ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

General bound of overfitting for MLP regression models

J. Rynkiewicz

Universite Paris 1 - SAMM
90 Rue de Tolbiac, 75013 Paris - France

Abstract. Multilayer perceptrons (MLP) with one hidden layer have been used for a long time to deal with non-linear regression. However, in some task, MLP's are too powerful models and a small mean square error (MSE) may be more due to overfitting than to actual modelling. If the noise of the regression model is Gaussian, the overfitting of the model is totally determined by the behavior of the likelihood ratio test statistic (LRTS), however in numerous cases the assumption of normality of the noise is arbitrary if not false. In this paper, we present an universal bound for the overfitting of such model under smooth assumptions, this bound is valid without Gaussian or identifiability assumptions. The main application of this bound is to give a hint about determining the true architecture of the MLP model when the number of data goes to infinite.

1 Introduction

Feed-forward neural networks are well known and popular tools to deal with non-linear regression models. We can describe MLP models as a parametric family of regression functions. White [7] reviews statistical properties of MLP estimation in detail. However he leaves an important question pending i.e. the asymptotic behavior of the estimator when the MLP in use has redundant hidden units. If we assume that the noise is Gaussian it is well known that the Least Square Estimator (LSE) and the Maximum Likelihood Estimator (MLE) are equivalent and Amari et al. [1] give several examples of the behavior of MLE in case of redundant hidden units. Moreover, if n is the number of observations, Fukumizu [2] shows that, for unbounded parameters and Gaussian noise, LRTS can have an order lower bounded by $O(\log(n))$ instead of the classical convergence property to a χ^2 law. In the same spirit, Hagiwara and Fukumizu [3] investigate relation between LRTS divergence and weights size in a simple neural networks regression problem.

Even if Gaussian assumption for the noise is standard, it may be not suitable for some models. This assumption is false, for example, when the range of observations is known to be bounded, since Gaussian variables can be arbitrary large in absolute value, even if the probability of such events is small. Hence, we need a theory which gives evaluation of the overfitting of MLP regression without knowing the density of the noise and which works even if the model is not identifiable.

In this paper, we prove an inequality bounding the MSE difference between the true model and an over-determined model, this inequality shows that, under suitable assumptions, the asymptotic overfitting of the MSE is bounded in probability. Moreover, this bound shows that suitable penalized MSE criteria allow to select asymptotically the true model. The paper is organized as follows: In section 2 we state the model, section 3 presents our main inequality and in section 4 we apply this inequality to select the optimal architecture of the MLP model.

2 The model

Let $x = (x(1), \cdots, x(d))^T \in \mathbb{R}^d$ be the vector of inputs and $w_i := (w_{i1}, \cdots, w_{id})^T \in \mathbb{R}^d$ be a parameter vector for the hidden unit i. MLP function with k hidden units can be written :

$$f_\theta(x) = \beta + \sum_{i=1}^{k} a_i \phi\left(w_i^T x + b_i\right),$$

with $\theta = (\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_{11}, \cdots, w_{1d}, \cdots, w_{k1}, \cdots, w_{kd})$ the parameter vector of the model and ϕ a bounded transfer function, usually a sigmoid function. Note that we consider only real functions, extension to vector-valued functions is straightforward but not discussed in this paper. Let $\Theta_k \subset \mathbb{R}^{k\times(d+2)+1}$ be a compact (i.e. closed and bounded) set of possible parameters, we consider regression model $\mathcal{S} = \{f_\theta(y, x),\ \theta \in \Theta_k\}$ with

$$Y = f_\theta(X) + \varepsilon \tag{1}$$

X is random input variable and ε is the noise of the model. Let n be a strictly positive integer, we assume that the observed data $(x_1, y_1), \cdots, (x_n, y_n)$ come from a true model $(X_i, Y_i)_{i\in\mathbb{N}, i>0}$ of which the true regression function is f_{θ^0}, for an θ^0 (possibly not unique) in the interior of Θ_k. In the sequel, we write P the probability distribution of (X_i, Y_i).

2.1 Estimation of MLP regression model

The main goal of non-linear regression is to give an estimation of the true parameter θ^0 thanks to observations $((x_1, y_1), \cdots, (x_n, y_n))$. This can be done by minimizing the MSE function:

$$E_n(\theta) := \frac{1}{n}\sum_{t=1}^{n} \left(y_t - f_\theta(x_t)\right)^2 \tag{2}$$

with respect to parameter vector $\theta \in \Theta_k$. The parameter vectors $\hat{\theta}_n$ realizing the minimum will be called Least Square Estimator (LSE). Note that parameters realizing the true distribution function may belong to a non-null dimension submanifold if number of hidden units is overestimated. Suppose, for example, we have a multilayer perceptron with two hidden units and the true function f_{θ^0} is given by a perceptron with only one hidden unit, say $f_{\theta^0} = a_0 \tanh(w_0 x)$, with $x \in \mathbb{R}$. Then, any parameter θ in the set:

$$\{\theta\, | w_2 = w_1 = w_0, b_2 = b_1 = 0, a_1 + a_2 = a_0\}$$

realizes the function f_{θ^0}. Hence, classical statistical theory for studying the LSE can not be applied because it requires the identification of the parameters (up to a permutation).

In the next section, we will compare MSE of over-parameterized models against MSE of the true model :

$$\frac{1}{n}\sum_{t=1}^{n}\left(y_t - f_\theta(x_t)\right)^2 - \frac{1}{n}\sum_{t=1}^{n}\left(y_t - f_{\theta^0}(x_t)\right)^2 = E_n(\theta) - E_n(\theta^0). \tag{3}$$

3 A general bound for the MSE

For an square-integrable function $g(X,Y)$ the L_2 norm is:

$$\|g(X,Y)\|_2 := \sqrt{\int g^2(x,y)dP(x,y)},$$

for $\lambda > 0$, let us define the generalized derivative function :

$$d_\theta^\lambda(X,Y) = \frac{\frac{e^{-\lambda(Y-f_\theta(X))^2} - e^{-\lambda(Y-f_{\theta 0}(X))^2}}{e^{-\lambda(Y-f_{\theta 0}(X))^2}}}{\|\frac{e^{-\lambda(Y-f_\theta(X))^2} - e^{-\lambda(Y-f_{\theta 0}(X))^2}}{e^{-\lambda(Y-f_{\theta 0}(X))^2}}\|_2} = \frac{e^{-\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta 0}(X))^2\right)} - 1}{\|e^{-\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta 0}(X))^2\right)} - 1\|_2} \tag{4}$$

and let us define $\left(d_\theta^\lambda\right)_-(x,y) = \min\left\{0, d_\theta^\lambda(x,y)\right\}$. For now, let us assume that d_θ^λ is well defined, this point will be discuss later. We can state the following inequality:

Inequality:
for $\lambda > 0$,

$$\sup_{\theta\in\Theta_k} n \times \left(E_n(\theta^0) - E_n(\theta)\right) \leq \frac{1}{2\lambda}\sup_{\theta\in\Theta_k}\frac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)} \tag{5}$$

Proof:
We have

$$\begin{aligned}
&n \times \left(E_n(\theta^0) - E_n(\theta)\right) = \\
&\tfrac{1}{\lambda}\textstyle\sum_{i=1}^n \log\left(1 + \|\frac{e^{-\lambda(Y-f_\theta(X))^2} - e^{-\lambda(Y-f_{\theta 0}(X))^2}}{e^{-\lambda(Y-f_{\theta 0}(X))^2}}\|_2 d_\theta^\lambda(x_i,y_i)\right) \\
&\leq \sup_{0\leq p\leq\|\frac{e^{-\lambda(Y-f_\theta(X))^2} - e^{-\lambda(Y-f_{\theta 0}(X))^2}}{e^{-\lambda(Y-f_{\theta 0}(X))^2}}\|_2} \tfrac{1}{\lambda}\textstyle\sum_{i=1}^n \log\left(1 + pd_\theta^\lambda(x_i,y_i)\right) \\
&\leq \sup_{p\geq 0}\tfrac{1}{\lambda}\left(p\textstyle\sum_{i=1}^n d_\theta^\lambda(x_i,y_i) - \tfrac{p^2}{2}\textstyle\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)\right).
\end{aligned}$$

Since for any real number u, $\log(1+u) \leq u - \frac{1}{2}u_-^2$. Finally, replacing p by the optimal value, we found

$$n \times \left(E_n(\theta^0) - E_n(\theta)\right) \leq \tfrac{1}{2\lambda}\tfrac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)}$$

■

This inequality allows to prove that $n \times \left(E_n(\theta^0) - E_n(\theta)\right)$ is bounded in probability under simple assumptions. It is used in the next section to prove consistency of an estimator of the number of hidden unit using penalized MSE criterion.

4 Estimation of the number of hidden units.

Let k^0 be the minimal number of hidden units needed to realize the true regression function f_{θ^0}. In this section, the set Θ of possible parameters will be set to

$$\Theta = \cup_{k=1}^{K} \Theta_k,$$

where K is a, possibly huge, fixed constant: The maximum number of hidden units for MLP models. We define the minimum-penalized MSE estimator of k^0, as the minimizer $\hat{k}$ of

$$T_n(k) = \min_{\theta\in\Theta} \left(E_n(\theta) + a_n(k)\right) \tag{6}$$

Let us assume the following assumptions:

(A1) $a_n(.)$ is increasing, $n \times (a_n(k_1) - a_n(k_2))$ tends to infinity as n tends to infinity, for any $k_1 > k_2$ and $a_n(k)$ tends to 0 as n tends to infinity for any k.

(A2) It exists $\lambda > 0$ so that $\left\{d_\theta^\lambda, \theta \in \Theta\right\}$ is a Donsker class (see van der Vaart [6]).

We now have:
Theorem:
Under **(A1)** *and* **(A2)**, $\hat{k}$ *converges in probability to the true number of hidden units* k^0.
Proof:
By applying the inequality,

$$P(\hat{k} > k^0) \leq \sum_{k=k^0+1}^{K} P\left(T_n(k) \geq T_n(k^0)\right) =$$
$$\sum_{k=k^0+1}^{K} P\left(n\left(\sup_{\theta\in\Theta_{k^0}} E_n(\theta) - \sup_{\theta\in\Theta_k} E_n(\theta)\right) \geq n\left(a_n(k) - a_n(k^0)\right)\right) \leq$$
$$\sum_{k=k^0+1}^{K} P\left(\frac{1}{\lambda} \sup_{\theta\in\Theta_k} \frac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)} \geq n\left(a_n(k) - a_n(k^0)\right)\right)$$

Now, under **(A2)**

$$sup_{\theta\in\Theta_k} \frac{1}{n}\left(\sum_{i=1}^{n} d_\theta^\lambda(x_i,y_i)\right)^2 = O_P(1)$$

where, $O_p(1)$ means bounded in probability. Moreover, under **(A2)** the set $\left\{\left(d_\theta^\lambda(x_i,y_i)\right)^2\right\}$ is Glivenko-Cantelli (the set admits an uniform law of large numbers). Hence

$$\inf_{\theta\in\Theta_k}\frac{1}{n}\sum_{i=1}^{n}\left(d_\theta^\lambda(x_i,y_i)\right)_-^2 \overset{n\to\infty}{\longrightarrow} \inf_{\theta\in\Theta_k}\|\left(d_\theta^\lambda(X,Y)\right)_-\|_2^2$$

But $\inf_{\theta\in\Theta_k}\|\left(d_\theta^\lambda(X,Y)\right)_-\|_2>0$, since the random variable $d_\theta^\lambda(X,Y)$ is centered and $\|d_\theta^\lambda(X,Y)\|_2=1$. Then, we get :

$$\frac{1}{\lambda}\sup_{\theta\in\Theta_k}\frac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n\left(d_\theta^\lambda\right)_-^2(x_i,y_i)}=O_P(1)$$

and $P(\hat{k}>k^0)$ tends to 0 as n tends to infinity.

Finally,

$$P(\hat{k}<k^0)\leq\sum_{k=1}^{k^0-1}P\left(\sup_{\theta\in\Theta_k}\frac{E_n(\theta)-E_n(\theta^0)}{n}\geq\frac{a_n(k)-a_n(k^0)}{n}\right)$$

and $\sup_{\theta\in\Theta_k}\frac{E_n(\theta)-E_n(\theta^0)}{n}$ converges in probability to

$$\sup_{\theta\in\Theta_k}E\left(E_n(\theta)-E_n(\theta^0)\right)<0$$

since $k<k^0$, so $\hat{k}\xrightarrow{P}k^0$ ■

The assumption **(A1)** is fairly standard for model selection, in the Gaussian case **(A1)** will be fulfilled by the BIC criterion. The assumption **(A2)** is more difficult to check. First we note:

$$\begin{array}{l}\left(e^{-\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)}-1\right)^2=\\ e^{-2\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)}-2e^{-\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)}+1\end{array}$$

So, d_θ^λ is well defined if $E\left[e^{-2\lambda\left((Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2\right)}\right]<\infty$, but

$$\begin{array}{l}(Y-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2=\\ (Y-f_{\theta^0}(X)+f_{\theta^0}(X)-f_\theta(X))^2-(Y-f_{\theta^0}(X))^2=\\ 2\varepsilon(f_{\theta^0}(X)-f_\theta(X))+(f_{\theta^0}(X)-f_\theta(X))^2\end{array}$$

where $\varepsilon=Y-f_{\theta^0}(X)$ is the noise of the model. Since an MLP function is bounded, d_θ^λ is well defined if $\lambda>0$ exists such that $e^{\lambda|\varepsilon|}<\infty$ i.e. ε admits exponential moments. Finally, using the same techniques of reparametrization as in Rynkiewicz [5], assumption **(A2)** can be shown to be true for MLP models with sigmoïdal transfer functions, if the set of possible parameters Θ is compact.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

5 Conclusion

Summary of the findings. This paper shows that the overfitting of MLP regression models is moderate for a large number of application without any Gaussian assumptions on the noise. Note that we assume in this paper that the norm of the weights of the MLP are a priori bounded by a possibly huge constant. In this framework, the user can select the true number of hidden units thanks to penalized means square criteria similar to BIC. So, if the user seeks to minimize

$$E_n(\theta) + D \times \frac{\log(n)}{n}$$

where D is proportional to the number of hidden units of the models and n the number of observations, then the true number of hidden units will be automatically selected if n is large enough.

As a conclusion MLP regression is widely used and always a very competitive method (see Osowski et al. [4]), however their is a lack of theoretical justification for determining the true architecture and especially the number of hidden units. Indeed, the classical asymptotic theory fails when the model is not identifiable. In this paper, we prove an inequality showing that overfitting of MLP is moderate if the noise admits exponential moments and the parameters of the model are a priori bounded. This bound justifies the use of penalized criterion similar to the BIC criterion in order to fit the architecture of MLP models in the framework of regression without knowing the density of the noise. Finally, a more challenging task may be to get a more precise tuning of penalization term which, according to our result, can be chosen among a wide range of functions.

References

[1] S. Amari, H. Park and T. Ozeki, Singularities affect dynamics of learning in Neuromanifolds, *Neural computation*, 18, 1007–1065, MIT Press, 2006.

[2] K. Fukumizu, Likelihood ratio of unidentifiable models and multilayer neural networks *The Annals of Statistics*,31, 833–851, IMS, 2003.

[3] K. Hagiwara and K. Fukumizu, Relation between weight size and degree of over-fitting in neural network regression. *Neural networks*, 21, 48–58, Elsevier, 2008.

[4] S. Osowski, K. Siwek and T. Markiewicz, MLP and SVM networks - a Comparative study. *proceedings of the* 6th *Nordic Signal Processing Symposium* (NSPS 2004), pages 37-40, June 9-11, Espoo (Finland), 2004.

[5] J. Rynkiewicz, Consistent estimation of the architecture of multilayer perceptrons. In M. Verleysen, editor, *proceedings of the* 14th *European Symposium on Artificial Neural Networks* (ESANN 2006), d-side pub., pages 149-154, April 28-30, Bruges (Belgium), 2006.

[6] A.W. van der Vaart, *Asymptotic statistics*, Cambridge university Press, Cambridge, 1998.

[7] H. White, *Artificial Neural Networks: Approximation and Learning Theory.*, Basil Blackwell, Oxford, 1992.

[8] J. Yao, On least square estimation for stable nonlinear AR processes. *The Annals of Institut of Mathematical Statistics*,52, 316–331, Springer, 2000.