Examples of basic Bayesian inference procedures

We have earlier considered representation and revision of beliefs as the basis of empirical learning.

Here we shall investigate some simple examples.

Example 1 Single observation from a normal distribution. *Let x have a normal distribution $N(\theta, v)$ with unknown mean θ and known variance v, and let the prior distribution for θ be $N(m, w)$. Let the precision parameters be $\lambda_0 = 1/v$ and $\lambda_1 = 1/w$. Then,*

$$\begin{aligned} p(x|\theta, v) &= \frac{1}{\sqrt{2\pi v}} \exp(-\frac{1}{2v}(x-\theta)^2) \qquad (1) \\ p(\theta|m, w) &= \frac{1}{\sqrt{2\pi w}} \exp(-\frac{1}{2w}(\theta - m)^2). \end{aligned}$$

By multiplying together the prior and the likelihood, and expanding the squares, we get the exponential

$$\exp\left(-\frac{1}{2}\lambda_0(x^2 - 2x\theta + \theta^2) - \frac{1}{2}\lambda_1(\theta^2 - 2\theta m + m^2)\right). \qquad (2)$$

The exponential can be further written as

$$
\begin{aligned}
& -\frac{1}{2}\lambda_0 x^2 + \lambda_0 x\theta - \frac{1}{2}\lambda_0\theta^2 - \frac{1}{2}\lambda_1\theta^2 + \lambda_1\theta m - \frac{1}{2}\lambda_1 m^2 \qquad (3)\\
= \; & -\frac{1}{2}(\lambda_0+\lambda_1)\theta^2 + \theta(\lambda_0 x + \lambda_1 m) - \frac{1}{2}(\lambda_0 x^2 + \lambda_1 m^2)\\
= \; & -\frac{1}{2}(\lambda_0+\lambda_1)\left(\theta - 2\theta\frac{\lambda_0 x + \lambda_1 m}{\lambda_0+\lambda_1} + \left(\frac{\lambda_0 x + \lambda_1 m}{\lambda_0+\lambda_1}\right)^2\right) + c
\end{aligned}
$$

where c does not depend on θ. Since the constants cancel in

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}, \qquad (4)$$

the posterior is recognized as the density function of the normal distribution

$$N\left(\frac{\lambda_0 x + \lambda_1 m}{\lambda_0+\lambda_1}, \lambda_0+\lambda_1\right), \qquad (5)$$

where the mean is a weighted average of prior mean m and observation x.

Therefore the posterior mean (as well as mode and median) is a compromise between the prior information and the sample information.

We see also that each source of information is weighted proportionately to its precision.

Consequently, the posterior mean will lie closer to whichever source has the stronger information.

If, for instance, prior information is very weak, expressed by λ_1 being close to zero, then the posterior mean will be close to x.

The posterior precision is the sum of the prior and data precisions, reflecting the combination of information from the two sources.

The posterior information is stronger than either source of information alone.

Example 2 Several observations from a normal distribution. *In the previous example we had only a single observation available for making inference about the mean of the distribution. However, typically, we would utilize several observations. Let* $x_1, ..., x_n$ *be conditionally independent observations from a normal distribution* $N(\theta, 1)$ *with unknown mean* θ *and known variance* 1*. Suppose the prior distribution for* θ *is again* $N(m, w)$*, i.e. the precision parameter is* $\lambda = 1/w$*. The likelihood function can be written as*

$$p(\mathbf{x}|\theta) = (2\pi)^{-n/2} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right), \qquad (6)$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ *is the sample mean. Multiplying the likelihood and prior together and simplifying yields the following expression for the numerator*

of the posterior formula,

$$\exp\left(-\frac{n+\lambda}{2}\left(\theta-\frac{\lambda m+n\bar{x}}{\lambda+n}\right)^2\right), \tag{7}$$

thus, the posterior is $N(\frac{\lambda m+n\bar{x}}{\lambda+n}, 1/(\lambda+n))$*. We see that the posterior variance decreases (i.e. the precision increases) as the sample size increases, and similarly that the dependence on the prior mean decreases as well.*

Example 3 Predictive distribution of a future observation. *Let us continue analysis of the previous example by considering the predictive density of a future observation* x_{n+1}

$$
\begin{aligned}
& p(x_{n+1}|\mathbf{x}) && (8)\\
= & \int p(x_{n+1}|\theta)p(\theta|\mathbf{x})d\theta && (9)\\
= & \int \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-\theta)^2\right)\frac{\sqrt{n+\lambda}}{\sqrt{2\pi}}\exp\left(-\frac{n+\lambda}{2}\left(\theta-\frac{\lambda m+n\bar{x}}{\lambda+n}\right)^2\right)d\theta\\
= & \frac{\sqrt{n+\lambda}}{\sqrt{2\pi(n+\lambda+1)}}\exp\left(-\frac{n+\lambda}{2(n+\lambda+1)}\left(y-\frac{\lambda m+n\bar{x}}{\lambda+n}\right)^2\right),
\end{aligned}
$$

which is the density of the normal distribution $N(\frac{\lambda m+n\bar{x}}{\lambda+n}, 1+1/(\lambda+n))$. *Thus, we see that the excess uncertainty in the predictive distribution, which is due to the "estimation" of the unknown parameter* θ*, vanishes as the sample size tends to infinity. This procedure is in perfect harmony with intuition ;)*

about how information is gathered and utilized.

Example 4 Observations from a Poisson distribution. *Let x have the Poisson distribution with unknown mean θ,*

$$p(x|\theta) = \frac{\theta^x}{x!}e^{-\theta}, \tag{10}$$

and suppose that the prior density has the Gamma(α, β) form

$$p(\theta) = \frac{\alpha^\beta \theta^{\beta-1}}{\Gamma(\beta)}e^{-\alpha\theta}, \theta > 0. \tag{11}$$

by combining the prior and the likelihood we enter into the Gamma(α+1, β+x) posterior. When the likelihood comprises n observations $x_1, ..., x_n$

$$p(x|\theta) = \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}e^{-n\theta},$$

use of the same prior as above, gives us the Gamma($\alpha + n, \beta + \sum_{i=1}^n x_i$) posterior. The mean of this distribution equals $\beta + \sum_{i=1}^n x_i/(\alpha + n)$ and the

variance $\beta + \sum_{i=1}^{n} x_i/(\alpha + n)^2$.

Example 5 Bayesian estimation of Shannon entropy. *We now consider a considerably more complicated inference situation than encountered in the previous examples, taken from Yuan and Kesavan (1997). Recall from the previous chapter the concept of the entropy for a discrete random quantity* x *taking values conveniently labeled by a finite set of integers* $\{1, ..., s\}$ *associated with a probability distribution* $\mathbf{p} = (p_1, ..., p_s)$ *satisfying* $p_i > 0, i = 1, ..., s$, *and* $\sum_{i=1}^{s} p_i = 1$. *The entropy is defined as*

$$h = -\sum_{i=1}^{s} p_i \log p_i. \tag{12}$$

Here we use the natural logarithm in the definition of entropy, however, other bases are also often used in the literature. If the true distribution is known, then the calculation of the entropy is straightforward. In practice, however, we often have to estimate h *from data under no or vague knowledge about the underlying probability distribution* $\mathbf{p}$. *Suppose we have frequency data generated from a*

multinomial distribution $\mathbf{p}$*, leading to the likelihood*

$$\binom{n}{n_1 \cdots n_s} p_1^{n_1} \cdots p_s^{n_s} \tag{13}$$

where $n = \sum_{i=1}^{s} n_i$ *and* $\binom{n}{n_1 \cdots n_s}$ *is the multinomial coefficient. We recall from earlier that the maximum likelihood estimate* $\hat{p}_i$ *of* p_i *is provided by the observed relative frequency* $n_i/n, i = 1, ..., s$*. Apparently, this procedure leads to the entropy estimate*

$$h_n = -\sum_{i=1}^{s} \hat{p}_i \log \hat{p}_i. \tag{14}$$

While the above estimate may be deemed satisfactory for large n *relative to* s*, its properties could be improved upon when the converse is true. From the definition of entropy we see that the values of* x *having zero observed frequencies make no contribution to the estimate* h_n*.*

Assume we have a prior guess about the unknown distribution $\mathbf{p}$, *say* $\boldsymbol{\pi} = (\pi_1, ..., \pi_s)$, *with* $\sum_{i=1}^{s} \pi_i = 1, \pi_i > 0$. *We could now use the Dirichlet* $D(\alpha\pi_1, ..., \alpha\pi_s)$ *distribution to describe our prior beliefs, where the parameter* α *is a measure of our confidence about our guess. A larger value of* α *implies more concentration of the prior around* $(\pi_1, ..., \pi_s)$. *If we do not have any prior knowledge, a uniform prior* $D(1, ..., 1)$ *could be used.*

Under the above Dirichlet prior we get an explicit expression for the posterior mean of the entropy, which equals

$$h_B = -\sum_{i=1}^{s} \frac{\alpha\pi_i + n_i}{\alpha + n} \left[\psi(\alpha\pi_i + n_i + 1) - \psi(\alpha + n + 1)\right], \qquad (15)$$

where $\psi(t) = \Gamma'(t)/\Gamma(t)$ *is the digamma function. When* α *is large compared with* n, h_B *is mainly determined by the prior, and consequently, the contribution of the data is small. With the increase of* n, *the behavior of* h_B *is as that of*

h_n. *When the prior is uniform we get the expression*

$$h_{B_0} = -\sum_{i=1}^{s} \frac{1+n_i}{s+n} \left[\psi(n_i+2) - \psi(s+n+1)\right]. \tag{16}$$