Course Introduction

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

January 13, 2021

Course Overview

Bayesian Modeling

- Represent our knowledge in the form of a *probability distribution*; probabilities measure confidence or belief (defined over one or more parameters
- Establish a model indicating the dependence of observations on parameters (and the dependence of those parameters on one another) – generative model – could simulate outcomes
- Use Bayes' theorem to update the distribution(s) of the model parameter(s) based on data

Inference

- Estimate values of the model parameters (point estimates, interval estimates, etc.) expected value analogue of conf. intervals.
- Predict future observations

Topics

We'll cover the following topics, in something like this order:

- Foundations: probability theory and the Bayesian interpretation
- Simple one- and multi-parameter models
- Hierarchical and graphical models
- Model checking and evaluation
- Computational methods:
  - Exact inference with belief propagation
  - Approximate inference with Markov chain Monte Carlo (MCMC)
- Mixture models and expectation-maximization algorithms
- Inference for dynamical systems using Kalman filters
- Other topics as time allows

expectation propagation

Gaussian processes, Dirichlet processes.

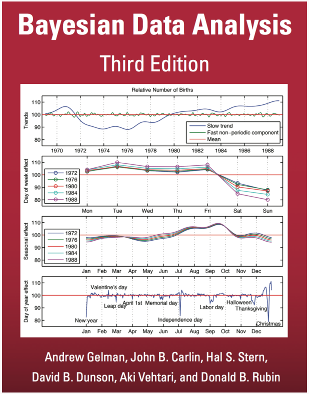
Prerequisites

What you need to know:

- Calculus (I and II are fine – we'll do some multi-variable calculus but no vector calculus)
- Linear algebra (at the level of an intro course)
- Probability and statistics
- Programming – preferably Python and preferably at least a year

Textbook

Bayesian Data Analysis, 3rd ed., by Gelman et al.



http://www.stat.columbia.edu/~gelman/book/

Free PDF here

Other sources as needed

Hard copy ~ $70

Software requirements

In order to complete the course, especially the programming components, you'll need the following software on your computer:

- Zoom (videoconference software).
- Python, with the following packages: NumPy, SciPy, matplotlib, **PyMC3**, and Jupyter (notebook app or JupyterLab)

Anaconda has all except PyMC3

(PyMC3 is highlighted because it's the one you're least likely to already have!)

Expectations

- Attend videoconference class meetings
- Prepare for meetings by completing assigned readings, attempting exercises
- Complete several homework assignments over the course of the semester 8 ± 2
- Complete a take-home midterm and final

Grading

Category	Weight
Homework Assignments	50%
In-class work, participation, etc.	15%
Midterm and Final	35%

15% midterm

20% final

Instructor and contact information

Dylan Murphy, Ph.D.

Lecturer in the iSchool

Please call me: Dylan or Dr. Murphy

Office: Harvill 444 (but I won't be there)

djmurphy@arizona.edu

Office hours: Th 1-3 PM (drop-in office hours on Zoom) or by ~~appointment~~ subject to change.

The Slack channel for this course will be available through a link on the D2L site.

A first example

Flipping a coin

A researcher is trying to determine whether a coin is fair (i.e., $p_H = 0.5$); she suspects it may be less likely to come up heads.

$p_H < 0.5$

The data consist of a history of 12 flips:

$$T, T, H, T, T, T, H, T, T, T, T, H$$

What do you do to reach a conclusion about p_H? Let's start by thinking from the classical (frequentist) statistical perspective.

p_H – frequency of heads in limit of infinitely many flips.

- Sample proportion $\hat{p}_H = \frac{3}{12} = 0.25$
- Idea: there is an underlying probability of heads, p_H

χ^2-test
z-test
equivalent

Flipping a coin

Idea:

- The result is drawn from a population of theoretically possible experiments, where we flip the coin $n = 12$ times and get r heads
- Under the assumption that $p_H = 0.5$ (called the null hypothesis), we can calculate the probability that $r \leq 3$
- If this probability (called the p-value) is small (conventionally, $p < 0.05$), then we conclude the null is inconsistent with the data

compute $\underbrace{\text{distribution of \# heads under the null}}_{\text{sampling distribution}}$

Flipping a coin

- flips are independent trials $\rightarrow$ r has a binomial distribution

Flipping a coin

- flips are independent trials $\rightarrow$ r has a binomial distribution
- If $r \sim \mathrm{Binomial}(\underbrace{\boxed{n}}_{\text{trials}}, \overbrace{\boxed{p_H}}^{\text{success prob.}})$, then

$$P(r = k) = \underbrace{\begin{pmatrix} n \\ k \end{pmatrix}}_{\text{\# of arrangements of k heads in n trials}} \underbrace{p_H^k}_{\text{prob of k heads}} \underbrace{(1 - p_H)^{n-k}}_{\text{prob. of n-k tails}}$$

Flipping a coin

- flips are independent trials $\rightarrow$ r has a binomial distribution
- If $r \sim \mathrm{Binomial}(n, p_H)$, then

$$P(r = k) = \begin{pmatrix} n \\ k \end{pmatrix} p_H^k (1 - p_H)^{1-k}$$

- So,

$$\left.\begin{aligned} P(r \leq k | p_H = 0.5) &= 0.5^{12} \left(\begin{pmatrix} n \\ 0 \end{pmatrix} + \begin{pmatrix} n \\ 1 \end{pmatrix} \right. \\ &\quad \left. + \begin{pmatrix} n \\ 2 \end{pmatrix} + \begin{pmatrix} n \\ 3 \end{pmatrix} \right) \end{aligned}\right\} \text{one sided alternative}$$

$$\approx 0.07$$

Flipping a coin

- flips are independent trials $\rightarrow$ r has a binomial distribution
- If $r \sim \mathrm{Binomial}(n, p_H)$, then

$$P(r = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p_H^k (1 - p_H)^{1-k}$$

- So,

$$\begin{aligned} P(r \leq k | p_H = 0.5) &= 0.5^{12} \left(\left(\begin{array}{c} n \\ 0 \end{array} \right) + \left(\begin{array}{c} n \\ 1 \end{array} \right) \right. \\ &\quad \left. + \left(\begin{array}{c} n \\ 2 \end{array} \right) + \left(\begin{array}{c} n \\ 3 \end{array} \right) \right) \\ &\approx 0.07 \end{aligned}$$

So we announce to our colleague that there is not statistically significant evidence of bias... right?

Flipping a coin

Our colleague comes back and examines the calculation. She says: "I'm sorry, but you've misunderstood my experimental design. r is not a random variable; I decided ahead of time that I would flip the coin until I got three heads, so n is the random variable.

What's different?

Flipping a coin

Our colleague comes back and examines the calculation. She says: "I'm sorry, but you've misunderstood my experimental design. r is not a random variable; I decided ahead of time that I would flip the coin until I got three heads, so n is the random variable.

What's different?

- The universe of possible experiments is different
- Now, the probability of n flips and r heads is given by

$$P(n|p_H = \overset{0.5}{\cancel{0.05}}) = \underbrace{\binom{n-1}{r-1}}_{\text{\# of arrangements of all but the last flip}} 0.5^n$$

- So,

$$P(n \geq 12|r = 3, p_H = 0.5) = \sum_{k=12}^{\infty} \binom{k-1}{r-1} 0.5^k \approx 0.03$$

Flipping a coin

Our colleague comes back and examines the calculation. She says: "I'm sorry, but you've misunderstood my experimental design. r is not a random variable; I decided ahead of time that I would flip the coin until I got three heads, so n is the random variable.

What's different?

- The universe of possible experiments is different
- Now, the probability of n flips and r heads is given by

$$P(n|p_H = 0.05) = \begin{pmatrix} n-1 \\ r-1 \end{pmatrix} 0.5^n$$

- So,

$$P(n \geq 12 | r = 3, p_H = 0.5) = \sum_{k=12}^{\infty} \begin{pmatrix} k-1 \\ r-1 \end{pmatrix} 0.5^k \approx 0.03$$

So, there is statistically significant evidence of bias... right?

Flipping a coin

My claim:

- This should make you uneasy
- There are two major problems:
  - The result of the analysis is dependent on a stopping rule outside of the data-generating process
  - The endpoint of the analysis is an artificial dichotomy between $p_H = 0.5, p_H \neq 0.5$

Flipping a coin

My claim:

- This should make you uneasy
- There are two major problems:
    - The result of the analysis is dependent on a stopping rule outside of the data-generating process
    - The endpoint of the analysis is an artificial dichotomy between $p_H = 0.5, p_H \neq 0.5$
- Both points can be addressed by the Bayesian approach

Thinking like a Bayesian

The Bayesian perspective:

- Treat p_H as if it were a random variable:
  - p_H has a probability distribution supported on $[0, 1]$
  - $\underbrace{Pr(a < p_H < b)}$ represents our belief that $a < p_H < b$
- Bayes' rule says that if we observe an outcome $y \in \{H, T\}$:

$$p(p_H|y) \underbrace{\boxed{\propto}}_{\text{proportional to}} p(p_H) \times p(y|p_H)$$

so we can update $p(p_H)$ each time we observe a flip.

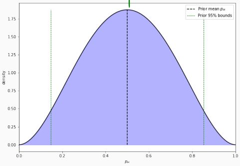
$$p(p_H | y) = \frac{p(p_H)\, p(y|p_H)}{p(y)}$$

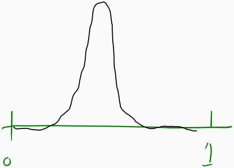
normalizing constant

Starting probability distribution for p_H:

prior distribution

0.5

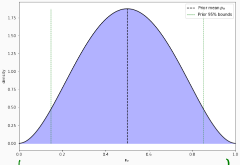




Prior says: p_H could be any value between 0 and 1,
but most likely is closer to 1/2

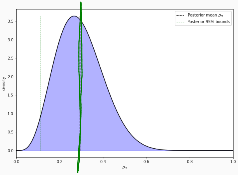
Thinking like a Bayesian

Starting probability distribution for p_H:



After updating:

posterior distribution of p_H



expected value

posterior distribution contains all we know about the coin after observing the data.

Assumptions and "objectivity"

A common criticism of Bayesian statistics is its lack of objectivity:

- We had to begin with a probability distribution
- This encodes information that has nothing to do with the observed data

This example shows that the requirement of assumptions is not unique to the Bayesian approach – we just made it more explicit.

For next time

Main task for the first week: get your software environment set up:

- I strongly recommend using the Anaconda distribution of Python
- `conda install pymc3`; then, if running Windows, you may need:
  (might need admin privileges
  - `conda install m2w64-toolchain`
  - `conda install -c anaconda libpython`

Confirm that you can `import pymc3` without errors;

Join the Slack workspace

Fill out intro survey.