 9 (Mon Oct 14 - Fri Oct 18): Statistical modeling I

- Fitting models to data
- Tidying model objects with `broom`
- Evaluating models

Week 10 (Mon Oct 21 - Fri Oct 25): Statistical modeling II

- Working with large numbers of models
- Resampling methods

Week 11 (Mon Oct 28 - Fri Nov 01): Maximum likelihood estimation

- General optimization
- Univariate and multivariate optimization
- Expectation-maximization (EM) algorithm

Week 12 (Mon Nov 04 - Fri Nov 08): Numerical integration

- Numerical quadrature
- Monte Carlo methods

Week 13 (Mon Nov 11 - Fri Nov 15): Markov chain Monte Carlo

No class Monday: Veterans Day Observed.

- Gibbs sampling
- Metropolis-Hastings methods

Week 14 (Mon Nov 18 - Fri Nov 22): Statistical prediction

- Training and test sets
- Parameter tuning
- Cross-validation

Week 15 (Mon Nov 25 - Fri Nov 29): Reproducible research

No class Wednesday and Friday: Thanksgiving Break.

- Version control, Git, GitHub
- R package

Week 16 (Mon Dec 02 - Fri Dec 06): Unsupervised analysis

- Principal component analysis (PCA)
- Singular value decomposition (SVD)
- Clustering
- **Fianl exam** (take-home) released on **Friday Dec 06**

Course operation

Each week, there will be lectures on Monday and Wednesday, and an in-class lab on Friday. The instructor will be available in the lab session for help. Unless otherwise noted, the lab will be due 11:59PM on Sunday (the end of the week). Attendance to labs is optional, but highly encouraged. There will also be an in-class midterm exam (on Wednesday Oct 9), and a final take-home exam (assigned on Friday Dec 6).

Grading

Grades will be calculated as follows:

- Labs: 60%
- Midterm exam: 20%
- Final exam: 20%

The final letter grades will follow the usual scale:

- 90-100 = A-range (i.e., A+, A or A-)
- 80-89 = B-range (i.e., B+, B or B-)
- 70-79 = C-range (i.e., C+, C or C-)
- 60-69 = D
- 0-59 = F

The cutoffs for "+" and "-" will be determined at the end of the semester, at the discretion of the intructor. This scale is subject to change at the discretion of the instructor

Labs

Each lab will consist of a set of hands-on exercises, and will be completed in R Markdown format (with Rmd extension). An Rmd file contains a combination of content with simple text and R code chunks. Labs will be turned in through Blackboard, due 11:59pm on Sunday (the end of the week). Each lab must be submitted as an R Markdown source file and the resulting PDF output (generated by calling "Knit to PDF").

Students may choose to discuss and collaborate with friends on the labs, but your submitted work must be your own. Sharing of solutions will not be tolerated and will be considered cheating.

No late work will be accepted. Extensions may be given individually if requested at least 48 hours in advance of the due date with a reasonable justification. The two lowest lab scores of the semester will be dropped.

Exams

One in-class midterm exam, and one take-home final exam. No collaboration with peers is allowed.