Bayesian Statistics Notes

Haolin Li

September 10, 2024

Abstract

The general purpose of Bayesian Statistics is to find/estimate the joint distribution $\mathbf{P(y_1, \ldots, y_n, \theta_1, \ldots, \theta_m)}$, from which we explore applications with the help of posterior and predictive distributions, and of course, the Bayes' Rule. It includes paramatric, from one-parameter distributions like Bernoulli and Poission to multiple-parameter like Binomial, and unparametric methods. This serves as an introductory course to the world of Bayesian.

Contents

Chapter 1

Introduction and Notation

The core difference between a Bayesian and a frenquencist is the belief on whether the latent or the parameter θ is a random variable or a constant. One of the major impacts is that iid assumptions changes to conditional independence instead of mutual independence due to the connection between θ and Y through the joint distribution $\mathbf{P}(\mathbf{Y_i}, \theta_\mathbf{j})$, where the marginal distribution of Y_i are parametrized by θ. Before everything, we should introduce some simple notations.

1.1 Notation

- $\mathcal{Y}$ denote the set of all possible observation values
- Y denote the random variable
- y denote the value of a single observation
- Θ is the space of parameters

Note. Let $\theta \in \Theta$, define $\pi(\theta)$ or $p(\theta)$ as the prior distribution.

Note. For any $\theta \in \Theta$, $y \in Y$, $\mathbf{P}(\mathbf{y}|\theta)$ describes the sampling model

Note. Let $\theta \in \Theta$, the posterior distribution $\mathbf{P}(\theta|\mathbf{y})$ describes our belief about the parameters based on samples.

Theorem 1.1.1 (Bayes' Rule)**.**

$$\mathbf{P}(\theta|\mathbf{y}) = \frac{\mathbf{P}(\mathbf{y}|\theta)\pi(\theta)}{\mathbf{P}(\mathbf{y})} \tag{1.1}$$

Example. Suppose $\theta \in [0, 1]$, and $Y_i|\theta \sim Bernoulli(\theta)$ with sample size 20. Then let $y|\theta = \sum Y_i \sim Binomial(20, \theta)$. We will see how the choice of the prior has on the posterior with $\theta \sim Beta(a, b)$, we have

$$\mathbb{E}[\theta] = \frac{a}{a+b}, \qquad mode[\theta] = \frac{a-1}{a-1+b-1}$$

Usually, from a prior sampling, we denote a be the number of events counted, and b as the total sample size. In this case, say the event happens twice, we have

$$\theta \sim Beta(2, 20)$$

We know that the pdf of $Beta(a, b)$ is

$$pdf(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},\ \theta \in [0,1]$$

Using Bayes' Rule, we have

$$\begin{aligned}
\mathbf{P}(\theta|\mathbf{Y_1},\dots,\mathbf{Y_{20}}) &\propto \mathbf{P}(\mathbf{Y_1},\dots,\mathbf{Y_{20}}|\theta)\cdot\pi(\theta) \\
&\propto \theta^{\sum Y_i}(1-\theta)^{n-\sum Y_i}\cdot\theta^{a-1}(1-\theta)^{b-1} \\
&= \theta^{a+y-1}(1-\theta)^{b+n-y-1} \\
&\sim Beta(a+y, b+n-y)
\end{aligned}$$

This means, amazingly, the posterior falls in the same family of distribution like the prior. For priors like this, we give them a special name: Conjugate Prior.

From the example, we can derive

$$\mathbb{E}[\theta] = \frac{a}{a+b}, \qquad mean(sample) = \frac{y}{n}, \qquad \mathbb{E}[\theta|y] = \frac{a+y}{a+b+n}$$

With a little massage, we have

$$\mathbb{E}[\theta|y] = \frac{a+b}{a+b+n}\cdot\frac{a}{a+b} + \frac{n}{a+b+n}\cdot\frac{\sum y_i}{n}$$

This break down is intriguing because the posterior mean is in fact the weighted sum of the prior mean and sample mean, implying insensitivity to the prior as $n \to \infty$ since the weight dominates. We can further conclude the above example with the following proposition:

Proposition 1.1.1. With $Y_i|\theta \sim Bernoulli(\theta)$ and $\theta \sim Beta(a,b)$ as the conjugate prior, we have the posterior $\theta|Y_1,\dots,Y_n \sim Beta(a+\sum Y_i, b+n-\sum Y_i)$

Remark.

$$Uniform[0,1] = Beta(1,1)$$

1.2 One-parameter Models

In this section, we talk about single-parameter models, where the example in Proposition 1.1.1 about Bernoulli with Beta priors was a perfect example. Let's start with a closer look at the posterior with uniform prior:

$$\mathbf{P}(\theta|\mathbf{Y_1},\dots,\mathbf{Y_n}) \ \propto\ \mathbf{P}(\mathbf{Y_1},\dots,\mathbf{Y_n}|\theta)\cdot\pi(\theta) \ =\ \theta^{\sum \mathbf{Y_i}}(\mathbf{1}-\theta)^{\mathbf{n}-\sum \mathbf{Y_i}}$$

By observing the last term, the posterior distribution is determined by the statistic $\sum Y_i$(we assume sample size known at all time). This means we don't need to examine the exact values of Y_i, but the sum would be enough/sufficient to find out the parameters of the posterior. As a result, we say the sum $\sum Y_i$ is the sufficient statistic of the posterior distribution.

Definition 1.2.1 (Sufficient Statistics)**.** Given any subject $\mathcal{S}$ we are trying to estimate, a distribution, a parameter, or even another statistic, a statistic $T(Y_i)$ is a sufficient statistic of $\mathcal{S}$ if $T(Y_i)$ contains enough information for us to determine that subject.

In the previous section, we also mentioned the rough idea of a conjugate prior, now we give it a formal definition:

Definition 1.2.2 (Conjugate Prior)**.** A class of prior distributions $\mathcal{P}$ for θ is called conjugate for a sampling model $\mathbf{P}(\mathbf{Y}|\theta)$ if

$$\pi(\theta) \in \mathcal{P} \Rightarrow \mathbf{P}(\theta|\mathbf{Y}) \in \mathcal{P}$$

Now we see how these two concepts play together with the following example.

Example. Previously we have talked about the posterior conditioned over the entire sequence, $\theta|Y_1, \ldots, Y_n$. What would happen if we instead condition the parameter with posterior's sufficient statistic? i.e. $\theta|y = \sum_{i=1}^{n} Y_i$

$$\begin{aligned}
\mathbf{P}(\theta|\mathbf{y}) &\propto \mathbf{P}(\mathbf{y}|\theta) \cdot \pi(\theta) \\
&= \binom{n}{y} \theta^{y}(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1} \\
&\sim Beta(a+y, b+n-y)
\end{aligned}$$

Surprisingly, it looks the same as $\theta|Y_1, \ldots, Y_n$! This is because y is the sufficient statistic for the posterior distribution with sampling model Bernoulli and prior Beta.