GENERAL ASSEMBLY

MODEL-BASED BAYES

Matt Brems

Data Science Immersive, GA DC

LEARNING OBJECTIVES

- Discuss model-based Bayesian inference and how it relates to Bayes' Theorem.
- Define improper prior, uninformative prior, informative prior, hierarchical modeling, and hyperparameter.
- Understand conjugacy and describe its benefits.

OPENING: RECALL BAYES' RULE

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

- $P(A)$ is the probability that A occurs given no supplemental information.
    - "Prior"
- $P(B|A)$ is the likelihood of seeing evidence (data) B assuming that A is true.
    - "Likelihood"
- $P(B)$ is what we scale $P(B|A)P(A)$ by to ensure we are only looking at A within the context of B occurring.
    - "Marginal Likelihood of B"

OPENING: BAYESIAN INFERENCE OF PARAMETERS

- Frequentist inference and Bayesian inference have different interpretations, and these interpretations give rise to different methods of analysis.
  - Example: The average height of women at Ohio State, denoted μ.
    - Frequentists treat μ as fixed: $\mu = 64 \text{ inches}$
    - Bayesians treat μ as a parameter with a distribution: $\mu \sim N(64, 2)$

OPENING: RECALL BAYES' RULE

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} = \frac{L(\theta|y)f(\theta)}{f(y)} \propto L(\theta|y)f(\theta)$$

- $f(\theta)$ is the distribution of θ given no supplemental information.
  - "Prior Distribution of θ"
- $L(\theta|y) = f(y|\theta)$ is the likelihood function relating y and θ.
  - "Likelihood"
- $f(y)$ is the normalizing constant to ensure $f(\theta|y)$ is a valid probability distribution.
  - "Marginal Likelihood of y"

OPENING

- In order to conduct Bayesian inference, we need to specify or estimate:
  - $f(\theta)$, the prior distribution of θ
  - $L(\theta|y) = f(y|\theta)$, the likelihood of the data y under a model that relates all parameters θ to the data y.
  - $f(y) = \sum_i f(\theta_i) f(y|\theta_i)$

- We can then find $f(\theta|y)$, the posterior distribution of θ given our data y, and now have a complete summary of our parameter of interest θ that takes into account our data y.

- But the question is, "How do we specify or estimate the prior and likelihood?"

MODEL-BASED BAYES

ESTIMATING A PRIOR DISTRIBUTION

DEFINITIONS

- We can think of our posterior distribution $f(\theta|y)$ as a combination of our data and our prior.
  - $f(\theta|y) \propto f(y|\theta) \times f(\theta) = likelihood \times prior$

- If our prior is too specific, then our posterior will be "dominated by" the prior.

- If our prior is too vague, then our posterior will be "dominated by" the data through the likelihood.

DEFINITIONS

- If our prior is too specific, then our posterior will be "dominated by" the prior.

- If our prior is too vague, then our posterior will be "dominated by" the data through the likelihood.

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

- If $P(A) = 0$, $P(A|B) = 0$.

- If $P(A) = 1$, $P(B|A) = P(B) \Rightarrow P(A|B) = P(A) = 1$.

DEFINITIONS

- Informative Priors
  - Includes prior knowledge about θ by taking past data and information into account. (i.e. scientific research, physical limits)
- Uninformative Priors
  - Includes minimal information about θ (i.e. flat priors)
- Improper Priors
  - Priors that are not valid probability functions.

BAYESIAN & FREQUENTIST STATISTICS

- Frequentist analysis makes no assumptions about the prior distribution of the parameter.

- You can think of a completely flat, improper prior distribution - this is frequentism!

SPECIFYING THE LIKELIHOOD

LIKELIHOOD PRINCIPLE

- Recall that the likelihood function relates our data y to parameters $\boldsymbol{\theta} = (\mu, \sigma)$.
  - Suppose that we believe $y \sim Normal(\mu, \sigma)$.
  - Then, $f(y|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$.
  - By definition, $L(\boldsymbol{\theta}|y) = f(y|\boldsymbol{\theta})$.
  - Maximizing the likelihood with respect to the parameters of interest is maximizing the probability density function with respect to the parameters.
    - The intuition here is that the likeliest values for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the ones associated with the highest probability.

LIKELIHOOD PRINCIPLE

- The likelihood principle states that if two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ provide likelihoods $L(\theta|\boldsymbol{x}) \propto L(\theta|\boldsymbol{y})$ for all values of θ, then our inferences gathered about θ should be identical whether we observed $\boldsymbol{x}$ or $\boldsymbol{y}$. (Statistical Inference, Casella and Berger, 2nd Ed.)
  - This, more simply, implies that the data influences our posterior distribution only through the likelihood function.

LIKELIHOOD PRINCIPLE

- Certain likelihood functions give rise to particularly elegant posterior distributions.
  - Normal prior, Normal likelihood ⇒ Normal posterior.
  - Beta prior, Binomial likelihood ⇒ Beta posterior.

- This is called conjugacy.
  - Priors are conjugate for likelihoods.
  - Prior and posterior follow the same parametric distribution.

CONJUGACY

- This requires a working knowledge of common statistical distributions, your data-generating process, and your subject area.
  - This will probably be more common in academia.
  - Generally, you will be able to conduct inference or describe the distribution through simulations without a need for conjugacy, but it's a good concept of which you should be aware.

EXAMPLE

- German tank problem.

- A railroad numbers its locomotives $1, \ldots, N$. You see a railcar with the number 60 painted on it.

- Estimate how many locomotives the railroad has.
  - Hypotheses
  - Data
  - Likelihood

EXAMPLE

- D&D dice problem.
- There are five dice: a 4-sided die, 6-sided die, 8-sided die, 12-sided die, 20-sided die.
- I roll a 6. What is the probability that I rolled each die?
  - Hypotheses
  - Data
  - Likelihood

MODEL-BASED BAYES

SEQUENTIAL UPDATING

UPDATING INFORMATION

- Prior: $p(\theta) \Rightarrow$ Posterior: $p(\theta|y_1)$
- Prior: $p(\theta|y_1) \Rightarrow$ Posterior: $p(\theta|y_1, y_2)$
- Prior: $p(\theta|y_1, y_2) \Rightarrow$ Posterior: $p(\theta|y_1, y_2, y_3)$

EXAMPLE

- M&M problem.

- Two bags of M&Ms:
  - Before 1995: B = 30%, Y = R = 20%, G = O = T = 10%
  - After 1995: Bl = 24%, G = 20%, O = 16%, Y = 14%, R = Br = 13%

- Pull yellow from bag 1.
  - Hypotheses
  - Data
  - Likelihood

EXAMPLE

- M&M problem.

- Two bags of M&Ms:
  - Before 1995: B = 30%, Y = R = 20%, G = O = T = 10%
  - After 1995: Bl = 24%, G = 20%, O = 16%, Y = 14%, R = Br = 13%

- Already pulled yellow from bag 1. Pull green from bag 2.
  - Hypotheses
  - Data
  - Likelihood

MODEL-BASED BAYES

HIERARCHICAL MODELS

EXAMPLE

- "Disentangling Bias and Variance in Election Polls"

$$y_i \sim N\left(\nu_{r[i]} + \alpha_{r[i]} + t_i\beta_{r[i]}, \sqrt{\frac{\nu_{r[i]}\left(1-\nu_{r[i]}\right)}{n_i} + \tau_{r[i]}}\right)$$

- $y_i =$ outcome of poll i
- $\nu_{r[i]} =$ final two-party vote share for Republican candidate
- $\alpha_{r[i]} + t_i\beta_{r[i]} =$ bias of ith poll with t in months
- $\sqrt{\frac{\nu_{r[i]}\left(1-\nu_{r[i]}\right)}{n_i}} =$ standard error of $\nu_{r[i]}$ under SRS
- $\tau_{r[i]} =$ election-specific variance

EXAMPLE

- "Disentangling Bias and Variance in Election Polls"

$$y_i \sim N\left(\nu_{r[i]} + \alpha_{r[i]} + t_i\beta_{r[i]}, \sqrt{\frac{\nu_{r[i]}\left(1 - \nu_{r[i]}\right)}{n_i} + \tau_{r[i]}}\right)$$

- $\alpha_r \sim N(\mu_\alpha, \sigma_\alpha); \mu_\alpha \sim N(0,0.05); \sigma_\alpha \sim N_+(0,0.05)$
- $\beta_r \sim N\left(\mu_\beta, \sigma_\beta\right); \mu_\beta \sim N(0,0.05); \sigma_\beta \sim N_+(0,0.05)$
- $\tau_r \sim N_+(0, \sigma_\tau); \sigma_\tau \sim N_+(0, 0.02)$

So ***how*** do we do Bayesian statistics?

- Goal: Find posterior probability of parameter θ given our data or evidence y.
  - This is written as $P(\theta|y)$.

- Needed:
  - Prior probability of parameter θ.
  - Likelihood of data y given parameter θ.
  - Marginal likelihood of data y with no knowledge of parameter.*