数理解析研究所講究録
第 723 巻 1990 年 82-86

An interpretation of a candidate's formula

電通大電子情報　久保木　久　孝（Hisataka Kuboki）

Abstract. A candidate's formula for the Bayesian predictive distribution of a future observation is a predictive version of Bayes' formula. However the formula connects the expected entropy, a measure of the goodness of prediction fit, with the expected information gain about a parameter.

1. Introduction

Let $\phi(y)$ be the density of a future observation y. Suppose that we can use the observation x which has the density $\psi(x)$. Assume that y and x are independent but that x provides information on y through the same indexing parameter (Aitchison & Dunsmore, 1975, p. 1). We consider the problem of estimating the true distribution $\phi(y)$. When the joint distribution $\phi(y)\psi(x)$ is a member of a parametric family $M = \{ f(y|\theta)g(x|\theta) : \theta \in \Theta \}$, such an estimate $f(y|x)$ is obtained from

$$f(y|x) = \int f(y|\theta)p(\theta|x)\, d\theta,$$

using a distribution $p(\theta|x)$ which specifies a parameter θ based on the observation x. Akaike (1978) termed $p(\theta|x)$ an inferential distribution.

Consider the situation where the distributions $\phi(y)$ and $\psi(x)$ are chosen randomly from the family M by a prior distribution $\pi(\theta)$. Then the Bayesian predictive distribution $f^*(y|x)$ is usually calculated from

$$f^*(y|x) = \int f(y|\theta)\pi(\theta|x)\, d\theta,$$

where $\pi(\theta|x)$ is the posterior distribution of θ. Recently, it is pointed out by Besag (1989) that $f^*(y|x)$ is expressible in the form

$$f^*(y|x) = f(y|\theta)\pi(\theta|x)/\pi(\theta|y,x) \tag{1}$$

without any need for integration, where $\pi(\theta|y,x)$ is the posterior distribution of θ, with x augmented by an additional observation y. This expression is termed a candidate's formula. However, we remark here that (1) is a predictive version of Bayes' formula; see also Leonard (1982). In fact, setting $h(y,x) = \int f(y|\theta)g(x|\theta)\pi(\theta)\,d\theta$ and $g(x) = \int h(y,x)\,dy$, we can easily see that the formula (1) follows from Bayes' formula

$$\pi(\theta|y,x)h(y,x) = f(y|\theta)g(x|\theta)\pi(\theta), \tag{2}$$

and the relation

$$f^*(y|x) = h(y,x)/g(x). \tag{3}$$

The purpose of the present note is to point out that the expression (1) is of theoretical importance, besides being useful in calculating $f^*(y|x)$. In the next section, we give an information-theoretic interpretation to this formula. The note closes with some comments on those three methods of specifying prior distributions which are proposed by Akaike (1978, 1983) and Bernardo (1979).

2. An interpretation

When a distribution $f(y)$ is used as an estimate of $\phi(y)$, an appropriate measure of the goodness of prediction fit is the entropy of $\phi(y)$ with respect to $f(y)$ defined by

$$B(\phi, f) = -\int \frac{\phi(y)}{f(y)} \log\left\{\frac{\phi(y)}{f(y)}\right\} f(y)\,dy;$$

see Akaike (1978). Thus the neg-entropy $-B(\phi, f)$, which is identical to the well-known Kullback-Liebler information, is a measure of the badness of prediction fit to $\phi(y)$. Since the distributions $\phi(y)$ and $\psi(x)$ are produced from the family M

according to the prior distribution $\pi(\theta)$, the goodness of a predictive distribution $f(y|x)$ is then evaluated by the expected entropy

$$E_\theta E_{x|\theta}[B\{\phi, f(\cdot|x)\}] = -\int \pi(\theta) \int g(x|\theta) \int f(y|\theta) \log\left\{\frac{f(y|\theta)}{f(y|x)}\right\} dy\, dx\, d\theta.$$

We are interested in finding $f(y|x)$ which will maximize the expected entropy or minimize the expected neg-entropy. As noted by Aitchison (1975), the desired maximum is attained with $f^*(y|x)$; i.e.

$$E_\theta E_{x|\theta}[B\{\phi, f^*(\cdot|x)\}] \geq E_\theta E_{x|\theta}[B\{\phi, f(\cdot|x)\}]. \tag{4}$$

On the other hand, if we follow Lindley (1956), the amount of information about θ provided by y after the value x is observed, with prior knowledge $\pi(\theta)$, can be defined to be

$$I\{\pi(\cdot|y,x), \pi(\cdot|x)\} = \int \pi(\theta|y,x) \log\left\{\frac{\pi(\theta|y,x)}{\pi(\theta|x)}\right\} d\theta.$$

Thus the expected information gain about θ is

$$E_{y,x}[I\{\pi(\cdot|y,x), \pi(\cdot|x)\}] = \iint I\{\pi(\cdot|y,x), \pi(\cdot|x)\} h(y,x)\, dy\, dx.$$

The following result describes the relation between the expected neg-entropy and the expected information gain.

THEOREM. *For any inferential distribution* $p(\theta|x)$,

$$E_{y,x}[I\{\pi(\cdot|y,x), \pi(\cdot|x)\}] \leq E_\theta E_{x|\theta}[-B\{\phi, f(\cdot|x)\}],$$

with equality if, and only if, $f(y|x)$ *agrees with the right hand side of* (1).

The proof is straightforward from the formula (1) and the inequality (4). This theorem shows that the expected information gain about θ provided by y after the value x is observed becomes a lower bound to the badness of the prediction fit to $\phi(y)$. Then the formula (1) states how the information gain about the parameter θ is converted into the best prediction of the future observation y.

3. Discussion

The theorem in § 2 suggests the use of the lower bound $E_{y,x}[I\{\pi(\cdot|y,x),\pi(\cdot|x)\}]$ for the comparison of various possible prior distributions; that is, we select a prior distribution $\pi(\theta)$ such that the future observation y adds 'minimum' amount of information to the corresponding posterior distribution $\pi(\theta|x)$. The impartial prior distribution introduced by Akaike (1978), although it is actually defined by a prior distribution $\pi(\theta)$ which maximizes the quantity $\min_{\theta\in\Theta} E_{x|\theta}[B\{\phi, f^*(\cdot|x)\}]$, is similar to this prior distribution. On the other hand, the definition of the reference posterior distribution introduced by Bernardo (1979) is based on a prior distribution $\pi(\theta)$ which 'maximizes' the expected information gain about θ provided by the observation x:

$$E_x[I\{\pi(\cdot|x),\pi\}] = \int g(x)\int \pi(\theta|x)\log\left\{\frac{\pi(\theta|x)}{\pi(\theta)}\right\}d\theta\,dx.$$

If interest is both in estimation of the parameter θ and in prediction of the future observation y, then to compare various possible prior distributions, we should use a criterion function which evaluates both the information gain about θ and the prediction fit to $\phi(y)$. As a natural choice of such a function we adopt

$$E_x[I\{\pi(\cdot|x),\pi\}] + E_\theta E_{x|\theta}[B\{\phi, f^*(\cdot|x)\}]. \tag{5}$$

We want then to find a prior distribution $\pi(\theta)$ which 'maximizes' this criterion function.

Let us now consider the special situation where $\phi(\cdot) = \psi(\cdot)$ and $f(\cdot|\theta) = g(\cdot|\theta)$. From (2) and (3),

$$\frac{h(y,x)}{g(y)g(x)} = \frac{\pi(\theta|y)f^*(y|x)}{\pi(\theta)g(y|\theta)}.$$

Here using (2), we have

$$\begin{aligned}&\pi(\theta|y,x)h(y,x)\log\left\{\frac{h(y,x)}{g(y)g(x)}\right\}\\ &\quad= g(x|\theta)g(y)\pi(\theta|y)\log\left\{\frac{\pi(\theta|y)}{\pi(\theta)}\right\} - \pi(\theta)g(x|\theta)g(y|\theta)\log\left\{\frac{g(y|\theta)}{f^*(y|x)}\right\}.\end{aligned}$$

Since $f(\cdot|\theta) = g(\cdot|\theta)$, it follows that $E_y[I\{\pi(\cdot|y), \pi\}] = E_x[I\{\pi(\cdot|x), \pi\}]$. Thus we obtain

$$\int\int h(y,x) \log\left\{\frac{h(y,x)}{g(y)g(x)}\right\} dy\, dx = E_x[I\{\pi(\cdot|x), \pi\}] + E_\theta E_{x|\theta}[B\{\phi, f^*(\cdot|x)\}].$$

This observation shows that the minimal information prior distribution introduced by Akaike (1983) agrees with that $\pi(\theta)$ which maximizes the criterion function (5).

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547–554.

Aitchison, J. & Dunsmore, I. R. (1975). *Statistical Prediction Analysis.* Cambridge University Press.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika* **65**, 53–59.

Akaike, H. (1983). On minimum information prior distributions. *Ann. Inst. Statist. Math.* **35**, 139–149.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. R. Statist. Soc.* B **41**, 113–147.

Besag, J. (1989). A candidate's formula: A curious result in Bayesian prediction. *Biometrika* **76**, 183.

Leonard, T. (1982). Comment on "A simple predictive density function". *J. Am. Statiat. Assoc.* **77**,657–658.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.