J4.6 BAYESIAN INFERENCE IN NEURAL NETWORKS

Caren Marzban*
National Severe Storms Laboratory, Norman, OK 73069
Cooperative Institute for Mesoscale and Meteorological Studies, and
Department of Physics, University of Oklahoma, Norman, OK 73019

1. INTRODUCTION

The rather lofty title of this article may suggest that the research reported herein may be more fundamental than it really is! The foundations of Bayesian techniques in neural networks have been already laid down in the works of Buntine and Weigend (1991), MacKay (1996), Neal (1996), and Wolpert (1993). Bishop (1996) devotes an entire chapter to the application of Bayesian techniques in neural networks. The present article aims to only illustrate some of the ideas developed in these works.

All nonlinear regression and classification models can over-fit data; over-fitting occurs when the flexibility/complexity of a model allows it to fit a data set or a decision boundary to such high accuracy that the fit is driven by the statistical fluctuations in the data. Consequently, such a model has no predictive capability. The true cause of this phenomenon is the finiteness of the sample size. To restrain overfitting, one traditional approach calls for splitting the data into several sets - a training set, employed to estimate the parameters of the model, a validation set for determining the complexity of the model, and a test set for estimating the unbiased performance of the model. ''Parameters of the model'' include the weights of a parametric model, or smoothing parameters for estimating probability density functions in nonparametric models; the ''complexity of the model'' refers to, for example, the order of a polynomial regression, or the number of hidden nodes and the magnitude of the weights in a neural network.

This split-sample procedure suffers from several faults the least of which is not employing the entire data set in each of the three phases. Novel Bayesian techniques have been developed that allow for the ''determination" of the complexity of a model, and the assignment of confidence intervals to the predictions of the model, given a *single* data set.

Sarle (1995) has examined several learning procedures in the context of regression problems and has found that the use of the Bayesian method is fully warranted if the underlying function is expected to be nonlinear. In the present article, some aspects of the Bayesian method will be illustrated in a 2-group classification problem, first, and then applied at a very rudimentary level to the development of a neural network for the prediction of tornados. Further details will be provided elsewhere.

2. THE BAYESIAN APPROACH

Conventional learning methods yield a unique set of weights, whereas Bayesian procedures produce a posterior distribution for the weights. This distribution manifests itself as a distribution for the outputs of the network which, in turn, allows for the computation of measures of confidence in the outputs. Therefore, a network that is used for prediction purposes has its outputs integrated over the entire distribution of the weights. Such integration is performed also when there exist other parameters in the model whose precise values are not known (though they depend on the weights). An example of such a (hyper) parameter is the coefficient of the ''weight decay" term in the error function, its purpose being to prevent the magnitude of the weights from becoming exceedingly large. Since most activation functions (e.g., tanh, logistic, etc.) are linear for small values of the weights, and highly nonlinear for large values of the weights, the value of the weight-decay coefficient in the error function affects the overall nonlinearity (complexity) of the network. Of course, the other quantity that measures the nonlinearity of a network is the number of hidden nodes. It is possible to approximate the relevant integrals and to express the results in terms of the most probable values of the parameters and the errors thereof. In that case, the ''optimal" value of the weight-decay coefficient and the number of hidden nodes can be ''inferred" through Bayesian reasoning.

There are two distinct approaches in the implementation of Bayesian ideas in neural networks. Neal (1996) uses exact simulations, and advocates the point of view wherein a large number of hidden nodes are to be selected and then controlled through hyperparameters. MacKay (1996), on the other hand, approximates the posterior distributions thereby allowing for analytic results. In MacKay's approach, it appears that the optimal number of hidden nodes can be addressed by the so-called ''evidence framework'';

*e-mail: marzban@nssl.noaa.gov

it has been conjectured that the maximum of the evidence approximately marks the number of hidden nodes at which overfitting occurs (MacKay 1996; Bishop 1996). If so, then one may use both the training set and the validation set as a single training set. A test set will still be necessary for gauging the performance of the network; the performance issue will not be addressed here. In the next section, this conjecture will be supported by means of a simulated data set, and then applied to a realistic classification problem.

For a classification problem, the error function is given by cross-entropy:

$$S = -\sum [t \log y(x,\omega) - (1-t)\log(1-y(x,\omega))],$$

where x is the vector of inputs, ω is the vector of the weights, and t is the target value that is to be produced by the output $y(x,\omega)$. The summation is over the number of cases in the relevant data set. In the 2-group case, with a single (binary) output node representing group membership, for this error function to be consistent it is important for the activation function to be the logistic activation function $f(x) = 1/(1+exp(-x))$. The output can then be interpreted as the posterior probability of group-membership, given the inputs (if a global minimum has been reached), and learning can then be thought of as inferring the weights, ω, given some data, D. In other words, in the learning phase a quantity of interest is the probability $P(\omega|D,\alpha,M)$, where α is a hyperparameter in the model M. By Bayes' theorem

$$P(\omega|D,\alpha,M) = \frac{P(D|\omega,M)P(\omega|\alpha,M)}{P(D|\alpha,M)} \quad ,$$

where

$$P(D|\omega,M) \sim exp^{-S} \quad ,$$

and

$$P(\omega|\alpha,M) \sim exp^{-\alpha S_W} \quad ,$$

where $S_W = \frac{1}{2}\sum \omega^2$ is the weight-decay term. These equations imply that the most likely ω is given by the maximum of $P(\omega|D,\alpha,M)$, or the minimum of $E = S + \alpha S_W$,where E is defined by $P(\omega|D,\alpha,M) = exp(-E)$.

In the training phase, the ω-independent part, $P(D|\alpha,M)$, is simply a normalization constant, but it is the important quantity in inferring the value of α, because again by Bayes' theorem

$$P(\alpha|D,M) = \frac{P(D|\alpha,M)P(\alpha|M)}{P(D|M)} \quad .$$

The quantity $P(D|\alpha,M)$ is called the evidence for α (and M). It is also possible to obtain an evidence for M by marginalization:

$$P(D|M) = \int P(D|\alpha,M)P(\alpha|M)\, d\alpha \quad .$$

This quantity is believed to incorporate the Occam factor which favors the model with the lowest complexity (MacKay 1996). Consequently, its maximum is expected to correspond to the minimum of generalization error. The computer codes for the computation of these quantities are available via ftp at 131.111.48.8 in the directory pub/mackay/bigback. A few other codes that are necessary for obtaining the results reported herein were written by the author himself.

3. A CLASSIFICATION PROBLEM

A data set was generated with 2 independent variables ranging from -1 to +1, and one dependent variable whose 0/1 values label two groups. This was done such that the decision boundary between the two groups, i.e., the inverted ''Mexican hat'' with the filled circles in Figure 1, corresponds to that of a network with 4 hidden nodes. Indeed, prior to the addition of noise to the data, a 4-hidden-node network would learn this boundary with zero error. The addition of some gaussian noise ($\sigma = 0.2$) produces the data appearing in Figure 1 with the lower-pointing filled triangles representing one group and the upper-pointing triangles representing the second group. The training set contained 300 cases and the validation set 200. The empty circles in Figures 1a-1c outline the decision boundaries produced by several networks with different number of hidden nodes whose weights and hyperparameters have been inferred by Bayesian means. Evidently, the network with 2 hidden nodes underfits the boundary, while a network with 15 hidden nodes overfits the boundary; in addition to the excessive meandering of the neural net's fit about the underlying boundary, note the false boundaries at the top and the bottom of Figure 1c as the network attempts to create separate decision regions for two individual data points. As expected, the network with 4 hidden nodes is optimal. (As mentioned above, an additional virtue of the Bayesian approach is the measure of confidence in the outputs that can be calculated. The computation of this very important quantity, however, will be postponed to a later time. The conclusions drawn here, therefore, must be interpreted with care.)

This behavior can be quantified by considering the training errors and the validation errors for each of the networks. Figure 2a displays the training and validation cross-entropy errors for a range of number of hidden nodes. Evidently, as the training error decreases with increasing number of hidden nodes, the validation error decreases for up to 4 hidden nodes and then begins to increase. Again, as expected, the network with 4 hidden nodes appears to be the optimal one.

Figure 2b displays the log of the evidence for the same networks. As seen, the evidence and the validation error behave similarly (oppositely) even though the former is computed from the training set alone. Therefore, it appears that the conjecture stated above - that the maximum of evidence marks the onset of overfitting - may indeed be true. There are several other issues that must be addressed before this conclusion

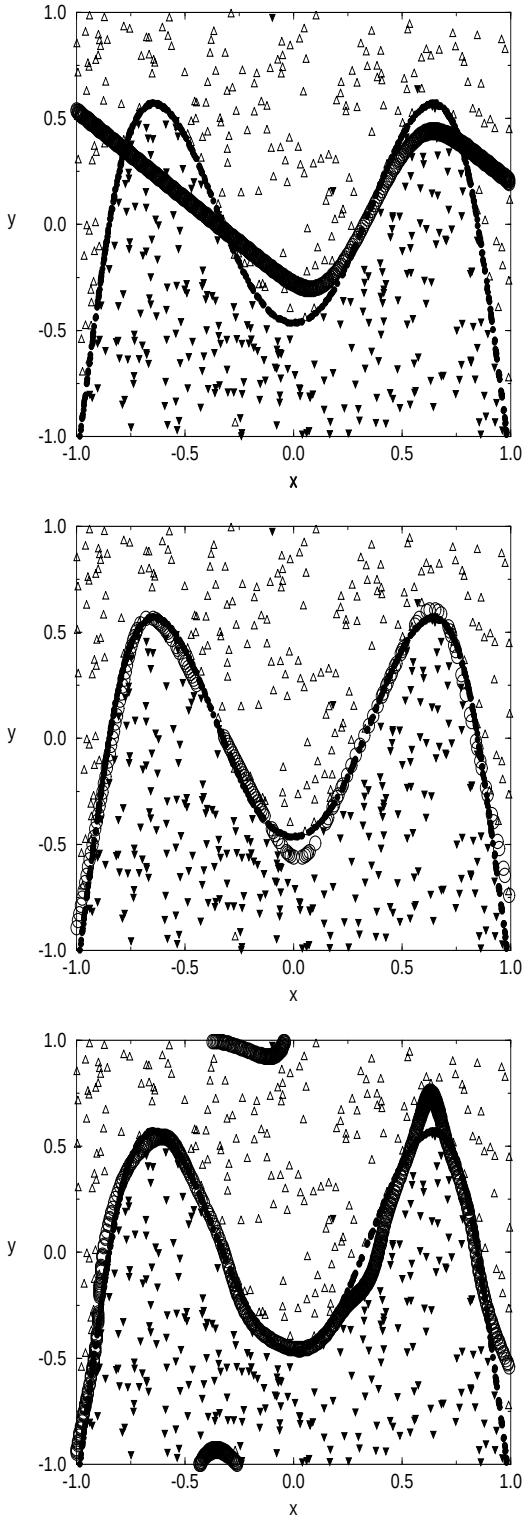

Figure 1: The underlying decision boundary (filled circles) between two groups (upper and lower pointing triangles), and the neural net estimates (empty circles) for a) 2, b) 4, and c) 15 hidden nodes, from top to bottom, respectively.

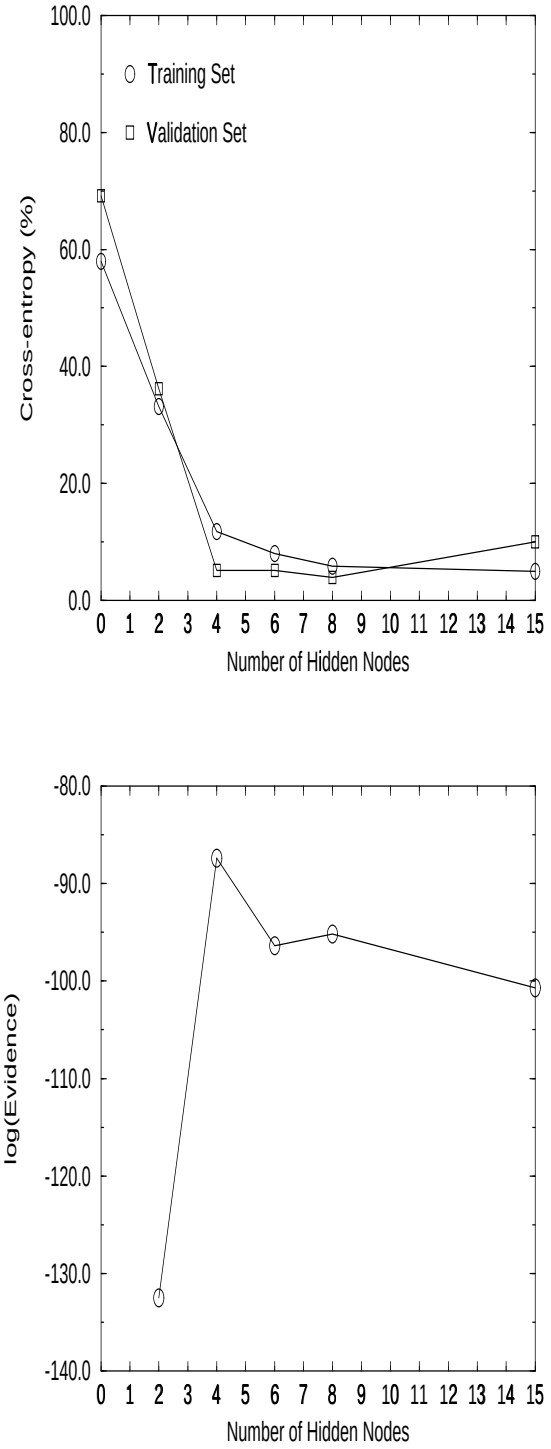

Figure 2: a) Cross-entropy error for the training set and the validation set (top plot), and b) the log of the evidence (lower plot), for a range of hidden nodes.

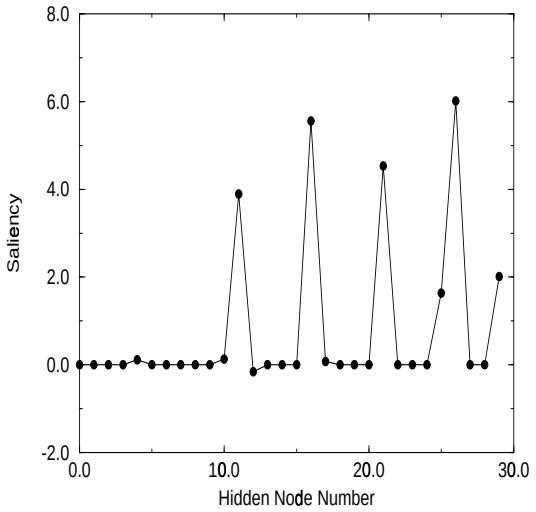

Figure 3: The saliency of a 30-hidden-node network

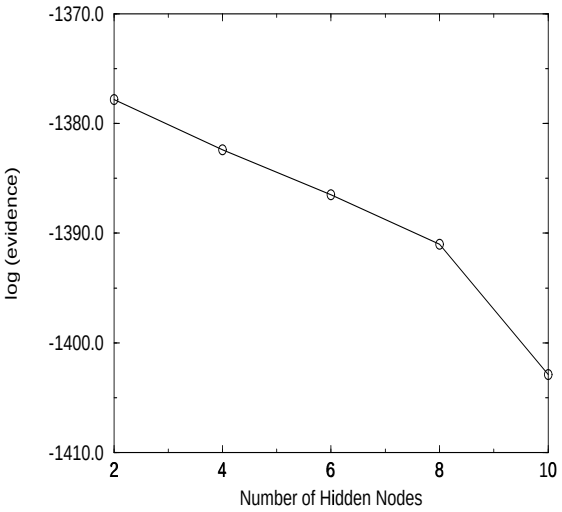

Figure 4: The evidence for a range of networks trained for tornado prediction.

can be stated with further certainty; some of these issues will be addressed in the discussion section.

Another feature of the Bayesian method is the ability to compute a quantity called saliency which for the i^{th} hidden node is defined as ω_i^2/H_{ii}, where H is the Hessian $H = \nabla\nabla \log P(\omega|D, \alpha, M)$ (MacKay 1996). This quantity gauges the ''activity'' of each hidden node. For instance, if in the present classification problem an extravagant researcher were to place, say, 30 hidden nodes in the network, then by computing the saliency of each hidden node, she would quickly note that mostly 4 of the hidden nodes are in fact active (Figure 3) - again the correct answer, since the boundary was designed to be represented by a network with 4 hidden nodes.

4. TORNADOS

Having shown that the evidence framework appears to yield the optimal number of hidden nodes, we can proceed to apply this method to a realistic data set. The National Severe Storms Laboratory has developed a number of algorithms for the diagnosis of circulations that have the potential of becoming tornadic. Neural networks have been developed for the diagnosis of tornadic circulations forming from mesocyclones (Marzban and Stumpf 1995, 1997; Marzban, Paik, and Stumpf 1997), and the performance of these networks has been gauged in terms of a number of performance measures (Marzban 1997). There exist, however, circulations that do not meet the characteristics of mesocyclones, and there exists an algorithm for the identification of such circulations (Mitchell, et al. 1997). That algorithm produces 21 attributes derived from Doppler radar, but for visual purposes only 3 will be employed here as inputs to the network; they are the base, depth and the low-level gate-to-gate velocity difference of the circulations detected. These quantities were transformed to z-scores for training purposes. The target values are 0 or 1, depending on whether or not a tornado really does exist corresponding to the input values.

A number of networks with different number of hidden nodes were trained with all (5791) circulations (with the prior probability of tornados being 0.09), and the evidence was computed for each network. The results are shown in Figure 4. It appears that the optimal number of hidden nodes is 2. Such a network represents a decision boundary that is approximately plainer. A 3-dimensional plot of this boundary is shown in Figure 5. Note that the boundary surface is parallel to the second variable for large values of that variable, and is almost perpendicular to it for small values. This implies that this variable (i.e., depth) is a good predictor of tornados when it is small, but it is a poor predictor of tornados when it is large. An example of a boundary surface that overfits the data is provided by a network with 8 hidden nodes (Figure 6); note that this boundary is in fact composed of two disconnected pieces.

5. CONCLUSIONS

Some empirical evidence is presented in support of the ''evidence approach'' to Bayesian inference, i.e., that the maximum of the evidence appears to mark the onset of overfitting, and as such there is no need for a validation data set. In particular, the evidence framework appears to allow for the identification of the optimal number of hidden nodes with only one data set. Additionally, the application of this methodology to a realis-

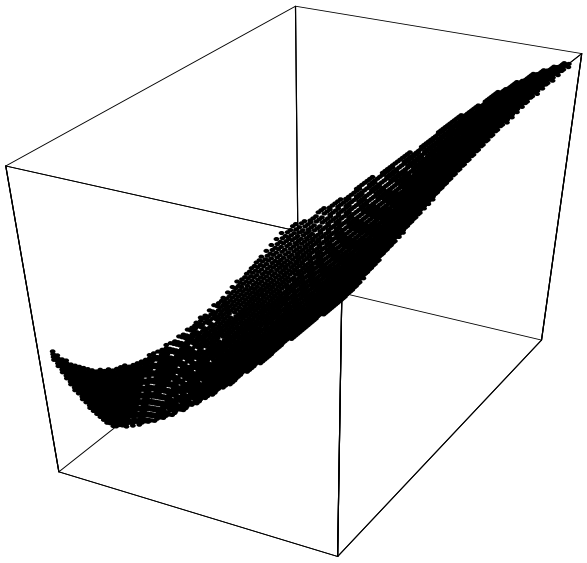

Figure 5: The decision surface between tornados and non-tornados according to a 2-hidden-node network.

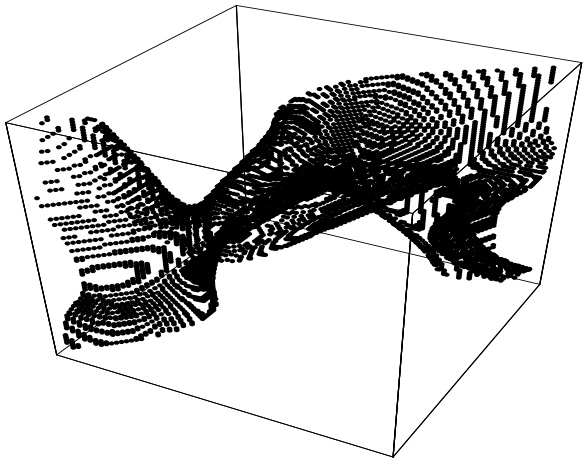

Figure 6: The decision surface between tornados and non-tornados according to an 8-hidden-node network.

tic problem involving tornado prediction suggests that the decision surface underlying the problem is approximately plainer. The orientation of the plain indicates that depth is a good (poor) predictor of tornados when it is small (large).

6. DISCUSSION

The findings in this report are preliminary and should be interpreted with care because a few contingencies have not been taken into account! For instance, in establishing the connection between generalization and evidence, it is important to

- acknowledge that the two measure different quantities, and so the connection may not be exact,
- note that both are prone to errors (the Hessian is difficult to compute and this can adversely affect the evidence),
- repeat the entire procedure from a different set of initial weights in order to account for different local minima,
- consider several different partitions of the data set into training and validation sets, and then average over the outcomes,
- examine different measures of performance in addition to cross-entropy.

There exist other contingencies as well, but they will be addressed elsewhere.

7. ACKNOWLEDGEMENTS

The author is grateful to V. Lakshmanan and D. Mitchell of the National Severe Storms Laboratory for valuable discussions and for providing the tornado data set, respectively. Bruce Mason of the physics department, university of Oklahoma, is acknowledged for assistance in the use of Mathematica. Partial support was provided by the FAA and the NWS/OSF.

8. REFERENCES

Bishop, C. M., 1996: *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, pp. 482.

Buntine, W. L., and Weigend, A. S. 1991: Bayesian back-propagation, Complex Systems, 5, 603-643.

MacKay, D. J. C., 1996: Bayesian methods for back-propagation networks, in *Models of Neural Networks III*, E. Domany, J. L. van Hemmen, K. Schulten (Eds.), Springer-Verlag, New York, physics of neural network series, pp. 309.

Marzban, C., 1997: Scalar measures of performance in rare-event situations. To appear in *Wea. Forecasting*.

Marzban, C., and G. Stumpf, 1995: A neural network for tornado prediction based on Doppler radar-derived attributes. *Journal of Applied Meteorology*, **35**, 617-626.

Marzban, C., and G. Stumpf, 1997: A neural network for damaging wind prediction. To appear in *Wea. Forecasting*.

Marzban, C., H. Paik, and G. Stumpf, 1997: Neural networks vs. gaussian discriminant analysis. *AI Applications*, **11**, No. 1, 49-58.

Mitchell, D., S. Vasiloff, M. Eilts, and A. Witt, 1997: The National Severe Storms Laboratory's tornado detection algorithm. Submitted to *Wea. Forecasting*.

Neal, R. M., 1996: *Bayesian learning for neural networks*. Cambridge, Cambridge University Press, pp. 183.

Sarle, W. S., 1995: Stopped training and other remedies for overfitting. Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics, 352-360, Cary, NC.

Wolpert, D. H., 1993: On the use of evidence in neural networks, in C. L. Giles , S. J. Hanson, and J. D. Gowan (Eds.), Advances in Neural Information Processing Systems 5, San Mateo, California, Morgan Kaufmann, 539-546.