STATS 503

Statistical Learning II: Multivariate Analysis

Winter 2024

Section 100

Instructor: Jeffrey Regier
Class time: Monday & Wednesday, 1:00 pm – 2:30 pm
Class location: 420 Central Campus Classroom Building (CCCB)
Course Canvas website: `https://umich.instructure.com/courses/670434`
Instructor email: regier@umich.edu (Please see the section below titled "How to get help" before emailing.)
Office hours: See Canvas.

Labs

Lab sections are held on Fridays beginning January 19, 2024 in 1360 East Hall. During labs, students will be taught programming skills by the GSIs and will work together on select homework problems.

Section 101
GSI: Gabriel Alfonso Patron Herrera
Section time: 11:30 am – 1 pm
GSI email: gapatron@umich.edu

Section 201
GSI: Gang Qiao
Section time: 1:00 pm – 2:30 pm
GSI email: qiaogang@umich.edu

GSI office hours are posted on Canvas. Please read the section below titled "How to get help" before emailing the GSIs.

Textbooks

G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor (2023). *An Introduction to Statistical Learning with Applications in Python.* Springer. `https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html`.

T. Hastie, R. Tibshirani, and 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.

Course description

This course covers methods for modern multivariate data analysis and statistical learning, including both their theoretical foundations and practical applications. Topics include principal component analysis and other dimension reduction techniques, classification (discriminant analysis, nearest neighbor classifiers, logistic regression, support vector machines, decision trees, ensemble methods, neural networks), clustering (k-means, hierarchical clustering, model-based methods, spectral clustering), graphical models, and some basics of reinforcement learning. The objective is to learn what

methods are available for modern multivariate data analysis, how to use them, and when they should and should not be applied.

Prerequisites

STATS 500 (Statistical Learning I: Regression) or equivalent.

Linear Algebra (at the level of MATH 214) and Theoretical Statistics (at the level of STATS 426).

Computing

We will use Python and Jupyter notebook throughout the course. The labs will show students Python code for the methods that are introduced in the lecture. For students without much previous experience in Python, a good reference is *Python for Everybody* by Charles Severance.

Grading

Course grades are based on homework (20%), the midterm exam (20%), the final exam (40%), and the group project (20%). I expect the distribution of final grades will be similar to what it has been for previous offerings of this course: 45% A/A+, 70% A-/A/A+, 98% B- and above.

Bonus points worth up to 2% will be awarded to students who participate verbally in class.

Homework

There will be weekly homework assignments consisting of data analysis, to be done in Python, 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. Any Python code submitted should run without errors.

Exams

The midterm exam will be administered in class on March 18, 2024. The final exam will be administered at the university-designated time: April 29, 2024 from 4 pm to 6 pm.

Both exams are closed book and do 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. Neither exam tests knowledge of Python functions, though either may require understanding Python output. A sample final exam will be released on Canvas beforehand. The final exam is cumulative.

If you have SSD-approved accommodations for exams, please submit your documentation at least two weeks in advance. There are 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. It will also include a prediction challenge that we administer through Kaggle. Students should form project groups before the midterm, ideally with three members, but two is also acceptable. Because

the final project is regarded as the final exam on the computational part of this course, cross-team collaboration is not allowed.

Academic integrity

Academic misconduct includes copying code from the web or from your fellow students, *providing code to your fellow students*, looking up solutions online, turning in assignments from other classes or previous iterations of this course, and hiring others to complete your work.

From the LSA Community Standards of Academic Integrity:

> Academic dishonesty may be understood as any action or attempted action that may result in creating an unfair academic advantage for oneself or an unfair academic advantage or disadvantage for any other member or members of the academic community. Conduct, without regard to motive, that violates the academic integrity and ethical standards of the College community cannot be tolerated.

See `https://lsa.umich.edu/lsa/academics/academic-integrity.html` for more information.

You are welcome to discuss homework with your classmates, but the work you submit must be yours and yours alone. There is one exception to the above policy: you may submit Python code supplied by either GSI during a lab section without attribution.

If you use external sources, you must cite and credit them.

Academic misconduct will be met with disciplinary action. Students who engage in academic misconduct will typically receive a failing grade in this course. Additional sanctions may be imposed by the college.

How to get help

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

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

For questions/concerns about grading, please first contact our GSI, either by attending GSI office hours or by email. If you wish to request a re-grading 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 our GSI's 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 email me to make an appointment.