Bayesian Statistics Notes

Haolin Li

September 11, 2024

Abstract

The general purpose of Bayesian Statistics is to find/estimate the joint distribution $\mathbf{P}(\mathbf{y_1}, \ldots, \mathbf{y_n}, \theta_\mathbf{1}, \ldots, \theta_\mathbf{m})$, 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

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

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 $\theta|Y_1,\ldots,Y_n \sim Beta(a+\sum Y_i, b+n-\sum Y_i)$

Remark.

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

1.2 One-parameter Models

1.2.1 Sufficient Statistics and Conjugate Prior

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

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

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, as the sufficient statistic for the posterior, is enough to determine the distribution.

1.2.2 Predictive Distribution

There are two main reasons why we model things with mathematics in general: to explain, and to predict. In this section, we explore how can we make predictions within the Bayesian framework.

Definition 1.2.3 (Predictive Distribution). Given data points $Y_1, \ldots, Y_n$, the predictive distribution refers to

$$Y_{n+1}|Y_1, \ldots, Y_n$$

Example. Say we have $Y_i|\theta \sim Bernoulli(\theta)$ and prior $\pi(\theta) \sim Beta(a, b)$, then the predictive distribution $Y_{n+1}|Y_1, \ldots, Y_n \sim Bernoulli(\hat{\theta})$, and we will try to find the exact value of $\hat{\theta}$ in order to determine the predictive distribution. Using marginalization, we have

$$\begin{aligned}
\hat{\theta} &= \mathbf{P}(\mathbf{Y_{n+1}} = \mathbf{1}|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \\
&= \int_0^1 \mathbf{P}(\mathbf{Y_{n+1}} = \mathbf{1}, \theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \mathbf{d}\mathbb{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \\
&= \int_0^1 \theta \cdot pdf(Beta(a + \sum Y_i, b + n - \sum Y_i)) d\theta \\
&= \mathbb{E}[\theta|Y_1, \ldots, Y_n] \\
&= \frac{a + \sum Y_i}{a + b + n}
\end{aligned}$$

As a result, we have our predictive distribution $Y_{n+1}|Y_1, \ldots, Y_n \sim Bernoulli(\frac{a+\Sigma Y_i}{a+b+n})$

1.2.3 Confidence Regions

After we have known how to estimate the next observation with predictive distribution, now let's find out how to estimate the parameters.

If we were frenquencists, we first assume the sample distribution, with which then determine the confidence interval(also random variables), after that we sample and see if the result lies in the interval, which determines whether we reject hypothesis H_0.

However, as a Bayesian, the confidence of our estimation on θ originates from the posterior $\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n})$, which means we have to sample before determining the "interval" or in essence, the distribution itself. Also, we are able to tell the odds very clearly because right now we know exactly how $\theta|Y_1, \ldots, Y_n$ distributes, at least that's what we hope for.

Moreover, in the Bayesian way, there are two major ways to determine the interval with the posterior distribution.

1. Highest Posterior Density(HPD): points with the highest pdf of the posterior distribution, also $\mathbb{P}(HDP_\alpha) = 1 - \alpha$.

2. Quantile based interval: old-fashion $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$

1.2.4 Case Study: Possion + Gamma Prior

Now we are in business, in this section, we introduce another pair of conjugate prior. A little recap, if $Y \sim Poission(\theta)$, then $\mathbf{P}(\mathbf{Y} = \mathbf{y}|\theta) = \frac{\mathbf{e}^{-\theta}\theta^{\mathbf{y}}}{\mathbf{y!}}$, $\mathbf{y} \in \underline{\mathbb{N}}$. We will leave the pdf of Gamma for now because in this case study, we present a way to actually find the distribution. Firstly, a little analysis on the joint sampling distribution:

$$\begin{aligned}\mathbf{P}(\mathbf{Y_1}, \ldots, \mathbf{Y_n}|\theta) &= \prod_{i=1}^{n} \mathbf{P}(\mathbf{Y_i}|\theta) \\ &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{Y_i}}{Y_i!} \\ &= \frac{e^{-n\theta}\theta^{\Sigma Y_i}}{\prod_{i=1}^{n} Y_i!}\end{aligned}$$

> **Note.** $Y = Y_1 + \cdots + Y_n \sim Possion(n\theta)$.

Right now we can work on the posterior.

$$\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \propto \pi(\theta) \cdot \mathbf{P}(\mathbf{Y_1}, \ldots, \mathbf{Y_n}|\theta) \tag{1.2}$$

$$\propto \pi(\theta) \cdot e^{c_1\theta}\theta^{c_2} \tag{1.3}$$

Let's guess what is the conjugate prior should look like. If $\pi(\theta) \propto e^{d_1\theta}\theta^{d_2}$, then the prior and the posterior will be proportionate to the same pattern, i.e. $\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \propto \mathbf{e}^{(\mathbf{c_1}+\mathbf{d_1})\theta}\theta^{(\mathbf{c_2}+\mathbf{d_2})}$. Then all we need to do now is to determine the constant depending on c_i, d_i, the result is

$$\pi(\theta) \sim Gamma(a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}$$

Therefore, we can push further with 1.3 so we have

$$\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) \propto \mathbf{Gamma}(\mathbf{a} + \mathbf{\Sigma Y_i}, \mathbf{b} + \mathbf{n})$$

And we conclude this case study with the following proposition

> **Proposition 1.2.1.** With $Y_i|\theta \sim Poission(\theta)$ and $\theta \sim Gamma(a, b)$ as the conjugate prior, we have the posterior $\theta|Y_1, \ldots, Y_n \sim Gamma(a + \Sigma Y_i, b + n)$.

1.3 Exponential Family and Monte Carlo Method

In previous examples, we have seen two conjugate distribution pairs that saves our lives from tedious computation. A natural question one may ask is, does there exist a pattern for us to find conjugate priors given any sampling distribution? The answer is almost yes, there's indeed a pattern for huge family of distributions: the exponential family.

> **Definition 1.3.1** (One-parameter Exponential Family)**.** A sampling model $\mathbf{P}(\mathbf{Y}|\theta)$ is an one-parameter exponential family model if we have the following decomposition:
>
> $$\mathbf{P}(\mathbf{Y}|\theta) = \mathbf{h}(\mathbf{Y})\mathbf{c}(\theta)\exp(\theta\mathbf{t}(\mathbf{Y})) \tag{1.4}$$
>
> where θ is a parameter, and $t(Y)$ is the sufficient statistic for the posterior.

In imitation to (1.4), with n_0, t_0 being the sample size and sufficient statistic of the prior sample, we suppose our prior takes the form:

$$\pi(\theta) = k(n_0, t_0)c(\theta)^{n_0} \exp(n_0 t_0 \theta)$$

We then can derive the posterior distribution using the Bayes' Rule:

$$\begin{aligned}\mathbf{P}(\theta|\mathbf{Y_1}, \ldots, \mathbf{Y_n}) &\propto \pi(\theta) \cdot \mathbf{P}(\mathbf{Y_1}, \ldots, \mathbf{Y_n}|\theta) \\ &\propto c(\theta)^{n_0+n} \exp\{(n_0+n)\frac{n_0 t_0 + n\bar{t(y)}}{n_0+n}\theta\}\end{aligned}$$

With the color, we can easily see that the prior and the posterior fall into the same family of distribution, which by definition is conjugate.

As for Monte Carlo, it's just a fancy name for simulation/experiment. Specifically, it applies to the field of Bayesian by allowing us to simulate the sampling process. Given a joint distribution $\mathbf{P}(\mathbf{Y}, \theta)$, we can rewrite it as $\pi(\theta)\mathbf{P}(\mathbf{Y}|\theta)$. There are 3 steps for us to follow, given sample Y_i:

1. Sample a list of parameter $\vec{\theta} = \{\theta_1, \ldots, \theta_n\}$ with $\theta_i \sim \pi(\theta|Y_1, \ldots, Y_{n_0})$
2. For each θ_i, sample $\tilde{Y_i} \sim \mathbf{P}(\mathbf{Y}|\theta, \mathbf{Y_1}, \ldots, \mathbf{Y_{n_0}})$
3. return $\{(\theta_i, \tilde{Y_i})\}_{i=1}^n$

Sometimes we may find our simulation does not align with our empirical data. Firstly, it's possible we were unlucky, so we can run this entire experiment k times, meaning a total of $k \cdot n$ pairs of (θ, Y) will be generated. It's also possible that we need to change our model, both the sampling model and the prior have room for adjustment.