Sparse LS-SVMs using Additive Regularization with a Penalized Validation Criterion

K. Pelckmans, J.A.K. Suykens, B. De Moor

K.U. Leuven, ESAT-SCD-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium
Email: {kristiaan.pelckmans,johan.suykens}@esat.kuleuven.ac.be

Abstract. This paper is based on a new way for determining the regularization trade-off in least squares support vector machines (LS-SVMs) via a mechanism of additive regularization which has been recently introduced in [6]. This framework enables computational fusion of training and validation levels and allows to train the model together with finding the regularization constants by solving a single linear system at once. In this paper we show that this framework allows to consider a penalized validation criterion that leads to sparse LS-SVMs. The model, regularization constants and sparseness follow from a convex quadratic program in this case.

1 Introduction

Regularization has a rich history which dates back to the theory of inverse ill-posed and ill-conditioned problems [12]. Regularized cost functions have been considered e.g. in splines, multilayer perceptrons, regularization networks [7], support vector machines (SVM) and related methods (see e.g. [5]). SVM [13] is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation which has also led to many other recent developments in kernel based learning methods in general [3, 8]. SVMs have been introduced within the context of statistical learning theory and structural risk minimization. In the methods one solves convex optimization problems, typically quadratic programs. Least Squares Support Vector Machines (LS-SVMs) [9, 10] are reformulations to standard SVMs which lead to solving linear KKT systems for classification tasks as well as regression and primal-dual LS-SVM formulations have been given for kFDA, kPCA, kCCA, kPLS, recurrent networks and control [10].

The relative importance between the smoothness of the solution and the norm of the residuals in the cost function involves a tuning parameter, usually called the regularization constant. If the relative importance is given by a single parameter in the classical sense, we refer to this trade-off scheme as a Tikhonov regularization scheme [12]. The determination of regularization constants is important in order to achieve good generalization performance with the trained model and is an important problem in statistics and learning theory [5, 8, 11]. Several model selection criteria have been proposed in literature to tune the model to the data. In this paper, the performance on an independent validation dataset is considered. The optimization of the regularization constant in LS-SVMs with respect to this criterion proves to be non-convex in general. In order to overcome this difficulty, a reparameterization of the regularization trade-off has been recently introduced in [6] referred to as *additive regularization* (AReg). The combination of model training equations of the AReg LS-SVM and the validation minimization leads to one convex system of linear equations from which the model parameters and the regularization constants follow at once. In order to explicitly restrict the degrees of freedom of the additive regularization constants (to avoid data snooping), a penalizing term is introduced here at the validation level leading to sparse solutions of AReg LS-SVMs with parameter tuning by solving a convex quadratic program.

This paper is organized as follows. In Section 2 the formulation of LS-SVMs and additive regularization are briefly reviewed. Section 3 discusses a criterion for tuning of the regularization constants leading to a sparse solution. In Section 4 a number of experiments on regression as well as classification tasks are given.

2 Model Training

Let $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ be the training data with inputs x_i and outputs y_i. Consider the regression model $y_i = f(x_i) + e_i$ where $x_1, \ldots, x_N$ are deterministic points (fixed design), $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown real-valued smooth function and $e_1, \ldots, e_N$ are uncorrelated random errors with $E[e_i] = 0$, $E\left[e_i^2\right] = \sigma_e^2 < \infty$. The n data points of the validation set are denoted as $\{x_j^v, y_j^v\}_{j=1}^n$. In the case of classification, $y, y^v \in \{-1, 1\}$.

2.1 Least Squares Support Vector Machines

The LS-SVM model is given as $f(x) = w^T\varphi(x) + b$ in the primal space where $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ denotes the potentially infinite ($n_h = \infty$) dimensional feature map. The regularized least squares cost function is given by [10]

$$\min_{w,b,e_i} \mathcal{J}_\gamma(w, e) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^N e_i^2 \text{ s.t. } w^T\varphi(x_i) + b + e_i = y_i, \quad \forall i = 1, ..., N \tag{1}$$

Note that the regularization constant γ appears here as in classical Tikhonov regularization [12]. The Lagrangian of the constraint optimization problem becomes $\mathcal{L}_\gamma(w, b, e_i; \alpha_i) = 0.5w^T w + 0.5\gamma\sum_{i=1}^N e_i^2 - \sum_{i=1}^N \alpha_i(w^T x_i + b + e_i - y_i)$. By taking the conditions for optimality $\partial\mathcal{L}_\gamma/\partial\alpha_i = \partial\mathcal{L}_\gamma/\partial b = \partial\mathcal{L}_\gamma/\partial e_i = 0$ and application of the kernel trick $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ with a positive definite (Mercer) kernel K, one gets $e_i\gamma = \alpha_i$, $w = \sum_{i=1}^N \alpha_i\varphi(x_i)$, $\sum_{i=1}^N \alpha_i = 0$ and $w^T\varphi(x_i) + b + e_i = y_i$. The dual problem is given by

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + I_N/\gamma \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \tag{2}$$

where $\Omega \in \mathbb{R}^{N\times N}$ with $\Omega_{ij} = K(x_i, x_j)$. The estimated function $\hat{f}$ can be evaluated at a new point x^* by $\hat{f}(x^*) = \sum_{i=1}^N \alpha_i K(x_i, x^*) + b$.

Optimization of the optimal γ with respect to the validation performance in the regression case can be written as

$$\min_\gamma \sum_{j=1}^n (y_j^v - \hat{f}_\gamma(x_j^v))^2 = \sum_{j=1}^n \left(y_j^v - \left[\frac{1}{\Omega^v}\right]^T \left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + I_N/\gamma \end{array}\right]^{-1} \left[\frac{0}{y}\right]\right)^2 \tag{3}$$

where $\Omega^v \in \mathbb{R}^{n\times N}$ with $\Omega_{ij}^v = K(x_i, x_j^v)$. The determination of γ becomes a non-convex optimization problem which is often also non-smooth such as in the case of cross-validation methods. For the choice of the kernel $K(\cdot, \cdot)$, see e.g. [2, 8, 3]. Typical examples are the use of a linear kernel $K(x_i, x_j) = x_i^T x_j$ or the RBF kernel $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$ where σ denotes the bandwidth of the kernel.

A derivation of LS-SVMs was given originally for the classification task [9]. The LS-SVM classifier $f(x) = \text{sign}(\varphi(x)^T w + b)$ is optimized with respect to

$$\min_{w,b,e_i} \mathcal{J}_\gamma(w, e) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^N e_i^2 \text{ s.t. } y_i(w^T\varphi(x_i) + b) = 1 - e_i, \quad \forall i = 1, \ldots, N. \tag{4}$$

Using a primal dual optimization interpretation, the unknowns α, b of the estimated classifier $\hat{f}(x) = \text{sign}(\sum_{i=1}^N \alpha_i y_i K(x_i, x) + b)$ are found by solving the dual set of linear equations

$$\left[\begin{array}{c|c} 0 & y^T \\ \hline y & \Omega^y + I_N/\gamma \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 1_N \end{array}\right] \tag{5}$$

where $\Omega^y \in \mathbb{R}^{N\times N}$ with $\Omega_{ij}^y = y_i y_j K(x_i, x_j)$. The remainder focuses on the regression case, although it is applicable just as well to the classification problem as illustrated in the experiments [6].

2.2 LS-SVMs with additive regularization

An alternative way to parameterize the regularization trade-off associated with the model $f(x) = w^T\varphi(x) + b$ is by means of the vector c [6]:

$$\min_{w,b,e_i} \mathcal{J}_c(w,e) = \frac{1}{2}w^Tw + \frac{1}{2}\sum_{i=1}^{N}(e_i - c_i)^2 \text{ s.t. } w^T\varphi(x_i) + b + e_i = y_i \quad \forall i = 1,\dots N \tag{6}$$

where the elements of the vector c serve as tuning parameters, called the additive regularization constants. After constructing the Lagrangian with multipliers α and taking the conditions for optimality w.r.t. w, b, e_i, α_i (being $e_i = c_i + \alpha_i$, $w = \sum_{i=1}^{N}\alpha_i\varphi(x_i)$, $\sum_{i=1}^{N}\alpha_i = 0$ and $w^T\varphi(x_i) + b + e_i = y_i$), the following dual linear system is obtained

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + I_N \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] + \left[\begin{array}{c} 0 \\ \hline c \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right]. \tag{7}$$

Note that at this point the value of c is not considered as an unknown to the optimization problem: once c is fixed, the solution of α, b is uniquely determined. The estimated function $\hat{f}$ can be evaluated at a new point x^* by $\hat{f}(x^*) = w^T\varphi(x^*) + b = \sum_{i=1}^{N}\alpha_i K(x_i, x^*) + b$. The residual $\hat{f}(x_j^v) - y_j^v$ is denoted by e_j^v such that one can write

$$y_j^v = w^T\varphi(x_j^v) + b + e_j^v = \sum_{i=1}^{N}\alpha_i K(x_i, x_j^v) + b + e_j^v. \tag{8}$$

We refer to this model as AReg LS-SVM where AReg stands for additive regularization.

By comparison of (5) and (7), LS-SVMs with Tikhonov regularization can be seen as a special case of AReg LS-SVMs with the following additional constraint on α, c, γ

$$\gamma^{-1}\alpha = \alpha + c \quad \text{s.t.} \quad 0 \leq \gamma^{-1}. \tag{9}$$

This means that solution to AReg LS-SVMs are also solutions to LS-SVMs whenever the support values α are proportional to the residuals $e = \alpha + c$.

3 Regularization determination for AReg LS-SVM

In this Section, the task of how to select an appropriate $c \in \mathbb{R}^N$ for the dataset at hand is discussed (tuning parameters determination for the AReg LS-SVM).

3.1 Fusion of additive regularization and validation

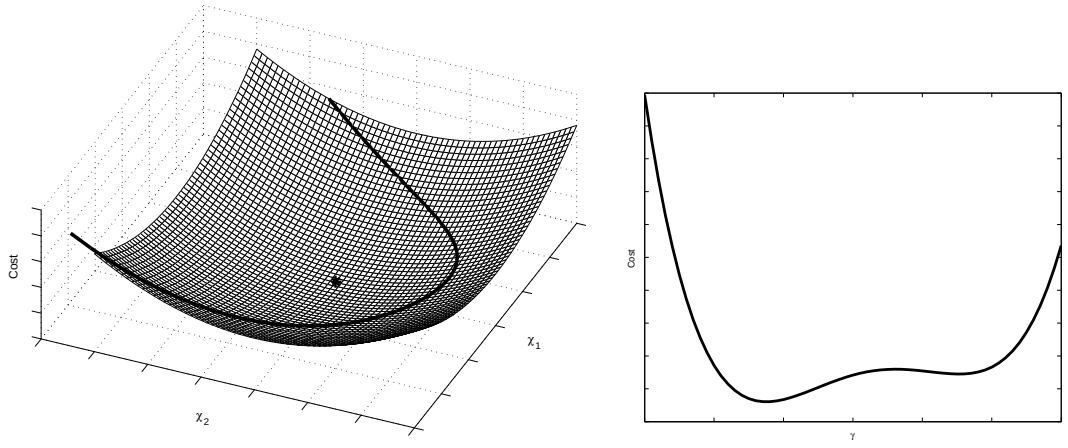
By combination of the training conditions (7) and validation equalities (8), a set of equations is obtained in the unknowns α, b, c and e^v, summarized in matrix notation as

$$\left[\begin{array}{cc|cc} 0_N^T & 0_n^T & 0 & 1_N^T \\ \hline I_N & 0_{N\times n} & 1_N & \Omega + I_N \\ \hline 0_{n\times N} & I_n & 1_n & \Omega^v \end{array}\right]\left[\begin{array}{c} c \\ e^v \\ \hline b \\ \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \\ \hline y^v \end{array}\right]. \tag{10}$$

We refer to this principle as *fusion* of the training and the validation. Equation (3) can also be seen as an appearance of fusion of training and validation as (3) is equivalent to minimizing $\|e^v\|$ in (8) fused to (7) and (9). ¿From this convex optimization problem with quadratic constraint (see Figure 1.a), it can be seen that the non-convex optimization problem (3) has exactly two local minima (see Figure 1.b).

Different schemes for finding a 'best' among the many candidate solutions of the underdetermined system (10) can be considered, e.g.

$$\min_{\alpha,b,c,e^v} \|e\|_2^2 + \|e^v\|_2^2 \text{ s.t. (10) holds} \tag{11}$$



(a) Validation Cost surface

(b) Cost over Tikhonov constraint

Figure 1: *Figure* **(a)** *visualizes the convex optimization of the additive regularization c with respect to a validation measure. The solid line gives those solutions of AReg LS-SVM corresponding with solutions of LS-SVMs for γ ranging from 0 to ∞ according to the quadratic constraint (9). Figure* **(b)** *displays only the costs associated with solutions satisfying this quadratic constraint resulting in a nonconvex optimization problem.*

where $e = \alpha + c$. This criterion can be motivated by the assumption that e_i as well as e_j^v are independently sampled from the same distribution. The criterion (11) leads to the unique solution with the following constrained least squares problem

$$\left\| \left[\begin{array}{c|c} 1_N & \Omega \\ \hline 1_N & \Omega^v \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] - \left[\begin{array}{c} y \\ \hline y^v \end{array}\right] \right\|_2^2 \quad \text{s.t.} \quad 1_N^T \alpha = 0. \tag{12}$$

Straightforward application of the criterion (11) should be avoided when the number of training data exceeds the number of validation points as overfitting will occur on the validation data as shown in [6]. One can overcome this problem in various way by confining the space of possible c values [6].

3.2 Penalized model selection leading to sparseness

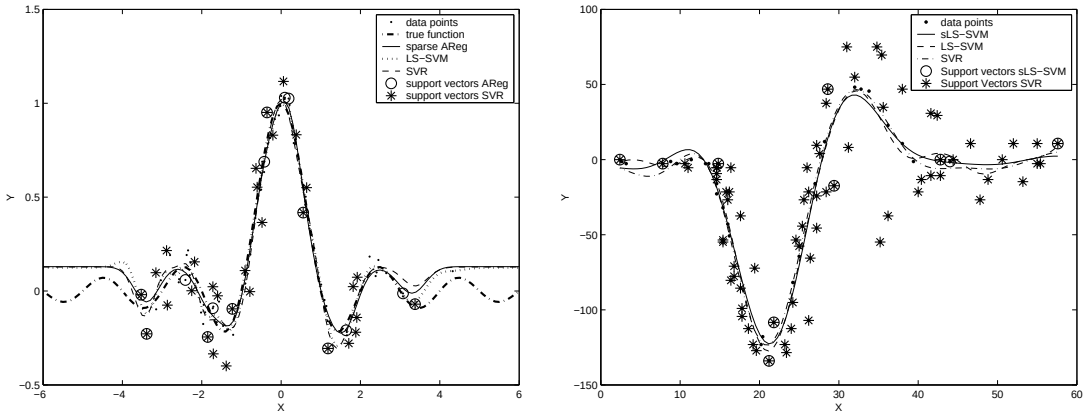
The effective degrees of freedom of the c-space can be restricted by imposing a norm on the solution of the final model [12, 13, 10]. The 1-norm is considered

$$\min_{\alpha,b,c,e^v} \|e\|_2^2 + \|e^v\|_2^2 + \xi \|e - c\|_1 \ \text{ s.t. (10) holds.} \tag{13}$$

This criterion leads to sparseness as $\|e - c\|_1 = \|\alpha\|_1$. Equivalently using (12):

$$\left\| \left[\begin{array}{c|c} 1_N & \Omega \\ \hline 1_n & \Omega^v \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] - \left[\begin{array}{c} y \\ \hline y^v \end{array}\right] \right\|_2^2 + \xi \|\alpha\|_1 \ \text{ s.t. } \ 1_N^T \alpha = 0. \tag{14}$$

This is a convex quadratic programming problem [1]. The tuning parameter ξ determines the relative importance of the model (validation) fit and the 1-norm (and thus the sparseness) of the solution α. We refer to this method as Sparse AReg LS-SVMs.



(a) Sinc

(b) Motorcycle dataset

Figure 2: *Comparison of the SVM, LS-SVM and Sparse AReg LS-SVM on (a) the artificial dataset* $sinc$ *and (b) the motorcycle dataset.*

	SVM		LS-SVM	Sparse AReg LS-SVM	
	Perf	**Sparse**	**Perf**	**Perf**	**Sparse**
Sinc	0.0052	68%	0.0045	0.0034	9%
Motorcycle	516.41	83%	444.64	469.93	11%
Ripley	90.10%	33.60%	90.40%	90.50%	4.80%
Pima	73.33%	43%	72.33%	74%	9%

Table 1: *Performances of SVMs, LS-SVMs and Sparse LS-SVMs expressed in Mean Squared Error (MSE) on a test set in the case of regression or Percentage Correctly Classified (PCC) in the case of classification. Sparseness is expressed in percentage of support vectors w.r.t. number of training data.*

4 Experiments

The performance of the proposed Sparse AReg LS-SVM was measured on a number of regression and classification datasets, respectively an artificial dataset sinc (generated as $X = \mathrm{sinc}(X) + e$ with $e \sim \mathcal{N}(0, 0.1)$ and $N = 100$, $d = 1$) and the motorcycle dataset [4] ($N = 100$, $d = 1$) for regression, the artificial Ripley dataset ($N = 250$, $d = 2$) and the PIMA dataset ($N = 468$, $d = 8$) from UCI for classification. The models resulting from Sparse AReg LS-SVMs were tested against SVMs and LS-SVMs where the kernel parameters and the other tuning-parameters (respectively C, ϵ for the SVM, γ for the LS-SVM and ξ for the Sparse Areg LS-SVM) were obtained from 10-fold cross-validation (see table 1). Some conclusions that can be made from these experiments are that the performance of Sparse Areg LS-SVMs is comparable to LS-SVMs, is better than for the standard SVM especially in the regression case and the degree of sparseness is significantly larger than for the standard SVM.

5 Conclusions

In this paper, a way is introduced to obtain sparseness of LS-SVMs with additive regularization by considering a penalized validation criterion. The fusion of the AReg LS-SVM training and regularization parameter tuning leads to a convex optimization problem from which the regularization and training parameters follow at once. The resulting Sparse AReg LS-SVMs were compared experimentally with SVMs and LS-SVMs on a number of classification as well as regression datasets with promising results.

Acknowledgements. This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven. It is supported by grants from several funding agencies and sources: Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP/40: Sustainibility effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS and BDM are an associate and full professor with K.U.Leuven Belgium, respectively.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[4] R.L. Eubank. *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York, 1999.

[5] T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, 2001.

[6] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Additive regularization: fusion of training and validation levels in kernel methods. *Internal Report 03-184, ESAT-SCD-SISTA, K.U.Leuven (Leuven, Belgium)*, 2003, submitted for publication.

[7] T. Poggio and F. Girosi. Networks for approximation and learning. In *Proceedings of the IEEE*, volume 78, pages 1481–1497. Proceedings of the IEEE, septenber 1990.

[8] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[9] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

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

[11] J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli, J. Vandewalle (Eds.) *Advances in Learning Theory: Methods, Models and Applications*. NATO Science Series III: Computer & Systems Sciences, 190, IOS Press Amsterdam, 2003.

[12] A.N. Tikhonov and V.Y. Arsenin. *Solution of Ill-Posed Problems*. Winston, Washington DC, 1977.

[13] V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.