Math 40024/50024-001: Computational Statistics

Tsung-Heng Tsai

Spring 2021

E-mail: ttsai1@kent.edu
Office Hours: MWF 10:00-11:00 a.m. (*online*)
Office: MSB 372

Web: tsunghengtsai.github.io/compstat-S21
Class Hours: MWF 8:50-9:40 a.m.
Class Room: *online*

Course Description

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 computational methods and tools for data wrangling, exploratory data analysis, Monte Carlo simulations, statistical inference, statistical modeling, 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.

Prerequisites

You should have completed both Math 20011 and Math 21001 with a C or better. If you are enrolled in 50024, you must have graduate standing in Mathematics. Students who do not have the proper prerequisites risk being deregistered from the class. Please contact instructor if you would like to take the course, but do not satisfy the prerequisite.

Textbook

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

1. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*, Rafael A. Irizarry, CRC Press, 2019
2. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, Hadley Wickham and Garrett Grolemund, O'Reilly Media, 2017

3. *Computer Age Statistical Inference*, Bradley Efron and Trevor Hastie, Cambridge University Press, 2016.
4. *An Introduction to Statistical Learning*, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, 2013.

Further reading will be recommended to support weekly class material.

Course Format

The course is offered remotely during January 19, 2021-May 4, 2021. Every week, there are recorded video lectures for classes on Monday and Wednesday. The videos and associated notes will be available on Blackboard. There is also a lab assignment on Friday. Unless otherwise noted, the lab will be due at 11:59 p.m. on next Thursday. The instructor will hold online meetings through Blackboard Collaborate Ultra at class hours (i.e., MWF 8:00-8:50 a.m.), to answer questions and/or discuss extra examples. Attendance to online sessions is optional, but highly encouraged.

There will also be a take-home midterm exam and a final project. More details about the final project will be posted and discussed later in the semester.

Labs

Every week, a lab assignment will be posted on Blackboard on Friday, and it will be due at 11:59 p.m. on next Thursday. Each lab consists of a set of hands-on exercises, and it needs be completed in R Markdown format (with Rmd extension). An Rmd file contains a combination of content with simple text and R code chunks. Each lab must be submitted to Blackboard as an R Markdown source file and the resulting HTML output.

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.

In general, **NO** late submissions will be accepted. In case of truly exceptional situations (e.g., family emergencies or illness), the instructor may make exceptions and allow late submission. The two lowest lab scores will be dropped at the end of the semester.

Midterm Exam

There will a take-home midterm exam. The exam will be posted on Blackboard at 8:00 a.m. on Wednesday March 17, and you have to upload your solutions as an R Markdown source file and the resulting HTML output to Blackboard by 11:59 p.m. on Friday March 19. Please note that you are **NOT** allowed to discuss with other students and the submitted work must be your own.

Course Policy

Important policy for this course is detailed below.

Grading

Grades will be calculated as follows:

- Labs: 50%
- Midterm exam: 20%
- Final project: 30%

The final letter grades will follow the usual scale:

- 90–100 = A-range (i.e., 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

Re-grades

All re-grading requests should be made in writing, within one week after receiving a grade. The request should state the specific question that needs to be re-graded, as well as a short explanation of why re-grading is necessary. The new grade may be lower than the original grade.

Academic Integrity

University policy 3-01.8 deals with the problem of academic dishonesty, cheating, and plagiarism. None of these will be tolerated in this class. The sanctions provided in this policy will be used to deal with any violations. If you have any questions, please read the policy at http://www.kent.edu/policyreg/administrative-policy-regarding-student-cheating-and-plagiarism and/or ask.

Accommodations for Students with Disabilities

Kent State University is committed to inclusive and accessible education experiences for all students. University Policy 3342-3-01.3 requires that students with disabilities be provided reasonable accommodations to ensure equal access to course content. Students with disabilities are encouraged to connect with Student Accessibility Services as early as possible to establish accommodations. If you anticipate or experience academic barriers based on a disability (including mental health, chronic medical conditions, or injuries), please let me know immediately.

Student Accessibility Services (SAS) Contact Information:

- Location: University Library, Suite 100
- Email: sas@kent.edu
- Phone: 330-672-3391; VP 330-968-0490
- Web: www.kent.edu/sas

Registration Requirement

The official registration deadline for this course is January 25, 2021. University policy requires all students to be officially registered in each class they are attending. Students who are not officially

registered for a course by published deadlines should not be attending classes and will not receive credit or a grade for the course. Each student must confirm enrollment by checking his/her class schedule (using Student Tools in FlashLine) prior to the deadline indicated. Registration errors must be corrected prior to the deadline.

Withdrawal

The last day to drop without a grade of "W" is February 1, 2021. The last day to withdraw this course is March 29, 2021. Other important Registrar dates can be found at http://www.kent.edu/registrar/registrar-dates-term.

Tentative Schedule

The schedule is subject to change and will be updated at the course website (https://tsunghengtsai.github.io/compstat-S21.html), so please check it regularly.

Week 01, 01/18 - 01/22: Introduction

Class begins on January 20.

Topics:

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

Week 02, 01/25 - 01/29: Fundamentals of R

Topics:

- Basic data structures
- Indexing and iteration
- Functions

Week 03, 02/01 - 02/05: Data Visualization

Topics:

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

Week 04, 02/08 - 02/12: Data Transformation

Topics:

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

Week 05, 02/15 - 02/19: Tidy Data

Topics:

- Principles of tidy data
- Data tidying with `tidyr`

Week 06, 02/22 - 02/26: Exploratory Data Analysis

Topics:

- EDA as an iterative process with data visualization, transformation, and modeling
- Data wrangling

Week 07, 03/01 - 03/05: Monte Carlo Simulation

Topics:

- Probability and random variable
- Law of large number
- Random number generator
- Monte Carlo simulation

Week 08, 03/08 - 03/12: Statistical Inference

Topics:

- Sampling distribution
- Central limit theorem
- Bootstrapping

Week 09, 03/15 - 03/19: Midterm Exam

Exam due 11:59 p.m. on March 19 (handed out 8:00 a.m. on March 17)

Week 10, 03/22 - 03/26: Statistical Modeling

Topics:

- Regression
- Fitting models to data
- Evaluating models
- Working with large numbers of models

Week 11, 03/29 - 04/02: Statistical Prediction

Topics:

- Statistical goals: inference vs. prediction
- Workflow of predictive analysis
- Training and test sets
- Cross-validation

Week 12, 04/05 - 04/09: Unsupervised Analysis

Topics:

- Principal component analysis (PCA)
- Singular value decomposition (SVD)
- Clustering

Week 13, 04/12 - 04/16: Spring Break

No class

Week 14, 04/19 - 04/23: Expectation-Maximization (EM) Algorithm

Topics:

- Maximum likelihood estimation
- Optimization
- Expectation-maximization (EM) algorithm

Week 15, 04/26 - 04/30: Markov Chain Monte Carlo

Topics:

- Gibbs sampling
- Metropolis-Hastings methods

Week 16, 05/03 - 05/07: Final Presentation

Student presentations on final projects on May 3, 2021