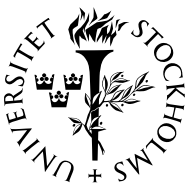
Royal Institute of Technology
Stockholm, Sweden

Stockholm
University



Department of Numerical Analysis and Computing Science
TRITA-NA-P9806 • ISSN 1101-2250 • ISRN KTH/NA/P–98/06SE

Statistics of the Information Component in Bayesian Neural Networks

Timo Koski and Roland Orre

Report from Studies of Artificial Neural Systems, (SANS)

NADA (Numerisk analys och datalogi)
KTH
100 44 Stockholm

Department of Numerical Analysis
and Computing Science
Royal Institute of Technology
S-100 44 Stockholm, Sweden

Statistics of the Information Component in Bayesian Neural Networks

Timo Koski and Roland Orre
timo@math.kth.se, orre@nada.kth.se

TRITA-NA-P9806

Abstract

In previous work approximate solutions have been used for expectation and variance of the *information component (IC)*. This report presents an analytical approach to calculate exact expressions for the expectation and variance of the information component (IC). The IC is used in a Bayesian neural network [3] as a weight between neurons representing discrete events. The IC relates the information possessed about one state of one variable with one state of another variable, and is used for calculation of a posterior probability distribution conditioned on a set of given input events. It is used as a measure of disproportionality in data mining [1]. The mutual information between two variables, as defined in information theory [4], can in its discrete form be regarded as a weighted sum of ICs. The expectation of the IC provides a measure of the strength of an association between two states and its variance a measure of the uncertainty, which is essential for low counter values.

1 Introduction

We are concerned with computing the expectation and variance of

$$IC_{ij} = \log \frac{\mathbf{p}_{ij}}{\mathbf{p}_i \cdot \mathbf{p}_j}.$$

We consider $\mathbf{p}_{ij}$, $\mathbf{p}_i$ and $\mathbf{p}_j$ as random variables. We know or assume that the marginal distributions of these variables are $Beta$-distributions with respective parameters.

2 Expectation of $\log \mathbf{p}$

To compute the expectation of IC_{ij} we observe that

$$E\left[IC_{ij}\right] = E \log \mathbf{p}_{ij} - E \log \mathbf{p}_i - E \log \mathbf{p}_j,$$

which means that the marginal distributions are sufficient to compute the required expectations. We do this in a generic manner, i.e. we consider $E \log \mathbf{p}$ when $\mathbf{p} \in B(a,b)$, where $B(a,b)$ is the $Beta$ -distribution with parameters a and b. This means that $\mathbf{p}$ has the probability density ([6, p. 173])

$$f_{\mathbf{p}}(p) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} p^{a-1} \cdot (1-p)^{b-1}, \tag{2.1}$$

where $\Gamma(z)$ is Euler's gamma function. The $Beta$ -function is

$$B(a,b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)} \tag{2.2}$$

so that

$$f_{\mathbf{p}}(p) = \frac{1}{B(a,b)} p^{a-1} \cdot (1-p)^{b-1}. \tag{2.3}$$

By a standard result about expectation of functions of random variables we have

$$E \log \mathbf{p} = \frac{1}{B(a,b)} \int_0^1 \log p \cdot p^{a-1} \cdot (1-p)^{b-1} dp. \tag{2.4}$$

Next we prove the following

Proposition 2.1 $$E \log \mathbf{p} = \frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)}.$$

Proof: In view of (2.4) we compute first the integral $\int_0^1 \log p \cdot p^{a-1} \cdot (1-p)^{b-1} dp$. This is done as follows. We recall that if $\frac{d}{da}$ denotes differentiation with respect to a then

$$\frac{d}{da} x^a = \frac{d}{da} e^{a \log x} = \log x \cdot x^a, \tag{2.6}$$

using the natural logarithms. Thus, since

$$\int_0^1 p^{a-1} \cdot (1-p)^{b-1} dp = B(a,b),$$

we have from (2.6) that

$$\int_0^1 \log p \cdot p^{a-1} \cdot (1-p)^{b-1} dp = \int_0^1 \frac{d}{da} p^a \cdot p^{-1} \cdot (1-p)^{b-1} dp = \frac{\partial}{\partial a} B(a,b). \tag{2.7}$$

assuming it is justified to exchange the order of the differentiation and the integral. Next we observe that if $f'(x)$ is the first derivative of $f(x)$, then

$$\frac{d}{dx} \log f(x) = \frac{f'(x)}{f(x)} \tag{2.8}$$

and

$$\begin{aligned} E \log \mathbf{p} &= \frac{1}{B(a,b)} \int_0^1 \log p \cdot p^{a-1} \cdot (1-p)^{b-1} dp \\ &= \frac{1}{B(a,b)} \cdot \frac{\partial}{\partial a} B(a,b) \qquad (2.9) \\ &= \frac{\partial}{\partial a} \log B(a,b). \qquad (2.10) \end{aligned}$$

By (2.2) we have

$$\frac{\partial}{\partial a} \log B(a,b) = \frac{\partial}{\partial a} \log \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}. \tag{2.11}$$

By straightforward differentiation this becomes

$$\frac{\partial}{\partial a} \log B(a,b) = \frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)}. \tag{2.12}$$

In other words we have obtained (2.5) above as claimed. □

2.1 A series formula

By a well known formula recapitulated and proved in [5, p. 467] it holds for z which is not a negative integer that

$$\frac{\Gamma'(z)}{\Gamma(z)} = -\gamma - \frac{1}{z} + \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{z+n} \right), \tag{2.13}$$

where $\gamma \approx 0.577215..$ is Euler's constant. We can also write this as that

$$\frac{\Gamma'(z)}{\Gamma(z)} = -\gamma - \frac{1}{z} + z \cdot \sum_{n=1}^{\infty} \left(\frac{1}{n \cdot (z+n)} \right). \tag{2.14}$$

If we now apply (2.14) in the left hand side of (2.5) we get by some elementary simplifications that

$$\frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)} = \frac{b}{a \cdot (a+b)} - b \cdot \sum_{n=1}^{\infty} \frac{1}{(a+n) \cdot (a+b+n)}. \tag{2.15}$$

Hence we have proved

Proposition 2.2 *If* $\mathbf{p}$ *is* $B(a,b)$ *- distributed and if* a *and* b *are not negative integers, then*

$$E \log \mathbf{p} = \frac{b}{a \cdot (a+b)} - b \cdot \sum_{n=1}^{\infty} \frac{1}{(a+n) \cdot (a+b+n)}. \tag{2.16}$$

□

3 Variance of $\log \mathbf{p}$

Proposition 3.1 *If* $\mathbf{p}$ *is* $B(a,b)$ *- distributed and if* a *and* b *are not negative integers, then* $Var\,[\log \mathbf{p}]$ *can be expressed as a convergent sum.*

Proof: Let us consider the second moment of $\log \mathbf{p}$. We have

$$E \log^2 \mathbf{p} = \frac{1}{B(a,b)} \int_0^1 \log^2 p \cdot p^{a-1} \cdot (1-p)^{b-1} dp, \tag{3.1}$$

where $\log^2 p = (\log p)^2$. But since from (2.6)

$$\frac{d^2}{da^2} x^a = \log x \cdot \frac{d}{da} x^a = \log^2 x \cdot x^a \tag{3.2}$$

we get by the above (3.2) that

$$E \log^2 \mathbf{p} = \frac{1}{B(a,b)} \cdot \frac{\partial^2}{\partial a^2} B(a,b). \tag{3.3}$$

On the other hand using $\frac{\partial}{\partial a} B(a,b) = B(a,b) \cdot \frac{\partial}{\partial a} \log B(a,b)$ we get

$$\frac{\partial^2}{\partial a^2} B(a,b) = \frac{\partial}{\partial a} B(a,b) \cdot \frac{\partial}{\partial a} \log B(a,b) + B(a,b) \frac{\partial^2}{\partial a^2} \log B(a,b)$$

so that

$$\frac{1}{B(a,b)} \cdot \frac{\partial^2}{\partial a^2} B(a,b) = \frac{1}{B(a,b)} \frac{\partial}{\partial a} B(a,b) \cdot \frac{\partial}{\partial a} \log B(a,b) + \frac{\partial^2}{\partial a^2} \log B(a,b). \tag{3.4}$$

This yields

$$\frac{1}{B(a,b)} \cdot \frac{\partial^2}{\partial a^2} B(a,b) = \left(\frac{\partial}{\partial a} \log B(a,b) \right)^2 + \frac{\partial^2}{\partial a^2} \log B(a,b). \tag{3.5}$$

From (2.12) we get

$$\frac{\partial^2}{\partial a^2} \log B(a,b) = \frac{\partial}{\partial a} \frac{\Gamma'(a)}{\Gamma(a)} - \frac{\partial}{\partial a} \frac{\Gamma'(a+b)}{\Gamma(a+b)}. \tag{3.6}$$

This is a useful formula due to the fact that derivatives of the natural logarithm of the Gamma function (*polygamma functions*) have a known series representation. In fact it holds that if $L'(z) = \Gamma'(z)/\Gamma(z)$, then for $k \geq 2$

$$\frac{d^k}{dz^k} L(z) = (-1)^k (k-1)! \sum_{n=0}^{\infty} (z+n)^{-k}, \tag{3.7}$$

see [5, p. 467]. Thus we have

$$\frac{\partial}{\partial a} \frac{\Gamma'(a)}{\Gamma(a)} = \sum_{n=0}^{\infty} (a+n)^{-2} \tag{3.8}$$

and

$$\frac{\partial}{\partial a}\frac{\Gamma'(a+b)}{\Gamma(a+b)} = \sum_{n=0}^{\infty}(a+b+n)^{-2}. \tag{3.9}$$

Thus

$$\frac{\partial^2}{\partial a^2}\log B(a,b) = \sum_{n=0}^{\infty}\left((a+n)^{-2} - (a+b+n)^{-2}\right), \tag{3.10}$$

which equals

$$\frac{\partial^2}{\partial a^2}\log B(a,b) = \sum_{n=0}^{\infty}\frac{b^2+2ab+2bn}{((a+n)\cdot(a+b+n))^2}, \tag{3.11}$$

Hence from (3.3), (3.4) and (2.9) we get

$$\begin{aligned} E\log^2\mathbf{p} &= \frac{1}{B(a,b)}\cdot\frac{\partial^2}{\partial a^2}B(a,b) \\ &= \left(\frac{\partial}{\partial a}\log B(a,b)\right)^2 + \frac{\partial^2}{\partial a^2}\log B(a,b) \\ &= (E\log\mathbf{p})^2 + \sum_{n=0}^{\infty}\frac{b^2+2ab+2bn}{((a+n)\cdot(a+b+n))^2}. \end{aligned} \tag{3.12}$$

Next, we recall that

$$Var[\log\mathbf{p}] = \left(E\log^2\mathbf{p}\right) - (E\log\mathbf{p})^2.$$

Thus we have proved

$$Var[\log\mathbf{p}] = \sum_{n=0}^{\infty}\frac{b^2+2ab+2bn}{((a+n)\cdot(a+b+n))^2}. \tag{3.13}$$

□

4 Information Component

In view of the preceding (2.16 and 3.13) we can compute expectation and variance of $IC_{ij} = \log\frac{\mathbf{p}_{ij}}{\mathbf{p}_i\cdot\mathbf{p}_j}$ if $\mathbf{p}_{ij}, \mathbf{p}_i$ and $\mathbf{p}_j$ have their respective Beta distributions. However, when calculating the variance we can not expect $\mathbf{p}_{ij}$, $\mathbf{p}_i$ and $\mathbf{p}_j$ to be independent which is why we need the joint distributions to compute the covariances $cov(\log\mathbf{p}_{ij}, \log\mathbf{p}_i)$, $cov(\log\mathbf{p}_{ij}, \log\mathbf{p}_j)$ and $cov(\log\mathbf{p}_i, \log\mathbf{p}_j)$. Expectation and variance of IC_{ij} can then be expressed as

$$E[IC_{ij}] = E\log\mathbf{p}_{ij} - E\log\mathbf{p}_i - E\log\mathbf{p}_j, \tag{4.14}$$

$$\begin{aligned} Var[IC_{ij}] &= Var\log\mathbf{p}_{ij} + Var\log\mathbf{p}_i + Var\log\mathbf{p}_j + \\ &\quad -2cov(\log\mathbf{p}_{ij}, \log\mathbf{p}_i) - 2cov(\log\mathbf{p}_{ij}, \log\mathbf{p}_j) + 2cov(\log\mathbf{p}_i, \log\mathbf{p}_j). \end{aligned} \tag{4.15}$$

References

[1] Andrew Bate, Marie Lindquist, I. Ralph Edwards, Sten Olsson, Roland Orre, Anders Lansner, and Rogelio Melhado De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, in press, 1998.

[2] Anders Holst. *The Use of a Bayesian Neural Network Model for Classification Tasks.* PhD thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden, September 1997. TRITA-NA-P9708.

[3] Anders Holst and Anders Lansner. A Bayesian neural network with extensions. Tech. Rep. TRITA-NA-P9325, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden, 1993.

[4] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[5] K. R. Stromberg. *An Introduction to Classical Real Analysis.* Wadsworth International Group, Bellmont, California, 1981.

[6] S. S. Wilks. *Mathematical Statistics.* John Wiley & Sons, New York, 1962.