An Incremental Construction of a Nonparametric Regression Model

Jan Smid
Department. of Mathematics
Morgan State University
Baltimore, MD 21239
smid@gsfc.nasa.gov

Petr Volf
UTIA
The Czech Academy of Sciences
Prague 8, CZ 182 08
volf@utia.cas.cz

Section: Stat. Strategy
AIStats97, Ft.Lauderdale, Fl, Jan 4-7, 97

Abstract

We study incremental (adaptive) algorithms for nonparametric regression models (one -hidden layer neural networks). Models are constructed from sets of localized functional units (e.g. radial basis functions). Both growing and pruning of a model is controlled by statistical tests and penalized error measures. The models are grown/pruned using different statistical/information criteria. Our numerical experiments show that the t-test and information type of criteria for growing/prunning for an incremental model perform at about the same level of effectivness.

1 Introduction

1.1 The Relevance to AI/Engineering Community

This paper deals with the problem of modeling of an unknown functional dependence between input and output variables of a system. The functional dependence modeling is highly relevant as a computer-aided decision tool in monitoring streams of data characterizing an instrument. For example, the Hubble Space Telescope (HST) batteries need to be monitored to assure their health and to detect a potential failure or degradation of the batteries. Functional dependencies parameters and their change in time can be incorporated into existing HST visual inspection tools. This tool can give the HST engineering deeper insight into the complex behavior of the batteries and the entire spacecraft system.

1.2 Existing Methods

One of the best known methods for fitting functions that includes interactions is Friedman's MARS, [3], and a similar method due to Breiman, [2]. MARS does not use localized units. Basis functions are added incrementally during learning, using the technique of sequential forward selection, which can be viewed as a tree-like technique, (Bishop [1]).

We deal in this paper with a class of incremental methods that use localized basis functions. Localized units offer more flexibility. For example, Weierstarss theorem guarantees approximation of any continuous function but other basis of localized functions are more effective in applications. Other drawback of global units is their inability to represent an abrupt change of the model structure. Incremental algorithms typically add a new basis function (unit) to the set of old units and then a set of parameters (one or more), typically in a lower dimensional parameter space, is adjusted.

Incremental algorithms perform better when we need to learn new features and keep the old ones as well, or if we only need to slightly modify the existing model. Incremental methods also allow us to exploit some abstract theorems and algorithms and the heuristic based on them. The advantage of an incremental procedure shows also when we deal with semi-data-driven model. What we mean by semi-data-driven model is the situation where some structural information, but not complete information, about a system is known ahead of time. When the structural information is missing we need to approach the data modeling incrementally and locally. Data-driven modeling and model-driven-modeling are extreme cases of the semi-data-driven modeling. In the model driven approach we assume a definitive functional structure of the model and look for unknown parameters. In the data-driven approach we assume no functional structure that is to be determined from the data.

1.3 Functional Dependence Modeling

The estimator is built from a set of localized basis functional units, e.g. B-splines or radial basis functions (RBF). The presence of random noise casts the problem in the context of statistical nonparametric regression. We propose a procedure for both growing and pruning of the estimator. The growing is based on incremental fitting of residuals. We deal with a number of stochastic and deterministic procedures. The presented methods are generalized for the case of likelihood models (e.g. GAM [4], logistic models for classification, Cox's nonparametric model). The selection of units and stopping rules employ criteria of statistical analysis, namely the criteria of the least penalized residual squares. There is a number of choices among these criteria (AIC, BIC, variants of cross–validation). Our numerical experiments and data case studies have been based on the BIC (Bayes Information Criterion) and the gamma criterion. We have also used standard methods of statistical inference, namely the t-test, and ANOVA (Analysis of Variance), respectively.

In the case of a multivariate input we use the additive model and models with interactions both based on univariate and multivariate units. The additive model allows us to use the same penalization methods and it also reduces complexity of the estimator.

Consider variables $\boldsymbol{X}$ and Y as the input (predictor) and output (response) of a system that are respectively p and 1 dimensional. i.e. $\boldsymbol{X}$ takes on values $\mathbf{x}$ from R^p, and Y takes on values y from R^1. We first assume that the data $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, $\mathbf{x}_i = (x_{i1}, x_{i2}, ...x_{ip})$, are mutually independent realizations of a pair of random variables $(\boldsymbol{X}, Y)$, and are generated by a nonparametric regression function r

$$y = r(\boldsymbol{x}) + e$$

where e stands for random noise with mean zero and a constant variance σ^2. Such a general model applies not only to standard regression problems, but also to recursive cases, including general autoregression, e. g. $y_t = r(\mathbf{x}_t) + e_t$, $y_t = r(y_{t-1}, y_{t-2,......,}y_{t-k}) + e_t$, respectively.

Our goal is to construct a function $\hat{r}(\boldsymbol{x})$ that approximates a regression function $y = r(\boldsymbol{x})$. We shall consider $\hat{r}$ constructed from localized basis functions. Such an estimator can also be viewed as a feedforward neural network with one hidden layer constructed form units (basis functions). There are two problems that need to be solved in this approach: First, how to choose units (basis functions) and of their number. Second, how to avoid data overfitting. In many cases the number of input variables (predictors) is unknown and also needs to be optimized.

Each procedure of the model construction has to answer three questions:

(i) How to select the unit which is the best candidate for innovation?

(ii) How to decide whether the innovation is effective?

(iii) When to stop the procedure?

2 Procedure of solution

We will consider units (localized basis functions) selected from a set P or a sequence of sets $\{P_j\}$ of units. There is a number of reasonable choices for a pool of units. For example a pool of units is a set of all radial basis functions $P_j = span\left\{B\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b}\right), b \in R;\ \mathbf{x}, \mathbf{c}_i \in R^d, i = 1, 2, ..., I; a \text{ constant } \alpha_o\right\}$ with the widths b and centers $\mathbf{c}$. The dimension of $\mathbf{x}$ can change, i.e. $d = d(i)$ is a function. Other choice of a set P is the additive model based on univariate units, and yet other one is the model with interactions, i.e. we consider products of univariate units.

2.1 Selecting Nonlinear Units

After we determined the pool of units a selection process is defined for an incremental process. We can make selection of a unit stochastically or we can optimize units' parameters deterministically. As a model situation we will consider a selection process of centers of radial basis functions. The basic strategy is to select a unit and an external parameter(s) that decrease RSS. The iteration is defined as

$$r_m = \beta_{m-1} r_{m-1} + \alpha_m B_m(\mathbf{c}_i)$$

or if we do want to modify the old parameters non-uniformly

$$r_m = r_{m-1} + \alpha_m B_m(\mathbf{c}_i), \; r_{m-1} = \sum_{j=1}^{m-1} \beta_{(m-1)j} B_j(\boldsymbol{x})$$

where r_m is the model after the m-th iteration, α, β are external parameters. This selection can be substantiated by Jones' theorem. Jones' theorem, [6, 7, 8], can be paraphrased as follows: iterations

$$r_m = (1-\alpha_m) r_{m-1} + \alpha_m B_m(\mathbf{c}_m), \text{ known as convex algorithm}$$

$$r_m = \beta_m r_{m-1} + \alpha_m B_m(\mathbf{c}_m), \text{known as linear algorithm}$$

converge to a target function $r(\mathbf{x}) : I^d \rightarrow R$, provided that B_m, α_m, β_m and $\mathbf{c}_m$ are chosen in an almost optimal way. To avoid nonlinear optimization problems we choose new units based on following heuristic rules.

1. Random selection of centers between existing adjacent centers, or at the center of gravity of a cluster of d-dimensional points.
2. Bisecting an interval given by existing adjacent centers (a deterministic variant of the random selection).

Remark 1 *The above procedures can be generalized for units defined by a set of some parameters , not necessarily in terms of centers and widths. An example of other set of parameters are knots of (B)splines.*

As the rule of thumb for the selection of the unit's diameter we set the width value to the distance of the center of a candidate unit to the center of the nearest neighboring unit.

2.2 Finding External Coefficients

Let us assume that $\{B_j(\mathbf{x}),\ j = 1, \ldots, m-1\}$ is a set of basis functions (old units), $B_m(\mathbf{x})$ is a new unit, from which the estimator is constructed. The estimator will then have the form of a linear combination $\hat{r}(\mathbf{x}) = \alpha_o + \sum_{j=1}^{m} \alpha_j\, B_j(\mathbf{x})$. Since we do not need to solve nonlinear problem the optimal values of external parameters $\hat{\alpha} = (\boldsymbol{\alpha}_\mathbf{o}, \boldsymbol{\alpha}_\mathbf{1}, \boldsymbol{\alpha}_{\mathbf{2},\ldots,} \boldsymbol{\alpha}_\mathbf{m})$ can be directly found

1. by solving the linear least squares problem

$$\hat{\alpha} = \arg \min_{\boldsymbol{\alpha}} \sum_{\mathbf{i}=1}^{\mathbf{m}} (\mathbf{y_i} - \hat{\mathbf{r}}(\mathbf{x_i}))^2.$$

2. or for each candidate B_m we solve a one dimensional least-squares problem to obtain a coefficient α_m. This approach defines a model incrementally and we can apply the following criteria for growing and pruning. This method has clear heuristic and is computationally very convenient.

2.3 Criteria for Model Growing

Let the residuals $e_i = y_i - \hat{r}_{m-1}(\boldsymbol{x}_i)$, correspond to an estimator

$$\hat{r}_{m-1}(\boldsymbol{x}) = \sum_{j=1}^{m-1} \alpha_j\ B_j(\boldsymbol{x}).$$

Now the data $(\boldsymbol{x}_i, e_i)$ is to be fitted by a new candidate unit, selected by one of the methods described above. Let choose the most successful method of our numerical experiments. For each candidate B_m, we solve a one dimensional least-squares problem to obtain a coefficient α_m.

$$\hat{r}_m(\boldsymbol{x}) = \hat{r}_{m-1}(\boldsymbol{x}) + \alpha_m\, B_m(\boldsymbol{x})$$

Now, the penalized error criterion

$$PE = \ln(MSE) + \text{complexity term}$$

determines, whether a new unit is effective. We have used several criteria, namely BIC of Schwarz with the complexity term $m\ln(n)/n$ and the γ–criterion, with the complexity term $mn^{-\gamma}$, where γ is selected from $(0.5, 1)$. They have different strengths and we can thus choose the more or less conservative strategy.

2.4 Use of the t–statistics for Model Growing and Pruning

In conventional statistics the t–test has been used for linear regression models to show whether certain linear coefficient can be set to zero without significant increase of the residual variance. The same procedure can be used for testing of external coefficients α of nonlinear units. In the growing phase, we accept a new unit B_m if the t–test indicates that α_m differs from zero significantly. In the pruning phase, we can perform the t–test for all external parameters (or for their subset) simultaneously. We simply delete that unit - i.e. we set to zero that coefficient - for which a t–statistics value lies below a critical value for the t–test (with $n - m - 1$ degrees of freedom, m is a number of units in the model, n is the number of data points). Simultaneously the penalizing criteria (e.g. gamma or BIC) are used as the additional indicators with the t–test.

2.5 Iteration of optimal additive model

Let us assume in this paragraph that basis functions $B(x_k)$ have one dimensional input x_k. By an additive model, we mean that (Breiman [2])

$$r(\boldsymbol{x}) = \sum_{k=1}^{p} g_k(x_k), \quad \mathrm{x} = (x_1, x_2, ...x_p)$$

Each function $g_k(x)$ can be modeled the sum of basis functions. The model then reads in terms of basis functions

$$r(\boldsymbol{x}) = \alpha_0 + \sum_{k=1}^{p}\sum_{j=1}^{m_k} \alpha_{jk}\, B_{jk}(x_k).$$

It is clear that growing and pruning of the additive model can be handled by the same procedure as the model with one dimensional input.

2.6 Models containing interactions of predictors

Let us now consider an estimator in the form of linear combination of units with multi-dimensional input vector. We assume that the units with multi-dimensional input are the tensor products of univariate unit functions. We follow C. J. Stone's "dimensionality reduction principle" [5] and keep the dimensionality of the model as low as possible. This means that we prefer models containing none or low–dimensional interactions. If necessary we choose the maximal dimension of interactions in advance and increase it during an adaptive process .

The procedure of model growing and pruning is essentially the same as in the case of strictly additive structure. The iterative selection of new basis functions runs through domains of components $x_1, x_2, ..., x_p$, then through all domains of couples of components (x_j, x_k), and so on.

3 Likelihood Models

In such models, the response function r , is a parameter of the conditional distribution of Y, given $\boldsymbol{X} = \boldsymbol{x}$. Let us consider a model $y = r(\boldsymbol{x};\boldsymbol{\alpha}) + \varepsilon$ for non-Gaussian noise, $r(\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{j=1}^{m} \alpha_j\, B_j(\boldsymbol{x})$. Denote by $f(y; r(\boldsymbol{x};\boldsymbol{\alpha}))$ the probability density of Y for an unknown parameter $\boldsymbol{\alpha}$. Then,

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln f(y_i; r(\boldsymbol{x}_i; \boldsymbol{\alpha})), \ \boldsymbol{\alpha} = (\alpha_1, \alpha_{2,...,}\alpha_m)$$

is the logarithm of likelihood function. The inference in such models is based on the maximum likelihood principle (i.e. to find an optimal vector $\boldsymbol{\alpha}$ we have to solve nonlinear equations $\partial\mathcal{L}/\partial\alpha_j = 0, \ j = 1, ..., m)$. In order to reduce the complexity of computations we propose incremental procedures for growing and pruning steps. Our numerical experiments showed that a several Newton–Raphson iterations suffice to solve one new parameter α_i of a nonlinear likelihood equation based on localized units.

3.1 Criteria of acceptance of a new model.

For Gaussian noise, the logarithm of likelihood is proportional to the averaged sum of squares of residuals, i.e. to the estimate of residual variance. As we have seen, this estimate, suitably penalized by model complexity, may serve as a criterion of acceptance. In the case of a likelihood model, we can employ the penalized log-likelihood in order to decide whether the update of a model is effective. We can also find an analogy of the t–tests for selecting the candidates for pruning. The maximum likelihood estimator has, under certain regularity conditions, favorable asymptotic properties, namely the estimator is consistent and asymptotically normally distributed. The variance of this distribution is given by the negative inverse of the matrix of the second derivatives of the log–likelihood. Therefore, we can compute approximately normal variables for the test of hypothesis and set coefficients α_j to zero correspondingly.

4 Applications

Using the methods proposed in this paper we have solved a number of examples, namely standard Gaussian nonparametric regression problems, nonparametric Cox's model of the hazard rate and logistic classification model. Numerical results of our experiments are encouraging. In this paper we present two case studies, a test problem and a real-world data problem. The first one deals with the approximation of a noisy sinusoidal function. A real-world problem deals with the analysis of the voltage and the current of Hubble Space Telescope nickel-hydrogen batteries.

4.1 Noisy Sinusoidal Example:

We have generated 200 points x_i distributed randomly uniformly over the interval $(0, 5)$. A function $y_i = x_i \sin(x_i^2) + e_i$ has been sampled (Fig.1). The components of Gaussian noise e_i were sampled independently from the normal distribution $N(\mu = 0, \sigma = 0.5)$. Our goal was to reconstruct, from the data (x_i, y_i), $i = 1, .., 200$, a function $r(x) = x \sin(x^2)$. Our algorithm employed Gausian radial basis functions, namely $B(x) = \frac{1}{\sqrt{2\pi}b} \exp(-\frac{(x-c)^2}{2b^2})$. The centers c were optimized by the BIC and Gamma penalization procedures, while the radii b were adapted automatically, namely set to the distance of the center of a candidate unit to the center of the nearest neighboring unit.

The procedure was initialized by 5 equidistantly located units. The candidates for the new units were generated randomly (i.e. following the method 1). After the growing phase, the model contained 15 units (Fig.5). The growing phase was stopped when 3 consequent global iterations did not change the model. We used the minimum of the BIC and Gamma (with $\gamma = 0.8$) criteria. Both criteria behave very similarly. The MSE of the model was $\hat{\sigma}^2 = 0.32666$.

Then we apply the pruning procedure, based on the t–test choice of units for deletion. This phase ended up with the final model of 10 units (Fig.3) The t–test criterion was compared with the penalizing criteria. There were no significant differences in performance between these three criteria. Final MSE was $\hat{\sigma}^2 = 0.37302$.

4.2 Hubble Space Telescope Nickel-Hydrogen Batteries Modelling

In a real–data case, we modeled the dependence of the-voltage (Fig.2) on the current (Fig.6), during repeated periods of charging and discharging. We considered the additive model of a three component function $V(t) = r^*(C, t, d) + e(t)$, where $r^*(C, t, d) = r_1(C(t)) + r_2(C(t - d)) + r_3(C(t - 2d))$, t denotes the time, d is the time lag constant. One dimensional functions r_j were modeled by RBF. The process of model growing started with 5 RBF units for each component, after 50 global iteration loops, the model had 10, 16, 18 units, respectively, the MSE was 0.0424. This phase was controlled by the gamma criterion, $\gamma = 0.7$. Then, the pruning phase, controlled by the t-tests, reduced the model to 6, 4, 10 units (Fig.4) and reached the final MSE = 0.0465.

5 Conclusion

We have experimented with a several incremental methods of building a nonparametric regression model and several growing and stopping criteria. The incremental method using localized units has some advantages over the batch modeling methods.

- It reduces computational complexity by allowing to solve a sequence of low dimensional optimization problems that are easier to handle than a high dimensional nonlinear problem.
- It allows us to implement heuristically based iteration that decrease RSS and avoid nonlinear optimization.
- It is a good starting point for solving optimal modeling problems with respect to other measures of deviation than RSS. For instance, for exponential families of models or logistic classification models rigorous estimation of linear parameters $\boldsymbol{\alpha}$ requires an iterative procedure. Therefore, each possible reduction of computations is desirable. For more discussion about this approach, see also Buja et al. in [3].
- Our studies indicated that the use of different growing/pruning strategies, namely the t-statistics criterion and penalization criteria provided approximately same results. The advantage of the t-test is in its theoretical justification.

References

[1] Bishop Ch.(1995): *Neural Networks for Pattern Recognition. Clarendon* Press, Oxford.

[2] Breiman L. (1993): Fitting additive models to data. *Computational Statistics and Data Analysis* 15, 13–46.

[3] Friedman J.H. (1991): Multivariate adaptive regression splines; with discussion. *Annals of Statistics*, vol. 19, pp. 1–141.

[4] Hastie T. and Tibshirani R. (1993): *Generalized Additive Models.* Chapmann and Hall, London.

[5] Stone C.J. (1994): The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics* 22, pp. 118–194.

[6] Jones, L. K.,(1992): A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, vol.20, pp. 608-613.

[7] Kurkova V. and Smid J.(1995): An incremental architecture algorithm for feedforward neural nets. *Preprints of the European IEEE Workshop Computer Intensive Methods in Control and Signal Processing*, Prague, 1994.

[8] Kurkova, V:Personal communication.

[9] Barron A.R. (1993): Universal approximation bounds for superposition of a sigmoidal function. *The IEEE Transactions on Information Theory*, vol. 39, 930-945, 1993.

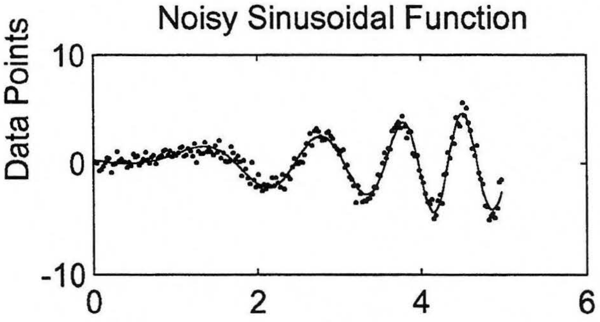

Fig.1

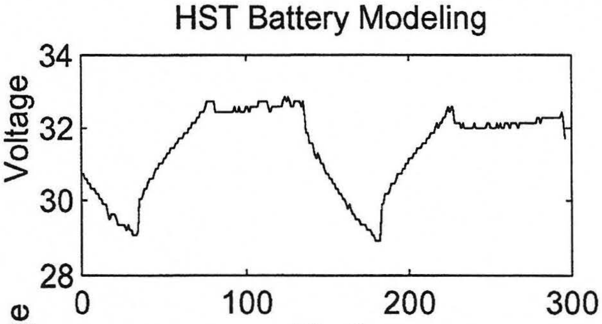

Fig.2

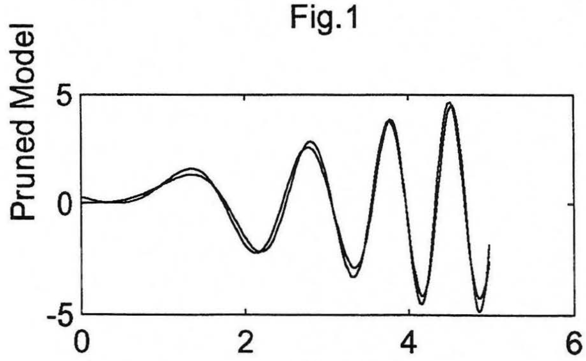

Fig.3

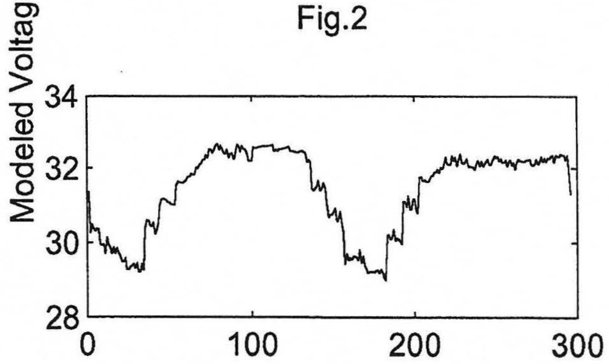

Fig.4

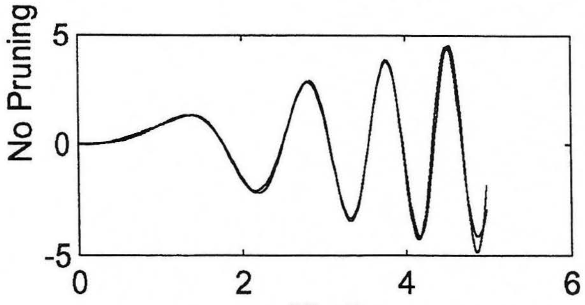

Fig.5

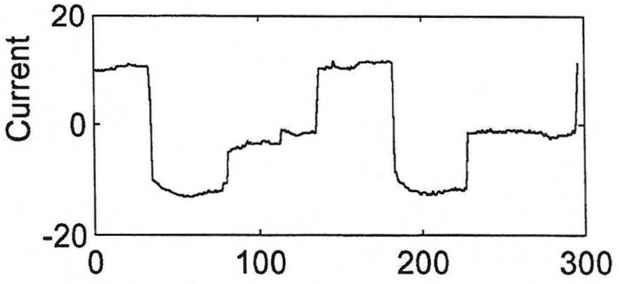

Fig.6