# LS-SVM Hyperparameter Selection with a Nonparametric Noise Estimator

Amaury Lendasse[1], Yongnan Ji[1], Nima Reyhani[1], and Michel Verleysen[2]

[1] Neural Network Research Centre,
Helsinki University of Technology, P.O. Box 5400,
02150 Espoo, Finland
{lendasse, yji, nreyhani}@.hut.fi
[2] Machine Learning Group,
Université catholique de Louvain, DICE, 3 place du Levant, 1348
Louvain-la-Neuve, Belgique
verleysen@dice.ucl.ac.be

**Abstract.** This paper presents a new method for the selection of the two hyperparameters of Least Squares Support Vector Machine (LS-SVM) approximators with Gaussian Kernels. The two hyperparameters are the width $\sigma$ of the Gaussian kernels and the regularization parameter $\lambda$. For different values of $\sigma$, a Nonparametric Noise Estimator (NNE) is introduced to estimate the variance of the noise on the outputs. The NNE allows the determination of the best $\lambda$ for each given $\sigma$. A Leave-one-out methodology is then applied to select the best $\sigma$. Therefore, this method transforms the double optimization problem into a single optimization one. The method is tested on 2 problems: a toy example and the Pumadyn regression Benchmark.

**Keywords:** Least Squares Support Vector Machines, Leave-one-out, Noise Estimation, Regression.

## 1 Introduction

The selection of hyperparameters is a important issue in the fields of Artificial Neural Networks, Machine Learning and System Identification. Many resampling techniques have been successfully used as Leave-One-Out (LOO), Bootstrap and Cross-Validation [1, 2].

Least Squares Support Vector Machines with Gaussian kernels are efficient regression models [3]. For example, they do not suffer from the problem of local minima. Unfortunately, two hyperparameters have to be tuned, for example using LOO [4]. The two hyperparameters are the width $\sigma$ of the Gaussian kernels and the regularization parameter $\lambda$. This problem leads to a grid search that is highly time consuming. In this paper, we propose the use of Nonparametric Noise Estimator (NNE) in order to select the regularization parameter as a function of the width $\sigma$.

The paper is organised as follows: LS-SVM are introduced in Section 2, NNE in Section 3 and the methodology in Section 4. In Section 5, the method is successfully tested on 2 problems: a toy example and the Pumadyn regression Benchmark.

W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 625 – 630, 2005.

## 2 Least Squares Support Vector Machines

LS-SVM are regularized supervised approximators, which has been proved to be efficient for function approximation. Only solving linear equation is needed in the optimization process, which not only simplifies the process, but also avoids the problem of local minima in SVM. In this section, a short summary of the LS-SVM model is given. The LS-SVM model [4, 5] is defined in its primal weight space by,

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

where $\varphi(x)$ is a function which maps the input space into a higher dimensional feature space, $x$ is the $M$-dimensional vector of inputs $x_j$, and $\omega$ and $b$ the parameters of the model. Given N input-output learning pairs $(x^i, y^i) \in R^M \times R$, Least Squares Support Vector Machines for function estimation formulate the following optimization:

$$\min_{\omega,b,e} J(\omega, e) = \frac{1}{2}\omega^T\omega + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \quad \text{subject to} \quad y^i = \omega^T \varphi(x^i) + b + e^i, i = 1, \ldots, N \tag{2}$$

The parameter set $\theta$ consists of vector $\omega$ and scalar $b$. Solving this optimization problem in dual space leads to finding the $\alpha_i$ and $b$ coefficients in the following solution:

$$h(x) = \sum_{i=1}^{N} \alpha_i K(x, x^i) + b \tag{3}$$

Function $K(\mathbf{x}, \mathbf{x}^i)$ is the kernel defined as the dot product between the $\varphi(x)^T$ and $\varphi(x)$ mappings. The meta-parameters of the LS-SVM model are the width of the Gaussian kernels (taken identical for all kernels) and the $\gamma$ regularization factor. The training method for the estimation of $\omega$ and $b$ can be found in [4].

## 3 Nonlinear Noise Estimator

The problem of function approximation consists in the determination of the relationship between a set $x$ of inputs and one single output $y$. Given $N$ inputs-output pairs $(x^i, y^i) \in R^M \times R$, the relationship between $x_i$ and $y_i$ can be expressed as $y_i = f(x_i) + \varepsilon_I$,.where $f$ is the unknown relationship and $\varepsilon_i$ the noise. Any estimation of model $f$ based on a finite number $N$ of learning data goes through a compromise between a low learning error (small bias) and a smooth model (small variance). In the case of LS-SVM, this compromise is implemented through the choice of an adequate value of $\gamma$. If the value of $\gamma$ is set too large, the model will overfit the data, including the noise. Still, the value of $\gamma$ should be set as large as possible; a too small value of $\gamma$ would simply mean that the model does not fit the learning data! It is therefore suggested to select the largest value of $\gamma$ so that the learning error does goes below the level of noise. Indeed it is unreasonable to expect that a model could lead to an error that is lower than the level of noise; if it was the case, the model would be in overfitting region.

Selecting $\gamma$ then means first to estimate the learning error of the model in function of $\gamma$, and secondly to estimate the variance of the noise. Of course, the noise estimator should not use the model itself, but only the data at disposal; it should be nonparametric.

An approach called "Delta Test" has been proposed for estimating the variance of the noise on the output [6]. It is based on the similarity of the noise behaviour between two closed data points. As the distance $\delta$ between two close points $x$ and $x'$ goes to zero, the average MSE between the corresponding outputs tends to var($\varepsilon$) [7]:

$$E\left\langle \frac{1}{2}(y'-y)^2 \,\middle|\, |x'-x| < \delta \right\rangle \to \mathrm{var}(\varepsilon) \quad as \quad \delta \to 0 \tag{6}$$

Despite this approach seems to be promising for noise estimation purposes, it fails when the size of the data set is small with respect to the complexity of underlying function and noise distribution. Jones *et al*. [6] improved the Delta test using the *k*-nearest neighbour distances between data in the input space and corresponding data in the output space. This leads to an approach called here Nonparametric Noise Estimator (NNE). Referring to [6], the estimate of noise variance is the intercept of the linear regression line which is drawn between the average of the *k* nearest distances in the inputs space and the corresponding average of the *k* nearest distances in the output space (see equation 7 below). A proof of NNE (which is also called Gamma Test in some papers) can be found in [7] and is based on a generalization of Chybechov inequality and the property of *k*-nearest neighbor structures. Moreover, it has been shown that NNE is useful too for evaluating the nonlinear correlation between two random variables, or input and output pairs realizations. In the proof, the following conditions are necessary:

- the first and second partial derivatives of the underlying function exist;
- the first to the fourth moments of the noise distribution exist;
- the noise is independent from the input.

Using this three conditions, the variance of noise is given by the intercept with the vertical line $\delta(k)=0$, of the regression line between $\gamma(k)$ and $\delta(k)$, where $1 \leq k \leq p$ and

$$\delta(k) = \frac{1}{N}\sum_{i=1}^{N} \left| x_{NN(x_i,k)} - x_i \right|^2 \quad \text{and} \quad \gamma(k) = \frac{1}{2N}\sum_{i=1}^{N} \left| y_{NN(x_i,k)} - y_i \right|^2 . \tag{7}$$

In (7), $NN(x_i,k)$ is the index of the $k^{\text{th}}$ neighbour of $x_i$. According to [6], $p$=10 is used in experiments presented in section 5.

This noise variance estimator based on [6] is similar to the variogram based estimator detailed in [8]. However, it differs from the fact that Jones' estimator only uses the *k* nearest neighbours of the data points. This reduces the computation time and makes the estimator efficient when the number of data points is large enough by concentrating on small values of $\delta(k)$.

## 4 Methodology

The goal of the presented methodology is to transform the double optimization of $\gamma$ and $\sigma$ in LS-SVM into a simple optimization procedure. The double optimization of the metaparameters using LOO presented in [3, 4] is very efficient but is highly time consuming.

Our methodology can be expressed as the following:

1) A range of σ is selected.
2) For each σ, the Nonparametric Noise Estimate is performed.
3) A bisection method is used to estimate the value of γ such that the training error of the LS-SVM is equal to the value of the Nonparametric Noise Estimate. The training error is strictly decreasing with respect to γ and then the solution is unique and its computation is very fast. Taking the largest γ value such that the training error does not exceed the noise variance leads to the more accurate mode without overfitting.
4) The LOO error (LOO MSE) is estimated for each value of σ.
5) The value of σ and corresponding γ minimizing the LOO error are selected.

## 5 Experiment

### 5.1 Toy Example

A toy example with 1000 samples is build using the following function:

$$y = \sin(x) + \sin(5x) + \sin(15x) + \varepsilon \tag{8}$$

with ε an uniform noise in [-0.5, 0.5]. The function is represented in Fig.2 The real value of the variance of the noise is 0.0822 and the estimate obtained with the NNE is also 0.0822. The methodology presented in section 4 is applied. The range of σ is between 0.01 and 0.4 by step of 0.005. For each value of σ, γ is calculated using the estimate of the NNE (see Fig. 1. a).

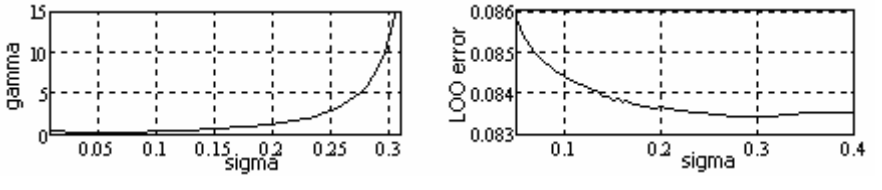


**Fig. 1.** Toy example results. **a -** γ with respect to σ. **b** – LOO error with respect to σ.

For each value of σ (using the corresponding γ), the LOO error is computed (see Fig. 1. b). The optimum is obtained for σ = 0.295 and the corresponding γ = 9.727.

The approximation obtained the selected LS-SVM is represented in Fig. 2.

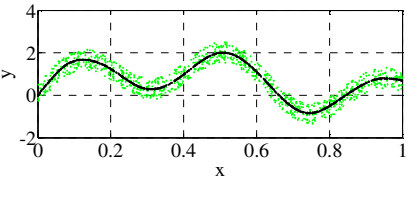


**Fig. 2.** The toy example and the approximation after the selection of the hyperparameters

### 5.2 Pumadyn Benchmark

The pumadyn datasets [9] are a family of datasets synthetically generated from a realistic simulation of the dynamics of a Puma robot arm. The tasks associated with these datasets consist of predicting the angular acceleration of one of the links of the robot arm given the angular positions, velocities, torques, and in some cases, other dynamic parameters of the robot arm. The dataset contains 8192 samples, 8 inputs and one output. The methodology presented in section 4 is applied. The range of $\sigma$ is between 5 and 110 by step of 5. For each value of $\sigma$, $\gamma$ is calculated using the estimate of the NNE (see Fig. 3. a).

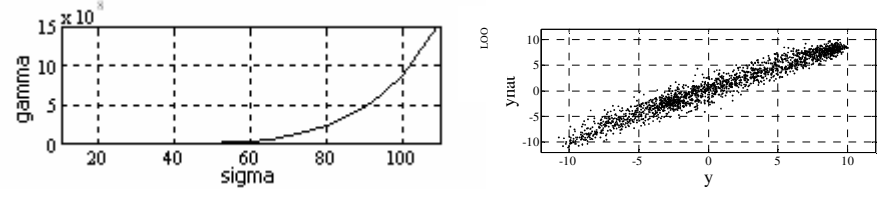


**Fig. 3. a** $\gamma$ with respect to $\sigma$. **b** – LOO error with respect to $\sigma$.

For each value of $\sigma$ (using the corresponding $\gamma$), the LOO error is computed; a smooth slope similar to the one in Fig. 1. b. is obtained. Its minimum is found for $\sigma$ = 95 and the corresponding $\gamma$ = 6.4749e+008. The approximation with respect to the target value *y* is represented in see Fig. 3. b. The LOO error that is obtained is 1.81.

## 6 Conclusion and Further Work

In this paper, a Nonparametric Noise Estimator has been introduced for the selection of the hyperparameters of LS-SVM. The proposed methodology transforms the double optimization problem of the selection of the hyperparameters into a single optimization one, therefore reducing drastically the computation time for similar results.

The method has been illustrated on two examples and gives accurate approximations. Further work includes the test of other methods for Nonparametric Noise Estimation (see for example [8]) and their embedding into the same methodology to select hyperparameters in LS-SVM and other learning schemes.

## Acknowledgements

LS-SVMlab [3, 4] has been used for the optimization of the LS-SVM models and to perform the Leave-one-Out procedures. Part of work of A. Lendasse, Y.N. Ji and N. Reyhani is supported by the project of New Information Processing Principles, 44886, of the Academy of Finland. M. Verleysen is a Senior Research Associate of the Belgian F.N.R.S. (National Fund For Scientific Research).

## References

1. Bishop, C. M., *Neural Networks for Pattern Recognition*. New York: Oxford, 1995.
2. Lendasse A., Wertz V., Verleysen M.: Model selection with cross-validations and bootstraps – Application to time series prediction with RBFN models. In: Artificial Neural Networks and Neural Information Processing – ICANN/ICONIP (2003), Kaynak O., Alpaydin E., Oja E., Xu L. (eds): Springer-Verlag Lecture Notes in Computer Science 2714, Berlin (2003) 573-580.
3. Suykens, J., A., Van Gestel, K., T., De Brabanter, J., De.Moor, B., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore, ISBN 981-238-151-1 (2002).
4. http://www.esat.kuleuven.ac.be/sista/lssvmlab/
5. Suykens, J., A., De brabanter, K., J., Lukas, L., Vandewalle, J.: Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing, Special Issue on fundamental and information processing aspects of neurocomputing.
6. Jones, A. J. New Tools in Non-linear Modeling and Prediction. Computational Management Science, Vol. 1, Issue 2, p.p. 109-149, 2004.
7. Evans, D. and Jones, A. J., A proof of the Gamma test, Proc. Roy. Soc. Lond. A, Vol. 458, pp. 1-41, 2002.
8. Pelckmans K., De Brabanter J., Suykens J.A.K., De Moor B., Variogram based noise variance estimation and its use in Kernel Based Regression, in *Proc. of the IEEE Workshop on Neural Networks for Signal Processing*, pp. 199-208.
9. Corke, P. I. (1996). A Robotics Toolbox for MATLAB. *IEEE Robotics and Automation Magazine,* **3** (1): 24-32.