Introduction

BIOS 7717 Bayesian Inference

John Hughes

Department of Biostatistics and Informatics
University of Colorado Denver

Laplace Said

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.

– Pierre Laplace (1749–1827)

Laplace Said

- Laplace was a physicist and mathematician. He is considered to have been the best mathematician of his day, and he is counted among the greatest scientists of all time.
- Laplace laid out probability theory.
- Laplace independently discovered Bayes' theorem.

1. To what science was Laplace referring?
2. Is probability (or statistics) a science?
3. Is mathematics a science?
4. What is science?
5. Laplace believed the universe to be completely deterministic. Does that belief contradict his interest in probability?
6. What is the universe?
7. Where in the universe is mathematics?
8. The universe is expanding. What is the universe expanding into?
9. Is probability (or statistics) the most important object of human knowledge?

Probability Theory

- Probability theory is a formalism.
- *formalism*: emphasizes form rather than content or meaning
- Two famous formulations of probability are the Kolmogorov and Cox formulations.
- The two formulations share the same laws of probability.

1. The probability of an event is a non-negative real number:

   $\mathbb{P}(E) \in \mathbb{R},\ \mathbb{P}(E) \geq 0\ \ \forall E \in \mathcal{F}$, where $\mathbb{P}$ denotes a probability function, E is an event, and $\mathcal{F}$ is the space of events.

2. At least one elementary event will occur:

   $\mathbb{P}(\Omega) = 1$, where Ω is the sample space.

3. Any countable sequence of mutually exclusive events $E_1, E_2, \ldots$ satisfies

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

Bayes' Rule

- Both Cox's theory and Kolmogorov's theory include Bayes' rule.
- Thomas Bayes (1702–1761) was an English statistician, philosopher, and Presbyterian minister.
- Bayes employed a special case of what is now called Bayes' rule. This work was published after Bayes' death.
- For us, Bayes' rule will most often take the form

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \frac{\ell(\boldsymbol{\theta} \mid \boldsymbol{Y})\, p(\boldsymbol{\theta})}{\int \ell(\boldsymbol{\theta} \mid \boldsymbol{Y})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \propto \ell(\boldsymbol{\theta} \mid \boldsymbol{Y})\, p(\boldsymbol{\theta}),$$

  where
  - π denotes the posterior distribution,
  - $\boldsymbol{\theta}$ are parameters,
  - $\boldsymbol{Y}$ are data,
  - ℓ is the likelihood, and
  - p is the prior distribution.

Bayes' Rule

- The likelihood is, of course, a sampling model turned on its head:

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{Y}) = f(\boldsymbol{Y} \mid \boldsymbol{\theta}),$$

  where f denotes a probability density function. This part of the model expresses our belief regarding the data-generating mechanism, i.e., how our evidence (the data) arose (conditional on a given value of $\boldsymbol{\theta}$).
- The prior distribution p expresses what we believe regarding the value of $\boldsymbol{\theta}$, prior to having considered the evidence $\boldsymbol{Y}$.
- The posterior distribution π represents a revision, in light of the evidence $\boldsymbol{Y}$ (viewed through the "lens" of the likelihood), of our prior belief regarding the value of $\boldsymbol{\theta}$.
- In other words, Bayes' rule provides a mathematical method of interpreting evidence in the context of previous experience.

An Example

- Suppose you develop painful blisters on the right side of your torso.

- Your doctor tells you that 90% of people who have shingles experience the same symptom you are experiencing, i.e.,

$$\mathbb{P}(\text{painful blisters on torso} \mid \text{disease is shingles}) = 0.9.$$

- Similarly, 80% of patients who have contact dermatitis present with painful blisters:

$$\mathbb{P}(\text{blisters} \mid \text{dermatitis}) = 0.8.$$

An Example

- Should your doctor use your symptoms alone to decide which disease you have?
- This seems less than ideal since you do not (or should not, at least) care about these two probabilities. Instead, you want to know

$$\mathbb{P}(\text{shingles} \mid \text{blisters})$$
$$\mathbb{P}(\text{dermatitis} \mid \text{blisters}).$$

- Your knowledgeable doctor consults public health statistics and learns that the prevalence of shingles in your age group is 0.1 while the prevalence of contact dermatitis is 0.8.

An Example

- Now, we have

$$\begin{aligned}\mathbb{P}(\text{shingles} \mid \text{blisters}) &\propto \mathbb{P}(\text{blisters} \mid \text{shingles})\, \mathbb{P}(\text{shingles}) \\ &= 0.9 \cdot 0.1 = 0.09 \\ \mathbb{P}(\text{dermatitis} \mid \text{blisters}) &\propto \mathbb{P}(\text{blisters} \mid \text{dermatitis})\, \mathbb{P}(\text{dermatitis}) \\ &= 0.8 \cdot 0.8 = 0.64.\end{aligned}$$

- This implies that your doctor should conclude that you have contact dermatitis.
- Why can we ignore the marginal likelihood $\mathbb{P}(\text{blisters})$?
- Dermatitis is the maximum a posteriori (or MAP) estimate of the disease.
- What is the maximum likelihood estimate of the disease?

Data vs Parameters

- In $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto f(\boldsymbol{Y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$, the data $\boldsymbol{Y}$ and the parameters $\boldsymbol{\theta}$ have the same "status" in the sense that each has been assigned a probability distribution.
- Contrast this with the method of maximum likelihood, in which p plays no role.
- Does this mean that Bayesians believe parameters to be random?
- Does it make sense to believe $\boldsymbol{\theta}$ is random?

On the Purposes of Priors

- Whatever meaning we attach to prior distributions, priors have uses besides incorporating previous experience.
- Indeed, some prior distributions have nothing to do with previous experience.
- For example, an appropriately chosen prior may
  - improve prediction,
  - aid computation,
  - yield a closed form for the posterior distribution,
  - improve inference for other parameters of interest, or
  - reflect our ignorance regarding the value of the parameter.

Bayes' Example

- Bayes considered the task of doing inference regarding the probability of "success" in a binomial experiment.
- More specifically, we wish to do inference regarding q based on evidence $X \sim \mathcal{B}(n, q)$, where $\mathcal{B}$ denotes the binomial distribution.
- Bayes put a standard uniform prior on q, i.e.,

$$p(q) = 1\{q \in (0, 1)\}.$$

  - This seems to reflect a state of total ignorance regarding q.
  - This prior also leads to a closed-form posterior.

Bayes' Example

- Now, we have

$$\pi(a < q < b \mid X = x) = \frac{\int_a^b \binom{n}{x} q^x (1-q)^{n-x} dq}{\int_0^1 \binom{n}{x} q^x (1-q)^{n-x} dq}$$
$$= \frac{\int_a^b q^x (1-q)^{n-x} dq}{B(x+1, n-x+1)},$$

  where B denotes the beta function.
- Recall that the pdf for the beta distribution is

$$f(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} 1\{y \in (0,1)\}.$$

- And so we see that the posterior distribution is the $\mathcal{B}eta(x+1, n-x+1)$ distribution.

On “Deliverables”

- When we employ the method of maximum likelihood, we obtain

$$\hat{\boldsymbol{\theta}} = \arg\sup \ \ell(\boldsymbol{\theta} \mid \boldsymbol{Y}),$$

  and we use properties of $\hat{\boldsymbol{\theta}}$ to construct a confidence region for $\boldsymbol{\theta}$.
- By contrast, the (preliminary) product of a Bayesian analysis is a distribution. More precisely, the product of a Bayesian analysis is
  - the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y})$, if such is available in closed form (it rarely is); or
  - a sample that is distributed approximately according to $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y})$.
- In other words, the method of maximum likelihood leads directly to a point estimate of $\boldsymbol{\theta}$ while point estimation is more complicated in a Bayesian context. Presumably, one uses the posterior distribution (or sample) to obtain a point estimate of $\boldsymbol{\theta}$. But how, exactly, should one do so?

Loss Functions

- Although loss functions are important in both paradigms, loss functions are crucial to Bayesian inference.
- This is so because we use the posterior distribution and some loss function to arrive at a (Bayesian) decision rule.
- That is, a Bayesian analysis has three key ingredients: likelihood, prior distribution, and loss function.

Loss Functions

- The most popular loss function is the squared-error loss function:

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \|\boldsymbol{\theta} - \boldsymbol{\delta}\|^2.$$

  Use of this loss function leads to the posterior mean as our point estimate:

$$\tilde{\boldsymbol{\theta}} = \mathbb{E}^{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \int \boldsymbol{\theta}\, \pi(\boldsymbol{\theta} \mid \boldsymbol{Y})\, d\boldsymbol{\theta} = \frac{\int \boldsymbol{\theta}\, \ell(\boldsymbol{\theta} \mid \boldsymbol{Y})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int \ell(\boldsymbol{\theta} \mid \boldsymbol{Y})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}.$$

- Other popular loss functions are
  - $L(\boldsymbol{\theta}, \boldsymbol{\delta}) = 1\{\boldsymbol{\theta} \neq \boldsymbol{\delta}\}$ (0-1 loss), and
  - $L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_j |\theta_j - \delta_j|$ (absolute loss).

  The former is useful for prediction while the latter offers robustness and leads to the posterior median as the point estimate.

Posterior Prediction

- Under the Bayesian paradigm, we use the posterior predictive distribution to predict a new value of the response.
- Let $\boldsymbol{Y}^*$ denote the new outcome. Then the posterior predictive distribution is given by

$$\pi(\boldsymbol{Y}^* \mid \boldsymbol{Y}) = \int f(\boldsymbol{Y}^* \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{Y})\, d\boldsymbol{\theta}.$$