Bayes Theorem and Bayesian Inference

Priors and Posteriors!

DS6040 Fall 2024
Teague R. Henry

University of Virginia | School of Data Science

Outline

- Bayes Theorem Redux
- Priors
- Our First Conjugate Test
- Types of Priors

Bayes' Theorem

model = parameters = θ_i

Likelihood
(Probability of Data Given Model)

Posterior Distribution
(Probability of Model Given Data)

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

Prior Distribution
(Probability of Model)

$P(X) = P(X|\theta)P(\theta) + P(X|\theta^c)P(\theta^c)$

this is a normalizing constant with respect to theta

discrete → prob distribution

$$P(X) = \sum_i P(X \cap \theta_i) = \sum_i P(X|\theta_i)P(\theta_i)$$

Probability of ***your data*** occurring for any choice of θ
Note: Possible θ_i are constrained by the model type.

Marginal Distribution
(Probability of Data)

The Likelihood $P(X|\theta)$

The likelihood is the model you are fitting.

- What is the probability of your data given specific values of parameters θ

Typically, you don't need to provide the model in likelihood form

- For example, $y = \boldsymbol{\beta x} + \varepsilon$ has a likelihood of $y \sim Normal(\boldsymbol{\beta x}, \sigma)$

for a single observation

↳matrices

↳mean

- Many likelihoods are not proper distributions.
  - For example, the likelihood of a neural network exists, but it's a complicated analytic expression without a specific distributional family.

The Likelihood $P(X|\theta)$

You can construct likelihood functions by thinking case by case and using basic probability rules.

LIKELIHOOD CORRESPONDS TO MODEL! EVERY MODEL HAS A LIKELIHOOD

- For independent observations….

Let y be a vector of observations of sample size n

- Model: $y = \boldsymbol{\beta x} + \varepsilon$, $\varepsilon \sim N(0, \sigma)$ (What are the parameters here?)

Parameters: β, σ

- Likelihood: $\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \boldsymbol{\beta x}_i}{\sigma}\right)^2}$

normal pdf

The likelihood is a probability, but we either don't know the distribution, or more accurately, we don't care.

care about prob dist of posterior really

The Marginal $P(X)$

The marginal is a weird quantity. It represents the probability of the data over all possible values of θ

How to calculate it:

- You don't, the software does…
- More specifically, you need to calculate the following multidimensional integral:

$$P(X) = \int_{\theta \in \Theta} P(X|\theta) d\theta$$

← prob of new obsv of data
prior predictive dist

- This is only possible in simple cases, we usually approximate!

All of Bayesian inference is about how to approximate the marginal.

The Prior $P(\theta)$

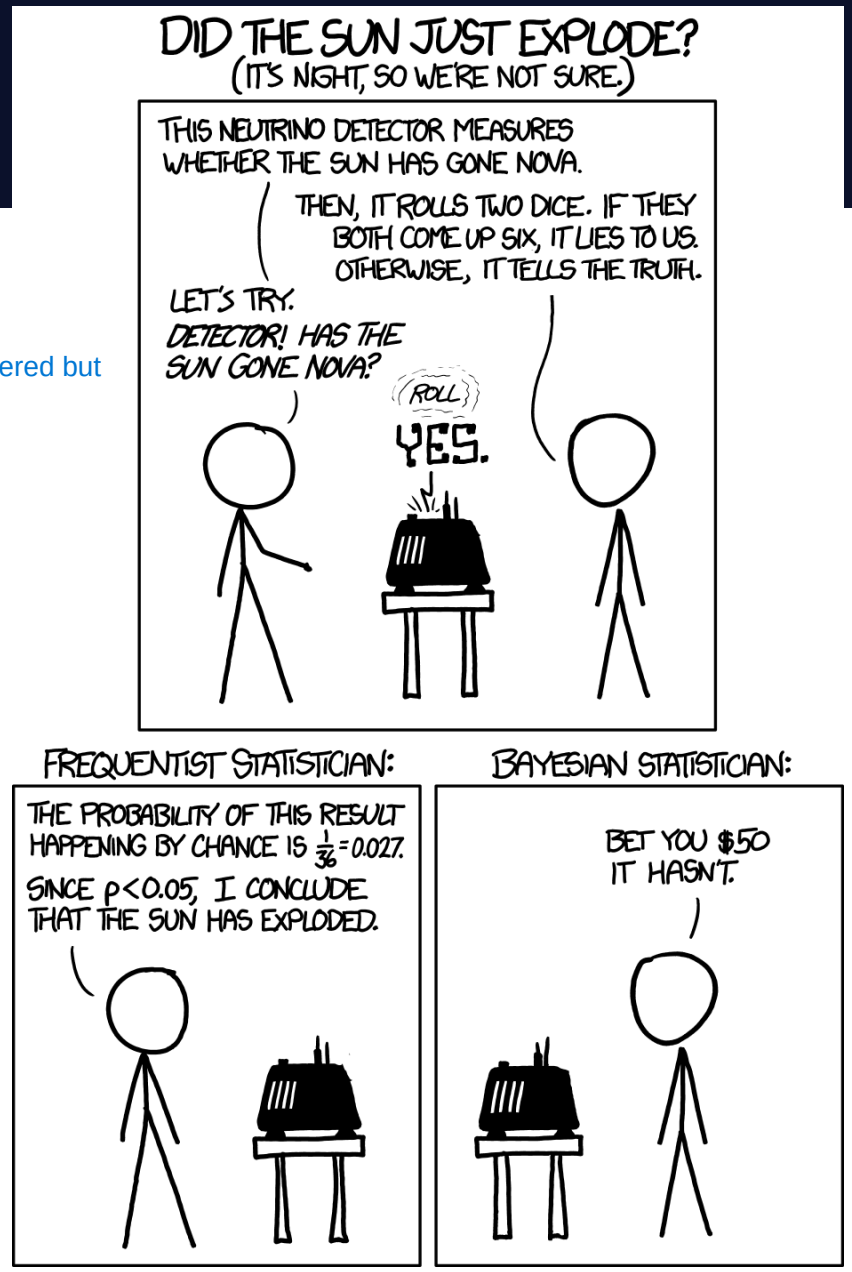
The prior represents your prior knowledge about the parameter values.

- You might not have any prior knowledge! That's alright, keep listening!

Where do priors come from?

- From the Council of Priors, which meets every 3rd weekend. BYOL.
- **We make them up** ★ in principled ways! Can force to est

many wrong answered but one right answer



The Prior $P(\theta)$

A prior is a guess about the parameter value, incorporating your uncertainty about your guess.

- A prior is a probability distribution you assign to the parameter.
- Priors can be *informative:*
  - You might know something about the base rate of a disease.
  - There might be previous experimental data on something.
  - You might have a previous dataset you could use to estimate priors ("Empirical Bayes")
- Priors can also be *uninformative*
  - Properly constructed, these priors result in very frequentist results…

low uncertainty means the more you need to justify it

might have some prior data/knowledge

The Prior $P(\theta)$

You can also do some very wild things with priors:

Regularization:

- For a regression equation with coefficients β not actual pdf, but is prop to
  - $\pi(\beta) \propto \exp\{-\lambda_1 ||\beta||_1 - \lambda_2 ||\beta||_2^2\}$ is the elastic net regularizer

↳ prop to

Model Averaging:

- Using the appropriate priors, you don't need to fit just one regression. You can fit all possible regressions and just average them together.

Priors help to pay the informational cost, so choosing priors with more information can help stabilize estimation of models.

we pay the price with priors — use more informative

The Prior $P(\theta)$

Important Concepts:

need prior per param but use

- All parameters need to have a prior assigned to them.
- Sets of parameters might have a multivariate prior assigned to them.

$$y = \boldsymbol{\beta}\boldsymbol{x} + \varepsilon, \varepsilon \sim N(0, \sigma)$$
$$\pi(\boldsymbol{\beta}) \sim MVN(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$$
$$\pi(\sigma) \sim Gamma(k, \theta)$$

multivariate normal

- The parameters of the prior distributions are called *hyperparameters.*
- *Hyperparameters can also have priors. It's priors all the way down…*
- More often, you are providing the specific values of hyperparameters.
  - They quantify the knowledge you have about the distribution.

The Posterior $P(\theta|X)$

So, we have the likelihood, somehow calculated the marginal, and specified the priors. So now we have the *posterior!*

- The posterior is the distribution of the parameters given the data and prior information.
- You can now start saying things like:
  - The expected value of the regression coefficients is blah blah blah.
  - There is a 50% chance that the effect of X on Y will be greater than 0

The posterior distribution is a full distribution. However, it might not be *proper*. improper has no name or analytical generalization of it

- A proper distribution is a known, named distribution, like a normal.

Conjugate Priors

Conjugate priors are ones we know how to calculate the marginal with!

A prior $P(\theta)$ is conjugate if the posterior $P(\theta|X)$ is of the same distribution.

A prior is proper if it is a well defined probability distribution.

- Advantages:
  - Analytically tractable – Calculating posterior probabilities is simple.
  - Computationally simple – No greedy estimation, everything is solvable directly
  - Very useful in more complex estimation – Simplifies parts of the problem…
- Disadvantages:
  - Very specific per distribution – In bespoke models, conjugate priors are difficult.
  - Requires some level of "subjective" information – This can be very small though.

Conjugate Priors – Normal Distribution

What is the conjugate prior of ***the mean*** of a normal distribution if we fix its variance?

If $x \sim N(\mu, \sigma^2), \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, where n is the sample size. sampling distribution of the mean from basic stat

Let σ^2 be fixed and known. We are now interested in $P(\mu \mid X, \sigma^2)$.

$$P(\bar{x}|\mu, \sigma^2) \propto \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right)$$

$$\mu \sim N(\mu_0, \sigma_0^2) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right)$$ prior distribution

Conjugate Priors – Normal Distribution

$$P(\bar{x}|\mu,\sigma^2) \propto \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right)$$

$$\mu \sim N(\mu_0,\sigma_0^2)\ implies\ P(\mu) \propto \exp(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2)$$

implies the prior on mu is that, says are two normally distributed variables

P(mu|X, sigma^2) is prop P(X|mu, signma^2)P(mu)/P(X)

but ignore the bottom P(X)

$$P(\mu|\bar{x},\sigma^2) = P(\bar{x}|\mu,\sigma^2)P(\mu)$$

$$\exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right)\exp(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2)$$

$$\exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

Expanding out and then completing the square…

Conjugate Priors – Normal Distribution

$$\exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

Expanding out and then completing the square...

$$\mu|\bar{x},\sigma^2 \sim N(\mu_n,\sigma_n^2)$$

$$\mu_n=\frac{\sigma^2\mu_0}{n\sigma_0^2+\sigma^2}+\frac{n\sigma_0^2\bar{x}}{n\sigma_0^2+\sigma^2} \qquad \sigma_n^2=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}$$

What happens as $n \to \infty$?

less certain about priors, / less data matters

this not → 0, stays similar no matter n

we can see that it becomes zero, very small

★ more data have, less prior matters ★

Conjugate Priors – Beta Distribution

Binomial: Distribution for k = the number of success ("heads") out of n Bernoulli trials each with probability p of success:

$$f(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Beta Distribution: — conj prior of p param for binomial

$$g(p|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

A Beta Distribution is the conjugate prior for the p parameter of a Binomial.

α, β overwhelmed by k, n if large

$$p|n,k,\alpha,\beta \sim Beta(k+\alpha, n-k+\beta)$$

Flat Priors

$$P(\theta_i|X) = \frac{P(X|\theta_i)P(\theta_i)}{\sum_k P(X|\theta_k)P(\theta_k)}$$

model selection
model = i
need prior on which model is correct

- Now, consider a selection between models θ_i
  - Here, we are abstracting away from parameters, but this logic holds there too.
- If you had no clue which θ_i is correct, what prior would you choose?
- A flat prior is ill defined, but has two major uses:
  - In the above case, let $P(\theta_i) = \frac{1}{k}$ (or equivalently, $P(\theta_i) = c$ for any constant c)
  - Sometimes, a flat prior refers to conjugate priors with high variance, i.e.

$$P(\mu|\bar{x}, \sigma^2) = N(0, 10000)$$

no good choice for params
→ flat priors

A flat prior is un-informative, but not **non-informative**

no prior is non-informative, no one cares though

“Non-informative” Priors

Spoiler Alert: There are no such things as non-informative priors.

Example: Let $y \sim N(\theta, 1), p(\theta) = c. We\ observe\ y_1 = 1.\ P(\theta > 0 | y_1 = 1) = .84.$

↳ prob any val in θ = constant

- “Non-informative” priors (flat, high variance) are “fine” when you have enough data to overwhelm the prior.
- “Non-informative” priors have problems when θ is high dimensional
- A “non-informative” prior under one parameterization can be highly informative under a different parameterization (the “model” remains the same.)
- Check if the posteriors “make sense” after using a non-informative prior.
  - Requires knowledge of the problem, and what is reasonable vs. unreasonable.

custom regression

Great, albeit a little technical, post: https://statmodeling.stat.columbia.edu/2013/11/21/hidden-dangers-noninformative-priors/

Jeffreys Priors

What is the probability that an arbitrarily biased coin lands on heads?

$$P(x = 1) = p$$

- $P(p) = c$ (for $p \in [0,1]$) – Reasonable "non-informative" prior
- Reparameterize the model into log-odds of heads - $\eta = \log \frac{p}{1-p}$
- Keep same prior on p …
- $P(\eta) \approx \frac{\exp \eta}{(1+\exp \eta)^2}$ - ***<u>This is not uninformative</u>***

it is highly informative, not necessarily good in case they are wrong

Jeffreys Priors

Jeffreys Priors are “non-informative” priors that are applicable under any monotone change of variables.

- $X \sim f(x|\theta)$ where θ is a scalar
- Jeffreys Prior - $P(\theta) \propto \sqrt{I(\theta)}$
- $I(\theta)$ - Fisher information - $I(\theta) = -E_{[X|\theta]}\ddot{L}(\theta)$ expected value of the theta of the likelihood at that value

Useful for when you absolutely must have a “non-informative” prior.

Note: Does not fix non-informative priors being over-informative with little data AMOUT OF DATA IS STILL THE SUPER IMPORTANT PART

monotone= only increases (or only decreases), he assumes increase

Empirical Bayes Priors

You, a clever data scientist: But Teague, we have information about the parameters… we collected data to estimate them, can't we just use that?

Empirical Bayes Priors

- **Idea:** Use our observed data to inform our prior selection. Note that this only applies to hierarchical problems (though, this covers the majority of analyses)

Example: Let $X_i \sim Poisson(\lambda_i), with\ \lambda_i \sim Gamma(\alpha, \beta)$ where α is known.

We want to know the posterior distribution of $\lambda_i | x_i, \alpha, \beta$ but we have no clue as to β

We can estimate β using a frequentist approach (MLE) and plug that into Bayes Theorem!

$$\hat{\beta} = \frac{\alpha}{\bar{x}}$$

match the moments

$$\lambda_i | x_i, \alpha \sim Gamma(x_i + \alpha, 1 + \frac{\alpha}{\bar{x}})$$

A fully Bayesian approach would put a distribution on β. The empirical Bayes approach doesn't fully account for the uncertainty in $\hat{\beta}$

In Class Exercise

Groups of 5!

Question: What are the three most important things to consider when choosing a prior?

- This exercise is to let me see how well the class is grasping the idea of priors, so discuss and volunteer your list, but don't worry about being "wrong!"

1. Start by choosing a non-informative prior
2. Background knowledge of data
3. Could use data from before

The distribution of the prior tells us whether it is informative or not