Inference in stochastic information processing

JELENKA SAVKOVIC-STEVANOVIC

Faculty of Technology and Metallurgy,
Belgrade University, Karnegijeva 4,
11000 Belgrade, SERBIA,
stevanoviccace@gmail.com

Abstract.- In this paper stochastic information processing was studied. Inference method by conditional distribution was examined. The Bayesian decision analysis for unknown event was performed. Prior and poserior distribution functions were studied and applied to information definition about the system state. Frequency meaning and subject probability were considered. Contribution of this paper are information processing method by likelihood function and information processing in conflict resolution.

Keywords: Conflict, processing, conditional statement, inference, posterior distribution, likelihood.

1 Introduction

The decision problems are nearly all concerened with the situation common in scientific inference where the prior distribution is dominated by the likelihood [1]-[4].

The fact that in such situations different decisions can results from different choice of prior distribution has worred some statisticians [5]-[6]. However, that making explicit the dependence of the decision on the choice of what is believed to be true is an advantage of Bayesian analysis rather than the reverse. Suppose four different executives, after careful consideration, produce four different prior distributions for the size of a potential market and separate analyses are made for each. Then either the decision will be the same in spite of differences in the priors, or the decission will be different. In either case the Bayesian decision analysis will be valuable. In the first case, the ultimate arbiter would be reassured that such differences in opinion did not logically lead to differences on what the appropriate action should be. In the second case, it would be clear to him the responsibility of ignoring the judgement of one or more of his executives, or of arranging that further data be obtained to resolve the conflict. Far from nulifying the value of Bayesian analysis, the fact that such analysis shows to what extent different decisions may not be appropriate when different prior opinions are held, seems to enhance it. For problems of this kind any procedure which took no account of such opinion would seem necessarily ill conceived.

In this paper inference for stochastic information processing was developed.

2 Conditional distribution

Conditional distributions and densisties will apear very often. The various definitions, if presented in their final form, could be very confusing. They seem artificial, unrelated, and at best difficult to remember. If, however, all definitions are expressed as conditional probabilities of events, then the confusion is eliminated and the concepts become almost self evident.

Given an event θ with nonzero probability,

$$P(\theta) > 0 \tag{1}$$

it can define the conditional probability of φ assuming θ, by

$$P(\varphi|\ \theta) = \frac{P(\varphi\theta)}{P(\theta)} \tag{2}$$

In words $P(\varphi|\ \theta) > 0$ equals to probability of the event $\varphi\theta$, the part of φ included in θ, divided by the probability of θ. Clearly, if φ and θ have no common elements (mutual, exclusively) then $P(\varphi|\ \theta) = 0$.

If recall from eq.(2) that, with θ an event such that

$$P(\varphi|\ \theta) \neq 0 \qquad (3)$$

The conditional probability of φ, assuming θ, is given by eq.(1).

In this section need to express φ and θ or both in terms of the random variable.

Definition of $F_x(x|\ \theta)$ *and* $f_x(x|\ \theta)$**.** The conditional distribution is defined as the conditional probability of the event $F_x(x|\ \theta)$ of the random variable X , assuming θ, defined as the conditional probability of the event $\{X \ \le x\}$:

$$F_x(x|\ \theta) = P\{X \le x \big| \theta\} = \frac{P(X \le x, \theta\}}{P(\theta)} \quad (4)$$

where $\{X \ \le x, \theta\}$is the event consisting of all outcomes ξ such that

$$X(\xi\ ) \le x \quad \xi \in \theta \qquad (5)$$

That is, the set product of the events { $X \le x$ } and θ.

The above is nothing else but definition in the experiment θ, φ, $P(\varphi|\ \theta)$. Therefore all properties of ordinary distributions apply also to $F_x(x|\ \theta)$**.** Then, just have dropping the subscript

$$F\ (x|\ \theta) = 1 \qquad F(-\infty|\theta) = 0 \quad (6)$$

and

$$F\ (x_2|\ \theta) - F(x_1|\theta) = P\{x_1 < X < x_2 \big| \theta = \frac{P\{x_1 \le X \le x_2, \theta\}}{P(\theta)} \quad (7)$$

If assumed that X continuous type. Its conditional density $f\ (x|\ \theta)$ is defined as the derivative of $F\ (x|\ \theta)$:

$$f\ (x|\ \theta) = \frac{dF(x|\theta)}{dx} = \lim_{\Delta x \to 0} \frac{P\{x \le X \le x + \Delta x|\theta)}{\Delta x} \quad (8)$$

and it has all properties of ordinary densities;

$$\int_{-\infty}^{\infty} f\ (x|\ \theta)dx = F(\infty|\theta) - F(-\infty|\theta) = 1 \ (9)$$

3 Conditional inference

In Bayesian decision analysis, it is supposed that a choice has to be made from a set of available actions $(a_1, a_2, ...a_n)$, where the payoff of utility of a given action depends on a state of nature, say θ, which is unknown.

3.1 Bayes principle in inference

Suppose that $Y' = (y_1, y_2, ...y_n)$ is a vector of n observations whose probability distribution $p(y|\theta)$ depends on the values of k parameters $\theta' = (\theta_1, \theta_2, ...\theta_n)$. Suppose also that θ itself has a probability distribution $P(\theta)$. Then,

$$p(y|\theta)p(\theta) = p(y, \theta) = p(\theta|y)p(y) \quad (10)$$

Given the observed data y, distribution of the conditional θ is

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \qquad (11)$$

Also, can write

$$p(y) = Ep(y|\theta) = c^{-1} = \begin{cases} \int p(y|\theta)p(\theta)d\theta \\ \theta \text{ continuous (12)} \\ \sum p(y|\theta)p(\theta) \\ \theta \text{ discrete} \end{cases}$$

where the sum or the integral is taken over the admissible range of θ, and where $E[f(\theta)]$ is the mathematical expectation of $f(\theta)$ with respect to the distribution $p(\theta)$. Thus may write eq.(12) alternatively as

$$p(\theta|y) = cp(y|\theta)p(\theta) \qquad (13)$$

The statement eq.(10) or its equivalent eq.(13), is usually referred to as *Bayes theorem.* In this expression, $p(\theta)$, which tells the quantity c is merely a normalizing constant neccessary to ensure that the posterior distribution what is known about θ without knowledge of the data, the distribution of θ a priori. Correspondingly, $p(\theta|y)$, which tells what is known about θ given knowledge of the data, is called the posterior distribution of given y, or the distribution of θ a posteriori. $p(\theta|y)$ integrates or sums to one.

Likelihood function. Now given the data y, $p(\theta|y)$ in eq.(13) may be regarded as a function not of y but of θ. When so regarded, following Fisher [1], it is called the likelihood function of θ for a given y and can be written $l(\theta|y)$. Thus can write Bayes formula as

$$p(\theta|y) = l(\theta|y)p(\theta) \qquad (14)$$

In order words, then, Bayes theorem tells that the probability distribution for θ posterior to the data y is proportional to the product of the distribution for θ prior to the data and the likelihood for θ given y. That is, θ coming from the data.

$$posterior \quad distribution \propto likelihood \quad x \qquad (14a)$$

$$prior\ distribution$$

The likelihood function $l(\theta|y)$ plays a very important role in this formula. It is the function through which the data y modifies prior knowledge of θ, it can therefore be regarded as representing the information about.

The likelihood function is defined up to a multiplicative constant, that is, multiplication by a constant leaves the likelihood unchanged. This is in accord with the role it plays in Bayes formula, since multiplying the likelihood function by an arbitrary constant will have no effect on the posterior distribution of θ. The constant will cancel upon normalizing the product on the write side of eq.(14). It is only the relative value of the likelihood which is of importance .

When the integral $\int l(\theta|y)p(\theta)$, taken over the admissible range of θ, is finite then occasionally it will be convenient to refer to the quantity

$$\frac{l(\theta|y)}{\int l(\theta|y)p(\theta)} \qquad (15)$$

It is so called standardized likelihood, that is, the likelihood scaled so that the area, volume, or hypervolume under the curve, surface, or hypersurface, is one. Eq.(14a)provides a mathematical formulation of how previous knowledge may be combined with new knowledge. Indeed, the theorem allows to continually update information about a set of parameters θ as more observations are taken.

Thus, suppose that have an initial sample of observations y_1, then Bayes formula gives

$$p(\theta|y) \propto l(\theta|y_1)p(\theta) \quad (16)$$

When, suppose that have a second sample of observations y_2 distribured independently of the first sample, then

$$p(\theta|y_2, y_1) \propto p(\theta)l(\theta|y_1)l(\theta|y_2) \qquad (17)$$
$$\propto p(\theta|y_1)l(\theta|y_2)$$

The expression eq.(17) is predisely of the same form as eq.(16) except that $p(\theta|y_1)$, the posterior distribution for θ given y_1, plays the role of the prior distribution for the second sample. Obviously this process can be repeated any number of times. In particular, there is n independent observations, the posterior distribution can, if desired, be recalculated after each new observation, so that at the m th stage the likelihood associated with the m th observation to give the new posterior distribution

$$p(\theta|y_1, y_2, ...y_m) \propto p(\theta|y_1, y_2, ...y_{m-1}) l(\theta|y_m, \quad m = 2,.3,......, n \qquad (18)$$

where $p(\theta|y_1) \propto p(\theta) l(\theta|y_1)$.

Thus Bayes theorem describes, in a fundamental way, the process of learning from experience, and shows how knowledge about the state of nature represented by θ is continually modified as new data becomes available.

4 Probability interpreted as frequencies

Applications of Bayes formula with probability interpreted as frequencies has been questioned about some difficulties. The difficulties concern

1. The meaning of probability, and
2. The choice of, and necessity for, the prior distribution.

Specific examples can be found of applications of Bayes formula where the probabilities involved may be directly interpreted in terms of frequencies and may therefore be said to be objective, and where the prior probabilities can be supposed exactly known. The validity of applications of this sort has not been in serious dispute.

Some examples of this situation were described in literature [2],[3]. Other application of this sort are to be found in the theory of design sampling inspection schemes[3]. In these exampes, all the probabilities, both prior and posterior, are objective in the sense that they may be given a direct limiting frequency interpretation and are, in principle, subject to experimental confirmation.

5 Subjective probability

Let consider probability as a mathematical expression of our degree of belief with respect to a certain proposition, in this context the concept of verification of probabilities by repeated experimental trials is regarded merely as a means of calibrating a subjective attitude.

The actual elucidation of what is believed by a particular person can be attempted in terms of betting odds. If, for example, the value of a continuous parameter θ is in question, may, in suitable circumstances, infer an experimenter's prior distribution by asking at what value θ_0 would be prepared to bet at particular odds that $\theta > \theta_0$. Given that a subjective probability distribution of this kind represents *a priori* what a person believes, then the posterior distribution obtained by combining this prior with the likelihood function shows how the prior beliefs are modified by information coming from the data.

6 Inference analysis

Statistical inference means inference about the state of nature made in terms of probability, and a statistical inference problem is refarded as solved as soon as can make an appropriate probability statement about the state of nature in question. Important as the topic is, concern will not be with ern. Usually the state of nature is described by the value of one or more parameters. Such a parameter could, for example, be the velocity of light or the thermal conductivity of a certain alloy. Thus, a solution to the inference problem is supplied by a posterior distribution $p(\theta|y)$ which shows what can be inferred about the parameters θ from the data y given a relevant prior state of knowledge represented by $p(\theta)$.

Let consider which concerning the estimation of the location parameter θ of a Normal distribution. In general, if the prior distribution is Normal $N(\theta_o, \sigma_0{}^2)$ and n independent observations with average $\bar{y}$ are taken from the distribution $N(\bar{\theta}, \bar{\sigma}^2)$, then from

$$\bar{\theta} = \frac{1}{w_0 + w_n}(w_0\theta_0 + w_n\bar{y}),$$
$$\frac{1}{\bar{\bar{\sigma}}^2} = w_0 + w_n \qquad (19)$$

with $w_0 = 1/\sigma_0^2$ and $w_n = n/\sigma_n^2$.
the posterior distribution of θ is

$$\theta \approx N(\bar{\theta}_n, \bar{\sigma}_n{}^2)\text{,with} \qquad (20)$$
$$\bar{\bar{\theta}}_n = \frac{1}{w_0 + w_n}(w_0\theta_0 + w_n\bar{y}), \text{and} \qquad (21)$$
$$\bar{\bar{\sigma}}^{-2} = w_0 + w_n, \qquad (22)$$

where $w_0 = \sigma_0^{-2}$ is the weight associated with the prior distribution and $w_n = n/\sigma_0^2$ is the weight associated with the likelihood. In this expression, if w_0 is small compared with w_n then approximately the posterior distribution is numerically equal to the standardized likelihood, and is

$$N(\bar{y}, \frac{\sigma^2}{n}) \qquad (23)$$

Strictly speaking, this result is attained only when the prior variance σ_0^2 becomes infinite so that w_0 is zero. Such a limiting prior distribution would. However, by itself make little theoretical or practical sense. For, when $\sigma_0^2 \to \infty$, in the limit the prior density becomes uniform over the entire line from $-\infty$ to $+\infty$, and is therefore not a proper density function. Furthermore, it represents a situation where all values of θ from $-\infty$ to $+\infty$ are equally accepable a priori. But it is difficult, if not impossible, to imagine a practical situation where sufficiently extreme values could not be virtually ruled out. The practical situation is represented not by the limiting case where $w_0 = 0$, but by the case where w_0 is small compared with w_n, that is, where the prior is locally flat so that the likelihood dominates the prior.

It is, therefore, important to note that the use of the limiting posterior in eq.(23) corresponding to $w_0 = 0$ to supply a numerical approximation to the practical situation is not the same thing as assuming w_0 is actually zero. Limiting cases of this kind are frequently used, but it must be remembered this is for the purpose of supplying a numerical approximation and for this purpose only.

Proper and Improper prior distribution. A basic property of a probability density function $f(x)$ is that it integrates or sums over its admissible range to 1, that is,

$$\left.\begin{matrix} \int f(x)dx \\ \sum f(x) \end{matrix}\right\} = 1 \left\{\begin{matrix} (x \;\; continuous) \\ (x \;\; discrete) \end{matrix}\right. \qquad (24)$$

If $f(x)$ is uniform over the entire line from $-\infty$ to $+\infty$

$$f(x) = k, \quad -\infty < x < \infty, \quad k > 0, \qquad (25)$$

then it is not a proper density since the integral

$$\int_{-\infty}^{\infty} f(x)dx = k\int_{-\infty}^{\infty} dx \qquad (26)$$

does not exist no matter how small k is. Density functions of this kind are sometimes called improper distributions. As another example, the function

$$f(x) = kx^{-2}, \quad 0 < x < \infty, \quad k > 0, \quad (27)$$

is also improper. Density functions of the types to eq.(24) and eq.(25) are frequently employed to represent the local benhavior of the prior distribution in the region where the likelihood is appreciable, but not over its entire admissible range. By supposing that to a sufficient approximation the prior follows the form eq.(26) and eq.(27) only over the range of appreciable likelihood and that it suitably tails to zero outside that range, to ensure that the priors actually used are proper. Thus, by employing the distribution in a way that makes practical sense need to relieve of a theoretical difficulty.

7 Conclusion

In this paper stochastic information processing was examined. Inference analysis was employed to information processing. Bayes formula states that the probabilities may be directly interpreted in terms of frequencies and may therefore be said to be objective.

Prior and posterior distributions include likelyhood function. Proper and improper prior distribution were considered. Prior belief function was studied and weighting factors were defined.

Notation

E -expectation

$l(\theta|y)$ -likelihood function

$P(x)-$ distribution probability function

$p(x)-$ density probability function

X - random variable

x - value of random variable X

Greek Symbols

φ -set

ϕ -experiment

σ -variance

θ -event

Θ -arbitrary event

References

[1] G.E.P Box, and G.C.Tiao, *Bayesian Inference in Statistical Analysis,* Addison-Wesley Publishing Company, MA, U.S.A.,1973.

[2] A.Papoulus, Probability, *Random variable and stochastic processes,* McGraw-Hill, Book Company,1965.

[3] J.Savkovic-Stevanovic, *Stochastic models in Analysis and Optimization*, University of Belgrade,1982.

[4] A.H.Bowker, and J. Lieberman ,*Engineering Statistics*, Prentice Hall Inc., Englewood Cliffs, New Jersey,1972.

[5] J. Savković-Stevanovič, T. Mosorinac , Complex system for cognitive products modelling, Proc.EUROSIM 2007(B.Zupančič R.Karba, S.Blazić) ,pp.393, 9-13 September, Ljubljana, Slovenia, 2007.

[6] J. Savkovic-Stevanovic, J., Chapter 6 in the book *Process modelling and simulation,* University of Belgrade,1995.