Bayesian Statistics

Siddhant Garg

January 2018

1 Bayesian Procedures

1.1 Introduction

In Bayesian Statistics, we try to determine something about a parameter of a distribution based on some prior knowlegde about that parameter and some data that depends on the parameter. In Bayesian inference, the parameter is treated as a random variable Θ.

1.2 Prior and Posterior Distributions

Consider a random variable X that has a distribution of probability that depends upon the symbol θ, where θ is an element of a well-defined set Ω. Let us now introduce a random variable Θ that has a distribution of probability over the set Ω. We now look upon θ as a possible value of the random variable Θ.

$$X|\Theta \sim f(x|\theta)$$
$$\Theta \sim h(\theta)$$

The pdf $h(\theta)$ is called the **prior** pdf of θ. Moreover, we now denote the pdf of X by $f(x|\theta)$ since we think of it as a conditional pdf of X, given $\Theta = \theta$.

Suppose that $X_1, \ldots, X_n$ is a random sample from the conditional distribution of $\mathbf{X}$ given $\Theta = \theta$ with pdf $f(x|\theta)$. Thus we can write the joint conditional pdf of $\mathbf{X}$, given $\Theta = \theta$, as

$$L(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta)$$

Thus the joint pdf of $\mathbf{X}$ and Θ is

$$g(\mathbf{x}, \theta) = L(\mathbf{x}|\theta)h(\theta)$$

If Θ is a random variable of the continuous type, the joint marginal pdf of $\mathbf{X}$ is given by

$$g_1(\mathbf{x}) = \int_{-\infty}^{\infty} g(\mathbf{x}, \theta)d\theta$$

The conditional pdf of Θ, given the sample $\mathbf{X}$, is

$$k(\theta|\mathbf{x}) = \frac{g(\mathbf{x}, \theta)}{g_1(\mathbf{x})} = \frac{L(\mathbf{x}|\theta)h(\theta)}{g_1(\mathbf{x})}$$

The above pdf is called **posterior pdf**. The prior distribution reflects the subjective belief of Θ before the sample is drawn, while the posterior distribution is the conditional distribution of Θ after the sample is drawn.

1.3 Bayesian Point Estimation

Suppose we want a point estimator of θ. From the Bayesian viewpoint, this really amounts to selecting a decision function δ, so that $\delta(x)$ is a predicted value of θ when both the computed value $\mathbf{x}$ and the conditional pdf $k(\theta|\mathbf{x})$ are known.

The choice of the decision function should depend upon a loss function $L[\theta, \delta(\mathbf{x})]$. A Bayes estimate is a decision function δ that minimizes

$$E\{L[\theta, \delta(x)]|\mathbf{X} = \mathbf{x}\} = \int_{-\infty}^{\infty} L[\theta, \delta(\mathbf{x})]k(\theta|\mathbf{x})d\theta$$

If Θ is a random variable of the continuous type. That is,

$$\delta(\mathbf{X}) = Argmin \int_{-\infty}^{\infty} L[\theta, \delta(\mathbf{x})] k(\theta|\mathbf{x}) d\theta$$

If $L[\theta, \delta(\mathbf{x})] = [\theta - \delta(\mathbf{x})]^2$, then $\delta(\mathbf{x}) = \mathbf{E}(\Theta|\mathbf{x})$
If $L[\theta, \delta(\mathbf{x})] = |\theta - \delta(\mathbf{x})|$, then median of the conditional distribution of Θ given $\mathbf{X} = \mathbf{x}$ is the Bayes Solution.

1.4 Bayesian Interval Estimation

1.4.1 Confidence Interval

Theorem 1.1 (Central Limit Theorem) *Let $X_1, \ldots, X_n$ denote the observations of a random sample from a distribution that has mean μ and finite variance σ^2. Then the distribution function of the random variable $W_n = (\bar{X} - \mu)/(\sigma/n)$ converges to Φ, the distribution function of the $N(0,1)$ distribution, as $n \to \infty$.*

Large Sample Confidence Interval for mean μ
Suppose $X_1, \ldots, X_n$ is a random sample on a random variable X with mean μ and variance σ^2.

$$Z_n = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Distribution of Z_n is approximately $N(0,1)$.
Let α be such that $\alpha/2 = P(Z_n > z_{\alpha/2})$.

$$1 - \alpha \approx P\Big(-z_{\alpha/2} < Z_n < z_{\alpha/2}\Big)$$

$$1 - \alpha \approx P\Big(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < Z_n < \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\Big)$$

Again, letting $\bar{x}$ and s denote the realized values of the statistics $\bar{X}$ and S, respectively, after the sample is drawn, an approximate $(1-\alpha)100\%$ confidence interval for μ is given by,

$$(\bar{x} - z_{\alpha/2} s\sqrt{n}, \bar{x} + z_{\alpha/2} s\sqrt{n})$$

This is called a **large sample** confidence interval for μ.
If an interval estimate of θ is desired, we can find two functions $u(\mathbf{x})$ and $v(\mathbf{x})$ so that the conditional probability

$$P[u(\mathbf{x}) < \Theta < v(\mathbf{x})|\mathbf{X} = \mathbf{x}] = \int_{u(\mathbf{x})}^{v(\mathbf{x})} k(\theta|\mathbf{x})$$

These intervals are often called **credible** or **probability intervals**, so as not to confuse them with confidence intervals.

1.5 More Bayesian Terminology and Ideas

Definition: A class of prior pdfs for the family of distributions with pdfs $f(x|\theta)$, $\theta \in \Omega$, is said to define a **conjugate family of distributions** if the posterior pdf of the parameter is in the same family of distributions as the prior.

1.6 Gibbs Sampler

Integration techniques play a significant role in Bayesian inference. Hence, we now touch on some of the Monte Carlo techniques used for integration in Bayesian inference.

Consider the Bayes Model
$Y|\theta \sim N(\theta, \sigma^2/n)$
$\Theta \sim h(\theta) \propto b^{-1} exp\{-(\theta - a)/b\}/(1 + exp\{-[(\theta - a)/b]\})^2, -\infty < \theta < \infty$, a and $b > 0$ are unknown.
Let $w(\theta) = f(y|\theta)$ then,

$$\delta(y) = \frac{E[\Theta w(\Theta)]}{E[w(\Theta)]}$$

The estimation can be carried out by simple Monte Carlo. Independently, generate $\Theta_1, \Theta_2, \ldots, \Theta_m$ from $h(\theta)$. Then form a random variable

$$T_m = \frac{m^{-1}\sum_{i=1}^{m}\Theta_i w(\Theta_i)}{m^{-1}\sum_{i=1}^{m} w(\Theta_i)}$$

By the Weak Law of Large Numbers and Slutsky's Theorem , $T_m \rightarrow \delta(y)$, in probability. The value of m can be quite large. Thus simple Monte Carlo techniques enable us to compute this Bayes estimate.

Theorem 1.2 *Suppose we generate random variables by the following algorithm:*

$$1.Generate Y \sim f_Y(y),$$

$$2.Generate X \sim f_{X|Y}(x|Y)$$

Then X has the pdf $f_X(x)$.

Example: Suppose the random variable X has pdf

$$f_X(x) = \left\{ \begin{array}{cc} 2e^{-x}(1-e^{-x}) & 0 < x < \infty \\ 0 & elsewhere \end{array} \right\}$$

$$f_Y(y) = \left\{ \begin{array}{cc} 2e^{-2y} & 0 < x < \infty \\ 0 & elsewhere \end{array} \right\}$$

$$f_{X|Y}(x|y) = \left\{ \begin{array}{cc} e^{-(x-y)} & y < x < \infty \\ 0 & elsewhere \end{array} \right\}$$

Generating random variables from the above algorithm for 10,000 samples (python code),
$\bar{x} = 1.50848164922$
$s = 1.12441764787$
Confidence Interval for $E(X)$ is $(1.48998662476, 1.52697667369)$

Gibbs Sampler Algorithm:Let m be a positive integer, and let X_0 , an initial value, be given. Then for $i = 1, 2, \ldots, m$,

$$1.Generate Y_i|X_{i-1} \sim f(y|x).$$

$$2.Generate X_i|Y_i \sim f(x|y).$$

as $i \rightarrow \infty$

$$Y_i \xrightarrow{D} Y \sim f_Y(y)$$

$$X_i \xrightarrow{D} X \sim f_X(x)$$

Example: Suppose $Y|X \sim \Gamma(\alpha + x, 1/2)$ and $X|Y \sim Poisson(Y)$. Then running gibbs sampler algorithm for $m = 6000$ simulations,

$$1.Generate Y_i|X_{i-1} \sim \Gamma(\alpha + x, 1/2).$$

$$2.Generate X_i|Y_i \sim Poisson(Y).$$

then for $n = 3000$

$$\bar{Y} = (m-n)^{-1}\sum_{i=n+1}^{m} Y_i \xrightarrow{P} E(Y)$$

$$\bar{X} = (m-n)^{-1}\sum_{i=n+1}^{m} X_i \xrightarrow{P} E(X)$$

You can find the python code here .

Parameter	Estimate	Sample Estimate	Sample Variance	Approximate 95% confidence Interval
E(Y) $= \alpha = 10$	$\bar{y}$	10.030150	10.191269	(9.9342,10.1260)
E(X) $= \alpha = 10$	$\bar{x}$	10.049333	19.86618	(9.91548,10.1831)

1.7 More Bayesian Methods

1.7.1 Hierarchical Bayes

The prior pdf has an important influence in Bayesian inference. One way of having more control over the prior is to model the prior in terms of another random variable. This is called the **hierarchical Bayes** model, and it is of the form

$$X|\theta \sim f(x|\theta)$$

$$\Theta|\gamma \sim h(\theta|\gamma)$$

$$\Gamma \sim \psi(\gamma)$$

With this model we can exert control over the prior $h(\theta|\gamma)$ by modifying the pdf of the random variable Γ.

The parameter γ can be thought of a nuisance parameter. It is often called a **hyperparameter**. As with regular Bayes, the inference focuses on the parameter θ; hence, the posterior pdf of interest remains the conditional pdf $k(\theta|\mathbf{x})$.

$$\begin{aligned} g(\theta,\gamma|\mathbf{x}) &= \frac{g(\mathbf{x},\theta,\gamma)}{g(\mathbf{x})} \\ &= \frac{g(\mathbf{x}|\theta,\gamma)g(\theta,\gamma)}{g(\mathbf{x})} \\ &= \frac{f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)}{g(\mathbf{x})} \end{aligned} \tag{1}$$

Therefore the posterior pdf is given by,

$$k(\theta|\mathbf{x}) = \frac{\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)d\theta}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)d\theta d\psi}$$

Furthermore, assuming squared-error loss, the Bayes estimate of $W(\theta)$ is

$$\delta_W(\mathbf{x}) = \frac{\int_{\infty}^{\infty}\int_{-\infty}^{\infty} W(\theta)f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)d\theta}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)\psi(\gamma)d\theta d\psi}$$

To obtain the Bayes Estimate of $W(\theta)$, we refer to Gibbs Sampler Algorithm. For $i = 1, 2, \ldots, m$, at the i^{th} step

$$\Theta_i|\mathbf{x},\gamma_{i-1} \sim g(\theta|\mathbf{x},\gamma_{i-1})$$

$$\Gamma_i|\mathbf{x},\theta_i \sim g(\gamma|\mathbf{x},\theta_i)$$

as $i \to \infty$

$$\Theta_i \xrightarrow{\text{D}} k(\theta|\mathbf{x})$$

$$\Gamma_i \xrightarrow{\text{D}} g(\gamma|\mathbf{x})$$

Furthermore the arithmetic average

$$\frac{1}{m-n}\sum_{i=n+1}^{m} W(\Theta_i) \xrightarrow{\text{P}} E[W(\Theta|\mathbf{x})] = \delta_w(\mathbf{x}) \text{ as } m \to \infty$$

Because of the **Monte Carlo** generation these procedures are often called **MCMC**, for **Markov Chain Monte Carlo** procedures.

1.7.2 Emperical Bayes

The empirical Bayes model consists of the first two lines of the hierarchical Bayes model; i.e.,

$$\mathbf{X}|\Theta \sim f(\mathbf{x}|\theta)$$

$$\Theta|\gamma \sim h(\theta|\gamma)$$

Instead of attempting to model the parameter γ with a pdf as in hierarchical Bayes, empirical Bayes methodology estimates γ based on the data as follows.

$$\begin{aligned} g(\mathbf{x}, \theta|\gamma) &= \frac{f(\mathbf{x}|theta)h(\theta|\gamma)\psi(\gamma)}{\psi(\gamma)} \\ &= f(\mathbf{x}|\theta)h(\theta|\gamma) \end{aligned} \tag{2}$$

Consider, then, the likelihood function

$$m(\mathbf{x}|\gamma) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)d\theta$$

Using the pdf $m(\mathbf{x}|\gamma)$, we obtain an estimate $\widehat{\gamma} = \widehat{\gamma(\mathbf{x})}$, usually by the method of maximum likelihood. For inference on the parameter θ, the empirical Bayes procedure uses the posterior pdf $k(\theta|\mathbf{x}, \widehat{\gamma})$.