ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 121-126

Equivalent Error Bars For Neural Network Classifiers Trained By Bayesian Inference

Peter Sykacek *

Austrian Research Institute For Artificial Intelligence,
Schottengasse 3, A-1010 Vienna, Austria

Abstract. The topic of this paper is the problem of outlier detection for neural networks trained by Bayesian inference. I will show that marginalization is not a good method to get moderated probabilities for classes in outlying regions. The reason why marginalization fails to indicate outliers is analysed and an alternative measure, that is a more reliable indicator for outliers, is proposed. A simple artificial classification problem is used to visualize the differences. Finally both methods are used to classify a real world problem, where outlier detection is mandatory.

1. Introduction

Neural networks are often used in safety-critical applications for regression or classification purpose. Since neural networks are unable to extrapolate into regions not covered by the training data (see [6]), one should not use their predictions in such regions. Consequently methods for outlier detection got a lot of attraction. Outliers may be detected by assigning a confidence measure to network decisions. Confidence should be high in regions well represented in the training data and low everywhere else. Different methods can be used to get a confidence measure for network decisions. In [6] S. Roberts et. al. use an artificial class outside training data subspace. In [3] D.J. MacKay uses Bayesian inference and marginalization to get moderated probabilities for classes in outlying regions. In conjunction with doubt levels this should prohibit classification of outliers. The aim of my paper is to discuss marginalization and compare it to a different method for outlier detection, that can be used within the Bayesian framework. The effects of both methods are visualized using a simple artificial classification problem. Results of a real world classification problem with outliers are presented in the final section.

This work was sponsored by the Austrian Federal Ministry of Science, Transport and the Arts. It was done in the framework of the BIOMED 1 concerted action ANNDEE, financed by the European Commission, DG. XII.

*I want to acknowledge the work of R. Neal from the Departments of Statistics and Computer Science at the University of Toronto, who's hybrid Monte Carlo implementation of Bayesian inference was used to calculate simulation examples.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 121-126

2. Marginalization, the current practice

In this chapter I want to review how confidence is incorporated into classification decisions within a Bayesian framework. I assume, that the neural network is designed to solve a two class problem, extension to 1 of c class problems is straightforward. In such a case, a neural network with one hidden layer with sigmoid activations and a single output unit is used. As shown in (1), the output is transformed by a final sigmoid, a method that can be proven to lead to outputs, that are posterior probabilities for classes. Details are presented in [1].

$$P(C_1 \mid \underline{x}) = g(a)$$
$$g(a) = \frac{1}{1+\exp(-a)} \tag{1}$$

If the targets for classes are 1 and 0 for C_1 and C_2 respectively, then $g(a)$ is the probability for class 1 and $1-g(a)$ is the probability for class 2. In (1) the input into the final sigmoid is represented by a.

A Bayesian solution for neural networks is a posterior distribution over weight space calculated via Bayes theorem using a prior over weights.

$$p(\underline{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \underline{w})p(\underline{w})}{p(\mathcal{D})} \tag{2}$$

In (2) $\underline{w}$ is the weight vector of the network and $\mathcal{D}$ represents the training data. Two different possibilities are known to calculate the posterior in (2). In [2] D.J. MacKay derives an analytical expression assuming a Gaussian distribution, in [4] R. Neal uses a hybrid Monte Carlo method to sample from the posterior. The posterior over weight space will lead to a distribution of network outputs for one input pattern. We expect a distribution with small deviation in regions well represented in the training data space, since the likelihood term $p(\mathcal{D} \mid \underline{w})$ forces the outputs to lie close to the target patterns and a distribution with large deviation everywhere else. For a classification problem, following MacKay [3], the network guess is calculated by marginalization over the output distribution as shown in (3).

$$P(C_1 \mid \underline{x}, \mathcal{D}) = \int P(C_1 \mid \underline{x}, \underline{w})p(\underline{w} \mid \mathcal{D})d\underline{w}$$
$$= \int y(\underline{x}, \underline{w})p(\underline{w} \mid \mathcal{D})d\underline{w} \tag{3}$$

According to MacKay [3], marginalization over weights leads to moderated outputs of the classifier in regions where "the ensemble is very uncertain about what class is". Suspicious cases can be detected by using doubt levels. Since the integral can not be solved directly due to the sigmoid activation used in the output node, he approximates the integral as shown in (4).

$$P(C_1 \mid \underline{x}, \mathcal{D}) = g(\kappa(\sigma_a)a_{MP})$$
$$\kappa(\sigma_a) = (1+\frac{\pi\sigma_a^2}{8})^{-1/2} \tag{4}$$

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 121-126

In (4) a_{MP} denotes the input into the sigmoid of the output unit according to the most probable weight vector and σ_a^2 is the corresponding variance resulting from the distribution over weights. As shown in [4], sampling from the posterior allows us to calculate (3) as a sum over finite network guesses directly.

For both methods, the question arises whether marginalization shows the desired effect of moderated probabilities for classes in regions not covered in the training set. Two important thoughts will guide us. The authors of [7] have shown, that a neural network, used for classification, performs a sort of nonlinear discriminant analysis. Data points close to the decision boundaries of input space are mapped to input values of the final sigmoid close to 0, corresponding to network outputs of approximately 0.5. Data points arbitrary far from the decision boundary are mapped to arbitrary large positive or negative input values of the final sigmoid, resulting in output values of approximately 1 or 0 respectively. On the other hand in [4] R. Neal has investigated the effects of Gaussian priors and shown, that the variance of the outputs of a two layer network with sigmoid activation in the hidden layer units and linear output activation is bounded under a Gaussian prior. Therefore, in our classification problem, the variance of the input of the sigmoid activation function of the output unit has to be bounded. Both facts together let us conclude, that there must exist regions in input space not covered by training data, where marginalization will not lead to moderated probabilities for classes. Outlier detection with marginalization will not be possible in those regions.

Visuaizations in this section are done with a two layer network with two inputs, 20 hidden units and one output unit. The number of 20 hidden units is motivated by the results reported by R. Neal in [4]. R.Neal studied the properties of multi layer perceptron networks with infinite number of hidden units under Gaussian priors. He concluded that in a correct Bayesian framework, there is no need to limit the number of hidden units.

The left sub plot in figure 1 shows an example of moderated outputs calculated by approximation of the integral in (3) by a sum over discrete weight vectors. The right sub plot shows MacKay's approximation of the marginalized output formulated in (4). I generated an artificial two class problem with bivariate Gaussian class conditional densities. Training data are two Gaussian distributions with means (1,0) and (-1,0) and standard deviations of the marginal distributions of $\sigma_x = 0.5$ and $\sigma_y = 2$. As expected marginalization over the posterior distribution has no moderating effect in regions far outside the subspace covered by training data, it is impossible to make outlier decisions.

3. Equivalent error bars as a measure of confidence

The last section showed that marginalization is not a good method for outlier detection. Obviously on one hand the mean of the distribution $p(a \mid \underline{x}, \mathcal{D})$ (a is the input to the final sigmoid) grows fast and unbounded as we move away from the decision boundary. On the other hand the variance σ_a^2 of that distribution

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 121-126

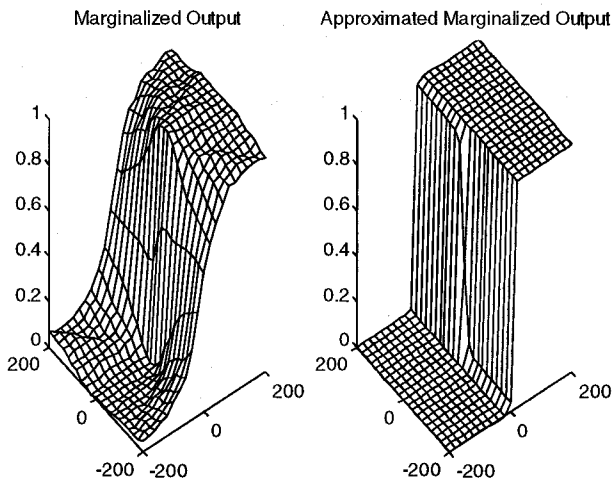

Figure 1: Marginalized output and its approximation for a classifier using inputs far beyond the training data subspace

stays bounded. The final sigmoid maps the resulting distribution to values close to 0 or 1. Nevertheless it is possible to get a measure of confidence, since the variance σ_a^2 is small in regions covered by the training data and larger everywhere else.

Such a measure of confidence can be calculated by mapping the standard deviation, σ_a, of $p(a \mid \underline{x}, \mathcal{D})$ into the range $[0, 1]$. A possible mapping, is shown in (5). It uses the logistic sigmoid to perform the mapping.

$$\begin{aligned} e(\underline{x}) &= g(k\sigma_a) - g(-k\sigma_a) \\ &= \frac{\exp(k\sigma_a) - 1}{\exp(k\sigma_a) + 1} \end{aligned} \tag{5}$$

The factor k allows to adjust the probability of $|m_a - a(\underline{x}, \mathcal{D})|$ to be within $k\sigma_a$. For the delimiting mapping in (5) we could have used any delimiting function. Using the logistic sigmoid, $g(k\sigma_a)$, gives us the possibility to interpret $e(\underline{x})$ as an "error bar" of a classification decision with equivalent variance, σ_a^2, close to a decision boundary. Since (5) is an "error bar", smaller values indicate higher confidence. Consequently a measure for confidence is given by $1 - e(\underline{x})$. The plots in figure 2 were produced with the same training data, already used to visualize marginalization effects in figure 1. In both plots training data is shown without showing class labels. The left illustration in figure 2 shows the "equivalent error bar" and training data. The right sub plot shows training data and a contour plot of the equivalent error bar. It is easy to see, that a threshold of 0.54 can be used to delimit the subspace covered by training data.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 121-126

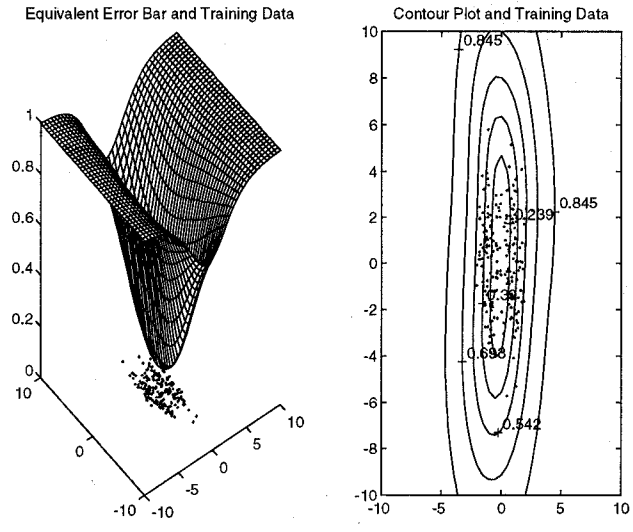

Figure 2: Equivalent error bar and training data

4. Outlier detection in cushing's syndrome data

Cushing's syndrome data is one of the datasets[0] used by B.D. Ripley in his recent book [5]. Each pattern is one of four classes, there are three types of syndrome and two samples marked as "other". The data set contains training and test data. I used two types of syndrome, "adenoma" and "bilateral hyperplasia", as training samples. The third type of syndrome, "carcinoma" and the test set were used for testing. The observations in the data set are the urinary excretion rates of tetrahydrocortisone and pregnanetriol (mg/24h).

The task of the trained network was to classify examples in the test set and refuse classification of suspicious patterns. In this case all samples from class "carcinoma" and "others" are suspicious cases. Simulations were done using a two layer network with 2 inputs 20 hidden units and one output. The large number of hidden units is again motivated by the results from R. Neal, reported in [4].

Outlier detection is done by assigning samples to class doubt within an interval from 0.3 to 0.7 when using marginalization. To detect outliers when using the equivalent error bar, we need an outlier threshold. It was set to a value

[0] I used Cushing's syndrome data generously provided by B.D. Ripley electronically via http://markov.stats.ox.ac.uk/pub/PRNN.

where about 5% of the training data are declared as outliers. The results are summarized by following table:

Method	Outliers	Cor. "a"	Wrg. "a"	Cor. "b"	Wrg. "b"
Marginalization	3	1	0	2	5
Eq. Error Bar	9	1	0	1	0

Using the "equivalent error bar" for outlier detection, one of the test samples of class "bilateral hyperplasia" was declared as an outlier. Using marginalization and a doubt level, three of the samples of class "carcinoma" and both "other" samples were classified as class "bilateral hyperplasia" and not declared as an outlier.

5. Conclusion

Visualizations of an artificial classification problem showed that the equivalent error bar of the classifier is a more reliable method for outlier detection than its marginalized output. Using Cushing's syndrome data I showed that it is likely that the reliability of the classifier can be enhanced by using the equivalent error bar as an indicator for outliers. The method is currently enhanced by investigating different training procedures that return a maximum threshold level for the equivalent error bar.

References

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

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

[3] D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4:720–736, 1992.

[4] R. M. Neal. *Bayesian Learning for Neural Networks.* Springer, New York, 1996.

[5] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[6] J. Pardey S. Roberts, L. Tarassenko and D. Siegwart. A confidence measure for artificial neural networks. In *International Conference Neural Networks and Expert Systems in Medicine and Healthcare*, pages 23–30, Plymouth, UK, 1994.

[7] A. R. Webb and D. Lowe. The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis. *Neural Networks*, 3:367–375, 1990.