UNIVERSITY *of* WASHINGTON

Data Science UW
Methods for Data Analysis

Bayesian models, Part 2

Steve Elston

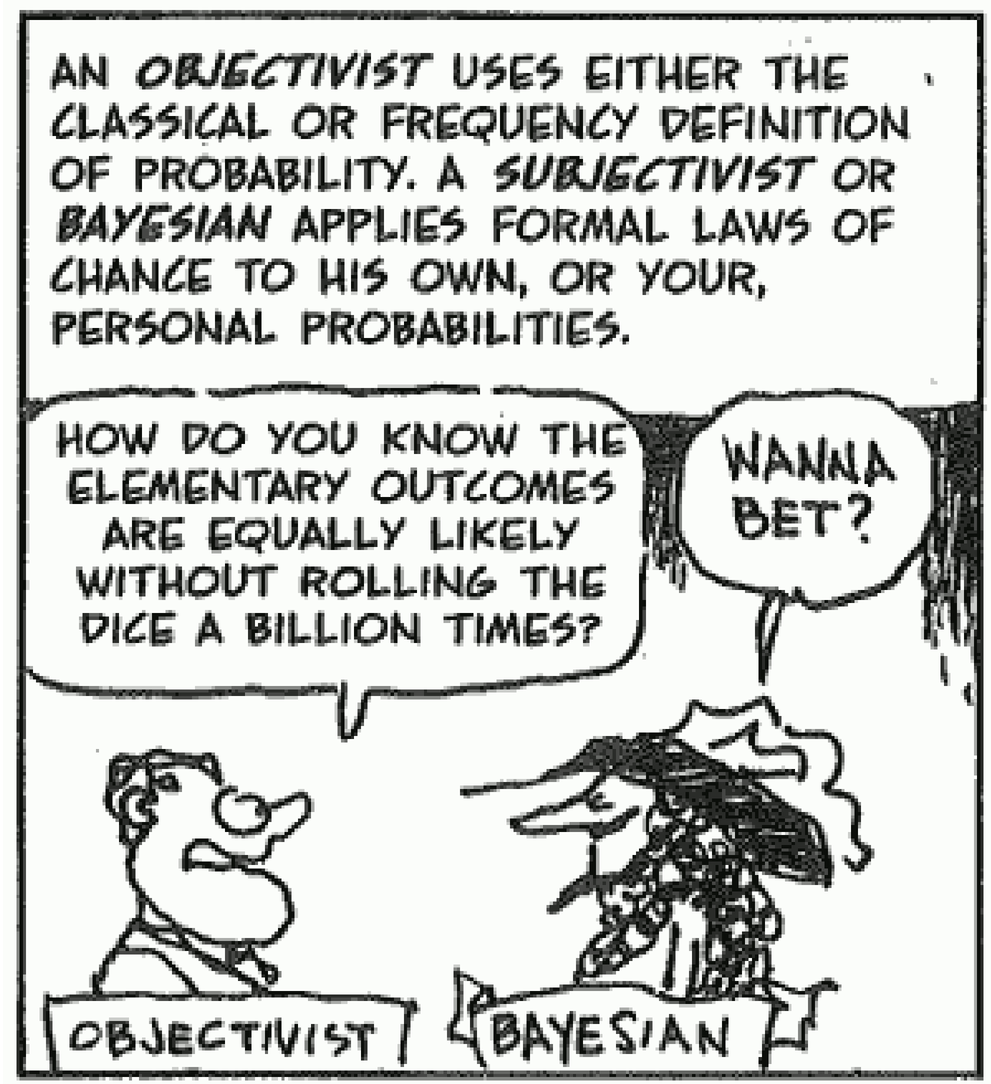
Required packages

To run the code for today's lecture, make sure you have the following installed.

- JAGS and the rjags Rbpackage. You must install the JAGS system from the downloads available at http://mcmc-jags.sourceforge.net/ rjags is an R API for JAGS
- LearnBayes
- mlbench
- ggplot2
- e1071
- coda

Deadline reminder!!!

All projects are due next Wednesday March 15. No exceptions can be granted!


AN OBJECTIVIST USES EITHER THE CLASSICAL OR FREQUENCY DEFINITION OF PROBABILITY. A SUBJECTIVIST OR BAYESIAN APPLIES FORMAL LAWS OF CHANCE TO HIS OWN, OR YOUR, PERSONAL PROBABILITIES.
HOW DO YOU KNOW THE ELEMENTARY OUTCOMES ARE EQUALLY LIKELY WITHOUT ROLLING THE DICE A BILLION TIMES?
WANNA BET?
OBJECTIVIST
BAYESIAN

Bayesian Model Summary

- Bayesian view of the world includes updating/changing beliefs new observations
- Bayesian view takes prior beliefs into account
- Based on Bayes theorem

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

- Can use simplified formulation with no P(B)

$$P(A|B) \propto P(B|A)P(A)$$

Posterior Distribution

Prior Distribution

The Likelihood

Bayes Model Summary

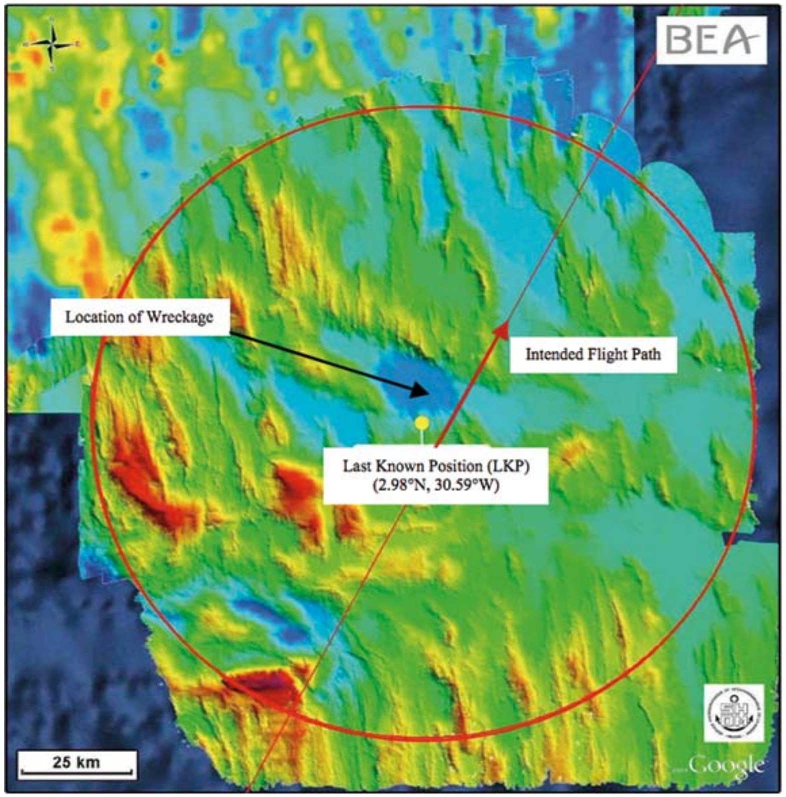
- > Use MCMC models to scale Bayesian analysis
  - – Metropolis-Hastings Algorithm
  - – Gibbs sampling for better convergence

Frequentist	Bayesian
Goal is a point estimate and confidence interval	Goal is posterior distribution
Start from observations	Start from prior distribution
Re-compute model given new observations	Update belief (posterior) given new observations
Examples: Mean estimate, t-test, ANOVA	Examples: posterior distribution of mean, overlap in highest density interval (HDI)

Reading assignment: Bayesian Inference Successes

$$P(parameters|data) \propto P(data|parameters)P(parameters)$$

> Bayesian inference used to successfully to find lost planes. E.g. Air France 447
> https://www.informs.org/ORMS-Today/Public-Articles/August-Volume-38-Number-4/In-Search-of-Air-France-Flight-447



Topics

- Bayesian Statistics
  - Markov Chain Monte Carlo (MCMC)
  - Multi-level (Hierarchical) models)
  - Bayes factor
  - Bayes hypothesis testing – Time permitting
  - MCMC diagnostics
- Naive Bayes

Bayesian Estimation of a Coin Flip Probability

$$P(parameters|data) = P(data|parameters)\frac{P(parameters)}{P(data)}$$

$$f(x) = x^{a-1}(1-x)^{b-1} \cdot (normalizing\ constant)$$

- After we choose a prior, we compute the posterior:

$$Posterior = Likelihood\frac{Prior}{P(data)}$$

- Always a problem estimating the P(data):

$$Posterior = Likelihood\frac{Prior}{P(data|all\ parameters)}$$

$$Posterior = Likelihood\frac{Prior}{\sum P(data|theta)}$$

Bayesian Estimation of Multiple Parameters

- We only had one parameter to estimate for the coin flip example, p(H).
- We created a grid to check (seq(0.01,0.99,length=100) and used this to calculate the p(data), by checking all the values.
- What if we had several parameters? If we had 6 parameters with a length 100 grid... = 100^6 = 1,000,000,000,000 = 1 trillion points to check.
- Maybe we don't have to sample everything, just enough points to understand and estimate the distribution of how p(data) behaves under the 6 parameters?

Markov Chain Monte Carlo

What is a Markov process?

> A Markov process makes a transition from one state other states with probability Π

- Π only depends on the current state
- Transition to one or more other states
- Can ‘transition’ to current state
- Π is a matrix of dim N X N for N possible states

> A Markov process is a random walk

Markov Chain Monte Carlo

Markov chain is a sequence of Markov transition processes:

$$P[X_{t+1} = x | X_t = x_t, \ldots, X_0 = x_0] = P[x_{t+1} = y | X_t = x_t]$$

'Memoryless' process

And

$\Pi =$

$P_{1,1}$	$P_{1,2}$	…	$P_{1,N}$
$P_{2,1}$	$P_{2,2}$	…	…
…	…	…	…
$P_{N,1}$			$P_{N,N}$

Introducing the Metropolis (Hastings) Algorithm

- The Metropolis algorithm is a specific MCMC algorithm.
- Algorithm:
  - 1. Pick a starting point in your parameter space and evaluate it according to your model. (find p(data)).
  - 2. Choose a nearby point randomly and evaluate this point.
    - If the p(data) of the new point is greater than your previous points, accept new point and move there.
    - If the p(data) of the new point is less than your previous point, only accept with probability according to the ratio: p(data new) / p(data old) .
  - 3. Repeat # 2 many times.

Introducing the Metropolis (Hastings) Algorithm

- M-H algorithm eventually converges to the underlying distribution.
- We only have to visit N points, not 1 Trillion points.
- There is high serial correlation in M-H chain, which slows convergence
- Need to 'tune' the state selection probability distribution used to find the next point
  - E.g. if we use Normal distribution need to pick σ.
  - If σ is too small chain will only search the space slowly.
  - If σ is too big, get large jumps and slow convergence

Gibbs Sampling

Improved version of M-H algorithm

> Uses systematic sampling of the parameter space
> Example: round-robin
  – With N dimensions
  – Sample 1, 2, …, N and then start over again
  – Transition still based on p(data)
> Reduces serial correlation and improves convergence

Remember Bayes Law:

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

- Tests are not the event. We have a disease test, which is different than the event of actually having the disease.
- Tests are flawed. Tests have false positives and false negatives.
- Tests return test probabilities, not the event probabilities.
- False positives skew results.
  - E.g. If fraud is rare, then the likelihood of a positive result of fraud is probably due to a false positive

Multi-level or Hierarchical Bayes Model

Simple Bayes models have all coefficients at same level

$$P(parameters|data) \propto P(data|parameters)P(parameters)$$

- Example: Recall the Beta distribution used as prior for Bernoulli likelihood

  $P(\theta | a, b) = \kappa\, \theta^{(a-1)} (1 - \theta)^{(b-1)}$

- But what if θ is not from a single population?

Multi-level or Hierarchical Bayes Model

How to model real-world hierarchies?

- Sub-populations may behave differently
- How to we partition the our model to account for sub-populations?
- Multi-level or hierarchical models accommodate this structure

Multi-level or Hierarchical Bayes Model

Examples

> Distinguish effect of individual player vs. team
> Performance of students vs. performance of school
> Product sales vs. store sale effect
> Species population vs. habitat

Multi-level or Hierarchical Bayes Model

Can use multi-level models to apply adjustments

- Individual player performance for team performance
- Individual students performance for school performance
- Sales for store effect
- Species population for habitat changes

Multi-level or Hierarchical Bayes Model

Extending Bayesian model

> Bayes rule becomes (chain rule for probabilities)

$$P(\theta, \omega | D) \propto P(D| \theta, \omega)\, p(\theta, \omega)$$

$$\propto P(D| \theta)\, p(\theta| \omega)\, p(\omega)$$

where

θ = parameters for each sub-group

ω = parameter for population

Multi-level or Hierarchical Bayes Model

Bayes rule for multi-level models

> Hierarchy of priors

$$P(\theta, \omega | D) \propto P(D | \theta)\, p(\theta | \omega)\, p(\omega)$$

Prior Distribution of ω

Posterior Distribution

Prior Distribution of θ given ω

The Likelihood

Multi-level or Hierarchical Bayes Model

Extending Bayesian model

- Bayes rule becomes

$$P(\theta, \omega | D) = P(D | \theta, \omega)\, p(\theta, \omega)$$
$$= P(D | \theta)\, p(\theta | \omega)\, p(\omega)$$

- Example: for beta prior and Bernoulli likelihood:

Prior of ω = Beta(A_ω, B_ω)

$$P(\theta, \omega | D) = \text{Bernoulli}(\theta)\ \text{Beta}(\omega\,(K-2) + 1, (1 - \omega)(K - 2) + 1)$$

Joint Prior

Multi-level or Hierarchical Bayes Model

Extending Bayesian model

> With Bayes rule:

$$P(\theta, \omega | D) = P(D | \theta)\, p(\theta | \omega)\, p(\omega)$$

> Example: for beta prior, the joint posterior probability is now:

$$p_j \sim \text{Beta}(y_j + K\eta,\ n_j - y_j + K(1 - \eta))$$

where

$\eta = a / (a+b)$

$K = a + b$

n_j = sample size

y_j = number of hits for player j

Multi-level or Hierarchical Bayes Model

The posterior is proportional to the product of individual probabilities

$$P(\theta, \omega | D) \propto \prod^{N}{}_{j=1} p_j$$

To simplify computation in example we reparameterize

$$\theta_1 = \log[\eta / (1 - \eta)]$$

$$\theta_2 = \log(K)$$

Bayesian Model Selection

How do we find the best model?

- Want model maximum apostiori probability
- Different likelihood distributions
- Different prior distributions
- Compare hierarchies of models

Compare Performance of Bayesian Models

Bayes Factor – identify the most likely model

> Hierarchy for models m:

$P[\Theta_1, \Theta_2, \ldots m|D] \propto P[\Theta_1, \Theta_2, \ldots, m]\ P[D|\Theta_1, \Theta_2, \ldots m]$

> Compare (hierarchy) of two models as a ratio:

$$\frac{p(m=1|D)}{p(m=2|D)} \propto \frac{p(D|m=1)}{p(D|m=2)} \frac{p(m=1)}{p(m=2)}$$

> Reduces to

$$\frac{p(m=1|D)}{p(m=2|D)} = \frac{p(D|m=1)}{p(D|m=2)} = Bayes\ Factor$$

Hypothesis Testing with Bayes Models

Use HCI to perform hypothesis tests

- Analogous to hypothesis tests on bootstrap resampled distributions
- Test conditions for **posterior** distribution
  - If HDI overlap; accept Null Hypothesis
  - If no HDI overlap reject Null Hypothesis
- HDI is different from Confidence Interval
  - HDI is for interval with greatest probability mass
  - Difference with CI is greatest for asymmetric prior
- Tests can be one-sided or two-sided

Diagnostics for MCMC

Multiple ways to look at convergence

- Summary statistics
  - Mean, median, se, time series se, quantiles
  - Plot cumulative mean and quantiles
  - Plot trace of each chain
  - Plot posterior distribution
- Plots based on convergence of multiple chains
  - Gelman-Rudin plot of chain convergence
  - Compares shrinkage of between chain and within chain variance
  - Should converge to 1.0

Diagnostics for MCMC

Detect convergence issues

- High rejection rate inhibits convergence
- High autocorrelation inhibits convergence
- Use ACF
- Effective Sample Size

$$ESS = N \,/\, \left(1 + 2 \sum_{K} ACF(k)\right)$$

Introduction to Naïve Bayes

Naïve Bayes is a remarkably good and flexible classifier

> Widely used classifier
  - Document classification
  - SPAM detection
  - Image classification

> Scales well
  - Does not require a prior
  - Computation linear in number of parameter/features
  - Requires minimal data
  - Simple regularization

Introduction to Naïve Bayes

Simplify the conditional probability calculation

> Start with Bayes Theorem: $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$

> The probability of class C_k is the joint distribution:

$$p(C_k, x_1, x_2, \ldots, x_n) = p(x_1, x_2, \ldots, x_n, C_k)$$
$$= p(x_1 | x_2, \ldots, x_n, C_k)\, p(x_2, \ldots, x_n, C_k)$$
$$= p(x_1 | x_2, \ldots, x_n, C_k)\, p(x_2 | x_3, \ldots, x_n, C_k)\, p(x_3, \ldots, x_n, C_k)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$= p(x_1 | x_2, \ldots, x_n, C_k)\, p(x_2 | x_3, \ldots, x_n, C_k) \ldots p(C_k)$$

> **But if $\{x_1, x_2, \ldots, x_n\}$ are independent:**

$$p(x_i | x_{i+1}, \ldots, x_n, C_k) = p(x_i | C_k)$$

Introduction to Naïve Bayes

Simplify the conditional probability calculation

- With $\{x_1, x_2, \ldots., x_n\}$ independent:

  $$p(x_i | x_{i+1}, \ldots., x_n, C_k) = p(x_i, | C_k)$$

- The probability of class C_k is the joint distribution:

$$p(C_k | x_1, x_2, \ldots., x_n) \propto p(C_k) \prod^{N}_{j=1} p(x_j | C_k)$$

- And the most likely class y_{hat} is:

  $$y_{hat} = \text{argmax}_k \left[p(C_k) \prod^{N}_{j=1} p(x_j | C_k) \right]$$

No Prior

Naïve Bayes Classifiers

Different distributions lead to different classifiers

- Different Naïve Bayes models are not the same!
- Normal naïve Bayes classifier
- Multinomial naïve Bayes classifier

$$\mathrm{Log}(p(C_k \mid x)) \propto \log[\, p(C_k) \prod^{N}_{j=1} p_{kj}^{\,Xi} \,]$$
$$= \log(\, p(C_k)\,) + \sum^{N}_{j=1} xi \log(\, p_{kj}\,)$$

- Bernoulli naïve Bayes classifier

$$p(x \mid C_k) = \prod^{N}_{j=1} p_{kj}^{\,Xi} \, (1 - p_{kj})^{(1 - Xi)}$$

Naïve Bayes Document Classification

Use 'bag of words' model

> Want the probability of topic C in document D given set of words in topic $\{\mathbf{w_1, w_2, \ldots., w_n}\}$:

$$p(C \mid D) = \prod_{j=1}^{N} p(w_j \mid C)$$

> Spam classification:

$$p(S+ \mid D) \propto p(S+) \prod_{j=1}^{N} p(w_j \mid S+)$$

> Test the hypothesis text is spam:

$$\ln(\, p(S+ \mid D) / p(S- \mid D)) =$$

$$\ln(p(S) / p(S-)) + \sum_{j=1}^{N} \ln(p(w_j \mid S+) / p(w_j \mid S-)) > 0$$

Naïve Bayes Pitfalls

A few words of caution

- Multiplication of small probabilities leads to floating point underflow
  - Compute with ln(p)
- If no samples/data get probability = 0
  - Product of probabilities = 0
  - Use Laplace smoother to ensure all p > 0
- Collinear features can be a problem
  - Do not exhibit independence
- Regularization is minor issue
  - Uninformative feature tends to uniform distribution

Final Projects

Only one week to go!

> This project gives you a chance to demonstrate your knowledge of the topics covered in the course

> You must create your report independently
  - Collaboration with others on the analysis is okay

> Report must contain:
  - Introduction and summary with clearly stated conclusions
  - Support your conclusions based on exploration of data and model results
  - See Florence Nightingale report for example

Final Projects, Continued

- Steps which you must show
  - Exploration of data from several views using graphics and summary statistics as appropriate
    - Demonstrate your understanding of the data relationships and properties
  - Comparison of several models
    - Compare difference classes of models and/or features as required
- R Code must in a professional style
  - Well structured
  - Clean comments
- **Due Monday August 29**
- **NO EXTENSIONS!** University policy