Maximum Likelihood Weights for a Linear Ensemble of Regression Neural Networks*

Michiel C. van Wezel, Walter A. Kosters and Joost N. Kok
{michiel,kosters,joost}@wi.leidenuniv.nl
Leiden Institute of Advanced Computer Science
Leiden University
P.O. Box 9512, 2300 RA Leiden, The Netherlands

Abstract

In this paper we study the problem of combining the outputs of the members of an ensemble of neural networks. We review the commonly used methods and thoroughly derive a cost function from a maximum likelihood perspective, that can be minimized in order to obtain maximum likelihood weights. The solution is shown to be closely related to a well-known statistical method. The various combination methods are tested on several datasets, with encouraging results.

1 Neural Network Ensembles

The output Q of a linear ensemble of K independent neural networks is given by $Q = \sum_k w_k z_k$, where z_k denotes the output of the k-th network ($k = 1, 2, \ldots, K$), and w_k denotes its weight. The problem is to determine a weight vector $W^T = (w_1, w_2, \ldots, w_K)$, in such a way that the total error is as small as possible. If all w_k's are taken to be equal, we arrive at a method sometimes referred to as "bagging"; if the lowest error network is selected (i.e., exactly one w_k is nonzero), we arrive at "bumping"; more sophisticated methods using quadratic programming, known as "balancing", try to find w_k's that are optimal in some respect. All three methods yield significant improvements over the individual performance of the ensemble members—but even better methods are possible.

2 Maximum Likelihood Approach

Under the assumption that the errors of the individual ensemble members have Gaussian distributions with mean 0, the distribution for the ensemble error is shown to be Gaussian too, again with mean 0. Here moment generating functions play a crucial role. The variance of the ensemble error equals $W^T \sigma W$ under the constraint $\sum_k w_k = 1$, where σ

*The full paper is published in the Proceedings of the International Conference on Neural Information Processing (ICONIP'98), Kitakyushu, Japan, October 1998.

is the covariance matrix of the individual networks. Good estimates for this matrix are available from the separate errors produced by the ensemble members. Note that the w_k's are not necessarily positive.

Now we search for a function that is minimal in case of optimal weights. Assuming independence between the samples in the dataset, and inspired by ideas from Bayesian analysis, we are led to the maximum likelihood cost function

$$E_{\mathrm{ml}} = \frac{1}{2}\left\{ N \log(W^T \sigma W) + \frac{1}{W^T \sigma W} \sum_{i=1}^{N} e^2_{Q(i)} \right\},$$

to be minimized under the constraint $\sum_k w_k = 1$. Here the dataset is supposed to have N elements, and $e^2_{Q(i)}$ denotes the square of the ensemble error on its i-th element: $e_{Q(i)} = \sum_k w_k z_{ik} - y_i$, where y_i is the desired output and z_{ik} is the output of the k-th network from the ensemble.

The constraint can be removed by explicit elimination of w_K, the weight of the last network. In all vectors and matrices we can remove the last row and column, leading to more complicated formulas. However, conjugate gradient methods—involving the partial derivatives of $W^T \sigma W$ and $e_{Q(i)}$ with respect to the w_k's ($k = 1, 2, \ldots, K-1$)—can be applied in order to find minima. Other methods that are appropriate include simulated annealing and the 30 years old Bates-Granger formula, a statistical technique developed for the combination of forecasts. Hybridizations of these methods are also possible.

3 Comparison of Results

We applied the methods from the previous sections to three different datasets: the first is a marketing dataset, where the output of the network is the market share of some fast moving consumer goods in terms of variables such as price and product availability; the second wants a prediction for the price of a particular share on the stock exchange, given the prices of some other shares during the preceding week; and the third is concerned with the prediction of the height of waves during storm situations in the Netherlands, as a function of the wave height and the weather in the preceding time period—a hot topic this time of year.

Maximum likelihood weighting yields the most convincing results on both test data and training data, and does not suffer from numerical instability, as seems to be the case with the Bates-Granger method (that involves matrix inversion). Especially for increasing ensemble sizes, the root mean squared error is the lowest for the two maximum likelihood methods (with conjugate gradient or simulated annealing) and the related Bates-Granger method, whereas bumping and balancing show results somewhere between these methods and naive methods like bagging. We conclude that our methods deserve further investigation.

For more details, including mathematical derivations of the maximum likelihood methods and the conjugate gradient approach, references to the literature and detailed results from experiments, the reader is referred to the full paper.