

Priors for Bayesian Neural Networks

Mark Robinson,
mark@stat.ubc.ca

September 9, 2001

Abstract

In recent years, Neural Networks (NN) have become a popular data-analytic tool in Statistics, Computer Science and many other fields. NNs can be used as universal approximators, that is, a tool for regressing a dependent variable on a possibly complicated function of the explanatory variables. The NN parameters, unfortunately, are notoriously hard to interpret. Under the Bayesian view, we propose and discuss prior distributions for some of the network parameters which encourage parsimony and reduce overfit, by eliminating redundancy, promoting orthogonality, linearity or additivity. Thus we consider more senses of parsimony than are discussed in the existing literature. We investigate the predictive performance of networks fit under these various priors.

1 Introduction

Say we wish to carry out a regression where the responses $y = (y_1, \dots, y_n)$ are probably a non-linear function of the p -dimensional explanatory variables $x = (x_1, \dots, x_n)$. Used for non-parametric regression, the model for a NN with a single layer of k hidden nodes can be written as:

$$y_i = \beta_0 + \sum_{j=1}^k \beta_j \psi(\alpha_j + \gamma_j^T x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where $\theta = (\beta, \alpha, \gamma)$ are the network weights and biases, $\psi(\cdot)$ is the activation function (e.g. logistic), and the errors ϵ_i can be assumed to have mean 0 and constant variance σ^2 . For convenience, normality of the errors is often assumed. We take the Bayesian point-of-view thus requiring a specification of a prior distribution over the parameter θ . Because our main focus here is prediction, we concentrate on the predictive distribution

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

where y^* is a new observation to predict based on x^* and $p(\theta|x, y)$ is the posterior distribution of θ .

Recent approaches for analysing such models under the Bayesian framework include Lee (2000), Rios Insua and Müller (1998), Müller and Rios Insua (1998), and Neal(1996) among others. Briefly, Neal, Rios Insua and Müller explore a hierarchical prior on each of α , β and γ , generally with priors centred at 0 and a hyperparameter specifying their variance. The hyperparameters themselves typically have a flat prior specified over them in the hope that the data will speak for reasonable values. Inevitably, Markov Chain Monte Carlo (MCMC)

methods are required for posterior inference. Neal uses the hybrid MCMC algorithm which is based on methods of dynamical simulation, whereas Rios Insua and Müller use a straight forward random walk Metropolis-Hastings (RWMH) algorithm, after integrating out the β 's. Lee introduces a data-dependent non-informative prior in which network parameters are restricted to a region where linear independence of the logistic basis functions is achieved; he uses the Rios Insua/Müller algorithm for sampling of the posterior.

The next section describes a new approach to the specification of prior distributions for neural networks. Posterior inference via the RWMH algorithm is examined in Section 3. A discussion of predictive performance under the new model is discussed in Section 4. Concluding remarks are given in the final section.

2 Prior Considerations

The typical prior on γ_j which achieves weight decay is:

$$\gamma_j \sim N(0, \tau^2 I) \text{ for } j = 1, \dots, p.$$

In any given problem, though, a reasonable value of τ^2 is difficult to specify. In the Bayesian framework, our uncertainty is best captured in a prior distribution on τ^2 and that is what is commonly done in the literature (Neal 1996, Rios Insua and Müller 1998). The standard approach is to take an inverted gamma (Γ^{-1}) prior on τ^2 where the parameters of this hyperprior disperse the density over a broad range. Though Lee advocates a prior which in effect has no weight decay, it has been our experience that weight decay is necessary for reliable prediction, especially in long runs of the Markov Chains.

For us, this motivates two further possibilities. First, we explore other senses of parsimony such as orthogonality of the weights γ_j or additivity of the nodes. Second, we consider in more detail the geometry of the problem and suggest priors which encourage the parameters to reside in an *effective domain of interest*.

2.1 Parsimony Priors

Note that a $\Gamma^{-1}(a, b)$ hyperprior on the τ^2 parameter is mathematically equivalent to an unconditional Multivariate- t on each γ_j . That is,

$$p(\gamma_j) \propto \left(1 + \frac{1}{b} \|\gamma_j\|^2\right)^{-\frac{a+p}{2}}. \tag{1}$$

Consider a reparameterization of $(\alpha, \gamma) \longleftrightarrow (\mu, w, \lambda^2)$ given by

$$w = \frac{1}{\|\gamma\|} \gamma, \quad \mu = \frac{-\alpha}{\|\gamma\|} \text{ and } \lambda^2 = \frac{1}{\|\gamma\|^2},$$

so that

$$\psi(\alpha + \gamma^T x) = \psi\left(\frac{w^T x - \mu}{\lambda}\right).$$

This allows one to think of desirable geometrical properties of (w_j, μ_j, λ_j) which promote model parsimony.

For example, because we wish different hidden nodes of the neural network to explain different components of the relationship, we may discourage w_i and w_j from pointing in the same

direction. One possibility is to add a penalty for larger values of the absolute inner product between all unit vectors w_i and w_j , perhaps in the same flavour as (1), such as

$$f(w_1, \dots, w_k) = \left(1 + \frac{1}{b \frac{k(k-1)}{2}} \sum_{i < j} |w_i^T w_j| \right)^{-\frac{k(a+p)}{2}}. \quad (2)$$

Yet another possibility is to encourage additivity of the hidden nodes, so as to penalize unit vectors of the form $(\frac{1}{\sqrt{p}}, \dots, \frac{1}{\sqrt{p}})$ in favour of vectors of the form $(0, \dots, 0, 1, 0, \dots, 0)$. This is achieved by defining a function $g(u) = \sum_i |u_i| - 1$, and defining the prior as

$$f(w_1, \dots, w_k) = \prod_{j=1}^k \left(1 + \frac{1}{b} g(w_j) \right)^{-\frac{(a+p)}{2}}. \quad (3)$$

Selection of the hyperparameters (a, b) may be done via a ‘‘factor of z ’’ calculation. That is, in the case of orthogonality, we favour total orthogonality over total linear dependence by a factor of z .

2.2 Effective Domain Prior

From the reparamaterization above, one can think of $w^T x$ resting most often within a so-called effective domain of interest as defined by $(-\Delta(w), \Delta(w))$. One logical possibility, based on the standard deviation of $w^T x$ may be

$$\Delta(w) = 2\sqrt{w^T R w},$$

where R is the correlation matrix of x . Note that, as is common procedure in these problems, the x ’s are precentred and prescaled.

Second, consider the logistic activation function, $\psi(z) = (1 + e^{-z})^{-1}$. ψ is well approximated by a function of the form

$$\psi_c(z) = \begin{cases} 0 & z < c \\ z & -c \leq z \leq c \\ 1 & z > c \end{cases}.$$

Selecting $c = 3$, for example, seems to work well. Large values of λ typically induce a more linear activation function, at least over the range of $\alpha_j + \gamma_j^T x$. The slope of this linear section is controlled to some extent by the λ parameter, but also by the β corresponding to the node. Hence, we have a redundancy.

By restricting $\mu \in (-\Delta(w), \Delta(w))$ and $\lambda < \frac{\Delta(w) + |\mu|}{c}$, one can show that a parameter value outside this region will have a representative inside the region which is able essentially to characterize the same functional form. Thus, we may restrict ourselves to this region with little or no effect on the model. In practical terms, though, we only encourage the parameters to lie in this region by specifying a prior distribution which has greater density in these areas.

In our approach, we advocate the desired region of the parameters by specifying the prior distributions on them as follows:

$$\begin{aligned} w &\sim \text{Uniform}, \\ \mu|w &\sim N\left(0, \left\{\frac{\Delta(w)}{2}\right\}^2\right) \\ \lambda^2|\mu, w &\sim \Gamma^{-1}\left(\frac{p}{2}, \left(\frac{p}{2} + 1\right) \left\{\frac{\Delta(w) + |\mu|}{c}\right\}^2\right) \end{aligned}$$

where the shape parameter of the Γ^{-1} mimics previous weight decay priors and the scale parameter gives a mode of $(\frac{\Delta(w)+|\mu|}{c})^2$. Essentially, we are downweighting large values of λ for being redundant and small values of λ as weight decay. Note also this is indeed a data-dependent prior via the correlation matrix R .

In effect, the prior is self-calibrating because the calculation of $\Delta(w)$ requires no subjective input.

The non-informative prior used in a Bayesian linear regression on (β, σ^2) is used.

3 Inference via MCMC

The MCMC algorithm for sampling the posterior follows that of Lee (2000) or the fixed-architecture version of Rios Insua and Müller (1998).

Though we have specified prior distributions on a reparameterization of the original parameters, we still can make the usual updates of the gamma and calculate the acceptance probabilities based on the densities of these new priors. Fortunately, the calculation of the Jacobian can be avoided because it would simply cancel out with the corresponding term in the probing density.

To explore the worth of the senses of parsimony discussed earlier we invoke importance sampling. Here, we are looking at the effect of the parsimony prior over and above the “Effective Domain” prior. To estimate the posterior mean (of the predictive density) under these alternative priors, we weight the MCMC sample by the densities defined in (2) and (3) above.

4 Predictive Performance: Boston Housing Data

The objective here is to predict median house prices for census tracts in the Boston area based on attributes of the area such as crime rate, pollution levels, distance to employment centres, etc. In all, the dataset consisted of 13 explanatory variables and included 506 census tracts.

Mean square errors (MSE) on the test set¹ are shown in Figure 1. Shown are 5 different random splits of the data and 5 different runs of the Markov Chain for networks with 4, 6, 8 and 10 hidden nodes. The standard performance is represented by a solid line. Each chain has a burnin of 20,000 iterations followed by a sampling phase of 10,000 iterations.

Also, we wish to explore the possible benefit of the parsimony priors discussed in Section 2. As mentioned, we use importance sampling to weight the MCMC sample as a means to imitate sampling under these priors. Figure 2 shows the percentage change in MSE from the standard prior (Effective Prior) to the new priors. Again, shown are 5 random splits of the Boston housing data and 5 runs from different starting points.

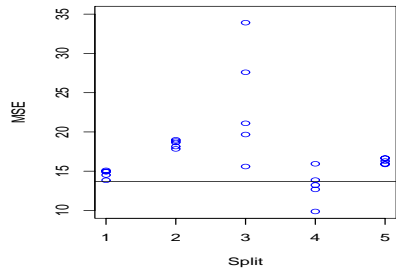
Negative values represent an improvement in prediction.

5 Concluding Remarks

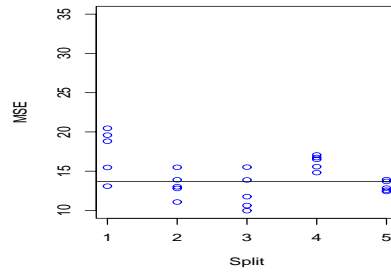
In this paper, we have introduced new priors for Bayesian neural networks from thoughts on desirable geometrical properties of the network parameters.

Although the prediction results do not overwhelm, it is encouraging to see these priors are able to contend with previous approaches. The research is ongoing with a view to including the architecture selection problem.

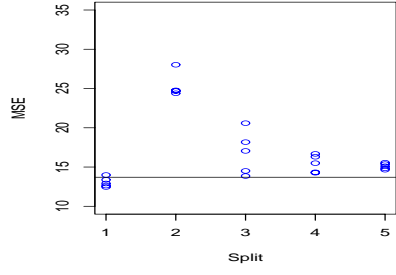
¹These results are highly dependent on the training and test sets used.



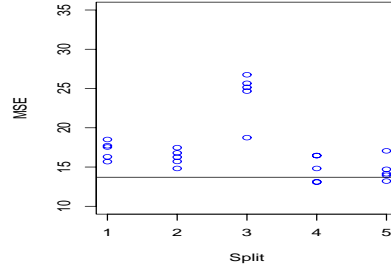
(a) 4 Hidden Nodes



(b) 6 Hidden Nodes

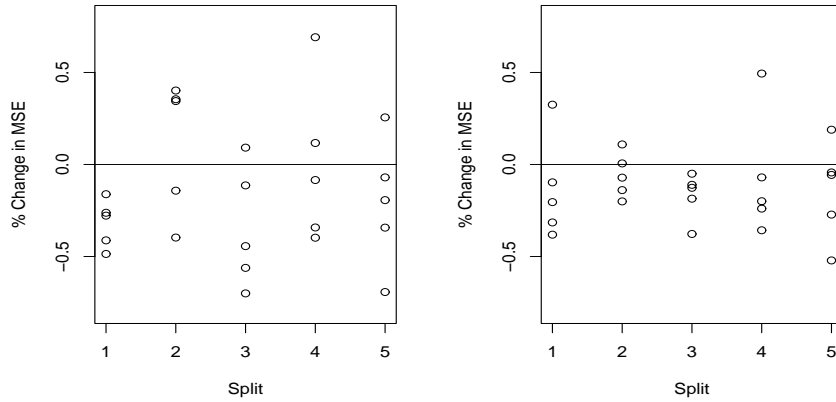


(c) 8 Hidden Nodes



(d) 10 Hidden Nodes

Figure 1: $MSEs$ on the Test Dataset (Boston Housing Data).



(a) Orthogonality Prior

(b) Additivity Prior

Figure 2: Gains in MSE under Parsimony Priors.

6 References

1. Lee, H.K.H. (2000). "A Framework for Nonparametric Regression Using Neural Networks." Technical Report 00-32, Duke University, Institute of Statistics and Decision Sciences.
2. Müller, P. and Rios Insua, D. (1998) "Issues in Bayesian Analysis of Neural Networks." *Neural Computation*, 10, 571-592.
3. Neal, R.M. (1996). *Bayesian learning for neural networks*, New York: Springer-Verlag.