ORIE6217/CS6384: Applied Bayesian Data Analysis for Research

Lecture 1: Introduction

Nikhil Garg

Plan for today

- Content overview
- Syllabus/Course structure
- Questions
- Hopefully end a few minutes early for specific questions

Please interrupt with questions at anytime

(but raise your hand)

Who am I?

Instructor: Nikhil Garg

Asst Professor, Cornell Tech, ORIE

Research on the application of algorithms, data science, and mechanism design to the study of democracy, markets, & societal systems

Past experiences/collabs: Uber, Upwork, other marketplaces, campaign data science, NYC Parks Department

What is Bayesian data analysis?

Bayes rule + Bayesian inference

$$P(\theta|x) = \frac{P(x|\theta)\,P(\theta)}{P(x)}$$

θ: parameters to learn about the world

x: data we have

$P(x|\theta)$: our "model" for how the world works "data generating process". how data is created **given** the world's parameters.

$P(\theta|x)$: what the data + model tell us about the world.

So why Bayesian data analysis?

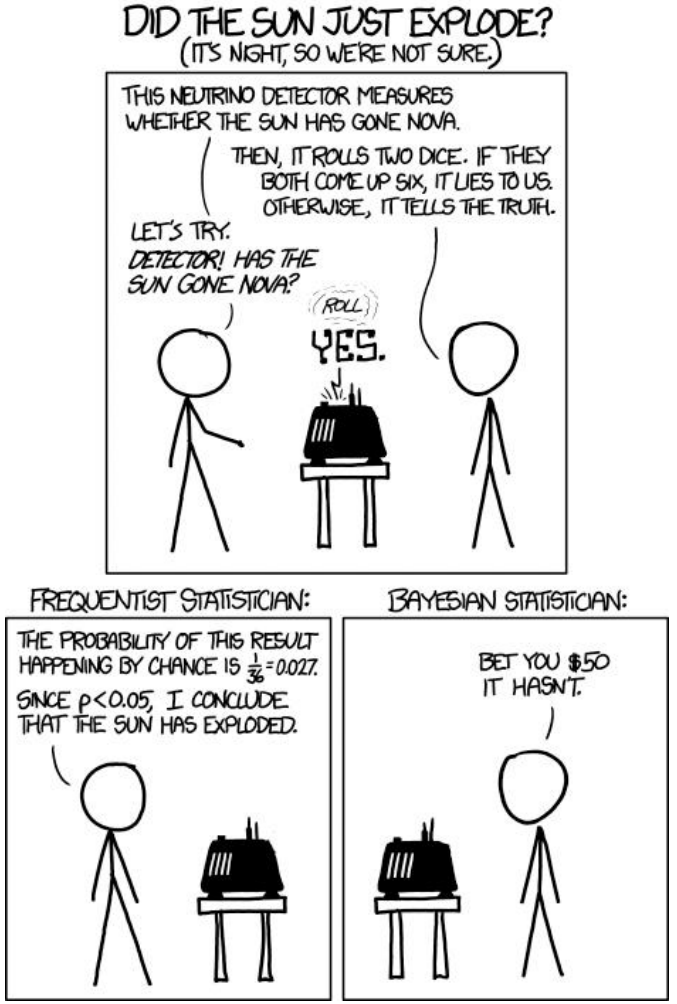
$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

If you ask a Bayesian statistician:

- Properly can incorporate “priors” (“expert information”) – what we know about the world
- Properly measure “uncertainty” that we have based on the data that is available.
  - Standard machine learning just gives us point estimate $\hat{\theta}$

If you also ask me:

It’s convenient for modern computational social science research, in a complicated world



Why BDA for CSS research?

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

- The world is a complicated place, and there are many things that cause modern machine learning paradigm a lot of heart-ache

  Data censoring, missing data, distribution shifts, spatio-temporal heterogeneity, strategic behavior

- These are complicated data generating processes (DGPs) $P(x|\theta)$
- Bayesian data analysis gives us:
  - A *statistical language* to incorporate models of complicated DGPs
  - A general *algorithmic* framework to calculate posteriors $P(\theta|x)$
  - Modern *programming languages* (Stan, Pyro, others) to actually fit models given data

Many applications of bayesian data analysis

Research + Industry

Public health, political science, ecology, biology, criminal justice, computational social science more generally

Content overview

A successful data analysis project has several components

- Careful understanding of the world and available data
- Careful description of your research question (“what is your estimand”)
- Translation of world/limitations to a statistical *model*

  “All models are wrong”, “The map is not the territory”

  Why did the authors include/exclude certain things in their model? What would change if they made different choices?
- Fitting the model with available data. Validating the model fit
- Communicating findings

Course components

Part 1: Basics of Bayesian Data analysis [4-5 weeks]

- Bayesian statistics primer
- Writing models in Stan (optional replacement: Pyro)
- Basics of fitting algorithms
- Examples of models (single-parameter models, regression, hierarchy, etc)
- Bayesian “workflow”: diagnosing model fit, identifiability, comparing models, etc
- Assignment: Homework assignment where you write, code, fit models
  - Peer review of another person’s assignment

Part 2: How have others used Bayesian methods? [4-5 weeks]

Part 3: Advanced topics [~2 weeks]

Class project [~2 weeks of classtime]

Course components

Part 1: Basics of Bayesian Data analysis [4-5 weeks]

Part 2: How have others used Bayesian methods? [4-5 weeks]

- Reading + discussing papers that apply Bayesian methods
- Assignment (in teams)
  - Give a (75 minutes) presentation on a paper to the class
    - Present the paper, focusing on statistical model + data
    - Ideally go through model code
    - Lead class discussion on the paper
  - Submit a report
    - Written summary and review of the paper
    - Simulate data from the paper's data generating process
    - Write + fit a Stan/Pyro model reflecting the paper's model
  - Read paper for 1 other team, participate in discussion, write peer review

Part 3: Advanced topics [~2 weeks]

Class project [~2 weeks of classtime]

Course components

Part 1: Basics of Bayesian Data analysis [4-5 weeks]

Part 2: How have others used Bayesian methods? [4-5 weeks]

Part 3: Advanced topics [~2 weeks, as time permitting]

- Advanced models: Gaussian processes/dynamic models, HMMs and Spatial models, conditional random fields, nonlinear and nonparametric models
- Sampling algorithms (Gibbs, Metropolis Hastings, HMC, Variational Inference)
- Computational concerns (parallelism, speeding up models, GPU usage, etc)

Class project [~2 weeks of classtime]

Course components

Part 1: Basics of Bayesian Data analysis [4-5 weeks]

Part 2: How have others used Bayesian methods? [4-5 weeks]

Part 3: Advanced topics [~2 weeks]

Class project [~2 weeks of classtime]

- Conduct your own (mini) research project using Bayesian methods. Define a research question, hypothesize DGP, write statistical model, program + fit it in Stan/Pyro/other such model
- Component 1: Project proposal + presentation
  - Component 1b: Peer review of someone else's proposal
- Component 2: Final report + presentation

Who is this course for?

PhD students across ORIE/CS/IS/Econ/CAM/Business/Statistics

The objectives of the class are: tldr – learn to do applied research using Bayesian data analysis techniques

- Start with a research question and estimand of interest, and (1) construct a data generating process for the setting, and (2) construct a Bayesian model reflecting that process.
- Determine whether (and in what cases) the estimand of interest is identifiable.
- Write down the model in a Bayesian programming language such as Stan and/or Pyro.
- Fit and evaluate model fit using data.

What is this class not?

This is not a Bayesian *theory* class

- David Ruppert is teaching ORIE 6780 this semester
- The courses are complementary
  - Some introductory material first few weeks will overlap

This is not an introduction to probability/statistics class, or an introduction to Python class

- I'm going to assume some level of familiarity
- I expect you to be able to pick things up on your own

Pre-requisites

Concepts [level of undergraduate probability/machine learning]

- Basic terms of probability theory
  - probability, probability density, distribution
  - sum, product rule, and Bayes' rule
  - expectation, mean, variance, median
- Some algebra and calculus
- R or **Python**
  - Data manipulation
  - Basic visualization : histogram, density plot, scatter plot

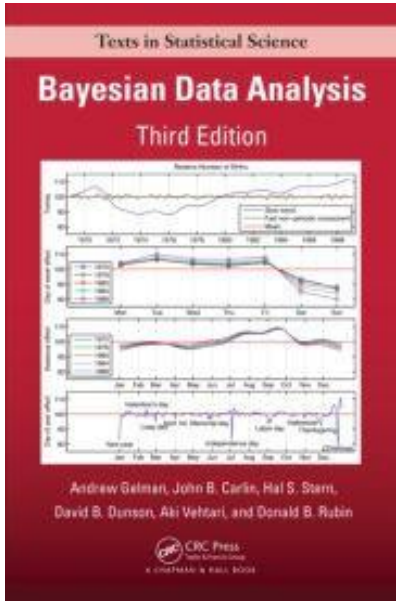
This is a PhD class…prerequisites not enforced but are your responsibility.

- Please ask questions in class!
- Ask for resources to learn material

Resource note: Adapting introductory lessons from following book/course

Bayesian data analysis

- Book: Gelman, Carlin, Stern, Dunson, Vehtari & Rubin: Bayesian Data Analysis, Third Edition. (online pdf available)
- The course website has more detailed information than these slides
  https://avehtari.github.io/BDA_course_Aalto/Aalto2022.html



Syllabus

ORIE6217 Syllabus -- Spring 2023 - Google Docs

Assignments + Grading

Homework assignment (1 assignment) [10%]

Research paper presentation + discussion lead (in teams) [30%]

Course project (in teams) [40%]

Participation [20%]

Course communication

Course Slack channel: First place for any question/comment

Office hours: Happy to chat about anything – sign up on link in syllabus

Email:

- Private issues
- Otherwise try to avoid; but preferred over private message on Slack.

Classroom norms

- Take space, make space: allow others to join the conversation, but please contribute as you feel comfortable.
- Embrace a growth mindset. Not understanding something in a paper is the default.
- Ask questions!
- Be willing to give and receive feedback respectfully.
- Zoom norms

  Feel free to take video-off breaks as necessary, and a couple lectures of video off the entire time. But I expect you to mostly keep video on and participate.

Announcements

- Watch out for the course pre-survey, posted on Slack soon

Questions?