Statistics 238: Bayesian Statistics

September 23, 2016

Catalog blurb

Bayesian methods and concepts: conditional probability, one-parameter and multiparameter models, prior distributions, hierarchical and multi-level models, predictive checking and sensitivity analysis, model selection, linear and generalized linear models, multiple testing and high-dimensional data, mixtures, nonparametric methods. Case studies of applied modeling. In-depth computational implementation using Markov chain Monte Carlo and other techniques. Basic theory for Bayesian methods and decision theory. The selection of topics may vary from year to year.

Course details

- Title: Bayesian Statistics
- Units: 3 units
- Format: Lec MW 9:10-10:30, Lab F 12:10 pm - 1 pm
- Semester: Fall 2016

Course Description

Bayes' theorem is a trivial identity in probability theory, but it has profound consequences for the theory and practice of statistical inference and in decision theory. The course will introduce you to the Bayesian approach and philosophy and give you experience in applying the Bayesian approach to statistical inference. During the first two-thirds of the course, we will focus on the core of Bayesian statistics, with topics including Bayes' theorem, general principles (likelihood, exchangeability, de Finetti's theorem), prior distributions, simple models, hierarchical models, methods of inference (exact, approximations, Monte Carlo strategies), model diagnostics and model selection. The second part of the course will introduce the Bayesian approach to a range of important statistical models and situations including GLMs and GLMMs, high-dimensional data and multiple testing, meta-analysis, nonparametrics, missing data, and causal inference.

The course is focused on concepts and methodology; some limited theory will be covered and I will present some of the material in the context of real data, with an occasional case study. Bayesian inference relies on the posterior distribution, derived through integration, which can be challenging in complicated (and even simple) models. Since 1990, advances in computational techniques have allowed statisticians to simulate from the posterior distribution; Markov chain Monte Carlo techniques are now a standard tool for fitting Bayesian models, and other techniques such as variational Bayes and Laplace approximation have gained wide use recently. As a result, the course will be heavily computational and there will be extensive discussion of computation, as well as a heavy dose of computing in the problem sets. Students are expected to be familiar with R, numerical Python, Julia, or Matlab, and will be introduced to specific software used for Bayesian computation as well.

An understanding of probability and statistics at the level of Statistics 201A and 201B (for example the Casella and Berger *Statistical Inference* text or the Wasserman *All of Statistics* text) is required; waivers from this requirement will require discussion with the instructor.

Objectives of the course

The goals of the course are that, by the end of the course, students be able to:

- understand and describe the Bayesian perspective and its advantages and disadvantages compared to classical methods;
- read and discuss Bayesian methods in the literature;
- select and build appropriate Bayesian models for data to answer research questions;
- implement Bayesian models and interpret the results;
- develop Bayesian models/methods for new types of data.

Course material

- Course textbook
  - Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. Bayesian Data Analysis, Third Edition. Chapman & Hall, ISBN: 978-1439840955.
  - The book is good and covers essentially all of the material I want to cover in the course. Given this and with the goal of having class periods focused on working through specific examples/problems/derivations, **I will expect you to do assigned reading in advance of many class periods**. Problem sets and the exam will include material I do not cover directly in class. I encourage questions about the reading, and we will discuss those questions in class.
- Other useful references are:
  - Carlin, Brad and Thomas Louis. 2009. Bayesian Methods for Data Analysis, Third Edition. Chapman & Hall/CRC Press, ISBN: 978-1584886976.
  - O'Hagan, Anthony. 1994. Kendall's Advanced Theory of Statistics: Bayesian Inference. Volume 2b. Edward Arnold Press.
  - Robert, Christian and George Casella. 2004. Monte Carlo Statistical Methods, Second Edition. Springer., ISBN: 978-0387212395. Electronic book: http://dx.doi.org/10.1007/978-1-4757-4145-2
- All books except Robert & Casella (available electronically) will be on reserve in the library, including the primary text (the latter for two hour periods).

Course requirements and grading

Assignments	**Time**	**% of total grade**
Problem sets	~5 throughout course	**25%**
In-class participation	ongoing, including attendance and questions on the readings	**10%**
Mid-term Exam		**35%**
Project		**30%**

Problem sets Students can expect roughly five problem sets during the semester, approximately every 2-3 weeks.

Much of statistical work is collaborative, therefore you may discuss your solutions with other students; in fact I encourage this. While sharing ideas on how to demonstrate something analytically or on computer syntax is allowed, you should make a full effort on your own initially, before discussing matters on which you are stuck or comparing methods. Your final writeup, including computer code (submitted as an appendix), MUST be your own, and no material may be copied from another student. I.e., you can talk about problems and work together at a board, but when you write up your solutions and code you must do it individually.

Problem sets are due on paper at the start of class on the due date. Assignments handed in later than the due time will be subjected to a partial reduction of the final score for each day (or part thereof) that it is late. No credit will be given after solutions are handed out. I may also require electronic submission in some cases.

Questions about problem sets should first be addressed by submission to the Piazza discussion board rather than emailing me directly.

The grading scheme for problem sets is as follows. In some cases I will use 1.5 and 2.5 to further refine the grade.

- 3: complete solutions with only minor errors
- 2: complete solutions but with serious errors
- 1: partial solutions

Labs Labs are an opportunity to work through problems and derivations with classmates and help from the GSI. In many cases you will be required to submit your lab problem solution as part of a problem set. In these cases (and only these cases), you are allowed to submit a group answer. Please talk with me if you have a conflict with some or all of the lab periods as we may be able to make some accommodations.

Readings and class participation My goal is to have classes be an interactive environment. Accordingly, part of your grade will be based on participation. I encourage you to ask questions (about topics I discuss in class and topics from the readings) and will pose questions to the class to think about and discuss. To increase time for discussion and assimilation of the material in class, and to decrease lecture time, **I will expect you to read chapters of the text in advance** and may have you respond to questions about the material in advance of class.

I may also implement a system for calling on students if participation is not sufficiently robust. This would likely involve designating a "panel" of 3-4 (random) students at the start of class who I have the option to call on during class to answer questions.

Exam The exam will be about 2/3 of the way through the semester and will include both an in-class portion and a take-home portion. The exam will cover the core material on Bayesian statistics covered in the first two-thirds of the course.

Project In lieu of a final exam, and in recognition that Bayesian methods are best learned by doing and of the computational focus of Bayesian methods, each student will carry out a final project. This can be 1.) development and implementation of a Bayesian analysis of their own data or data obtained elsewhere, 2.) implementation of a method/algorithm in the literature, possibly within the NIMBLE platform, or possibly as an R or Python package, 3.) contribution to a project involved in surveying and assessing the use of

simulation/bootstrapping/Bayesian computation in the scientific literature, or 4.) (possibly, depending on available time) in-depth presentation of a topic to the class. I will ask for project proposals midway through the term so that I can make sure the content of the project is acceptable for the course.

Coding and software

I will present demonstrations and template code using R, as well as using MCMC in a new software platform called NIMBLE (r-nimble.org), of which I am one of the developers. However you are welcome to use other languages in labs, problem sets, the exam, and the project, of which Python and Julia are good choices. If you'd like to use another language (e.g., MATLAB, Scala, Java, C++), please check with me first.

Course materials

Materials including the syllabus, problem set assignments, and lab assignments will be available through the course Github repository: https://github.com/berkeley-stat238/stat238-fall-2016. You're not required to use Git but it's an important skill to have. If you'd like to learn more about it, one resource is a tutorial that Jarrod Millman and I developed.

If you're not (yet) familiar with Git, you can simply download materials from https://github.com/berkeley-stat238/stat238-fall-2016/archive/master.zip.

Logistics

- Accomodations for Students with Disabilities: Please see me as soon as possible if you need particular accommodations, and we will work out the necessary arrangements.
- Scheduling conflicts: Please notify me by email by the third week of the term about any known or potential extracurricular conflicts (such as religious observances, graduate or medical school interviews, or team activities) that will affect (1) attendance for more than a week, (2) problem set submission, or (3) the exam. I will try to help you with making accommodations but cannot promise them.

(Approximate) Schedule of topics

- Introduction to Bayesian statistics (1 week)
- Prior distributions (1 week)
- Simple models (1.5 weeks)
- Bayesian asymptotics (0.5 week)
- Hierarchical, random effects, and multi-level modeling (1 week)
- Bayesian computation (3 weeks)
- Model assessment and comparison (1 week)
- Decision theory and hypothesis testing (0.5 week)
- Design and data collection considerations (0.5 week)
- Regression modeling (1 week)
- Multiple testing and high-dimensional data (1 week)

- Missing data and causal inference (1 week)
- Bayesian nonparametrics (1.5 weeks)

Academic integrity

The student community at UC Berkeley has adopted the following Honor Code: "As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." My expectation is that you will adhere to this code. Below I outline a few specifics related to academic integrity.

Collaboration and Independence: I encourage you to review lecture and reading materials and studying for the exam together. With regard to labs and problem sets, please see the guidance in the above sections. Anyone found to be copying problem set solutions (either from another student or providing solutions to another student) will receive a zero on the problem set and may face additional disciplinary action.

Cheating and plagiarism: Anyone caught cheating on the exam will receive a failing grade on the exam and will also be reported to the University Office of Student Conduct.