Chapter2
Bayesian single-parameter models

By:夏立&张博

LOGO

In this chapter, we consider four fundamental and widely used one-dimensional models—the binomial, normal, Poisson, and exponential—and at the same time introduce important concepts and computational methods for Bayesian data analysis.

2.1Estimating a probability from binomial data

Binomial sampling model：

$$p(y\mid\theta)=\mathrm{Bin}(y\mid n,\theta)=\binom{n}{y}\theta^{y}(1-\theta)^{n-y} \qquad (2.1)$$

Prior distribution：

$$\theta \sim U(0,1)$$

Posterior distribution：

$$p(\theta\mid y)\propto\theta^{y}(1-\theta)^{n-y}$$

In the present case, we can recognize Posterior distribution as the unnormalized form of the beta distribution：

$$\theta\mid y \sim \mathrm{Beta}(y+1,n-y+1)$$

But why is θ assumed to have a （prior）uniform distribution on $[0,1]$

In his famous paper, published in 1763, Bayes sought, in our notation, the probability $\Pr(\theta \in (\theta_1,\theta_2) \mid y)$；his solution was based on a physical analogy of a probability space to a rectangular table (such as a billiard table):

1. (Prior distribution) A ball W is randomly thrown (according to a uniform distribution on the table). The horizontal position of the ball on the table is θ, expressed as a fraction of the table width.
2. (Likelihood) A ball O is randomly thrown n times. The value of y is the number of times O lands to the right of W .

Thus, θ is assumed to have a (prior) uniform distribution on [0, 1].

Using direct probability calculations which he derived in the paper, Bayes then obtained

$$\Pr(\theta \in (\theta_1, \theta_2) \mid y) = \frac{\Pr(\theta \in (\theta_1, \theta_2), y)}{p(y)}$$

$$= \frac{\int_{\theta_1}^{\theta_2} p(y \mid \theta)\, p(\theta)\, d\theta}{p(y)}$$

$$= \frac{\int_{\theta_1}^{\theta_2} \binom{n}{y} \theta^{y} (1-\theta)^{n-y}\, d\theta}{p(y)}$$

Bayes succeeded in evaluating the denominator, showing that

$$p(y) = \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \, d\theta$$

$$= \frac{1}{n+1} \quad \text{for} \quad y = 0, \ldots, n$$

This calculation shows that all possible values of y are equally likely a priori.

In analyzing the binomial model, Laplace also used the uniform prior distribution. His first serious application was to estimate the proportion of girl births in a population. A total of 241,945 girls and 251,527 boys were born in Paris from 1745 to 1770. Letting θ be the probability that any birth is female, Laplace showed that

$$\Pr(\theta \geq 0.5 \mid y = 241{,}945, n = 251{,}527 + 241{,}945) \approx 1.15 \times 10^{-42}$$

and so he was ‘morally certain’ that $\theta < 0.5$

The currently accepted value of the proportion of female births in large European-race populations is 0.485.

Prediction

Letting y denote the result of a new trial, exchangeable with the first n ,

$$\Pr(y = 1 \mid y) = \int_0^1 \Pr(y = 1 \mid \theta, y)\, p(\theta \mid y) d\theta$$

$$= \int_0^1 \theta\, p(\theta \mid y) d\theta = E(\theta \mid y) = \frac{y + 1}{n + 2}$$

from the properties of the beta distribution.

2.2 Posterior as compromise between data and prior information

The process of Bayesian inference involves passing from a prior distribution, $p(\theta)$,to a posterior distribution, $p(\theta \mid y)$,and it is natural to expect that some general relations might hold between these two distributions. For example, we might expect that, because the posterior distribution incorporates the information from the data, it will be less variable than the prior distribution. This notion is formalized in the second of the following expressions:

$$E(\theta) = E(E(\theta \mid y))$$

and

$$\mathrm{var}(\theta) = E(\mathrm{var}(\theta \mid y)) + \mathrm{var}(E(\theta \mid y))$$

Prior mean: $\frac{1}{2}$

Posterior mean: $\frac{y+1}{n+2}$

Sample proportion: $\frac{y}{n}$

The posterior mean, is a compromise between the prior mean,and the sample proportion, where clearly the prior mean has a smaller and smaller role as the size of the data sample increases. This is a general feature of Bayesian inference: the posterior distribution is centered at a point that represents a compromise between the prior information and the data, and the compromise is controlled to a greater extent by the data as the sample size increases.

$$\frac{1}{2} < \frac{y+1}{n+2} < \frac{y}{n} \quad \text{or} \quad \frac{y}{n} < \frac{y+1}{n+2} < \frac{1}{2}$$

2.3Informative prior distributions

Binomial example with different prior distributions

In the binomial example, we have so far considered only the uniform prior distribution for θ. How can this specification be justified, and how in general do we approach the problem of constructing prior distributions?

We first pursue the binomial model in further detail using a parametric family of prior distributions that includes the uniform as a special case. For mathematical convenience, we construct a family of prior densities that lead to simple posterior densities. Considered as a function of θ, the likelihood (2.1) is of the form,

$$p(y \mid \theta) \propto \theta^{a}(1-\theta)^{b}$$

Thus, if the prior density is of the same form, with its own values a and b, then the posterior density will also be of this form. We will parameterize such a prior density as

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

which is a beta distribution with parameters α and β : θ □ $\mathrm{Beta}(\alpha, \beta)$

The parameters of the prior distribution are often referred to as hyperparameters. The beta prior distribution is indexed by two hyperparameters, which means we can specify a particular prior distribution by fixing two features of the distribution, for example its mean and variance.

For now, assume that we can select reasonable values α and β .The posterior density for θ is

$$p(\theta \mid y) \propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$

$$= \mathrm{Bate}(\theta \mid \alpha + y, \beta + n - y)$$

The property that the posterior distribution follows the same parametric form as the prior distribution is called conjugacy; the beta prior distribution is a conjugate family for the binomial likelihood.

To continue with the binomial model with beta prior distribution, the posterior mean of θ , which may be interpreted as the posterior probability of success for a future draw from the population, is now

$$E(\theta \mid y) = \frac{\alpha + y}{\alpha + \beta + n}$$

which always lies between the sample proportion, y/n and the prior mean $\alpha/(\alpha + \beta)$.

The posterior variance is

$$\mathrm{var}(\theta \mid y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \frac{E(\theta \mid y)[1 - E(\theta \mid y)]}{\alpha + \beta + n + 1}$$

As y and $n - y$ become large with fixed α and β , $E(\theta \mid y) \approx y/n$ and $\mathrm{var}(\theta \mid y) \approx \frac{1}{n}\frac{y}{n}(1 - \frac{y}{n})$ which approaches zero at the rate $1/n$. In the limit, the parameters of the prior distribution have no influence on the posterior distribution.

Conjugate prior distributions

Conjugacy is formally defined as follows. If F is a class of sampling distributions $p(y \mid \theta)$, and P is a class of prior distributions for θ , then the class P is conjugate for F if

$$p(\theta \mid y) \in P \text{ for all } p(\cdot \mid \theta) \in F \text{ and } p(\cdot) \in P$$

Nonconjugate prior distributions

Although they can make interpretations of posterior inferences less transparent and computation more difficult, nonconjugate prior distributions do not pose any new conceptual problems. In practice, for complicated models, conjugate prior distributions may not even be possible.

Conjugate prior distributions, exponential families, and sufficient statistics

Probability distributions that belong to an exponential family have natural conjugate prior distributions, so we digress at this point to review the definition of exponential families; for complete generality in this section, we allow data points y_i and parameters θ to be multidimensional. The class F is an exponential family if all its members have the form,

$$p(y_i \mid \theta) = f(y_i) g(\theta) e^{\phi(\theta)^T u(y_i)}$$

likelihood

$$p(y \mid \theta) = (\prod_{i=1}^{n} f(y_i)) g(\theta)^n \exp(\phi(\theta)^T \sum_{i=1}^{n} u(y_i))$$

For all n and y, this has a fixed form (as a function of θ):

$$p(y \mid \theta) \propto g(\theta)^n e^{\phi(\theta)^T t(y)} \text{ ,where } t(y) = \sum_{i=1}^{n} u(y_i)$$

The quantity $t(y)$ is said to be a sufficient statistic for θ , because the likelihood for θ depends on the data y only through the value of $t(y)$.Sufficient statistics are useful in algebraic manipulations of likelihoods and posterior distributions. If the prior density is specified as

$$p(\theta) \propto g(\theta)^{n} e^{\phi(\theta)^{T} v}$$

then the posterior density is

$$p(\theta \mid y) \propto g(\theta)^{\eta + n} e^{\phi(\theta)^{T} (v + t(y))}$$

which shows that this choice of prior density is conjugate.

2.4Estimating a normal mean with known variances

Likelihood of one data point

As the simplest first case, consider a single scalar observation y from a normal distribution parameterized by a mean θ and variance σ^2, where for this initial development we assume that σ^2 is known. The sampling distribution is

$$p(y \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}$$

Conjugate prior and posterior distributions

Considered as a function of θ , the likelihood is an exponential of a quadratic form in θ ,so the family of conjugate prior densities looks like

$$p(\theta) = e^{A\theta^2 + B\theta + C}$$

We parameterize this family as

$$p(\theta) \propto \exp(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2)$$

$$\theta \sim N(\mu_0, \tau_0^2)$$

As usual in this preliminary development, we assume that the hyperparameters are known.

The conjugate prior density implies that the posterior distribution for θ is the exponential of a quadratic form and thus normal, but some algebra is required to reveal its specific form. In the posterior density, all variables except θ are regarded as constants, giving the conditional density,

$$p(\theta \mid y) \propto \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right)$$

$$p(\theta \mid y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

$$\theta \mid y \;\square\; N(\mu_1, \tau_1)$$

here $\mu_1 = \dfrac{\dfrac{1}{\tau_0^2}\mu_0 + \dfrac{1}{\sigma^2}y}{\dfrac{1}{\tau_0^2} + \dfrac{1}{\sigma^2}}$ and $\dfrac{1}{\tau_1^2} = \dfrac{1}{\tau_0^2} + \dfrac{1}{\sigma^2}$

Precisions of the prior and posterior distributions

In manipulating normal distributions, the inverse of the variance plays a prominent role and is called the precision. The algebra above demonstrates that for normal data and normal prior distribution (each with known precision), the posterior precision equals the prior precision plus the data precision.

we can express μ_1 as the prior mean adjusted toward the observed y ,

$$\mu_1 = \mu_0 + (y - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

$$\mu_1 = y - (y - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

Each formulation represents the posterior mean as a compromise between the prior mean and the observed value.

Posterior predictive distribution

$$p(y \mid y) = \int p(y \mid \theta)\, p(\theta \mid y) d\theta$$

$$\propto \int \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right) \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right) d\theta$$

$$E(y \mid \theta) = \theta, \operatorname{var}(y \mid \theta) = \sigma^2$$

$$E(y \mid y) = E(E(y \mid \theta, y) \mid y) = E(\theta \mid y) = \mu_1$$

$$\operatorname{var}(y \mid y) = E(\operatorname{var}(y \mid \theta, y) \mid y) + \operatorname{var}(E(y \mid \theta, y) \mid y)$$

$$= E(\sigma^2 \mid y) + \operatorname{var}(\theta \mid y)$$

$$= \sigma^2 + \tau_1^2$$

Posterior predictive distribution

This development of the normal model with a single observation can be easily extended to the more realistic situation where a sample of independent and identically distributed observations $y = (y_1, \ldots, y_2)$ is available. Proceeding formally, the posterior density is

$$
\begin{aligned}
p(\theta \mid y) &\propto p(\theta)\, p(y \mid \theta) \\
&= p(\theta) \prod_{i=1}^{n} p(y_i \mid \theta) \\
&\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta - \mu_0)^2\right) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right)\Bigg)
\end{aligned}
$$

Algebraic simplification of this expression shows that the posterior distribution depends on y only through the sample mean, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; that is, y is a sufficient statistic in this model.

In fact, since $\bar{y} \mid \theta, \sigma^2 \sim N(\theta, \sigma^2/n)$ the results derived for the single normal observation apply immediately (treating $\bar{y}$ as the single observation) to give

$$p(\theta \mid y_1, \dots, y_n) = p(\theta \mid \bar{y}) = N(\theta \mid \mu_n, \tau_n^2)$$

where $$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$ and $$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

2.5 Other standard single-parameter models

- Normal distribution (known mean and unknown variance)

- Poisson model
  - Negative binomial distribution
  - Possion model parameterized in terms of rate and exposure

- Exponential model

Normal distribution (known mean and unknown variance)

- Likelihood from iid Normal sample y=(y_1, y_2 , … y_n)

$$p(y \mid \sigma^2) \propto \sigma^{-n} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2)$$

- The conjugate prior density is the inverse-gamma

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

We use the convenient but nonstandard notation , $\sigma^2 \sim Inv-\chi^2(v_0, \sigma_0^2)$

- The posterior distribution $\sigma^2 \mid y \sim Inv-\chi^2(v_0+n, \frac{v_0\sigma_0^2+nv}{v_0+n})$

Poisson model

- Likelihood from iid Poisson sample y=(y_1, y_2 , … y_n)

$$p(y \mid \theta) = \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta}$$

$$\propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}$$

- The conjugate prior density $p(\theta) \propto e^{-\beta\theta} \theta^{\alpha-1}$

- The posterior distirbution $\theta \mid y \sim Gamma \quad (\alpha + n\bar{y}, \beta + n)$

. The negative binominal distribution

- We can find the marginal distribution p(y) using the formula

$$p(y) = \frac{p(y \mid \theta)P(\theta)}{p(\theta \mid y)}$$

- For a single distribution,the Possion has prior predictive distribution

$$p(y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y}$$

which is known as the negative binominal density:

y~Neg-bin(α，β）

The above derivation shows that the negative binominal distribution is a mixture of Possion distribution and Gamma distribution.

.Possion model parameterized in terms of rate and exposure

- To extend the Poisson model for data points $y_{1,}$……y_n to the form

y_i~Poisson(x_iθ）

Here x and θis the unknown parameter of interest.In epidemiology,the parameter θis ofen called the rate, and x is called the exposure of ith unit.The likelihood in the extended Possion model is

$$p(y \mid \theta) \propto \theta^{\sum_{i=1}^{n} y_i} e^{-\left(\sum_{i=1}^{n} x_i\right)\theta}$$

- Similar to what we have said ever,the conjugate prior distribution is θ~Gamma(α，β)
- The posterior distribution is

$$\theta \mid y \sim Gamma \quad (\alpha + \sum_{i=1}^{n} y_i, \beta + \sum_{i=1}^{n} x_i)$$

.Possion model parameterized in terms of rate and exposure

A Poisson sampling model is often used for epidemiological data of this form. The Poisson model derives from an assumption of exchangeability among all small intervals of exposure.

➢Possion model:

y—the number of deaths in a city of 200,000 in one year, and

$y \sim \text{Poisson}(2.0\theta)$

θ— the true underlying long-term asthma mortality rate in our city (measured in cases per 100,000 persons per year).

x— exposure $x = 2.0$,since θ is defined in units of 100,000 people

➢Prior distribution:According to the trial-and-error exploration of the properties of the gamma distribution,

$$\theta \sim \text{Gamma}(3.0, 5.0) \qquad E(\theta)=0.6$$

Possion model parameterized in terms of rate and exposure

- Posterior distribution: because of a conjugate Gamma(α, β) prior distribution ,$\theta|y \sim \text{Gamma}(\alpha + y, \beta + x)$ in this case.With the prior distribution and data described,

$$\theta|y \sim \text{Gamma}(6.0, 7.0)$$

E(θ|y) =0.86, it means substantial shrinkage has occurred toward the prior distribution.

Exponential model

- Exponential distribution:appropriate for model ‘waiting times’ and other continuous, positive, real-valued random variables, often measured on a time scale. For an outcome y, given parameter θ,

$$p(y|\ \theta) = \theta\exp(-y\theta), \text{ for } y > 0$$

- Likelihood:for n independent exponential observations, y = (y_1, . . . , y_n), given constant rate θ

$$p(y \mid \theta) = \theta^{n} \exp(-n\bar{y}\theta) \qquad \text{for } y \geq 0$$

- Prior distribution: θ~ Gamma(α, β) ,which can be viewed as α-1 exponential observations with total waiting time β .

- Posterior distribution:Due to conjugate prior distribution for the exponential parameter θ,the posterior distribution is

$$\theta|y \sim \text{Gamma}(\alpha+n, \beta+n\bar{y})$$

2.6 Noninformative prior distributions

- Noninformative and weakly informative prior distribution

- Proper and improper priordistributions

- Jeffery's invariance principle

- Pivotal quantities

- Difficulties with noninformative prior distribution

Noninformative and weakly informative prior distribution

- Noninformative prior distributions :prior distributions that can be guaranteed to play a minimal role in the posterior distribution and the prior density is described as vague, flat, diffuse or noninformative.

- Weakly informative prior distribution:distributions which contains some information — enough to ‘regularize’ the posterior distribution,that is, to keep it roughly within reasonable bounds.

Proper and improper prior distributions

➢Proper prior density:In general, we call a prior density p(θ) proper if it does not depend on data and integrates to 1.

➢We return to the problem of estimating the mean θ of a normal model with known variance, with a N(μ_0, τ_0^2) prior distribution on θ. If the prior precision, 1/τ_0^2, is small relative to the data precision, n/σ^2, then the posterior distribution is approximately as if $\tau_0^2 = \infty$

$$p(\theta \mid y) \approx N(\theta \mid \bar{y}, \sigma^2 / n)$$

Such a distribution isnot strictly possible, since the integral of the assumed p(θ) is infinity, which violates theassumption that probabilities sum to 1.

➢Consider the normal model with known mean but unknown variance, with the conjugate scaled inverse-χ^2 prior distribution. If the prior degrees of freedom, ν0, are small relative to the data degrees of freedom,n, then the posterior distribution is approximately as if ν_0= 0:

Proper and improper prior distributions

$$p(\sigma^2 \mid y) \approx Inv - \chi^2(\sigma^2 \mid n, v)$$

- This limiting form of the posterior distribution can also be derived by defining the prior density for σ^2 as p(σ^2)$\propto$1/σ^2, which is improper, having an infinite integral over the range(0,∞).

Jeffery's invariance principle

- Let $\phi = h(\theta)$,the following is prior density on φ:

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right|$$

- Jeffreys’ general principle:any rule for determining the prior density p(θ) should yield an equivalent result if applied to the transformed parameter.

Jeffreys’ principle leads to defining the noninformative prior density as $p(\theta)\propto[J(\theta)]^{1/2}$,where J(θ) is the Fisher information for θ:

$$J(\theta) = -E\left(\frac{d^2\log p(y|\theta)}{d\theta^2}\middle|\theta\right)$$

To see that Jeffreys’ prior model is invariant to parameterization, evaluate J(φ) at θ $=h^{-1}(\varphi)$:

Jeffery's invariance principle

$$J(\phi) = J(\theta) \left| \frac{d\theta}{d\phi} \right|^2$$

thus $J(\phi)^{1/2} = J(\theta)^{1/2} \left| \frac{d\theta}{d\phi} \right|$,as required

$$p(\phi) \propto [J(\phi)]^{\frac{1}{2}}$$

Jeffert's principle can be extended to multiparameter models,but the results are more controversial.

Pivotal quantities

For the binomial and other single-parameter models, different principles give slightly different noninformative prior distributions. But for two cases—location parameters and scaleparameters—all principles seem to agree.

➢ If the density of y is such that p($y-\theta \mid \theta$) is a function that is free of θ and y, say,f (u), where u = y-θ, then y-θ is a pivotal quantity, and θ is called a pure location parameter. In such a case, it is reasonable that a noninformative prior distribution for θwould give f (y-θ) for the posterior distribution, $.p(y-\theta \mid y)$. Under this condition, using Bayes' rule, p($y-\theta \mid y$)∝p(θ)p($y-\theta \mid \theta$), thereby implying that the noninformative prior density is uniform on θ; that is, p(θ)∝constant over the range (-∞,∞).

Pivotal quantities

- If the density of y is such that p(y/θ |θ) is a function that is free of θ and y—say, g(u), where u = y/θ—then u = y/θis a pivotal quantity and θ is called a pure scale parameter. In such a case, it is reasonable that a noninformative prior distribution for θ would give g(y/θ)for the posterior distribution, p(y/θ |y). By transformation of variables, the conditionaldistribution of y given θ can be expressed in terms of the distribution of u given θ,

$$p(y \mid \theta) = \frac{1}{\theta} p(u \mid \theta)$$

and similarly

$$p(\theta \mid y) = \frac{y}{\theta^{2}} p(u \mid y)$$

After letting both p(u|θ) and p(u|y) equal g(u), we have the identity p(θ|y) = y/θp(y|θ).Thus, in this case, the reference prior distribution is p(θ)∝1/θor, equivalently, p(log θ)∝1/

θ^{2}

Difficulties with noninformative prior distribution

- Searching for a prior distribution that is always vague seems misguided

- For many problems, there is no clear choice for a vague prior distribution, since a density that is flat or uniform in one parameterization will not be in another.

- Further diffculties arise when averaging over a set of competing models that have improper prior distributions.