Baltzer Journals September, 1995

An Upper Bound on the Bayesian Error Bars for Generalized Linear Regression*

Cazhaow S. Qazaz
Christopher K. I. Williams
Christopher M. Bishop
Neural Computing Research Group
Dept. of Computer Science and Applied Mathematics
Aston University, Birmingham, UK.
E-mail: ncrg@aston.ac.uk
http://www.ncrg.aston.ac.uk/

Technical Report: NCRG/96/005

In the Bayesian framework, predictions for a regression problem are expressed in terms of a distribution of output values. The mode of this distribution corresponds to the most probable output, while the uncertainty associated with the predictions can conveniently be expressed in terms of error bars. In this paper we consider the evaluation of error bars in the context of the class of generalized linear regression models. We provide insights into the dependence of the error bars on the location of the data points and we derive an upper bound on the true error bars in terms of the contributions from individual data points which are themselves easily evaluated.

1 Introduction

Many applications of neural networks are concerned with the prediction of one or more continuous output variables, given the values of a number of input variables. As well as predictions for the outputs, it is also important to provide some measure of uncertainty associated with those predictions.

The Bayesian view of regression leads naturally to two contributions to the error bars. The first arises from the intrinsic noise on the target data, while the second comes from the uncertainty in the values of the model parameters as a consequence of having a finite training data set [1, 2]. There may also be a third contribution which arises if the true function is not contained within the space of models under consideration, although we shall not discuss this possibility further.

In this paper we focus attention on a class of universal non-linear approximators constructed from linear combinations of fixed non-linear basis functions, which we shall refer to as *generalized linear regression* models. We first review the Bayesian treatment of learning in such models, as well as the calculation of error bars [3]. Then, by considering

*This work was supported by EPSRC grant GR/J75425.

the contributions arising from individual data points, we provide insight into the nature of the error bars and their dependence on the location of the data in input space. This in turn leads to the key result of the paper which is an upper bound on the true error bars expressed in terms of the single-data-point contributions. Our analysis is very general and is independent of the particular form of the basis functions.

2 Bayesian error bars

We are interested in the problem of predicting the value of a noisy output variable t given the value of an input vector $\boldsymbol{x}$. Throughout this paper we shall restrict attention to regression for a single variable t since all of the results can be extended in a straightforward way to multiple outputs. To set up the Bayesian formalism we begin by defining a model for the distribution of t conditional on $\boldsymbol{x}$. This is most commonly chosen to be a Gaussian function in which the mean is governed by the output $y(\boldsymbol{x};\boldsymbol{w})$ of a network model, where $\boldsymbol{w}$ is a vector of adaptive parameters (weights and biases). Thus we have

$$p(t|\boldsymbol{x},\boldsymbol{w}) = \frac{1}{(2\pi\sigma_\nu^2)^{1/2}} \exp\left\{-\frac{(y(\boldsymbol{x};\boldsymbol{w})-t)^2}{2\sigma_\nu^2}\right\} \tag{1}$$

where σ_ν^2 is the variance of the distribution.

In the Bayesian framework, our state of knowledge of the weight values is expressed in terms of a distribution function over $\boldsymbol{w}$. This is initially set to some *prior* distribution, from which a corresponding *posterior* distribution can be computed using Bayes' theorem once we have observed the training data. A common choice of prior is a Gaussian distribution of the form

$$p(\boldsymbol{w}) = \frac{1}{(2\pi)^{M/2}} |\boldsymbol{S}|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{w}\right\} \tag{2}$$

where M is the total number of weight parameters, $\boldsymbol{S}$ is the inverse covariance matrix of the distribution, and $|\boldsymbol{S}|$ denotes the determinant of $\boldsymbol{S}$. Since the parameters in $\boldsymbol{S}$ control the distribution of other parameters they are often referred to as *hyperparameters*. The noise variance σ_ν^2 is commonly also called a hyperparameter since, in a Bayesian framework, it can be treated using similar techniques to $\boldsymbol{S}$. Here we shall assume that the values of σ_ν^2 and $\boldsymbol{S}$ are fixed and known.

The training data set D consists of N pairs of input vectors $\boldsymbol{x}^n$ and corresponding target values t^n where $n = 1, \ldots, N$. From this data set, together with the noise model (1), we can construct the *likelihood* function given by

$$p(D|\boldsymbol{w}) = \prod_{n=1}^{N} p(t^n|\boldsymbol{x}^n,\boldsymbol{w}) = \frac{1}{(2\pi\sigma_\nu^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\nu^2}\sum_{n=1}^{N}\{y(\boldsymbol{x}^n;\boldsymbol{w})-t^n\}^2\right\} \tag{3}$$

We can then combine the likelihood function and the prior using Bayes' theorem to obtain the posterior distribution of weights given by $p(\boldsymbol{w}|D) = p(D|\boldsymbol{w})p(\boldsymbol{w})/p(D)$. The predictive distribution of t given a new input $\boldsymbol{x}$ can then be written in terms of the posterior distribution in the form

$$p(t|\boldsymbol{x},D) = \int p(t|\boldsymbol{x},\boldsymbol{w})p(\boldsymbol{w}|D)\,d\boldsymbol{w} \tag{4}$$

where $p(t|\boldsymbol{x}, \boldsymbol{w})$ is given by (1).

Throughout this paper we consider a particular class of non-linear models of the form

$$y(\boldsymbol{x}; \boldsymbol{w}) = \sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{w} \tag{5}$$

which we shall call generalized linear regression models. Here the $\phi_j(\boldsymbol{x})$ are a set of fixed non-linear basis functions, with generally one of the basis functions $\phi_1 = 1$ so that w_1 plays the role of a bias parameter. Such models possess universal approximation capabilities for reasonable choices of the $\phi_j(\boldsymbol{x})$, while having the advantage of being linear in the adaptive parameters $\boldsymbol{w}$.

Since (5) is linear in $\boldsymbol{w}$, both the noise model $p(t|\boldsymbol{x}, \boldsymbol{w})$ and the posterior distribution $p(\boldsymbol{w}|D)$ will be Gaussian functions of $\boldsymbol{w}$. It therefore follows that, for a Gaussian prior of the form (2), the integral in (4) will be Gaussian and can be evaluated analytically to give a predictive distribution $p(t|\boldsymbol{x}, D)$ which will be a Gaussian function of t. The mean of this distribution is given by $y(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{MP}})$ where $\boldsymbol{w}_{\mathrm{MP}}$ is found by minimizing the regularized error function

$$\frac{1}{2\sigma_\nu^2} \sum_{n=1}^{N} \{\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}^n)\boldsymbol{w} - t^n\}^2 + \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{w} \tag{6}$$

and is therefore given by the solution of the following linear equations

$$\boldsymbol{A}\boldsymbol{w}_{\mathrm{MP}} = \sigma_\nu^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{t} \tag{7}$$

where $\boldsymbol{t}$ is a column vector with elements t^n, $\boldsymbol{A}$ is the Hessian matrix given by

$$\boldsymbol{A} = \frac{1}{\sigma_\nu^2} \sum_{n=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}^n)\boldsymbol{\phi}(\boldsymbol{x}^n)^{\mathrm{T}} + \boldsymbol{S} = \frac{1}{\sigma_\nu^2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \boldsymbol{S} \tag{8}$$

and $\boldsymbol{\Phi}$ is the $N \times M$ *design matrix* with elements $\Phi_{nj} = \phi_j(\boldsymbol{x}^n)$. Solving for $\boldsymbol{w}_{\mathrm{MP}}$ and substituting into (5) we obtain the following expression for the corresponding network output

$$y_{\mathrm{MP}}(\boldsymbol{x}) = \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{w} = \sigma_\nu^{-2}\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{t} \tag{9}$$

The covariance matrix for the posterior distribution $p(\boldsymbol{w}|D)$ is given by the inverse of the Hessian matrix. Together with (4) this implies that the total variance of the output predictions is given by

$$\sigma^2(\boldsymbol{x}) = \sigma_\nu^2 + \sigma_w^2(\boldsymbol{x}) = \sigma_\nu^2 + \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{A}^{-1}\boldsymbol{\phi}(\boldsymbol{x}) \tag{10}$$

Here the first term represents the intrinsic noise on the target data, while the second term arises from the uncertainty in the weight values as a consequence of having a finite data set.

3 An upper bound on the error bars

We first consider the behaviour of the error bars when the data set consists of a single data point. As well as providing important insights into the nature of the error bars, it also leads directly to an upper bound on the true error bars.

In the absence of data, the variance is given from (8) and (10) by

$$\sigma^2(\boldsymbol{x}) = \sigma_\nu^2 + \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{S}^{-1}\boldsymbol{\phi}(\boldsymbol{x}) \tag{11}$$

where the second term, due to the prior, is typically much larger than the noise term σ_ν^2. If we now add a single data point located at $\boldsymbol{x}^0$ then the Hessian becomes $\boldsymbol{S} + \sigma_\nu^{-2}\boldsymbol{\phi}(\boldsymbol{x}^0)\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}^0)$. To find the inverse of the Hessian we make use of the identity

$$\left(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}\right)^{-1} = \boldsymbol{M}^{-1} - \frac{\left(\boldsymbol{M}^{-1}\boldsymbol{v}\right)\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}^{-1}\right)}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}} \tag{12}$$

which is easily verified by multiplying both sides by $(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}})$. The variance at an arbitrary point $\boldsymbol{x}$ for a single data point at $\boldsymbol{x}^0$ is then given by

$$\sigma^2(\boldsymbol{x}) = \sigma_\nu^2 + C(\boldsymbol{x},\boldsymbol{x}) - \frac{C(\boldsymbol{x},\boldsymbol{x}^0)^2}{\sigma_\nu^2 + C(\boldsymbol{x}^0,\boldsymbol{x}^0)} \tag{13}$$

where we have defined the *prior covariance function*

$$C(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{S}^{-1}\boldsymbol{\phi}(\boldsymbol{x}') \tag{14}$$

The first two terms on the right hand side of (13) represent the variance due to the prior alone, and we see that the effect of the additional data point is to reduce the variance from its prior value, as illustrated for a toy problem in Figure 1. From (13) we see that the length scale of this reduction is related to the prior covariance function $C(\boldsymbol{x},\boldsymbol{x}')$.

If we evaluate $\sigma^2(\boldsymbol{x})$ in (13) at the point $\boldsymbol{x}^0$ then we can show that the error bars satisfy the upper bound $\sigma^2(\boldsymbol{x}^0) \leq 2\sigma_\nu^2$. Since the noise level is typically much less than the prior variance level, we see that the error bars are pulled down very substantially in the neighbourhood of the data point. Again, this is illustrated in Figure 1.

We now extend this analysis to provide an upper bound on the error bars. Suppose we have a data set consisting of N data points (at arbitrary locations) and we add an extra data point at $\boldsymbol{x}^{N+1}$. Using (8) the Hessian $\boldsymbol{A}_{N+1}$ for the $N+1$ data points can be written in terms of the corresponding Hessian $\boldsymbol{A}_N$ for the original N data points in the form

$$\boldsymbol{A}_{N+1} = \boldsymbol{A}_N + \sigma_\nu^{-2}\boldsymbol{\phi}(\boldsymbol{x}^{N+1})\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}^{N+1}) \tag{15}$$

Using the identity (12) we can now write the inverse of $\boldsymbol{A}_{N+1}$ in the form

$$\boldsymbol{A}_{N+1}^{-1} = \boldsymbol{A}_N^{-1} - \frac{\boldsymbol{A}_N^{-1}\boldsymbol{\phi}(\boldsymbol{x}^{N+1})\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}^{N+1})\boldsymbol{A}_N^{-1}}{\sigma_\nu^2 + \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}^{N+1})\boldsymbol{A}_N^{-1}\boldsymbol{\phi}(\boldsymbol{x}^{N+1})} \tag{16}$$

Substituting this result into (10) we obtain

$$\sigma_{N+1}^2(\boldsymbol{x}) = \sigma_N^2(\boldsymbol{x}) - \frac{\left[\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}^{N+1})\boldsymbol{A}_N^{-1}\boldsymbol{\phi}(\boldsymbol{x})\right]^2}{\sigma_\nu^2 + \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}^{N+1})\boldsymbol{A}_N^{-1}\boldsymbol{\phi}(\boldsymbol{x}^{N+1})} \tag{17}$$

From (8) we see that the Hessian $\boldsymbol{A}_N$ is positive definite, and hence its inverse will be positive definite. It therefore follows that the second term on the right hand side of (17) is non-negative, and so we obtain

$$\sigma_{N+1}^2(\boldsymbol{x}) \leq \sigma_N^2(\boldsymbol{x}) \tag{18}$$

This represents the intuitive result that the addition of an extra data point cannot lead to an increase in the magnitude of the error bars. Repeated application of this result shows that the error bars due to a set of data points will never be larger than the error bars due to any subset of those data points.

It can also be shown that the average change in the error bars resulting from the addition of an extra data point satisfies the bounds

$$\langle \Delta\sigma^2(\boldsymbol{x})\rangle \equiv \frac{1}{N}\sum_{n=1}^{N}\left[\sigma_{N+1}^2(\boldsymbol{x}^n) - \sigma_N^2(\boldsymbol{x}^n)\right] \geq -\frac{\sigma_\nu^2}{N} \tag{19}$$

A further corollary of the result (18) is that, if we consider the error bars due to each of a set of N data points individually, then the *envelope* of those error bars constitutes an *upper bound* on the true error bars. This is illustrated with a toy problem in Figure 1. The contributions from the individual data points are easily evaluated using (13) and (14) since they depend only on the prior covariance function and do not require evaluation or inversion of the Hessian matrix.

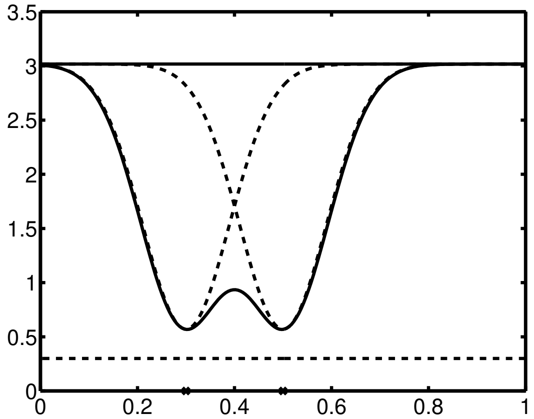

Figure 1: A simple example of error bars for a one-dimensional input space and a set of 30 equally spaced Gaussian basis functions with standard deviation 0.07. There are two data points at $x = 0.3$ and $x = 0.5$ as shown by the crosses. The solid curve at the top shows the variance $\sigma^2(x)$ due to the prior, the dashed curves show the variance resulting from taking one data point at a time, and the lower solid curve shows the variance due to the complete data set. The envelope of the dashed curves constitutes an upper bound on the true error bars, while the noise level (shown by the lower dashed curve) constitutes a lower bound.

4 Summary

In this paper we have explored the relationship between the magnitude of the Bayesian error bars and the distribution of data in input space. For the case of a single isolated data point we have shown that the error bar is pulled down close to the noise level, and that the length scale over which this effect occurs is characterized by the prior covariance function. From this result we have derived an upper bound on the error bars, expressed in terms of the contributions from individual data points.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[2] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[3] C. K. I. Williams, C. Qazaz, C. M. Bishop, and H. Zhu. On the relationship between Bayesian error bars and the input data density. In *Proceedings Fourth IEE International Conference on Artificial Neural Networks*, pages 160–165, Cambridge, UK, 1995. IEE.