Notes for Bayesian Data Analysis 3

Yuanxing Cheng

August 24, 2022

Contents

1 Probability and inference

1.1 The three steps of Bayesian data analysis

- Full probability model: a joint probability distribution of all observable and unobservable, *remember the underlying knowledge and data collection process*
- Conditioning on observed data: get posterior distribution, i.e. the conditional probability distri of the unobserved quantities, *given the observed data*
- Evaluating the fit of the model, and posterior. *How good? Sensitivity to assumptions?*

1.2 General notation for statistical inference

Population, sample, estimates, parameters, etc.

Parameters, data, and predictions

Denote θ as unobservable parameter vector, y as the observed data. $\tilde{y}$ as unknown but observable data.

Observational units and variables

Data, of n objects. Write $y = (y_1, \ldots, y_n)$ or $y^\top$. Notice y_i itself could be a vector, then the entire y is a n row matrix.

Exchangeability

n values y_i may be regarded as exchangeable. Then the joint pdf $p(y_1, \ldots, y_n$ is invariant to permutations of indexes.

Explanatory variables

Or *covariates*. Use X to denote the entire set of explanatory variables for all n units. If there're k explanatory variables, then X is a matrix of $n \times k$.

Hierarchical modeling

Or *multilevel models*. It's possible here to assume the exchangeability at each level of units.

1.3 Bayesian inference

Conclude about a parameter vector θ or unobserved data $\tilde{y}$ in probability statements, usually denoted as $p(\theta \mid y)$ or $p(\tilde{y} \mid y)$. And also implicitly condition on the known values x.

Probability notation

$p(\cdot \mid \cdot)$ denotes a conditional pdf w/ the arguments determined by the context. $p(\cdot)$ usually denotes a marginal distribution. And if for example $\theta \sim \mathcal{N}(\mu, \sigma^2)$, we also write $p(\theta) = \mathcal{N}(\theta \mid \mu, \sigma^2)$.

The geometric mean is $\exp\left(\mathrm{E}\left[\log \theta\right]\right)$

Bayes' rule

Of prior $p(\theta)$ and sample distribution $p(y \mid \theta)$, we have

$$p(\theta, y) = p(\theta)p(y \mid \theta).$$

Then by Bayes' rule we have the *posterior*:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{p(y)}, \tag{1.1}$$

where $p(y) = \sum_\theta p(\theta)p(y \mid \theta) = \int p(\theta)p(y \mid \theta)\, \mathrm{d}\theta$ is the total probability. Usually we write above in the following form

$$p(\theta \mid y) \propto p(\theta)p(y \mid \theta). \tag{1.2}$$

Prediction

The *prior predictive distribution* is

$$p(y) = \sum_\theta p(y, \theta) = \sum_\theta p(\theta)p(y \mid \theta) = \int p(y, \theta)\, \mathrm{d}\theta = \int p(\theta)p(y \mid \theta)\, \mathrm{d}\theta. \tag{1.3}$$

Then we predict an observable $\tilde{y}$. Then its distribution is *posterior predictive distribution*, with formula

$$\begin{aligned} p(\tilde{y} \mid y) &= \int p(\tilde{y}, \theta \mid y)\, \mathrm{d}\theta \\ &= \int p(\tilde{y} \mid \theta, y)p(\theta \mid y)\, \mathrm{d}\theta \quad \text{Given } \theta,\ y \text{ and } \tilde{y} \text{ are independent} \\ &= \int p(\tilde{y} \mid \theta)p(\theta \mid y)\, \mathrm{d}\theta \end{aligned} \tag{1.4}$$

Likelihood

From above 1.4, data y affect the posterior only through $p(y \mid \theta)$, i.e., the likelihood function when y is fixed. This is the likelihood principle.

Likelihood and odds ratio

Define *posterior odds* for two parameters θ_1 and θ_2 to be

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \frac{p(\theta_1)p(y \mid \theta_1)/p(y)}{p(\theta_2)p(y \mid \theta_2)/p(y)} = \frac{p(\theta_1)p(y \mid \theta_1)}{p(\theta_2)p(y \mid \theta_2)}, \tag{1.5}$$

The later part is *likelihood ratio* thus we have: *posterior odds=prior odds times likelihood ratio*

1.4 Discrete examples: genetics and spell checking

2 examples,

1.5 Probability as a measure of uncertainty

Basically, the idea is the bayesian methods are more subjective due to the reliance on a prior distribution.

1.6 Example: probability from football point spreads

1.7 Example: calibration for record linkage

1.8 Some useful results from probability theory

Regarding the joint density, we have the following

$$p(u) = \int p(u, v)\, \mathrm{d}v$$
$$p(u, v, w) = p(u \,|\, v, w)p(v \,|\, w)p(w)$$
$$p(u, v \,|\, w) = p(v \,|\, u, w)P(u \,|\, w) = p(u \,|\, v, w)p(v \,|\, w)$$

In vector calculus, we define covariance matrix as

$$\mathrm{Cov}\,[u] = \int (u - \mathrm{E}\,[u])(u - \mathrm{E}\,[u])^\top p(u)\, \mathrm{d}u$$

And conditional expectation is a function of conditioned variables. For example $\mathrm{E}\,[u \,|\, v]$ is a function of v. And we have the following formula

$$\mathrm{E}\,[u] = \mathrm{E}\,[\mathrm{E}\,[u \,|\, v]] \tag{1.6}$$
$$\mathrm{E}\,[u] = \int\int u \cdot p(u, v)\, \mathrm{d}u\, \mathrm{d}v = \int\int u \cdot p(u \,|\, v)\, \mathrm{d}u\; p(v)\, \mathrm{d}v \tag{1.7}$$
$$= \int \mathrm{E}\,[u \,|\, v]\, p(v)\, \mathrm{d}v \tag{1.8}$$
$$\mathrm{Var}\,[u] = \mathrm{E}\,[\mathrm{Var}\,[u \,|\, v]] + \mathrm{Var}\,[\mathrm{E}\,[u \,|\, v]] \tag{1.9}$$

Transformation of variables

Denote $p_u(u)$ the density for u and transformation is $v = f(u)$. If p_u is discrete and f is one-to-one, then $p_v(v) = p_u(f^{-1}(v))$. And if f is many-to-one, then we need to sum those probabilities of same value of $f(u)$.

And if p_u is continuous, and f is one-to-one, then $p_v(v) = |J|\, p_u(f^{-1}(v))$ where $|J|$ is the absolute value of the determinant of Jacobian, and can be denoted as $\frac{\partial u}{\partial v}$ even in vector form.

A useful 1-d function, the logarithm

$$\mathrm{logit}(u) = \log(\frac{u}{1-u}) \tag{1.10}$$

with the inverse $\mathrm{logit}^{-1}(v) = \frac{e^v}{1+e^v}$.

Another useful function is the probit transformation $\Phi^{-1}(u)$ where Φ is the standard normal cdf.

1.9 Computation and software

Summarizing inferences by simulation

Sampling using the inverse cumulative distribution function

For 1-d distribution $p(v)$ with cdf $F(v)$, the inverse cdf F^{-1} can be used to obtain random samples from the distribution p.

1. Draw a random value U from standard uniform
2. $v = F^{-1}(U)$ and this v will be a random draw from p.

Simulation of posterior and posterior predictive quantities

1.10 Bayesian inference in applied statistics

1.11 Selected Exercises

2 Single-parameter models

2.1 Estimating a probability from binomial data

Appendices

A Standard probability distribution

A.1 Continuous distribution

Uniform

Standard uniform $U(0,1)$, equal possibilities. If $u \sim U(0,1)$, then $\theta = a + (b-a)u \sim U(a,b)$. A noninformative distribution is obtained in the limit as $a \to \infty$ and $b \to \infty$.

Univariate normal

Standard normal $\mathcal{N}(0,1)$. If $z \sim \mathcal{N}(0,1)$ then $\theta = \mu + \sigma z \sim \mathcal{N}(\mu, \sigma^2)$. A noninformative (flat distribution) is obtained in the limit as $\sigma \to \infty$. And $\sigma = 0$ corresponds to point mass at θ.

Useful properties: If two independent $\theta_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\theta_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $\theta_1 + \theta_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. And mixture property states that if $\theta_1 \mid \theta_2 \sim \mathcal{N}(\theta_2, \sigma_1^2)$ and $\theta_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $\theta_1 \sim \mathcal{N}(\mu_2, \sigma_1^2 + \sigma_2^2)$.

Lognormal

When $\log\theta \sim \mathcal{N}(\mu, \sigma^2)$, θ is log normal. Using transformation, its density is

$$p(\theta) = \left(\sqrt{2\pi}\sigma\theta\right)^{-1} \exp\left(\frac{-1}{2\sigma^2}\left(\log\theta - \mu\right)^2\right).$$

Its mean is $\exp(\mu + \frac{1}{2}\sigma^2)$ and variance is $\exp(2\mu)\exp(\sigma^2)(\exp(\sigma^2 - 1))$, and mode is $\exp(\mu - \sigma^2)$

Multivariate normal

Standard Multi-normal $z = (z_1, \ldots, z_d) \sim \mathcal{N}(0, I_d)$ where I_d is $d \times d$ identity matrix. If $z \sim \mathcal{N}(0, I_d)$ then $\theta = \mu + Az \sim \mathcal{N}(\mu, AA^\top)$