Statistics: Principles, Methods and R (II)

Gao Fengnan[1 2]

27.02.2017

[1] School of Data Science, Fudan University

[2] Shanghai Center for Mathematical Sciences

Overview

Bayesian Inference

The Bayesian Method

Bayesian inference is usually carried out in the following steps.

1. Choose a probability density $\pi(\theta)$—the *prior distribution*—to express our beliefs about a parameter θ before any data
2. Choose a statistical model $f(x|\theta)$ that reflects our belief about x given θ
3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the *posterior distribution* $\pi(\theta|X_1, \ldots, X_n)$

Recall Bayes' theorem

Theorem (Bayes' Theorem)

For two events A and B with $\mathbb{P}(B) \neq 0$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Bayesian Inference

Bayesian Procedure

Keep in mind that the parameter θ is random!

- Θ–the parameter, X—data
- Suppose θ only takes discrete values,

$$\mathbb{P}(\Theta = \theta | X = x) = \frac{\mathbb{P}(X = x, \Theta = \theta)}{\mathbb{P}(X = x)}$$
$$= \frac{\mathbb{P}(X = x | \Theta = \theta)\mathbb{P}(\Theta = \theta)}{\sum_\theta \mathbb{P}(X = x | \Theta = \theta)\mathbb{P}(\Theta = \theta)}$$

- Suppose continuous θ, we use density function

$$\pi(\theta | x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

- Suppose n IID observations $X^{(n)} := \{X_1, \ldots, X_n\}$ and write non-random $x^{(n)} = \{x_1, \ldots, x_n\}$, then the likelihood function is

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = L_n(\theta).$$

Bayesian Procedure Continued

- We get

$$\pi(\theta|x^{(n)}) = \frac{f(x^{(n)}|\theta)\pi(\theta)}{\int f(x^n|\theta)\pi(\theta)d\theta} = \frac{L_n(\theta)\pi(\theta)}{c_n} \propto L_n(\theta)\pi(\theta)$$

  where $c_n = \int L_n(\theta)\pi(\theta)d\theta$ is called the normalizing constant.
- Posterior is proportional to Likelihood times Prior.
- With $L_n(\theta)\pi(\theta)$, c_n can always be recovered.
- Compare with normal distribution, the density is proportional to $\exp\big(-x^2/(2\sigma^2)\big)$, we can recover the full density by calculating the integral

$$\int \exp\big(-x^2/(2\sigma^2)\big)dx.$$

Examples

Example (Bernoulli Experiment)

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, how to estimate p?

- The MLE gives $\hat{p}_n = \overline{X}_n$
- The Bayesian way—specify a prior π on p first—a density taking value on all possible p's
- We take uniform prior on $[0, 1]$, i.e., $\pi(p) = 1_{[0,1]}(p)$
- Any other possible prior for p?

Bayesian Point Estimators

How to obtain an estimator from the posterior distribution?

- The Bayes estimator $\hat{\theta}_n^{\mathrm{B}} = \mathbb{E}_{\pi(\cdot|X^{(n)})}[\theta]$—the posterior mean
- The posterior mode—the maximizer of the posterior

$$\hat{\theta}^{\mathrm{PO}} = \arg\max_{\theta} \pi(\theta|X^{(n)})$$

Example (Normal Experiment)

Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with known σ^2, how to estimate μ?

- The MLE gives, again, sample mean $\hat{\mu}_n = \overline{X}_n$
- What possible priors can we put on $\mu \in \mathbb{R}$?
- Take a Normal prior $N(a, b^2)$ on μ, what is the posterior?
- Any other possible prior for μ?

Conjugate priors

Definition

When the prior and posterior are in the same family, we say the prior is *conjugate* with respect to the model.

- Beta prior is conjugate WRT the Bernoulli model.
- Normal prior is conjugate WRT the Normal model.
- Laplace prior is not conjugate WRT the Normal model.
- Laplace prior is conjugate WRT the Laplace model. Verify this in exercise.

Credible Intervals

- For frequentists, a *confidence interval* for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X^{(n)}$ and $b = b(X^{(n)}$ are functions of the data such that

$$\mathbb{P}(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta.$$

- For Bayesian, a *credible interval* for a parameter θ is an interval (a, b) such that $\int_{\infty}^{a} \pi(\theta | X^{(n)}) d\theta = \int_{b}^{\infty} \pi(\theta | X^{(n)} d\theta = \alpha/2$, then

$$\mathbb{P}\big(\theta \in (a, b)\big) = \int_{a}^{b} \pi(\theta | X^{(n)}) d\theta = 1 - \alpha.$$

Let $X^{(n)} = \{X_1, \ldots, X_n\}$ be IID with density $f(x;\theta)$ under parameter θ.

Definition

The **likelihood function** is defined by

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

The **log-likelihood function** is defined by $l_n(\theta) = \log L_n(\theta)$.

Definition (maximum likelihood estimator)

The **maximum likelihood estimator** MLE $\hat{\theta}_n$ is the value of θ that maximizes $L_n(\theta)$

MLE Example

Example

Let $X un = \{X_1, \ldots, X_n\}$ be iid Bernoulli(p).

- The probablity function is $f(x; p) = p^x(1-p)^{1-x}$ for $x = 0,\ 1$.

Example

Let $X^{(n)} = \{X_1, \ldots, X_n\}$ be IID $N(\mu, \sigma^2)$. How to calculate the MLE of μ and σ^2?

- Recall the normal density $f(x; \mu, \sigma^2) = \exp\big(-(x-\mu)^2/(2\sigma^2)\big)/\sqrt{2\pi\sigma^2}$

Fisher Information

Definition

The **score function** is

$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}.$$

The **Fisher information** is

$$I_n(\theta) = \mathrm{Var}\Big(\sum_{i=1}^{n} s(X_i,\theta)\Big) = \sum_{i=1}^{n} \mathrm{Var}\Big(s(X_i;\theta)\Big).$$

Theorem

$I_n(\theta) = nI(\theta)$. Also

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right) = -\int \left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right) f(x;\theta)dx.$$

Bernoulli Example

Example

For $X^{(n)}$ IID Bernoulli(p)

- the score function is

$$s(X;p) = \frac{\partial}{\partial p}\big(x \log p + (1-x)\log(1-p)\big)$$

- The derivative of the score function is

$$s'(X;p) = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2}.$$

- The Fisher information is

$$I_1(p) = -\mathbb{E}_\theta[s'(X;p)] = \frac{1}{p(1-p)}.$$

Theorem (Asymptotic Normality of the MLE)

Let $se = \sqrt{Var(\hat{\theta}_n)}$. *Under some regularity conditions, the following hold:*

1. $se \approx \sqrt{1/I_n(\theta)}$ *and*

$$\frac{\hat{\theta}_n - \theta}{se} \rightsquigarrow N(0, 1).$$

2. *Let* $\hat{se} = \sqrt{1/I_n(\theta_n)}$. *Then,*

$$\frac{\hat{\theta}_n - \theta}{\hat{se}} \rightsquigarrow N(0, 1).$$

Theorem

For the MLE $\hat{\theta}_n$ *and* $z_{\alpha/2}$ *is the* $1 - \alpha/2$ *quantile and* $\hat{se} = \sqrt{1/I_n(\hat{\theta}_n)}$. *Let*

$$C_n = \left(\hat{\theta}_n - z_{\alpha/2}\hat{se}, \hat{\theta}_n - z_{\alpha/2}\hat{se}\right).$$

Then $\mathbb{P}_\theta(\theta \in C_n) ra 1 - \alpha$ *as* $n \to \infty$.

Theorem

Let $\hat{\theta}_n$ be the MLE and let $\hat{se} = 1/\sqrt{nI(\hat{\theta}_n)}$. Under appropriate regularity conditions, the posterior is approximately Normal with mean $\hat{\theta}_n$ and standard error $\hat{se}$. Hence, the Bayes estimator $\theta_n^B \approx \hat{\theta}_n$. Also, if $C_n = (\hat{\theta}_n - z_{\alpha/2}\hat{se}, \hat{\theta}_n + z_{\alpha/2}\hat{se})$ is the asymptotic frequentist $1-\alpha$ confidence interval, then C_n is also an approximate credible interval such that

$$\Pi(\theta \in C_n | X^{(n)}) \to 1 - \alpha,$$

where $\Pi(\theta | X^{(n)})$ is the corresponding distribution function of the posterior $\pi(\theta | X^{(n)})$.

The above theorem essentially tells us under some regularity conditions, the credible interval is asymptotically the **same** as the frequentist confidence interval.

frequentist	Bayesian
Probability Refers to limiting relative frequencies. Probabilities are objective properties of the real world.	Probability describes degrees of belief, not limiting frequency.
Parameters are fixed, unknown constants.	We can make probability statements about parameters, even though they are fixed constants.
Statistical procedure should be designed to have well-defined long run frequency properties.	We make inferences about a parameter θ by producing a probability distribution for θ.

Frequentist v.s. Bayesian