ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 525-530

Fast Bootstrap for Least-square Support Vector Machines

A. Lendasse[1], G. Simon[2], V. Wertz[3], M. Verleysen[2]

[1]Helsinki University of Technology
Laboratory of Computer and Information Science
Otaniemi 02014, Finland, lendasse@cis.hut.fi.
[2]Université catholique de Louvain, Electricity Dept., 3 pl. du Levant,
B-1348 Louvain-la-Neuve, Belgium, {simon, verleysen}@dice.ucl.ac.be.
[3]Université catholique de Louvain, CESAME, 4 av. G. Lemaître
B-1348 Louvain-la-Neuve, Belgium, wertz@auto.ucl.ac.be.

Abstract. The Bootstrap resampling method may be efficiently used to estimate the generalization error of nonlinear regression models, as artificial neural networks and especially Least-square Support Vector Machines. Nevertheless, the use of the Bootstrap implies a high computational load. In this paper we present a simple procedure to obtain a fast approximation of this generalization error with a reduced computation time. This proposal is based on empirical evidence and included in a simulation procedure.

1. Introduction

Model design has raised a considerable research effort since decades, on linear models, nonlinear ones, artificial neural networks, and many others. Model design includes the necessity to compare models (for example of different complexities) in order to select the “best” model among several ones. For this purpose, it is necessary to obtain a good approximation of the generalization error of each model (the generalization error being the average error that the model would make on an infinite-size and unknown test set independent from the learning one). Nowadays there exist some well-known and widely used methods able to fulfill this task [1-6] and the Bootstrap is one of the more performing methods. The main problem when using the Bootstrap is the computation of the results that is really time consuming. In this paper we will show that, under reasonable and simple hypotheses usually fulfilled in real world applications, it is possible to provide a good estimate of the Bootstrap results with a considerably reduced number of modeling stages, thus saving a considerable amount of computation time. Previously, the Fast Bootstrap has been developed for the Radial Basis Functions Networks (RBFN). In this paper, the Fast Bootstrap is extended to Least-square Support Vector Machines (LS-SVM) [7-9]. The LS-SVM is presented in section 2. Then, the Bootstrap is presented in section 3 and finally, the Fast Bootstrap is introduced in section 4 using a toy example.

2. Least-Square Support Vector Machines (LS-SVM)

Consider a given training set of N data points $\{x_k, y_k\}$ with x_k a n-dimensional input and y_t a 1-dimensional output. In feature space SVM models take the form:

$$y(x) = \omega^T \varphi(x) + b \, , \tag{1}$$

where the nonlinear mapping φ(.) maps the input data into a higher dimensional feature space. In least squares support vector machines for function estimation, the following optimization problem is formulated:

$$\min_{\omega, e} J(\omega, e) = \frac{1}{2}\omega^T \omega + \gamma \frac{1}{2}\sum_{k=1}^{N} e_k^2 \, , \tag{2}$$

subject to the equality constraints

$$y(x) = \omega^T \varphi(x) + b + e_k \, , \; k = 1, ..., N. \, , \tag{3}$$

This corresponds to a form of ridge regression. The Lagrangian is given by

$$L(\omega, b, e; \alpha) = J(\omega, e) - \sum_{k=1}^{N} \alpha_k \left\{ \omega^T \varphi(x_k) + b + e_k - y_k \right\} , \tag{4}$$

with Lagrange multipliers α_k. The conditions for optimality are

$$\begin{array}{ll} \frac{\partial L}{\partial \omega} L = 0 \rightarrow \omega = \sum_{k=1}^{N} \alpha_k \varphi(x_k) & \frac{\partial L}{\partial b} L = 0 \rightarrow \sum_{k=1}^{N} \alpha_k = 0 \\ \frac{\partial L}{\partial e_k} L = 0 \rightarrow \alpha_k = \gamma e_k & \frac{\partial L}{\partial \alpha_k} L = 0 \rightarrow \omega^T \varphi(x_k) + b + e_k - y_k \end{array} \, , \tag{5}$$

for k = 1..N. After elimination of e_k and ω, the solution is given by the following set of linear equations

$$\begin{bmatrix} 0 & \vec{1}^T \\ \vec{1} & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} , \tag{6}$$

where $y = [y_1; \ldots; y_N]$, $\vec{1} = [1; \ldots; 1]$, $\alpha = [\alpha_1; ...;\alpha_N]$ and the Mercer condition

$$\Omega_{kl} = \varphi(x_k)^T \varphi(x_l) = \psi(x_k, x_l) \quad k, l = 1, ..., N \, , \tag{7}$$

has been applied. This finally results into the following LS-SVM model for function estimation

$$y(x) = \omega^T \varphi(x) + b \, , \tag{8}$$

where α and b are the solution to (6). For the choice of the kernel function ψ(.,.) one has several possibilities [7-9]. In this paper, Gaussian kernels are used: $\psi(x, x_k) = \exp\{-\|x-x_k\|^2/\sigma^2\}$ and the remaining unknowns are σ and γ. These model hyperparameters will be selected according to a model selection procedure detailed in the following of this paper.

3. Bootstrap for Model Structure Selection

The bootstrap [4] is a resampling method that has been developed in order to estimate some statistical parameters (like the mean, the variance, etc). In the case of model structure selection, the parameter to be estimated is the generalization error (i.e. the average error that the model would make on an infinite-size and unknown test set). When using the bootstrap, this error is not computed directly. Rather the bootstrap estimates the difference between the generalization error and the training error calculated on the initial data set. This difference is called the optimism. The estimated generalization error will thus be the sum of the training error and of the estimated optimism. The training error is computed using all data from the training set. The optimism is estimated using a resampling technique based on drawing within the training set with replacement. Using notation where the first exponent A_j denotes the training set while the second exponent A_j indicates the set used to estimate the model error, the Bootstrap method can be decomposed in the following stages:

1. From the initial set *I*, one randomly draws *N* points with replacement. The new set A_j has thus the same size that the initial set and constitutes a new training set. This stage is called the resampling.
2. The training of the various model structures *q* is done on the same training set A_j. One can compute the training error on this single set:

$$E_j^{A_j,A_j}(q,\theta_j^*(q)) = \frac{\sum_{i=1}^{N}\left(h^q(x_i^{A_j},\theta_j^*(q)) - y_i^{A_j}\right)^2}{N}, \tag{9}$$

with θ_j^* the model parameters after learning, h^q the q^{th} model that is used, x_i^{Aj} the i^{th} input vector from set A_j, y_i^{Aj} the i^{th} output and *N* the number of elements in this set. Index *j* means that the error is evaluated on the j^{th} new sample.

3. One can also compute the validation error on the initial sample which now plays the role of the validation set *V*=*I*:

$$E_j^{A_j,V}(q,\theta_j^*(q)) = \frac{\sum_{i=1}^{N}\left(h^q(x_i^{V},\theta_j^*(q)) - y_i^{V}\right)^2}{N}. \tag{10}$$

Here again index *j* means that the error is evaluated on the j^{th} new sample.

4. The difference between these two errors (9) and (10) is calculated and defined as the *optimism* by Efron [6]:

$$optmism_j(q,\theta_j^*(q)) = E_j^{A_j,V}(q,\theta_j^*(q)) - E_j^{A_j,A_j}(q,\theta_j^*(q)) \ . \qquad (11)$$

5. Steps 1 to 4 are repeated *J* times. The estimate of the optimism is then calculated as the average of the *J* values from (11):

$$opt\hat{m}ism(q) = \frac{\sum_{j=1}^{J} optmism_j(q,\theta_j^*(q))}{J} \ . \qquad (12)$$

6. The training of the *q* model structures is done on the initial data set *I* and the training error is calculated on the same set. Two exponents *I* are used to indicate that the initial data set is used for both training and error estimation:

$$E^{I,I}(q,\theta^*(q)) = \frac{\sum_{i=1}^{N}\left(h^q(x_i^I,\theta^*(q)) - y_i^I\right)^2}{N} \ . \qquad (13)$$

7. An approximation of the generalization error is finally obtained by:

$$\hat{E}_{gen}(q) = opti\hat{m}ism(q) + E^{I,I}(q,\theta^*) \ . \qquad (14)$$

$\hat{E}_{gen}(q)$ is an approximation of the generalization error for each model structure *q*. The best structure that will be selected is the one that minimizes this estimate of the generalization error.

4. Fast Bootstrap and Toy Example

In this section, an improvement of the Bootstrap methods is presented. This method is called Fast Bootstrap and allows reducing the computational time of the traditional Bootstraps [10-11]. This method is based on experimental observations and is presented on a function approximation example (represented in Figure 1). In this example, 200 inputs *x* has been drawn using a uniform random law between 0 and 1. The output *y* has been generated by the function:

$$y = \sin(5x) + \sin(15x) + \sin(25x) + \varepsilon \ , \qquad (9)$$

with ε a uniform random law between [-0.5 -0.5].

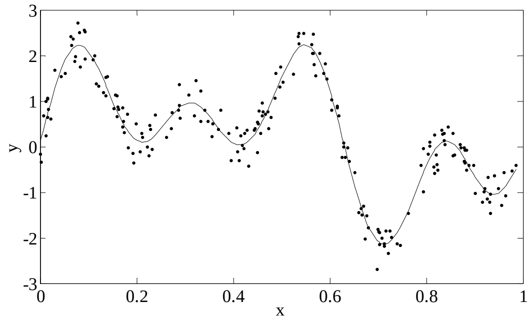

Figure 1: Example of function (dots) and its approximation (solid line).

A LS-SVM is used to approximate this function. Two parameters still have to be determined, namely σ and γ. For a fixed $\sigma = 0.1$, the optimal γ is determined using the Bootstrap method. The set of γ that is tested ranges from 0 to 100 with a 0.1 step. The number of resamplings in (12) is equal to 100.

The apparent error defined in (13) is computed and represented in Fig.2A. The optimism is computed using (12) and represented in Fig.2B. The generalization error is computed using (15) and represented in Fig.3. The value of γ that minimizes the generalization error is equal to 11.

In Fig.2B, the optimism is very close from an exponential function of γ. This fact has been observed on other examples and benchmarks. Then, using this information, the number of values of γ to be tested can be considerably reduced. In this example, this set is indeed reduced to 5 to 100 with an incremental step of 5. An exponential approximation of the optimism is used. Thanks to the approximation, the number of Bootstraps is also reduced by a factor 10 in (12). The new optimism and generalization error are represented as dotted lines in Fig.2B and Fig.3 respectively. The optimum is close to the one that has been selected by the Bootstrap method. This new method, denoted Fast Bootstrap, is in this toy example 500 times quicker than the traditional Bootstrap. In other examples, the Fast Bootstrap is at least 100 times quicker than traditional Bootstrap for the selection of the γ parameter for a LS-SVM, without loss of precision.

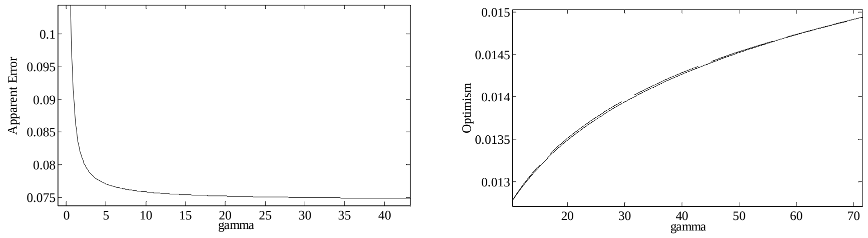

Figure 2A: Apparent Error with respect to γ. Figure 2B: Optimism with respect to γ using Bootstrap (solid line) and Fast Bootstrap (dashed line).

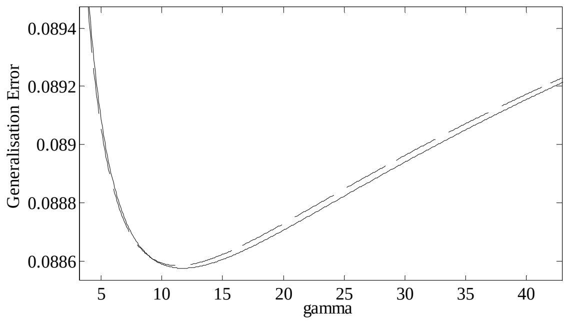

Figure 3: Generalization Error with respect to γ using Bootstrap (solid line) and Fast Bootstrap (dashed line).

5. Conclusions

In this paper we have shown that the optimism term of the Bootstrap estimator of the prediction error is approximatively an exponential with respect to the γ parameter of a LS-SVM. According to the results shown here and to other ones not illustrated in this paper, we recommend a conservative value of 10 for the number of Bootstrap replications before stopping the approximation computation. We would like to emphasize on the fact that the very limited loss of accuracy is balanced by a considerable saving in computation load, this last fact being the main disadvantage of the Bootstrap resampling procedure in practical situations. This saving is due to the reduced number of tested models and to the limited number of Bootstrap replications.

Acknowledgements

Michel Verleysen is Senior Research Associate of the Belgian National Fund for Scientific Research (FNRS). G. Simon is funded by the Belgian F.R.I.A. Part the work of V. Wertz is supported by the Interuniversity Attraction Poles (IAP), initiated by the Belgian Federal State, Ministry of Sciences, Technologies and Culture. The scientific responsibility rests with the authors.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, 2nd Int. Symp. on information Theory, 267-81, Budapest, 1973.

[2] G. Schwarz, “Estimating the dimension of a model”, Ann. Stat. 6, 461-464, 1978.

[3] L. Ljung, System Identification - Theory for the user, 2nd Ed, Prentice Hall, 1999.

[4] B. Efron, R. J. Tibshirani, An introduction to the Bootstrap, Chapman & Hall, 1993.

[5] M. Stone, An asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion, J. Royal. Statist. Soc., B39, 44-7, 1977.

[6] R. Kohavi, A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Proc. of the 14th Int. Joint Conf. on A.I., Vol. 2, Canada, 1995.

[7] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002.

[8] J.A.K. Suykens, J. Vandewalle, “Least squares support vector machine classifiers”, Neural Processing Letters, vol. 9, no. 3, Jun. 1999, pp. 293-300.98-86.

[9] J.A.K. Suykens, J. De Brabanter, L. Lukas, J. Vandewalle, “Weighted least squares support vector machines: robustness and sparse approximation'”, Neurocomputing, vol. 48, no. 1-4, Oct. 2002, pp. 85-105.

[10] G. Simon, A. Lendasse, V. Wertz, M. Verleysen, Fast Approximation of the Bootstrap for Model Selection, ESANN 2003, European Symposium on Artificial Neural Networks, Bruges (Belgium), 23-25 April 2003, pp. 99-106.

[11] G. Simon, A. Lendasse, M. Verleysen, Bootstrap for Model Selection: Linear Approximation of the Optimism, IWANN 2003, International Work-Conference on Artificial and Natural Neural Networks, Mao, Menorca (Spain), June 3-6, 2003. Computational Methods in Neural Modeling, J. Mira, J.R. Alvarez eds, Springer-Verlag, Lecture Notes in Computer Science 2686, 2003, pp. I182-I189.