ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 187-192

On fields of nonlinear regression models

Bruno Pelletier[1], Robert Frouin[2]

[1]Université du Havre, Laboratoire de Mathématiques Appliquées
76063 Le Havre, France
[2]Scripps Institution of Oceanography, La Jolla Shores Drive
La Jolla, California, USA
Bruno.Pelletier@univ-lehavre.fr, RFrouin@ucsd.edu

Abstract. In the context of nonlinear regression, we consider the problem of explaining a variable y from a vector $\mathbf{x}$ of explanatory variables and from a vector $\mathbf{t}$ of conditionning variables, that influences the link function between y and $\mathbf{x}$. A neural based solution is proposed in the form of a field of nonlinear regression models, by which it is meant that the relation between those variables is modeled by a map from some space to a function space. This approach results in a broader class of neural models than that of perceptrons, which therefore inherits the interesting approximation theoretical properties of the latter. The interest of such a modeling is illustrated by a real-world geophysical application, namely ocean color remote sensing.

1 Introduction

Statistical models, such as linear and, more generally, nonlinear regression models, aim at explaining an exogenous variable y from several explanatory, or endogenous, variables $x_1, ..., x_n$. Neural networks such as multilayer perceptrons and radial basis functions networks, falling in the class of so-called ridge constructions, achieve this goal with several well-known interesting properties. Let us just mention the density property, aka universal approximation property[3][5], and the results related to the approximation rate (the dimension independent upper bound[1][2], and the asymptotic expression obtained by Maiorov[6] for instance).

In this vein, we focus on a slightly different regression problem, for which we propose a neural based solution inheriting the interesting mathematical properties mentioned above. This problem still consists in explaining y from $x_1, ..., x_n$, but with the difference that, in fact, only some of the x_i, say $x_1, ..., x_d$ ($d < n$), convey information about y, while the remaining variables act as parameters, or conditionning variables, in the sense that they influence the link function between y and the real informative variables $x_1, ..., x_d$.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 187-192

Typical examples of this kind of problem are to be found in the field of geosciences, where the observed data may depend on several angular variables that define the geometry of the observation process. Let us briefly describe the ocean color remote sensing problem. It consists in estimating the concentrations of several oceanic constituants, such as phytoplankton chlorophyllous pigments, from radiometric measurements from space $x_1, ..., x_d$. In fact, those radiometric measurements depend continuously on angular variables that are used to characterize the positions of the observing satellite and of the sun relatively to the target point on the Earth' surface. Hence those angular variables, which obviously do not carry any information about the phytoplankton concentration, have to be taken into account, for the link function between the phytoplankton concentration and the measurements $x_1, ..., x_d$ depends on them. A much more academic example is the case of the reconstruction of continuously parametrized hypersurfaces from scattered noisy data.

For this kind of problem, it seems natural to separate the variables being effectively informative with respect to y, from the conditioning variables. We shall denote by $\mathbf{x}$ the d-dimensional vector of informative variables, and by $\mathbf{t}$ the p-dimensional vector of conditionning variables. The proposed solution consists in attaching to $\mathbf{t}$ a nonlinear regression model explaining y from $\mathbf{x}$, and where we demand that the attachment vary smoothly in $\mathbf{t}$. This approach yield a field of nonlinear regression models over the set of permitted values for $\mathbf{t}$. As will be explained further, it also conduces to a broader class of neural models that, consequently, inherits their approximation theoretical properties.

The paper is organized as follows. In the next section, the problem of interest is stated more formally, and fields of nonlinear regression models are defined. In section 3, stochastic learning algorithms are presented for the construction of such a model from scattered data. In section 4 are presented results obtained by applying this methodology to the ocean color problem. Finally, concluding remarks are given.

2 Function fields and nonlinear regression models fields

Let $\mathbf{x}$ be a vector of explanatory variables, let $\mathbf{t}$ be a vector of conditionning variables, and let y be the real variable to be explained. Let X and T be the sets of permitted values for $\mathbf{x}$ and $\mathbf{t}$, respectively. We consider statistical models of the following form:

$$y = f_{\mathbf{t}}(\mathbf{x}) + \epsilon \tag{1}$$

where for each $\mathbf{t} \in T$, $f_{\mathbf{t}}$ is an element of a subset $\mathcal{M}$ of $\mathcal{C}(X)$, the set of continuous real valued functions on X, and ϵ is a random variable of null mean and finite variance σ^2 that is not correlated with $\mathbf{x}$. Hence in this model, $\mathbf{x}$ carries information about y, while $\mathbf{t}$ does not, but the link function between y and $\mathbf{x}$ depends on $\mathbf{t}$. The definition of the set $\mathcal{M}$ will be stated later.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 187-192

To study the dependence of $f_{\mathbf{t}}$ on $\mathbf{t}$, we introduce the notion of a function field over T. We shall assume that X is locally compact and Hausdorff, and that T is compact, metric and Hausdorff. We define a *function field* over T as being a map $T \to \mathcal{C}(X)$. The set of all continuous function fields over T will be denoted by $(\mathcal{C}(X))^T$. The natural topology on $(\mathcal{C}(X))^T$ is the compact-open topology, which is equivalent to the topology of uniform convergence on compact sets, under the above assumptions on the sets X and T. Furthermore, there is the homeomorphism $\mathcal{C}(X \times T) \stackrel{\approx}{\to} (\mathcal{C}(X))^T$. Hence we introduce the following notation. For each $\zeta \in (\mathcal{C}(X))^T$, we define the map $\zeta_* : X \times T \to \mathbf{R}$ by letting $\zeta_*(\mathbf{x}, \mathbf{t}) = \zeta(\mathbf{t})(\mathbf{x})$. Similarly, the set of all $\mathcal{M}$-valued continuous function fields over T will be denoted by $\mathcal{M}^T$.

Returning to the initial problem, Eq. (1) may be rewritten equivalently as

$$y = \zeta(\mathbf{t})(\mathbf{x}) + \epsilon \tag{2}$$

or as

$$y = \zeta_*(\mathbf{x}, \mathbf{t}) + \epsilon \tag{3}$$

where ζ belongs to $\mathcal{M}^T$. Hence Eq. (2) defines a *field of regression models* over T. One may show that if $\mathcal{M}$ is dense in $\mathcal{C}(X)$ and if T is as above, then $\mathcal{M}^T$ is dense in $(\mathcal{C}(X))^T$.

Herein, we shall be interested in the case where the model set $\mathcal{M}$ is the set of one-hidden layer perceptrons or, more generally, the set spanned by functions of the ridge form. Hence we consider the set $\mathcal{M} = \cup_n \mathcal{M}_n$ where

$$\mathcal{M}_n = \{\sum_{i=1}^{n} c_i h(\mathbf{a}_i \mathbf{x} + b_i), b_i, c_i \in \mathbf{R}, \mathbf{a}_i \in \mathbf{R}^d\} \tag{4}$$

We address now the construction of fields of nonlinear regression models, a topic closely related to their parametrization. Let $\zeta \in \mathcal{M}^T$. Since T is compact, we may assume, without loss of generality, that ζ belongs to $\mathcal{M}_n^T$, for some integer n. Each element of $\mathcal{M}_n$ depends on parameters $c_i, \mathbf{a}_i, b_i$, for $i = 1, ..., n$, that we shall summarize by a vector θ_n. Let Θ_n be the set of allowable values for θ_n, i.e., $\Theta_n = \prod_{i=1}^{n} \mathbf{R} \times \mathbf{R}^d \times \mathbf{R}$, and let $i_n : \Theta_n \to \mathcal{M}_n$ be the continuous map carrying a parameter vector θ_n to the corresponding model of $\mathcal{M}_n$. We intend to build a continuous function field $\zeta \in \mathcal{M}_n^T$ through a parameter map $\xi : T \to \Theta_n$ such that $\zeta = i_n \circ \xi$. Let us mention the following difficulties, arising because the map i_n is only a continuous surjection. First for each $\zeta \in \mathcal{M}_n^T$, there might not exist a continuous map $\xi : T \to \Theta_n$ such that $\zeta = i_n \circ \xi$. Secondly, if we proceed conversely by building ζ according to $\zeta = i_n \circ \xi$, where ξ is continuous, we are not sure to get all of $\mathcal{M}_n^T$ when ξ is allowed to vary in all of $\mathcal{C}(T, \Theta_n)$. However one may show first, that the set of continuous function fields $\zeta \in \mathcal{M}^T$ such that

$$\zeta_*(\mathbf{x}, \mathbf{t}) = \sum_{i=1}^{n} c_i(\mathbf{t}) h\left(\mathbf{a}_i(\mathbf{t})\mathbf{x} + b_i(\mathbf{t})\right) \tag{5}$$

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 187-192

for some integer n, $c_i \in \mathcal{C}(T)$, $\mathbf{a}_i \in \mathcal{C}(T, \mathbf{R}^d)$ and $b_i \in \mathcal{C}(T)$ is dense in $(\mathcal{C}(X))^T$, and secondly, that for producing in this way a dense set of continuous function fields, it is sufficient that the c_i and $\mathbf{a}_i$ lie in subsets of $\mathcal{C}(T)$ and $\mathcal{C}(T, \mathbf{R}^d)$, respectively, containing the constant functions, and that the b_i lie in some subset of $\mathcal{C}(T)$ containing the affine functions. This results, roughly speaking, from the density in $\mathcal{C}(X \times T)$ of perceptrons set on $X \times T$. Clearly we see that we obtain a broader class of models which, therefore, inherits by construction the interesting approximation theoretical properties of those networks.

3 Algorithms

Let $\mathcal{D}$ be a data set of N samples $(\mathbf{x}_i, \mathbf{t_i}, y_i)$. Based on $\mathcal{D}$, we are willing to represent the link between y, $\mathbf{x}$ and $\mathbf{t}$ through a field of nonlinear regression models of the following form

$$y = \zeta(\mathbf{t})(\mathbf{x}) + \epsilon \tag{6}$$

where $\zeta \in \mathcal{M}_n^T$ and where $\mathcal{M}_n$ is as in (4). In light of the results stated in the previous section, we present below two methods for constructing ζ via a parameter map $\xi : T \to \Theta_n$. In both of them, we assume $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and use the averaged sum of the squared errors $\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \zeta(\mathbf{t}_i)(\mathbf{x}_i))^2$ as the natural criterion to be minimized. The main difference with traditional perceptrons is that here, the parameters for ζ are maps $T \to \Theta_n$.

The first method consists in taking a parametrized subset $\mathcal{F}_\rho$ of $\mathcal{C}(T, \Theta_n)$, that contains the constant and affine functions of $\mathbf{t}$ (for the reason previously mentioned), where ρ is the parameter vector. Hence the problem reduces to the one of minimizing $\mathcal{E}$ with respect to ρ, which may be achieved, for instance, by means of a stochastic gradient descent algorithm or simulated annealing. However this method may suffer from an inappropriate choice of $\mathcal{F}_\rho$, which may yield a much more larger n than necessary. The second method described below allows one to cope with this issue.

This second method consists in building a sample of a continuous map $\xi : T \to \Theta_n$ such that the induced field $\zeta := i_n \circ \xi$ minimizes $\mathcal{E}$. We proceed as follows. Assume T is a compact subset of $\mathbf{R}^p$. Let $\mathbf{t}_1^\Xi, ...\mathbf{t}_K^\Xi$ be K points of $\mathbf{R}^p$, being the vertices of a regular grid of $\mathbf{R}^p$, such that T is included in the smallest p-dimensional cube Ξ containing all of the $\mathbf{t}_k^\Xi$. Note that K is the product of p integers $k_i \geq 2$. Hence $T \subset \Xi$, and $\mathbf{t}_k^\Xi \in \Xi$ for all $k = 1, ..., K$. Let $\gamma_1, ..., \gamma_K$ be K real numbers and consider those continuous and piecewise-differentiable maps $g \in \mathcal{C}(\Xi)$ such that $g(\mathbf{t}_k^\Xi) = \gamma_k$ for all $k = 1, ..., K$, and defined for all $\mathbf{t} \in \Xi$ such that $\mathbf{t} \neq \mathbf{t}_k^\Xi \forall k$ by:

$$g(\mathbf{t}) = \sum_{i=1}^{2^p} \alpha_i(\mathbf{t}) g(\mathbf{t}_{k_i}^\Xi) \tag{7}$$

In this equation, the $\mathbf{t}_{k_i}^\Xi$ are the 2^p immediate neighbours of $\mathbf{t}$ on the grid, i.e., they are the vertices of a p-cube containing $\mathbf{t}$, and the coefficients $\alpha_i(\mathbf{t})$ are the

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 187-192

coefficients of the standard p-dimensional interpolation procedure on a p-cube. We shall denote by $\mathcal{F}_K$ the set of all such maps. Next consider those function fields ζ over T being the restrictions to T of function fields $\tilde{\zeta}$ over Ξ which satisfy to

$$\tilde{\zeta}_*(\mathbf{x}, \mathbf{t}) = \sum_{i=1}^{n} c_i(\mathbf{t}) h\left(\mathbf{a}_i(\mathbf{t})\mathbf{x} + b_i(\mathbf{t})\right) \tag{8}$$

where the b_i, the c_i and the components of the $\mathbf{a}_i$ belong to $\mathcal{F}_K$. Hence such a function field is parametrized by $n(d+2)K$ real numbers γ_i^j, where $1 \leq i \leq K$ and $1 \leq j \leq n(d+2)$, since $\dim(\Theta_n) = n(d+2)$. The minimization of $\mathcal{E}$ with respect to them may be performed as follows. First, pick randomly a sample $(\mathbf{x}_i, \mathbf{t}_i, y_i)$ from $\mathcal{D}$ and compute the error e_i of the model for that sample. Next, if $\mathbf{t}_i$ fall on one of the vertices of the grid, say on $\mathbf{t}_{k_i}^{\Xi}$, then modify the $\gamma_{k_i}^j$ for $j = 1, ..., n(d+2)$ by the amount $-\eta \frac{\partial e_i}{\partial \gamma_{k_i}^j} e_i$, where $\eta > 0$ is the learning rate. Otherwise, $\mathbf{t}_i$ is different from all the vertices. Let $\mathbf{t}_{k_l}^{\Xi}$ be the 2^p immediate neighbours of $\mathbf{t}_i$ on the grid ($l = 1, ..., 2^p$). Then modify all the $\gamma_{k_l}^j$ by the amounts $-\eta \frac{\partial e_i}{\partial \gamma_{k_l}^j} e_i$, respectively, the expanded expressions of which may be easily derived. Finally, those steps are repeated until convergence. By this method, we obtain from the sets $\mathcal{D}_j^{\Xi} := \{(\mathbf{t}_k, \gamma_k^j); k = 1, ..., K\}$ a sample of a map ξ inducing the function field $\zeta = i_n \circ \xi$. Its advantages with respect to the first method are i) that the grid may be refined during the execution of the learning algorithm, and ii) that the resulting sample may be used in a second time to choose an appropriate model set for ξ offering, for example, a higher degree of regularity.
By analogy with the case of nonlinear regression with multilayer perceptron when the number N of samples tends towards infinity, the resulting field ζ of nonlinear regression models is expected to be a good approximation to the field $E_{\mathbf{t}}[y|\mathbf{x}]$ over T of the ($\mathbf{t}$ dependent) conditional means of y given $\mathbf{x}$.

4 Application to ocean color remote sensing

For this problem, the vector $\mathbf{t}$ consists of three angular variables, the vector $\mathbf{x}$ is composed of reflectances at wavelengths located in the visible and near-infrared (typically a number of 8), and the variable y is the near surface chlorophyll-a concentration ([Chl-a]), physically related to the phytoplankton concentration. A statistically significant data set of about 62,000 samples, encompassing all the sources of variability (mostly due to the atmosphere) has been generated via intensive use of simulation (multiple runs of a radiative transfer code), and has been randomly split into data sets $\mathcal{D}_l^0$ and $\mathcal{D}_v^0$, used for learning and validation, respectively. Two fields F and F^ν of nonlinear regression models have been built according to the second method, on a 2x2x3 regular grid, where the nonlinear regression models attached to $\mathbf{t}$ are one-hidden layer perpceptrons with 10 neurons. They have been trained both on $\mathcal{D}_l^0$, but in the case of F^ν, some amount of realistic noise has been added to the data during the execution

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 187-192

of the stochastic learning algorithm. As shown in Table 1, the resulting error in the chlorophyll-a concentration estimation is at the order of 4.2% over the range $0.3 - 30$mg/m3, in the case of non-noisy data, and 10% in the case of realistic noisy data, which illustrates the efficiency and the robustness of this modelling. These results represent a significant improvement, since actual processing techniques[4] yield theoretical errors that may reach over 20% in the absence of noise, and larger values in the presence of noise.

	F	F^ν
$\mathcal{D}_l^0$	0.042	0.068
$\mathcal{D}_v^0$	0.042	0.070
$\mathcal{D}_l^1$	0.151	0.104
$\mathcal{D}_v^1$	0.151	0.105

Tab 1. This table gives the mean relative error in [Chl-a] estimation evaluated on data sets $\mathcal{D}_l^0$, $\mathcal{D}_v^0$, $\mathcal{D}_l^1$, $\mathcal{D}_v^1$, for models F (trained on non noisy data) and F^ν (trained on noisy data), where $\mathcal{D}_l^1$ and $\mathcal{D}_v^1$ are realistic noisy versions of $\mathcal{D}_l^0$, $\mathcal{D}_v^0$, respectively.

5 Concluding Remarks

Fields of nonlinear regression models allows one to deal with composite data, where some variables are effectively explanatory, while the others are conditionning, and without having recurse to the product space $X \times T$, which in some cases, such as the ocean color remote sensing problem, may be meaningless. The methodology developped in this work is rather general, and one could take for $\mathcal{M}_n$ any arbitrary set of models, such as a set being homeomorphic to some subset of a finite dimensional euclidean space, which would facilitate the construction of those fields. However of particular interest is the case where those models are taken as neural networks, since as shown above, they inherit the interesting approximation theoretical properties of neural networks.

References

[1] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Inform. Theory*, 39(3):930–945, 1993.

[2] M. Burger and A. Neubauer. Error bounds for approximation with neural networks. *J. Approx. Theory*, 112:235–250, 2001.

[3] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2:303–314, 1989.

[4] H.R. Gordon. Atmospheric correction of ocean color imagery in the earth observing system era. *J. Geophys. Res.*, 102:17081–17118, 1997.

[5] V. Ya. Lin and A. Pinkus. Fundamentality of ridge functions. *J. Approx. Theory*, 75:295–311, 1993.

[6] V.E. Maiorov. On best approximation by ridge functions. *J. Approx. Theory*, 99:68–94, 1999.