Tutorial Problems 1
MATH3871/MATH5970
Bayesian Inference and Computation

July 23, 2018

Recall that in the Bayesian framework it is assumed that the data vector, $\boldsymbol{x}$ say, has been drawn from a conditional pdf $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. With $\boldsymbol{\theta}$ is associated a (density) function $f(\boldsymbol{\theta})$ that conveys the *a priori* (existing beforehand) information about $\boldsymbol{\theta}$. Observing the data $\boldsymbol{x}$ will affect the knowledge about $\boldsymbol{\theta}$, and the way to update this information is to use Bayes' formula. The main concepts are summarized in the following definition.

Definition 1 (Prior, Likelihood, and Posterior)**.** *Let $\boldsymbol{x}$ and $\boldsymbol{\theta}$ denote the data and parameters in a Bayesian statistical model.*

- *The pdf $f(\boldsymbol{\theta})$ of the parameter $\boldsymbol{\theta}$ is called the* prior *pdf.*
- *The conditional pdf $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ is called the Bayesian* likelihood *function.*
- *The central object of interest is the* posterior *pdf $f(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ which, by Bayes' theorem, is proportional to the product of the prior and likelihood:*

$$f(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto f(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, f(\boldsymbol{\theta}) \,.$$

The posterior pdf thus conveys the knowledge of $\boldsymbol{\theta}$ after taking into account the information $\boldsymbol{x}$. Bayesian learning often requires Monte Carlo methods to evaluate the posterior; In practice, the choice of the prior is governed by two considerations. First, the prior should be simple enough to facilitate the computation or simulation of the posterior pdf. Second, the prior distribution should be general enough to model complete ignorance of the parameter of interest. Priors that do not convey any pre-knowledge of the parameter are said to be *uninformative*. Here are two simple examples of Bayesian models.

Example 1 (Normal Model)**.** Let $x_1, \ldots, x_n$ be a random sample from the $\mathsf{N}(\mu, \sigma^2)$ distribution. Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$. In classical statistics the model can be written as $\boldsymbol{x} \sim \mathsf{N}(\mu\mathbf{1}, \sigma^2\mathbf{I})$, where $\mathbf{1}$ is the n-dimensional vector of 1s and $\mathbf{I}$ the n-dimensional identity matrix. To formulate the corresponding Bayesian model, we start with a similar likelihood as in the classical case; that is,

$$(\boldsymbol{x} \,|\, \mu, \sigma^2) \sim \mathsf{N}(\mu\mathbf{1}, \sigma^2\mathbf{I}) \ .$$

In the Bayesian setting both μ and σ^2 are treated as random, and we need to specify their prior distributions to complete the model. A possible prior for μ is

$$\mu \sim \mathsf{N}(0, \sigma_0^2) \ , \tag{1}$$

where $\sigma_0^2 > 0$ is a constant. The larger σ_0^2 is, the more uninformative is the prior. Instead of giving directly a prior for σ^2 (or σ), it turns out to be convenient to give the following prior distribution to $1/\sigma^2$:

$$\frac{1}{\sigma^2} \sim \mathsf{Gam}(\alpha_0, \lambda_0) \ . \tag{2}$$

The smaller the α_0 and λ_0 are, the less informative is the prior. It is further assumed that μ and σ^2 are independent. Under this prior σ^2 is said to have an *inverse gamma* distribution. We use its pdf below. The joint pdf of $\boldsymbol{x}, \mu$ and σ^2 is now

$$\begin{aligned} f(\boldsymbol{x}, \mu, \sigma^2) &= f(\mu) \times f(\sigma^2) \times f(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \sigma^2) \\ &= \left(2\pi\sigma_0^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\frac{\mu^2}{\sigma_0^2}\right\} \\ &\times \frac{\lambda_0^{\alpha_0} (\sigma^2)^{-\alpha_0 - 1} \exp\left\{-\lambda_0\,(\sigma^2)^{-1}\right\}}{\Gamma(\alpha_0)} \\ &\times \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2}\frac{\sum_i (x_i - \mu)^2}{\sigma^2}\right\} \ . \end{aligned}$$

It follows that the posterior pdf is given by

$$f(\mu, \sigma^2 \,|\, \boldsymbol{x}) \propto \left(\sigma^2\right)^{-n/2 - \alpha_0 - 1} \exp\left\{-\frac{1}{2}\frac{\sum_i (x_i - \mu)^2}{\sigma^2} - \frac{1}{2}\frac{\mu^2}{\sigma_0^2} - \frac{\lambda_0}{\sigma^2}\right\} \ . \tag{3}$$

It is interesting to note that in the limit $\sigma_0^2 \to \infty$, $\alpha_0 \to 0$, and $\lambda_0 \to 0$, the posterior pdf becomes

$$f(\mu, \sigma^2 \,|\, \boldsymbol{x}) \propto \left(\sigma^2\right)^{-n/2 - 1} \exp\left\{-\frac{1}{2}\frac{\sum_i (x_i - \mu)^2}{\sigma^2}\right\} \ . \tag{4}$$

The posterior pdf is the main object of interest.

We now have a number of worked examples, followed by simple exercises.

Example 2 (Mixture Model)**.** Let $f_1, \ldots, f_k$ be probability densities (discrete or continuous) on some set $\mathscr{X}$, and let $w_1, \ldots, w_k$ be positive numbers summing up to 1. Then

$$f(\boldsymbol{x}) = w_1 f_1(\boldsymbol{x}) + \cdots + w_k f_k(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathscr{X}, \tag{5}$$

is another density of $\mathscr{X}$. It is called *mixture density* of $f_1, \ldots, f_k$ with *weights* $w_1, \ldots, w_k$. The classical model for a random variable from a mixture distribution could thus be written as

$$X \sim \sum_{i=1}^{k} w_i f_i \ .$$

A corresponding Bayesian model could be given as

$$\begin{aligned} \mathbb{P}(z = i) &= w_i, \quad i = 1, \ldots, k, \\ (\boldsymbol{x} \,|\, z) &\sim f_z \ . \end{aligned}$$

Apart from using a lower-case letter $\boldsymbol{x}$, the Bayesian model introduces an extra variable, z, wich takes values $1, \ldots, k$ with probability $w_1, \ldots, w_k$. We can interpret the $\{w_k\}$ as the prior probabilities that the data $\boldsymbol{x}$ comes from "class" z. Conditional on z, $\boldsymbol{x}$ is drawn from pdf f_z. That this gives the same pdf for $\boldsymbol{x}$ as in (5) can be seen as follows:

$$f(\boldsymbol{x}) = \sum_{z=1}^{k} f(\boldsymbol{x}, z) = \sum_{z=1}^{k} f(\boldsymbol{x} \,|\, z) f(z) = \sum_{z=1}^{k} f_z(\boldsymbol{x}) w_z \ .$$

The advantage of the Bayesian formalism is that it now makes sense to consider the conditional density $f(z \,|\, \boldsymbol{x})$. That is, given the data $\boldsymbol{x}$, we can find out what the probability is that it comes from pdf $f_z, z = 1, 2, \ldots, k$. As an example, Figure 1 gives the pdf of a mixture of normal, uniform and exponential pdfs, with weights $w_1 = w_2 = w_3 = 1/3$. How likely is it that an outcome $\boldsymbol{x} = 0.5$ comes from the uniform distribution? The answer is given by the posterior pdf

$$f(z \,|\, \boldsymbol{x}) \propto f(z) f(\boldsymbol{x} \,|\, z) \propto f(\boldsymbol{x} \,|\, z).$$

For the normal pdf, $f(0.5 \,|\, z = 1) = \exp(-\frac{1}{2}(0.5)^2)/\sqrt{2\pi} = 0.3521$, for the uniform pdf, $f(0.5 \,|\, z = 2) = 1$, and for the exponential pdf $f(0.5 \,|\, z = 3) = \exp(-0.5) = 0.6065$. The posterior probabilities are thus 0.1789, 0.5106, and 0.3097. It is thus almost 3 times more likely that the outcome 0.5 has come from the uniform distribution than from the normal one.

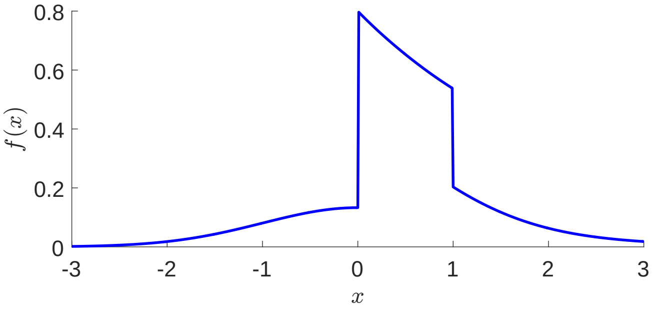


Figure 1: The pdf of an equal-weight mixture of $\mathsf{N}(0,1)$, $\mathsf{U}(0,1)$ and $\mathsf{Exp}(1)$ distributions.

Example 3 (Coin Flipping and Bayesian Learning)**.** Consider the random experiment where we toss a biased coin n times. Suppose that the outcomes are $x_1, \ldots, x_n$, with $x_i = 1$ if the i-th toss is heads and $x_i = 0$ otherwise, for $i = 1, \ldots, n$. A possible Bayesian model for the data is

$$p \sim \mathsf{U}(0,1)$$
$$(x_1, \ldots, x_n \,|\, p) \overset{\text{iid}}{\sim} \mathsf{Ber}(p) \ .$$

The likelihood is therefore

$$f(\boldsymbol{x} \,|\, p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^s \, (1-p)^{n-s} \ ,$$

where $s = x_1 + \cdots + x_n$ is the total number of heads. Since $f(p) = 1$, the posterior pdf is

$$f(p \,|\, \boldsymbol{x}) = c \, p^s \, (1-p)^{n-s} \ , \quad p \in [0,1] \ ,$$

which is the pdf of the $\mathsf{Beta}(s+1, n-s+1)$ distribution. The normalization constant is $c = (n+1)\binom{n}{s}$. The maximum a posteriori estimate of p is s/n, which coincides with the classical maximum likelihood estimate. The expectation of the posterior pdf is $(s+1)/(n+2)$. The graph of the pdf for $n = 100$ and $s = 1$ is given in Figure 2. For this case a left one-sided 95% credible interval for p is $[0, 0.0461]$, where 0.0461 is the 0.95 quantile of the $\mathsf{Beta}(2, 100)$ distribution.

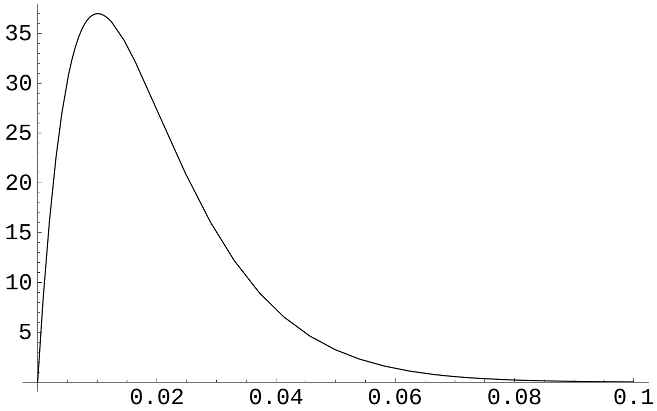
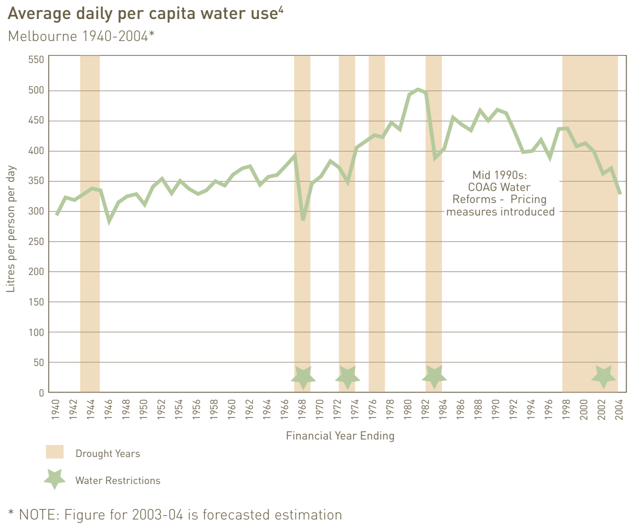


Figure 2: Posterior pdf for p, with $n = 100$ and $s = 1$.

1) Water consumption



In the Melbourne average daily per capita water use analyis, we modelled the discrete observations $x_1, \ldots, x_n$ as independent draws from a Poisson(θ) distribution. Assuming a Gamma(α, β) prior, which has a density function of

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta), \qquad \text{for } \alpha, \beta, \gamma > 0,$$

we computed the posterior as a Gamma $\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)$ distribution.

(a) Show that the posterior mean of θ is given by $\frac{\alpha+\sum_i x_i}{\beta+n}$.

(b) Show that the posterior variance of θ is given by $\frac{\alpha+\sum_i x_i}{(\beta+n)^2}$.

(e) Show that the predictive distribution for a future observation, y, is $\text{NegBin}\left(y \mid \alpha + \sum_i x_i, \frac{1}{\beta+n+1}\right)$, where the probability mass function of a Negative Binomial random variable with parameters $a > 0$, and $0 \leq p \leq 1$, is given by

$$\pi(y \mid a, p) = \left(\begin{array}{c} y + a - 1 \\ y \end{array} \right) (1-p)^a p^y.$$

2) Rock Strata

(a) Rock strata A and B are difficult to distinguish in the field. Through careful laboratory studies it has been determined that the only characteristic which might be useful in aiding discrimination is the presence or absence of a particular brachiopod fossil. In rock exposures of the size usually encountered, the probabilities of fossil presence are found to be as in the table below. It is also known that rock type A occurs about four times as often as type B in this area of study.

Stratum	Fossil present	Fossil absent
A	0.9	0.1
B	0.2	0.8

If a sample is taken, and the fossil found to be present, calculate the posterior distribution of rock types.

(b) If the geologist always classifies as A when the fossil is found to be present, and classifies as B when it is absent, what is the probability she will be correct in a future classification?