Support Vector Regression with Automatic Accuracy Control

B. Schölkopf[*,†], P. Bartlett[*], A. Smola[*,†], R. Williamson[*]
[*] FEIT/RSISE, Australian National University, Canberra 0200, Australia
[†] GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
bs@first.gmd.de, http://svm.first.gmd.de

Abstract

A new algorithm for Support Vector regression is proposed. For a priori chosen ν, it automatically adjusts a flexible tube of minimal radius to the data such that at most a fraction ν of the data points lie outside. The algorithm is analysed theoretically and experimentally.

1 Introduction

Support Vector (SV) machines comprise a new class of learning algorithms, motivated by results of statistical learning theory [4]. Originally developed for pattern recognition, they represent the decision boundary in terms of a typically small subset [2] of all training examples, called the Support Vectors. In order for this property to carry over to the case of SV Regression, Vapnik devised the so-called ε-insensitive loss function [4] $|y - f(\mathbf{x})|_\varepsilon = \max\{0, |y - f(\mathbf{x})| - \varepsilon\}$, which does not penalize errors below some $\varepsilon > 0$, chosen a priori. His algorithm, which we will henceforth call ε-**SVR**, seeks to estimate functions

$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b, \qquad \mathbf{w}, \mathbf{x} \in \mathbf{R}^N, b \in \mathbf{R}, \tag{1}$$

based on data

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell) \in \mathbf{R}^N \times \mathbf{R}, \tag{2}$$

by minimizing the regularized risk functional $\frac{1}{2}\|\mathbf{w}\|^2 + C \cdot R^\varepsilon_{emp}$, where C is a constant determining the trade-off between minimizing training errors and minimizing the model complexity term $\|\mathbf{w}\|^2$, and $R^\varepsilon_{emp} := \frac{1}{\ell}\sum_{i=1}^{\ell} |y_i - f(\mathbf{x}_i)|_\varepsilon$.

The parameter ε can be useful if the desired accuracy of the approximation can be specified beforehand. In some cases, however, we just want the estimate to be as accurate as possible. We present a modification of the ε-SVR algorithm which automatically minimizes ε, in a manner much like how SV classifiers automatically maximize the margin of separation.

2 The Soft ε-Tube — ν-SV regression

To estimate functions (1) from empirical data (2) we proceed as follows. At each point $\mathbf{x}_i$, we allow an error of ε. Everything above ε is captured in slack

variables $\xi_i^{(*)}$ ($^{(*)}$ is a shorthand implying both the variables with and without asterisks), which are penalized in the objective function via a regularization constant C, to be chosen a priori [4]. The size of ε is traded off against model complexity and slack variables via a constant ν:

$$\text{minimize} \quad \tau(\mathbf{w}, \boldsymbol{\xi}^{(*)}, \varepsilon) = \frac{1}{2}\|\mathbf{w}\|^2 \quad + \quad C \cdot \left(\nu\varepsilon + \frac{1}{\ell} \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \right) \tag{3}$$

$$\text{subject to} \quad ((\mathbf{w} \cdot \mathbf{x}_i) + b) - y_i \quad \leq \quad \varepsilon + \xi_i \tag{4}$$

$$y_i - ((\mathbf{w} \cdot \mathbf{x}_i) + b) \quad \leq \quad \varepsilon + \xi_i^* \tag{5}$$

$$\xi_i^{(*)} \quad \geq \quad 0, \quad \varepsilon \quad \geq \quad 0. \tag{6}$$

Here and below, it is understood that $i = 1, \ldots, \ell$, and that bold face greek letters denote ℓ-dimensional vectors of the corresponding variables. For the constraints, we introduce multipliers $\alpha_i^{(*)}, \eta_i^{(*)}, \beta \geq 0$, and obtain the Lagrangian $L(\mathbf{w}, b, \boldsymbol{\alpha}^{(*)}, \beta, \boldsymbol{\xi}^{(*)}, \varepsilon, \boldsymbol{\eta}^{(*)}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\nu\varepsilon + \frac{C}{\ell}\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) - \sum_{i=1}^{\ell}\alpha_i(\xi_i + y_i - (\mathbf{w}\cdot\mathbf{x}_i) - b + \varepsilon) - \sum_{i=1}^{\ell}\alpha_i^*(\xi_i^* + (\mathbf{w}\cdot\mathbf{x}_i) + b - y_i + \varepsilon) - \beta\varepsilon - \sum_{i=1}^{\ell}(\eta_i\xi_i + \eta_i^*\xi_i^*)$. To minimize (3), we have to find the saddle point of L, i.e. minimize over the primal variables $\mathbf{w}, \varepsilon, b, \xi_i^{(*)}$ and maximize over the dual variables $\alpha_i^{(*)}, \beta, \eta_i^{(*)}$. Setting the derivatives with respect to the primal variables equal to zero leads to

$$\mathbf{w} \quad = \quad \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i)\mathbf{x}_i \qquad \textit{(SV expansion)} \tag{7}$$

$$C \cdot \nu - \sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) - \beta \quad = \quad 0, \tag{8}$$

$\sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0$, and $C/\ell - \alpha_i^{(*)} - \eta_i^{(*)} = 0$. Substituting these conditions into L leads to the Wolfe dual. Moreover, as in [1], we substitute a kernel k for the dot product, corresponding to a dot product in some reproducing kernel feature space related to input space via some possibly nonlinear map Φ,

$$k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})). \tag{9}$$

Rewriting the constraints, noting that $\beta, \eta_i^{(*)} \geq 0$ do not appear in the dual, we arrive at the **ν-SVR Optimization Problem:** for $\nu \geq 0, C > 0$, maximize

$$W(\boldsymbol{\alpha}^{(*)}) = -\frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i - \alpha_i^*) k(\mathbf{x}_i, \mathbf{x}_j)(\alpha_j - \alpha_j^*) - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) y_i \tag{10}$$

$$\text{subject to} \quad \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \quad = \quad 0 \tag{11}$$

$$0 \quad \leq \quad \alpha_i^{(*)} \quad \leq \quad \frac{C}{\ell} \tag{12}$$

$$\sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) \quad \leq \quad C \cdot \nu. \tag{13}$$

The regression estimate then takes the form (cf. (1), (7), (9))

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) k(\mathbf{x}_i, \mathbf{x}) + b, \tag{14}$$

where b and ε can be computed by taking into account (4) and (5).

Before we give theoretical results explaining the significance of the parameter ν, the following observation concerning ε is helpful. If $\nu > 1$, then $\varepsilon = 0$, since it does not pay to increase ε (cf. (3)). If $\nu \leq 1$, it can happen that $\varepsilon = 0$, e.g. if the data are noise-free and can perfectly be interpolated with a low capacity model. The case $\varepsilon = 0$, however, is not what we are interested in; it corresponds to plain L_1 loss regression. Below, we will use the term **errors** to refer to training points lying outside of the tube,[1] and the term **fraction** of errors/SVs to denote the relative numbers of errors/SVs, i.e. divided by ℓ.

Proposition 1 *Assume $\varepsilon > 0$. The following statements hold:*

(i) ν is an upper bound on the fraction of errors.

(ii) ν is a lower bound on the fraction of SVs.

(iii) Suppose the data (2) were generated iid from a distribution $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$ with $P(y|\mathbf{x})$ continuous. With probability 1, asymptotically, ν equals both the fraction of SVs and the fraction of errors.

Proof Ad **(i)**: The constraints (12) and (13) imply that at most a fraction ν of all examples can have $\alpha_i^{(*)} = C/\ell$. All examples with $\xi_i^{(*)} > 0$, i.e. those outside the tube, do certainly satisfy $\alpha_i^{(*)} = C/\ell$ (if not, $\alpha_i^{(*)}$ could grow further to reduce $\xi_i^{(*)}$). Ad **(ii)**: By the Kuhn-Tucker conditions (e.g. [4]), $\varepsilon > 0$ implies $\beta = 0$. Hence, (13) becomes an equality (cf. (8)). Since SVs are those examples for which $0 < \alpha_i^{(*)} \leq C/\ell$, the result follows (using $\alpha_i \cdot \alpha_i^* = 0$ for all i [4]). Ad **(iii)**: Continuity of the conditional distribution $P(y|\mathbf{x})$ implies that for all f, all $t \in \mathbf{R}$, and all $\gamma > 0$, $\lim_{\gamma \to 0} P(|f(\mathbf{x}) - y + t| < \gamma) = 0$. Since the class of SV regression estimates f has well-behaved covering numbers (e.g. [4]), we get uniform convergence, so for all $\gamma > 0$, $\sup_f |P(|f(\mathbf{x}) - y + t| < \gamma) - \hat{P}_\ell(|f(\mathbf{x}) - y + t| < \gamma)|$ converges to zero in probability, where $\hat{P}_\ell$ is the sample-based estimate of P (that is, the proportion of points that satisy $|f(\mathbf{x}) - y + t| < \gamma$). But then for all $\alpha > 0$, $\lim_{\gamma \to 0} \lim_{\ell \to \infty} P(\sup_f \hat{P}_\ell(|f(\mathbf{x}) - y + t| < \gamma) > \alpha) = 0$. Hence, $\sup_f \hat{P}_\ell(|f(\mathbf{x}) - y + t| = 0)$ converges to zero in probability. Using $t \in \{\pm\varepsilon\}$ thus shows that almost surely the fraction of points exactly on the tube tends to zero, hence the fraction of SVs equals that of errors. Combining (i) and (ii) then shows that both fractions converge almost surely to ν. ∎

Hence, $0 \leq \nu \leq 1$ can be used to control the number of errors (note that for $\nu \geq 1$, (12) implies (13)). Moreover, since the constraint (11) implies that (13)

[1] For $N > 1$, the "tube" is actually a slab, the region between two parallel hyperplanes.

is equivalent to $\sum_i \alpha_i^{(*)} \leq C\nu/2$, we conclude that Proposition 1 actually holds for the upper and the lower edge of the tube separately, with $\nu/2$ each.

In the ε-SVR machine [4], the objective function (cf. (10)) contains an additional term $-\varepsilon \sum_{i=1}^{\ell}(\alpha_i^* + \alpha_i)$, which, for fixed $\varepsilon > 0$, encourages that some of the $\alpha_i^{(*)}$ will turn out to be 0. Accordingly, the constraint (13) is not needed, and indeed it does not occur there. The primal problems (cf. (3)) differ in the term $\nu\varepsilon$. In the following sense, ν-SVR includes ε-SVR. Note that in the general case, using kernels, $\bar{\mathbf{w}}$ is a vector in feature space.

Proposition 2 *If ν-SVR leads to the solution $\bar{\varepsilon}, \bar{\mathbf{w}}, \bar{b}$, then ε-SVR with ε set a priori to $\bar{\varepsilon}$, and the same value of C, has the solution $\bar{\mathbf{w}}, \bar{b}$.*

Proof If we minimize (3), then fix ε and minimize only over the remaining variables, the solution does not change. ∎

3 Experiments and Discussion

In the experiments, we minimized (10), using the interior point optimizer LOQO.[2] This has the serendipitous advantage that the primal variables b and ε can be recovered as the dual variables of the dual problem (10) (i.e. the double dual variables) that is fed into the optimizer. The task was to estimate a regression of a noisy sinc function, given ℓ examples (x_i, y_i), with x_i drawn uniformly from $[-3, 3]$, and $y_i = \sin(\pi x_i)/(\pi x_i) + \upsilon_i$, where υ_i were drawn from a Gaussian with zero mean and variance σ^2. We used the RBF kernel $k(x, x') = \exp(-|x - x'|^2)$, and, if not stated otherwise, $\ell = 50, C = 100, \nu = 0.2, \sigma = 0.2$. Whenever standard deviation error bars are given, the results were obtained from 100 trials. Finally, the **risk** (or test error) was computed with respect to the sinc function without noise, as $\frac{1}{6}\int_{-3}^{3} |f(x) - \mathrm{sinc}(x)|\, dx$. Results are given in Table 1 and Figures 1 and 2.

Table 1: The ε found by ν-SV regression is largely independent of the sample size ℓ. The fraction of SVs and errors approach $\nu = 0.2$ from above and below, respectively, as the number of training examples ℓ increases (Proposition 1).

ℓ	10	50	100	200	500	1000	1500	2000
ε	0.27	0.22	0.23	0.25	0.26	0.26	0.26	0.26
fraction of errors	0.00	0.10	0.14	0.18	0.19	0.20	0.20	0.20
fraction of SVs	0.40	0.28	0.24	0.23	0.21	0.21	0.20	0.20

The theoretical analysis and experimental results suggest that ν is a computationally efficient way to control an upper bound on the number of errors which is tighter than the one used in the soft margin hyperplane [4]. In many

[2] see R. Vanderbei's web page http://www.princeton.edu/~rvdb/

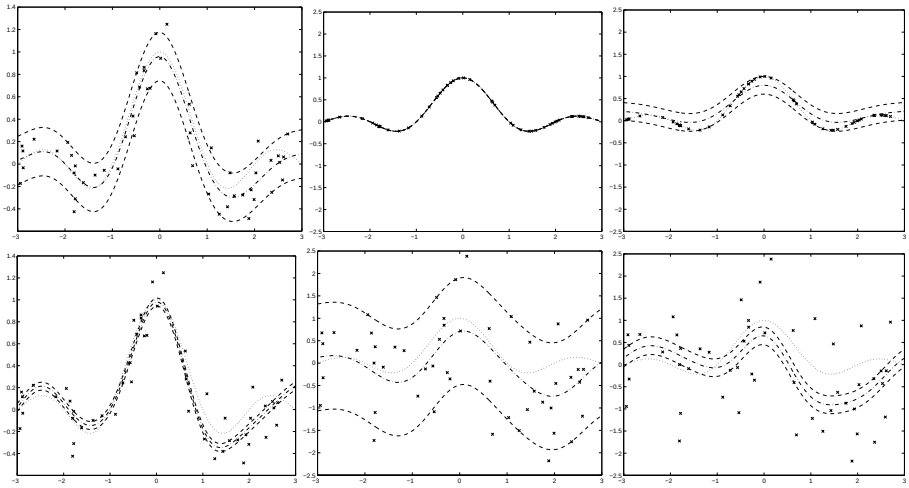

Figure 1: *Left:* ν-SV regression with $\nu = 0.2$ (top) and $\nu = 0.8$ (bottom). The larger ν allows more points to lie outside the tube (see Sec. 2). The algorithm automatically adjusts ε to 0.22 (top) and 0.04 (bottom). Shown are the sinc function (dotted), the regression f and the tube $f \pm \varepsilon$. *Middle:* ν-SV regression on data with noise $\sigma = 0$ (top) and $\sigma = 1$ (bottom). In both cases, $\nu = 0.2$. The tube width automatically adjusts to the noise (top: $\varepsilon = 0$, bottom: $\varepsilon = 1.19$). *Right:* ε-SV regression [4] on data with noise $\sigma = 0$ (top) and $\sigma = 1$ (bottom). In both cases, $\varepsilon = 0.2$. This parameter choice, which has to be specified a priori, is ideal for neither case. In the top figure, the regression estimate is biased; furthermore, in the bottom figure, ε does not match the external noise [3].

cases, this makes it a parameter which is more convenient than the one in ε-SVR. Asymptotically, it directly controls the number of Support Vectors, and the latter can be used to give a leave-one-out generalization bound [4]. In addition, ν characterizes the compression ratio: it is sufficient to train the algorithm only on the SVs, leading to the same solution [2]. In ε-SVR, the tube width ε must be specified a priori; in ν-SVR, it is computed automatically. Nevertheless, desirable properties of ε-SVR, including the formulation as a definite quadratic programming problem, and the sparse representation of the solution in terms of SVs, are retained.

We conclude with two open questions. We found that the automatic ε scales linearly with σ (Fig. 2). The *optimal* ε, leading to the best generalization, also scales linearly with σ [3]. Does there exist a value of C such that the present algorithm always (for all σ) finds the optimal ε? Secondly, is it possible to exploit the close resemblance of ε and the risk (as functions of the regularization constant C, cf. Fig. 2) in order to devise a method of selecting C without the need for cross-validation techniques?

Acknowledgement This work was partly supported by the Australian Research Council and the DFG (# Ja 379/71).

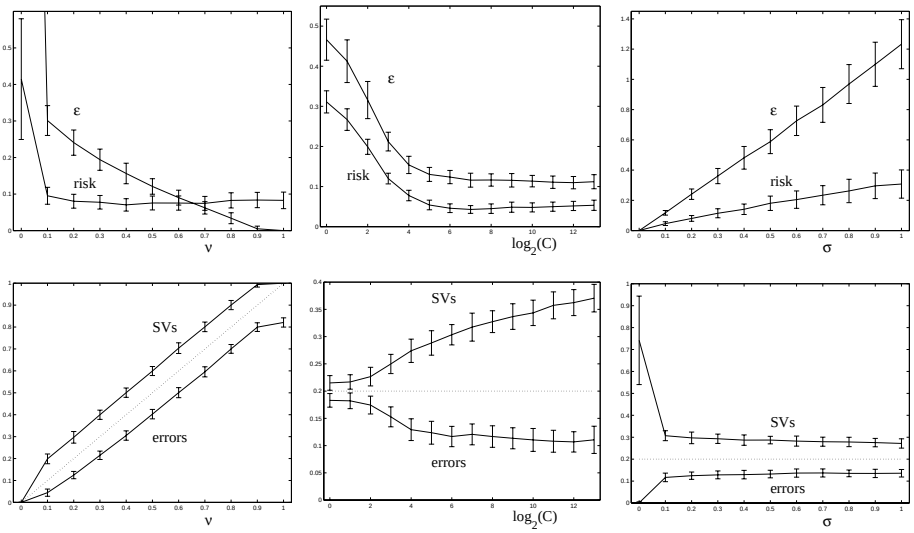

Figure 2: *Left:* ν-SVR for different values of the error constant ν. Notice how ε decreases when more errors are allowed (large ν). *Middle:* ν-SVR for different values of the regularization constant C. The top figure shows that ε decreases when the regularization is decreased (large C). Only very little, if any, overfitting occurs. In the bottom figure, note that ν upper bounds the number of errors, and lower bounds the number of SVs (cf. Proposition 1). The bound gets looser as C increases — this corresponds to a smaller number of examples ℓ relative to C (cf. Table 1). *Right:* ν-SVR for different values of the noise σ. The tube radius ε increases linearly with σ (this is largely due to the fact that both ε and the $\xi_i^{(*)}$ enter the cost function linearly). Due to the automatic adaptation of ε, the number of SVs and of points outside the tube (errors) is, except for the noise-free case $\sigma = 0$, largely independent of σ.

References

[1] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proc. 5th Ann. ACM Workshop on COLT*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

[2] B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings, First International Conference on Knowledge Discovery & Data Mining.* AAAI Press, Menlo Park, CA, 1995.

[3] A. Smola, N. Murata, B. Schölkopf, and K.-R. Müller. Asymptotically optimal choice of ε-loss for support vector machines. ICANN'98.

[4] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.