ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 267-272

To stop learning using the evidence

Domenico Perrotta

ISIS, Joint Research Centre, European Commission,
T.P. 361, Via E.Fermi, 21020 Ispra, ITALY

Abstract.
Theoretical and practical aspects of Multi-Layer Perceptron (MLP) learning methods from the bayesian perspective were addressed for the first time by David MacKay in 1991. In this framework, the learning algorithm is an iterative process which alternates optimization of weights and estimation of hyperparameters, such as weight decay parameters. Moreover, trained MLPs that generalize better have higher "evidence", a probability which quantifyes how well a MLP is adapted to a problem. This paper suggests and motivates a new methodology that computes the evidence during learning for different MLP configurations. Such estimations and the confidence intervals on test set are used to rank MLP configurations and, then, to stop learning. This learning strategy is illustrated on classification problems.

1. Introduction

In the last seven years various works have demonstrated that the bayesian paradigm can improve Multi-Layer Perceptron (MLP) learning methods. The work presented in this paper is based upon the framework originally developed by David MacKay [2]. Such framework is of particular interest because, contrary to other options, it has been successfully experimented on various real problems.

An important practical point of MacKay's setting is that among different MLP models M^i $(i = 1, 2, \ldots)$, the most adapted to a given problem is chosen by maximizing the conditional probability $P(D|M^i, H)$, called "evidence of M^i", where H is the knowledge that we have and the assumptions we have made about the problem before seeing data D. Two issues motivate this work (Section 3.): (1) for real problems, evidence estimation can be difficult and influenced by numerical inaccuracy; (2) the best configuration for the parameters (including hyperparameters) of a given MLP model is chosen using an iterative process whose convergence towards fixed limiting values is not theoretically demonstrated.

We suggest a new methodology to rank different MLP configurations based on the evidence estimation (Section 4.), experiment the methodology on classification problems and discuss results (Section 5.). Comments and perspectives conclude the paper.

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 267-272

2. Background on bayesian MLPs

We assume that the weights of a given MLP form a set of k indexed elements $W = (w_1, \ldots, w_k)$ and that the given training data set is $D = \{(x_1, t_1), \ldots, (x_N, t_N)\}$. The goal is to link x_i and t_i through the classical statistical model $t_i = R(x_i) + \epsilon_i$, where ϵ_i is an expression of the various uncertainties, e.g. noise. The learning problem is to approximate $R(x)$ by a function $F(x, W)$ from the set of non linear functions of the given MLP architecture. Typically, this sort of learning problem is expressed as a minimization problem, that we outline in the case of MacKay's setting.

Using MacKay's bayesian learning requires that some assumptions be true: (1) the prior distributions of noise and MLP weights follow two Gaussians, with 0 mean and standard deviations σ and σ_W; (2) the posterior weight distribution, obtained using Bayes' rule, is sharply peaked at some point W^*; (3) in a neighborhood of W^* the peak is approximated by a Gaussian; (4) the additional parameters σ and σ_W, also called hyperparameters, are "well determined" by the data D, in the sense that their joint posterior distribution has a sharp peak around values σ^* and σ_W^*. Under such conditions, it can be shown that the minimization problem is to choose W^* that minimizes:

$$\frac{1}{2\sigma^{*2}} \sum_{i=1}^{N} (F(x_i, W) - t_i)^2 + \frac{1}{2\sigma_W^{*\,2}} \sum_{j=1}^{k} w_j^2 \qquad (1)$$

Then, tuning a bayesian MLP is an iterative process which works as follows. At step t, starting from weights W^t and given parameters σ^t and σ_W^t:

a) Find W^{t+1} which minimizes $\frac{1}{\sigma^{t\,2}} \sum_{i=1}^{N} (F(x_i, W) - t_i)^2 + \frac{1}{\sigma_W^{t\,2}} \sum_{j=1}^{k} w_j^2$

b) Given W^{t+1}, find the most probable values σ^{t+1} and σ_W^{t+1} using data D.

c) Go back to (a) now using W^{t+1}, σ^{t+1} and σ_W^{t+1}.

The procedure is started by making some initial guesses for the values of σ^0 and σ_W^0, and different possibilities for W^0 exist. The key point of the process is the iterative re-estimation of the hyperparameters (step (b)).

MacKay's main result are the formulae for the values of σ^* and σ_W^*. If function (1) is written as $\beta E_D + \alpha E_W$, with

$$\beta = \frac{1}{\sigma^{*2}} \qquad \alpha = \frac{1}{\sigma_W^{*\,2}} \qquad E_D = \frac{1}{2} \sum_{i=1}^{N} (F(x_i, W) - t_i)^2 \qquad E_W = \frac{1}{2} \sum_{j=1}^{k} w_j^2 \quad (2)$$

then MacKay's formulae are

$$\alpha = \frac{\gamma}{2E_W} \qquad \beta = \frac{N - \gamma}{2E_D} \qquad \gamma = k - \sum_{a=1}^{k} \frac{\alpha}{\lambda_a + \alpha} \qquad (3)$$

where λ_a is the a^{th} eigenvalue of the Hessian matrix of βE_D, computed using weights W, in the natural basis of E_W. Such expressions are used as

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 267-272

re-estimation formulae for the above iterative process: at step t, γ^{t+1} is first computed for current parameters W^t, β^t and α^t and, then, new parameters β^{t+1} and α^{t+1} are estimated using γ^{t+1} and current errors E_D^t and E_W^t.

MacKay's second important result is an expression for the logarithm of $P(D \mid M, H^t)$, the evidence of the MLP configuration with $H^t = \{W^t, \sigma^t, \sigma_W^t, H\}$. The formula is quite long and we have omitted it: the important point here is that it requires estimation of the determinant of the Hessian matrix. This term depends on the product of the eigenvalues of the matrix, and is consequently much more sensitive to errors than the estimation of γ, which is based on the sum of the eigenvalues.

3. Motivations for a new methodology

We advise to study issues related to the convergence of "the bayesian process", the procedure of alternating optimization of MLP weights and re-estimation of MLP hyperparameters using formulae (3) every few cycles. Proof of the following is an open problem.

Conjecture 3..1 There exists $\hat{t}$ such that, for $t > \hat{t}$, values W^t, σ^t and σ_W^t minimize function

$$\frac{1}{{\sigma^t}^2}\sum_{i=1}^{N}(F(x_i, W) - t_i)^2 + \frac{1}{{\sigma_W^t}^2}\sum_{j=1}^{k} w_j^2$$

and maximize the evidence $P(D \mid M, H^t)$, being $H^t = \{W^t, \sigma^t, \sigma_W^t, H\}$.

In practice this conjecture, or at least the empirical and numerical stability of the bayesian process, is taken for granted by MacKay. In fact, in his experiments, MacKay chooses the weight configuration obtained after a heuristically decided number of iterations. In other words, his idea is to iterate "many times" and to take the last configuration.

A learning procedure always starts using random weights and, obviously, it is not reasonable to assume that the posterior distribution on weights is sharp-peaked around these random weights, as required by the bayesian process. Thus we must, in practice, start the learning with some non bayesian algorithm (such as stochastic and/or conjugate gradient) in order to first reach a "good" weight configuration where the bayesian process can be started. Our recent experience [3] has shown that if the bayesian process begins "too early" before having reached a "good" weight configuration, the MLP will not overfit, but it will never reach an acceptable solution, whilst if it begins "too late" at a point where (having used the conjugate gradient) the MLP has begun to overfit, the bayesian MLP will be unable to reach an acceptable solution either. In other words, the bayesian process must start when the conjugate gradient has found a solution that is "not too far" from its best possible configuration. It is then possible to observe that the bayesian MLP can slightly improve the MLP performance while not overfitting.

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 267-272

4. Ranking MLP configurations using the evidence

Analysis of above points suggests that bayesian MLPs should be used with this new methodology:

0. Use stochastic gradient descent for a initial rough search for a minimum.
1. Use conjugate gradient algorithm for various iteration steps $t_1, t_2, \ldots$
2. For each t_i start a bayesian learning process.
3. At various steps of these bayesian processes, compute the MLP evidence.
4. At the end of the computation, choose the configuration for which the evidence is maximum.

The peculiarity of this methodology is in the use of the MLP evidence to rank different MLP configurations obtained during a given bayesian learning process. This should take to reasonable solutions even in the case of numerical instability or when the best configuration is not the last of the bayesian process.

The evidence measures how well a MLP configuration is adapted to the problem. Therefore, we must ascertain whether the capacity of a bayesian MLP to have good predictive ability is a monotonic function of the evidence: the configuration with the largest evidence is expected to give the best results on future and unseen data. In such case, our methodology can also be used as a practical method to stop learning. For the moment, there is no theoretical link between predictive ability and evidence. We report the experimental results obtained with our methodology for two classification problems.

5. Application to classification problems

The first problem is based on the "Waveforms recognition" model constructed by Breiman et al. [1] and used to illustrate various parts of their 'tree classification' methodology. Though artificial, this problem is not trivial and is frequently utilized by machine learning practitioners. It is a 3 class and 21 inputs problem. According to Breiman, the learning set is respectively made of 100, 85 and 115 measurement vectors for classes 1, 2 and 3. Five thousand test vectors have been independently generated with equal proportions for the three classes. An analytical expression can be derived for the optimal Bayes classifier error rate. For the test sample size of 5000 this expression gives a recognition rate of 86%. The second problem is about medical diagnosis applied to breast cytology, "Cancer", from the University of Wisconsin Hospitals. It is available at the UCI Repository of Machine Learning Databases (`http://www.ics.uci.edu/mlearn`). It is a two class problem: benign or malignant based on cell descriptions. Each example is a real vector of size 9. There are 699 examples in which the first 350 are used as learning set.

Table 1 shows, for a MLP architecture with 78 weights in the Waveforms problem, the values of performances, quadratic errors E_D (on test set) and value of log-evidence at different steps of a bayesian procedure. We chose the fifth configuration, which has maximum logarithm of evidence (-491.67), as

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 267-272

W:	Performances	78.20%	84.68%	84.08%	84.38%	84.36%
W:	Quadratic Error E_D	3762	2843	2826	2830	2803
W:	Log of Evidence	−545.42	−496.45	−492.61	−491.84	−491.67
C:	Performances	98.85%	98.85%	98.56%	98.56%	98.85%
C:	Quadratic Error E_D	28.42	21.83	20.82	20.66	20.21
C:	Log of Evidence	−376.19	−337.64	−325.83	−322.04	−310.34

Table 1: Waveforms (**W**) and Cancer (**C**) problems: log-evidence function of performances and quadratic errors on test set.

the "best" MLP configuration. On the other hand, the configuration with maximum performance is the second one (84.68% vs. 84.36%). Let us examine this point.

First at all, there is a big difference between the first column of results and the others, whilst the values are very close for the other columns. Consider now the performance values of these last four columns. If we take into account the confidence interval or uncertainty on the performances obtained on the test set, which is 1.73% as given by the Hoeffding formula, these values may be considered essentially equivalent. Therefore, whether we choose the last of these four configurations, which has maximum log-evidence, or the first one, which has maximum performance, makes no difference. The important point is that, among the five configurations, the methodology is able to distinguish the first from the others. This suggests that the evidence as a method in ranking the different configurations "works". Note also that the correlation between evidence and quadratic error on test set is almost perfect, which suggests that our methodology may work even better on interpolation or regression problems.

We can observe the same results on other configurations of the architectures for the Cancer problem. Table 1 refers to a MLP with 122 weights. The confidence interval on test set is here 6.5%.

The tables are only a synthesis of our results on evidence. Figures 1 & 2 report the empirical correlation between generalization accuracy and log-evidence of two nets applied to the Waveform problem: the smaller net (Figure on the left) produced the results reported in Table 1. Note that the correlation between generalization accuracy and log-evidence is quite good for the bigger net. Moreover, with respect to Table 1, we now have a more organic and visible demonstration that some configurations of the best (and smaller) net violate the desired correlation. The nature of such outcome is probably due to the very small values that typically has the evidence, with consequent numerical problems: in Table 1 the values of the log-evidence give values of $P(D \mid M, H)$ below the precision of the computer!

It is important to have an idea of the computational complexity of the evidence estimation: the most complex and realistic problem on which we experimented our methodology and the evidence estimation with reasonable results is NASA "Satellite image problem" [3] (6 class, 36 inputs, ≈4 thousands

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 267-272

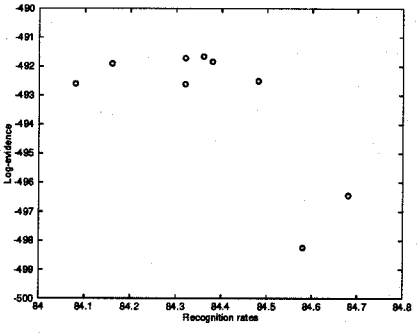

Figure 1: Waveform problem. log-evidence vs. generalization accuracy for a net with 78 weights.

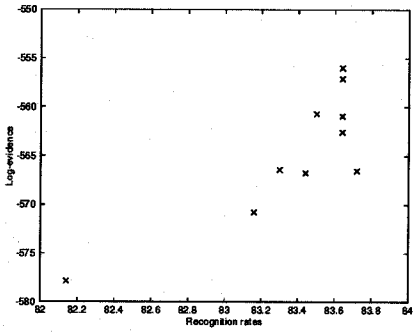

Figure 2: Waveform problem. log-evidence vs. generalization accuracy for a net with 103 weights.

training patterns), for which a MLP containing $\approx$ 1000 weights and 24h time on a SPARC10 for each evaluation of the evidence were needed.

Note finally that, in Figures 1 & 2, the log-evidence values are essentially bigger for the MLP configurations of the smaller net (Figure on the left), and that the performances on the test set also have this property. This result confirms that the evidence can actually rank different MLP architectures, as originally claimed and experimentally demonstrated by MacKay.

In conclusion, we advanced the use of a new methodology to rank MLP configurations during learning. It is crucial, however, to have some confidence interval estimation on the data in the test set, which allows to stop learning when different MLP configurations with similar evidence values (possibly influenced by numerical inaccuracy) have performances that can be considered "essentially equivalent". We underlined that evidence estimation can be numerically difficult and computational demanding, especially for practical applications. Even if results seem to be encouraging, further investigations must be undertaken to verify the accuracy of our (any!) evidence-based methodology, with particular care to the specific optimization algorithms used to find MLP weights.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth International Group, 1984.

[2] D.J.C. MacKay. *Bayesian Methods for Adaptive Models.* PhD dissertation, California Institute of Technology, Pasadena, California, December 1991.

[3] D. Perrotta. *Contribution of Bayesian Inference on Multi-Layer Perceptron Learning.* PhD thesis, École Normale Supérieure de Lyon, 46, allée d'Italie, 69364 Lyon Cedex 07, France, April 1997.