STATS/DATASCI 451: Bayesian Data Analysis

Yixin Wang
University of Michigan

August 29, 2022

Overview

Description. The course is an introduction to both principles and practice of Bayesian inference for data analysis.

We will focus on building probabilistic models, algorithms for approximate Bayesian inference, and methods for checking, criticizing, and revising models. Some of the models we will study include classic Bayesian mixture and regression models, hierarchical models, factor models, topic models, and deep generative models. Alongside these models, we will study algorithms for approximate Bayesian inference including Markov Chain Monte Carlo and variational inference algorithms. Finally, we will discuss methods for checking, criticizing, and revising models in an iterative manner, completing a virtuous cycle of applied Bayesian statistics.

At the end of this course, students will be familiar with the Bayesian paradigm, and will be able to analyze different classes of statistical models. The course gives an introduction to the computational tools needed for Bayesian data analysis and develops statistical modeling skills through a hands-on data analysis approach.

Prerequisites. The prerequisites are: (STATS 412: Introduction to Probability & Statistics or STATS/MATH 425: Introduction to Probability) and (STATS/DATASCI 306: Introduction to Statistical Computing or EECS 280: Programming and Intro Data Structures).

Textbooks. Class readings will mainly come from the following books.

- Bishop. Pattern Recognition and Machine Learning. New York: Springer, 2006. PDF
- Murphy. Probabilistic Machine Learning: Advanced Topics. MIT Press, 2023. PDF
- Gelman et al. Bayesian Data Analysis. Chapman and Hall, 2005. PDF
- McElreath. Statistical Rethinking: A Bayesian Course with Examples in R and STAN. Chapman and Hall/CRC, 2020. PDF

Piazza and participation bonus. All communications with the teaching team should be conducted on piazza; please do not email. If you wish to ask a question privately to the teaching team, please post a private note on Piazza; see instructions here; you can expect an answer within 24 hours during weekdays (except holidays). The GSIs and the instructor will be monitoring piazza, endorsing correct student answers, and answering questions that remain after a discussion.

As a bonus, **up to 3 percentage points will be added to your final course grade based on piazza participation.** You will get $x \cdot 3\%$ bonus points if the number of your total Piazza contributions lies in the top $(x \cdot 100)$-th quantile among all students. The number of Piazza contributions will be determined by Piazza class statistics.

Requirements and Grades

The requirements are just-in-time teaching (JiTT) questions (5%), weekly quizzes (24%), weekly homework assignments (50%), and one final project (21%).

1. **Just-in-time teaching (JiTT) questions (5%).** Before each class, you will answer short just-in-time teaching (JiTT) questions that prepare you for the class. The JiTTs test background knowledge, check that you've done the required reading, and give you a chance for feedback. They are required. To get a perfect grade on the JiTTs you do not need to get the right answers but you need to take them seriously.

   The JiTT questions are due two hours before each class. Late submissions are not accepted.

2. **Weekly quizzes (24%).** Every week, we will hold a 10-minute quiz on Gradescope.

   Only your top eight quiz scores will be counted; this policy is expected to accommodate circumstances where students could not complete the quizzes due to the add/drop period, registration matters, and/or personal reasons.

   The quizzes are released on Monday evening and are due at 11:59 pm EST each Wednesday.

3. **Weekly homework assignments (50%).** There will be weekly homework assignments involving problems, programming, and data analysis. We encourage you to prepare all written work using the LaTeX templates we provide.

   Only your top ten homework scores will be counted; this policy is expected to accommodate circumstances where students could not complete the homework assignments due to the add/drop period, registration matters, and/or personal reasons.

   The homework assignments are due at 11:59 pm EST each Monday.

   Submission requirements. Homework will be submitted electronically as pdfs, along with any notebook or markdown used to generate results appearing in the pdf. You must run all cells in your notebook to receive credit; we will not rerun your notebook. Any code submitted should run without errors. Note that the homework assignments may involve coding up the model and algorithm and applying it to a given dataset. You can code in Julia, Python, or R (i.e. as long as it runs in a Jupyter notebook).

   Homeworks should be written up clearly and succinctly; you may lose points if your answers are unclear or unnecessarily complicated.

Late days. Homework due dates are strict, and you may turn in work late only with the use of "late days." *You have seven late days to use over the course of the semester.* For each late day you spend, you extend the deadline for homework by 24 hours. You may spend multiple late days per homework. Once you have turned in your homework you may not spend more late days to turn in your homework again.

Once you run out of late days, you will incur a 25% penalty for each extra late day you use. Each late homework should be clearly marked as "Late" on the first page.

The purpose of this late-day policy is to enable you to deal with unexpected circumstances (e.g., illness, family emergencies, job interviews) without having to come to me. If dire circumstances arise (e.g., long-term illness that causes you to miss multiple weeks of lectures), please contact me as soon as possible. Due to the university grading schedule, you may not use late days to extend the deadline of the last homework assignment.

Regrade Policy. You may submit a regrade request if you believe that the course staff made an error in grading. Any regrade requests should be submitted through Gradescope within ten days of receiving your grade. Please try to be as specific as possible with your regrade request.

4. **Final project (21%).** The final project is an opportunity to use and develop Bayesian models to analyze real-world data.

   Proposal (1%). The project proposal is an abstract that imagines the completed project. Start thinking about what datasets you'd like to study and what questions you'd like to answer early! These will inform your choices about modeling tools. As a forcing function, part of your grade will be based on your proposal.

   <u>The project proposal is due at 11:59 PM EST on Oct 3.</u>

   Milestone (4%). The project milestone describes the problem you are addressing and discusses some preliminary results. Include what you have completed and what you plan to finish by the end of the semester. By this point, you should have a pretty clear idea about the dataset and question, and some initial thoughts about the types of models you will explore and experiments you will run.

   <u>The project milestone is due at 11:59 PM EST on Nov 7.</u>

   Report (16%). The final report will present your theoretical work and experimental results.

   <u>The project report is due at 11:59 PM EST on Dec 12.</u> No late days are allowed for deadlines related to the course project.

   We grade your project proposal and report on both content and writing quality. Please prepare all written work using the LaTeX templates we provide.

5. **Final letter grade.** The final grade will be set so that the distribution of final grades approximately matches that of previous offerings of the course. That said, your final grade is

guaranteed to be the same as or better than the following assignment mechanism of letter grades:

[95, 100]: A+;

[90, 95): A;

[85, 90): A-;

[80, 85): B+;

[70, 80): B;

[60, 70): B-;

Schedule

The course is organized around "Box's Loop," a conceptual framework for approaching applied data analysis (Blei, 2014). It's an iterative process of modeling, inference, criticism, and refinement. We'll take multiple "laps" around this loop throughout the quarter, each time introducing new probabilistic models, inference algorithms, and model checking procedures. (The schedule is subject to change.)

Introduction

1. Introduction and the "Box's Loop"
2. Probability: A Review of Basic concepts and Bayes' Theorem
3. The ingredients of probabilistic models I
4. The ingredients of probabilistic models II
5. The Exchangeable Data Model and Conjugate Priors I
6. The Exchangeable Data Model and Conjugate Priors II
7. Evaluating Probabilistic Models I
8. Evaluating Probabilistic Models II
9. Conditional Models: Linear and Logistic Regression I
10. Conditional Models: Linear and Logistic Regression II
11. Conditional Models: Linear and Logistic Regression III
12. Bayesian Mixture Models and an Introduction to Markov Chain Monte Carlo I
13. Bayesian Mixture Models and an Introduction to Markov Chain Monte Carlo II
14. Bayesian Mixture Models and an Introduction to Markov Chain Monte Carlo III
15. Bayesian Mixture Models and an Introduction to Markov Chain Monte Carlo IV
16. Mixed-Membership Models and an Introduction to Variational Inference I
17. Mixed-Membership Models and an Introduction to Variational Inference II
18. Mixed-Membership Models and an Introduction to Variational Inference III
19. Matrix Factorization Models I
20. Matrix Factorization Models II
21. Matrix Factorization Models III
22. The Exponential Family and Generalized Linear Models I
23. The Exponential Family and Generalized Linear Models II
24. The Exponential Family and Generalized Linear Models III
25. Deep Generative Models and Black Box Variational Inference I
26. Deep Generative Models and Black Box Variational Inference II
27. Deep Generative Models and Black Box Variational Inference III
28. Summary (and wiggle room)

Support Resources

Course recordings. Course lectures may be audio/video recorded and made available to other students in this course. As part of your participation in this course, you may be recorded. If you do not wish to be recorded, please contact the instructor during the first week of class (or as soon as you enroll in the course, whichever is latest) to discuss alternative arrangements.

Academic integrity. The University of Michigan community functions best when its members treat one another with honesty, fairness, respect, and trust. The college promotes the assumption of personal responsibility and integrity and prohibits all forms of academic dishonesty and misconduct. All cases of academic misconduct will be referred to the LSA Office of the Assistant Dean for Undergraduate Education. Being found responsible for academic misconduct will usually result in a grade sanction, in addition to any sanction from the college. For more information, including examples of behaviors that are considered academic misconduct and potential sanctions, please 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 that you turn in must be yours and yours alone, and you must disclose the names of those you spoke with in your homework, including both classmates and others outside the class. This disclosure applies whether a student has helped someone else or has received help. However, it is not necessary to disclose any discussion you have with the course instructor or the course GSIs.

Accommodations for students with disabilities. The University of Michigan recognizes disability as an integral part of diversity and is committed to creating an inclusive and equitable educational environment for students with disabilities. Students who are experiencing a disability-related barrier should contact Services for Students with Disabilities https://ssd.umich.edu/; 734-763-3000 or ssdoffice@umich.edu). For students who are connected with SSD, accommodation requests can be made in Accommodate. If you have any questions or concerns please contact your SSD Coordinator or visit SSD's Current Student webpage. SSD considers aspects of the course design, course learning objects and the individual academic and course barriers experienced by the student. Further conversation with SSD, instructors, and the student may be warranted to ensure an accessible course experience.

Mental Health and Well-Being. University Students may experience stressors that can impact both their academic experience and their personal well-being. These may include academic pressures and challenges associated with relationships, mental health, alcohol or other drugs, identities, finances, etc. If you are experiencing concerns, seeking help is a courageous thing to do for yourself and those who care about you. If the source of your stressors is academic, please contact me so that we can find solutions together. For personal concerns, U-M offers a variety of resources, many of which are listed on the Resources for Student Well-being webpage. You can also search for additional well-being resources here.

Acknowledgments.

The course materials are adapted from the related courses offered by David Blei, Yang Chen, Andrew Gelman, and Scott Linderman.

References

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application.*