Brief Introduction of Bayesian Estimation

Yao Weitong

Bayesian statistics is a huge system. Different authors have slightly different interpretations of Bayesian estimation. This article is a concise synthesis of many textbooks, contains the recommended chapters in those books, and is a brief introduction to Bayesian estimates from the perspective of non-statistical major. You can follow the textbooks listed in the Reference for in-depth study.

1 Importance of prior probability

As an example, when a guy purchases a life insurance, the insurance company will estimate the probability of the insured's risk and determine the premium. According to the classical frequentists, the person is considered as a certain individual of the sample with same features of age, income, etc.. However, information asymmetry always arises, and that’ s the limitation of excluding the personal prior information of the insured. If the insurance company fails to conduct a prior investigation, the insured may have individual differences compared with the sample; and the investigation of the medical records and living habits of the insured in the past is capable to update the estimated probability. Another example stresses the critical role played by prior probabilities in estimation. The mathematician named John Craven proposed a search scheme for a successful search of lost USS Scorpion (SSN-589), using the Bayesian formula. He used the probability map of the sea area, which was divided into many small squares with two probability p and q. p is the probability that the submarine lies in this square. q is the probability that it was searched if the submarine is in this grid. If a grid is searched and no trace of the submarine is found, then according to the Bayesian formula, the probability of existence of the lattice submarine will decrease$p' = \frac{p(1-q)}{(1-p)+p(1-q)}$; since the sum of probabilities is 1, then the probability value of other grid will be rise$r' = \frac{r}{1-pq} > r$ (change in prior probability)[1].

1.1 Frequentists and Bayesians

- For standard Frequentists, probability requires a lot of repeated sampling and the parameters are considered fixed but unknown. Frequency indicates probability. The Bayesians believe that the data should be fixed, the parameters are random, the frequency is a description of the degree of uncertainty, and the probability should be based on our understanding of the world or previous studies.

- In practice, Frequentists determine the specific values (e.g. expectations, variances, etc.) of the parameters by optimization criteria (e.g. maximize likelihood function), and derive conclusions from a single data set. In the Bayesian view, given the observation data, the first step is to model a prior distribution, and then repeatedly apply Bayes' theorem, combine the data set (as a prior probability distribution), and continuously update the probability to calculate posterior probability distribution. The result is a distribution rather than a specific value, so it needs sampling for further statistical inference.

Take an example; suppose we want to study the influence of Trump's Twitter content on the exchange rate of US dollars against RMB. Firstly, the text analysis method is used to extract the keyword frequency of the emotional judgment and intention, and regressed on the exchange rate fluctuation of the corresponding time period. This is what the Frequentists usually do. However, Bayesians believe that all other factors that affect exchange rate movements should be considered to construct a prior probability distribution (using the same data set of other factors used by Frequentists), and the content of the Twitter is used to update beliefs[2].

2 Bayes's Theorem and Bayesian Estimation

For a random vector θ (as parameters) and random vector D (as data set), the Bayes's Theorem indicates:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

$p(\theta \mid D)$is the posterior probability distribution; $p(D \mid \theta)$ is considered the function of θ, always denoted as likelihood function$L(\theta;\, D)$, indicating the probability of observation under certain θ; and $p(\theta)$ is prior probability distribution. The denominator $p(D)$ is a normalized constant, which makes sure posterior probability integral equal to 1, and posterior distribution can be directly proportional to the numerator(or said "density kernel"):

$$posterior \propto likelihood \times prior$$

2.1 Example of Gauss Distribution Estimation[3][4]

Suppose the random sample is $D = (d1, d2, ..., dn)'$, where D is the Gaussian distribution with known variance σ^2 and unknown mean μ, find the meanμ. Frequentists use MLE to solve the problem, whilst Bayesians prefer to applying prior distribution and implementing with three steps.

1. Model Prior Distribution $p(\theta)$:

Since target distribution is normal and conjugative, it's reasonable to set $\theta \sim N(\mu_0, \sigma_0^2)$; define ***presion*** of θ as follows, the larger the *presion* h is, the smaller the σ_0^2 will be.

$$h \equiv \frac{1}{\sigma_0^2}$$

the prior distribution of θ is:

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left\{-(\theta - \mu_0)^2/2\sigma_0^2\right\} \propto \ exp\left\{-h(\theta - \mu_0)^2/2\right\}$$

2. Model Likelihood Function $L(\theta;\ D)$(i.i.d):

$$L(\theta;\ D) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2} exp\left\{-(d_i - \theta)^2/2\sigma^2\right\}$$

$$= (2\pi\sigma^2)^{-n/2} exp\left\{-\sum_{i=1}^{n}(d_i - \theta)^2/2\sigma^2\right\}$$

$$\propto exp\left\{-\sum_{i=1}^{n}(\bar{d} - \theta)^2/2\sigma^2\right\} \ \propto \ exp\left\{-h^*(\bar{d} - \theta)^2/2\right\}$$

where $h^* = n/\sigma^2$. Obviously, the likelihood function (rather a PDF of θ) has similar formation as prior distribution, they are conjugate distribution. Since the multiplication of Gaussian functions is the summation of exponents, thus posterior probability is also a Gaussian distribution.

3. Model Posterior Distribution $p(\theta \mid D) = N(\theta \mid \mu_N, \sigma_N^2)$

$$p(\theta \mid D) \ \propto \ L(\theta;\ D)p(\theta)$$

$$\propto exp\left\{-h^*(\bar{d} - \theta)^2/2\right\} \cdot exp\left\{-h(\theta - \mu)^2/2\right\}$$

$$= exp\left\{-\frac{1}{2}\left[h^*(\bar{d} - \theta)^2 + h(\theta - \mu)^2\right]\right\} \ \propto \ \bar{h}(\theta - \bar{\mu})^2$$

where $\bar{h} = h + h^*$, $\bar{\mu} \equiv (h\mu + h^*\bar{d})/\bar{h}$, another forms are:

$$\mu_N = \bar{\mu} = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$$

$$\bar{h} = \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

From the inference shown above, it's clear that $\mu_N \in [\ \mu_0, \mu_{ML}]$, when$N \to 0, \mu_N \to \mu_0$; when $N \to \infty, \mu_N \to \mu_{ML}$, at the same time, $\bar{h} \to \infty$ and $\sigma_N^2 \to 0$. When N tends to be infinite, it connotes the presion of the sample is more and more important for the posterior presion, and the posterior distribution is almost not affected by the prior distribution, which helps to offset the shortcomings of subjective prior distribution. Another question is how we can calculate the posterior probability distribution.

2.2 Estimation of Posterior Probability Distribution[5][6]

2.2.1 *Monte Carlo Integral*

Let $\theta^{(s)}$ for $s = 1, ..., S$ be a random sample from $p(\theta \mid D)$, and define $\hat{I}_{MC} = \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)})$, then $\hat{f}_s$ converges to $E[f(\theta) \mid D]$ as S goes to infinity.
When the posterior distribution has an analytical formula, the Monte Carlo integral method is commonly used for the posterior mean calculation.Using the theorem, it allows to approximate $E[f(\theta) \mid D]$, and the numerical standard error

$$\sqrt{S}\left\{\hat{f}_s - E[f(\theta) \mid D]\right\} \to N(0, \sigma_f^2)$$

can ensure the difference is sufficiently small.

2.2.2 *Markov chain Monte Carlo (Gibbs sampler)*

When the integrals of posterior distribution is so complex or it has no analytical formula, it's necceary to derive a random sample from posterior distribution for further inference. One of popular sampler is ***Markov chain Monte Carlo sampling***.
Suppose the distribution of $X = (X_1, X_2, ...,X_n)$ is $f(x)$, for any fixed $T \in N\left\{1, 2, ..., n\right\}$, given $X_{-T} = x_{-T}$,define $\tilde{X} = (\tilde{x}'_1, \tilde{x}_2, ...,\tilde{x}_n)$, for any testible B:

$$P(\tilde{X} \in B) = \int_B \pi(\tilde{x}_{-T})\pi(\tilde{x}_{-T} \mid \tilde{x}_{-T})d\tilde{x} = \int_B \pi(\pi(\tilde{x})d\tilde{x} = \pi(B)$$

thus, from X to X', they have stationary distributions (the PDF is unchanged), and it's called ***Gibbs sampler.***

References

[1] 邓一硕,关菁菁,刘辰昂,邱怡轩,施涛,熊熹,周祺. 统计之都创作小组: 失联搜救中的统计数据分析.

[2] Jakevdp(Jake Vanderplas) .2014.*Frequentism and Bayesianism: A Practical Introduction.*GitHub.

[3] 陈强编.2010.《高级计量经济学及stata应用》,Chapter19,31.

[4] Bishop, C. M. 2011. *Pattern Recognition and Machine Learning.* Springer.

[5] Gary Koop. 2003. *Bayesian Economics.*

[6] 朱明惠, 林静著.《贝叶斯计量经济学模型》