MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1649 November 1998
C.B.C.L Paper No. 166

From Regression to Classification in Support Vector Machines

Massimiliano Pontil, Ryan Rifkin and Theodoros Evgeniou

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1XXX.ps

Abstract

We study the relation between support vector machines (SVMs) for regression (SVMR) and SVM for classification (SVMC). We show that for a given SVMC solution there exists a SVMR solution which is equivalent for a certain choice of the parameters. In particular our result is that for ϵ sufficiently close to one, the optimal hyperplane and threshold for the SVMC problem with regularization parameter C_c are equal to $\frac{1}{1-\epsilon}$ times the optimal hyperplane and threshold for SVMR with regularization parameter $C_r = (1-\epsilon)C_c$. A direct consequence of this result is that SVMC can be seen as a special case of SVMR.

This report describers research done at the Center for Biological & Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. This research was sponsored by the Office of Naval Research under contract No. N00014-93-1-0385 and contract No. N00014-95-1-0600. Partial support was also provided by Daimler-Benz AG, Eastman Kodak, Siemens Corporate Research, Inc., and AT&T.

1 Introduction

We assume that the reader has some familiarity with Support Vector Machines. In this section, we provide a brief review, aimed specifically at introducing the formulations and notations we will use throughout this paper. For a good introduction to SVMs, see [1] or [2].
In the support vector machine classification problem, we are given l examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$, with $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ for all i. The goal is to find a hyperplane and threshold $(\mathbf{w}, b)$ that separates the positive and negative examples with maximum margin, penalizing points inside the margin linearly in a user-selected regularization parameter $C > 0$. The SVM classification problem can be restated as finding an optimal solution to the following quadratic programming problem:

$$
\begin{array}{llll}
(\mathcal{C}) & \displaystyle\min_{\mathbf{w}, b, \xi} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\ell}\xi_i & \\
& & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \ \geq 1 - \xi_i & i = 1, \ldots, l \\
& & \xi \ \geq 0 &
\end{array}
$$

This formulation is motivated by the fact that minimizing the norm of w is equivalent to maximizing the margin; the goal of maximizing the margin is in turn motivated by attempts to bound the generalization error via structural risk minimization. This theme is developed in [2].
In the support vector machine regression problem, the goal is to construct a hyperplane that lies "close" to as many of the data points as possible. We are given l examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$, with $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ for all i. Again, we must select a hyperplane and threshold $(\mathbf{w}, b)$[1]. Our objective is to choose a hyperplane $\mathbf{w}$ with small norm, while simultaneously minimizing the sum of the distances from our points to the hyperplane, measured using Vapnik's ϵ-insensitive loss function:

$$
|y_i - (\mathbf{w}\cdot\mathbf{x}_i + b)|_\epsilon = \begin{cases} 0 & \text{if } |y_i - (\mathbf{w}\cdot\mathbf{x}_i + b)| \leq \epsilon \\ |y_i - (\mathbf{w}\cdot\mathbf{x}_i + b)| - \epsilon & otherwise \end{cases} \tag{1}
$$

The parameter ϵ is preselected by the user. As in the classification case, the tradeoff between finding a hyperplane with small norm and finding a hyperplane that performs regression well is controlled via a user selected regularization parameter C. The quadratic programming problem associated with SVMR is:

$$
\begin{array}{llll}
(\mathcal{R}) & \displaystyle\min_{\mathbf{w}, b, \xi, \xi^*} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) & \\
& & y_i - (\mathbf{w}\cdot\mathbf{x}_i + b) \ \leq \epsilon + \xi_i & i = 1, \ldots, l \\
& & -y_i + (\mathbf{w}\cdot\mathbf{x}_i + b) \ \leq \epsilon + \xi_i^* & i = 1, \ldots, l \\
& & \xi, \xi^* \ \geq 0 &
\end{array}
$$

The main aim of this paper is to demonstrate a connection between support vector machine classification and regression.
In general, SVM classification and regression are performed using a nonlinear kernel $K(\mathbf{x}_i, \mathbf{x}_j)$. For simplicity of notation, we chose to present our formulations in terms of a linear separating hyperplane $\mathbf{w}$. All our results apply to the nonlinear case; the reader may assume that we are trying to construct a linear separating hyperplane in a high-dimensional feature space.

[1] Observe that now the hyperplane will reside in $n + 1$ dimensions.

2 From Regression to Classification

In the support vector machine regression problem, the y_i are real-valued rather than binary-valued. However, there is no prohibition *against* the y_i being binary-valued. In particular, if $y_i \in \{-1, 1\}$ for all i, then we may perform support vector machine classification or regression on the same data set.

Note that when performing support vector machine regression on $\{-1,1\}$-valued data, if $\epsilon \geq 1$, $\mathbf{w} = \mathbf{0}$, $\xi = \mathbf{0}$, $\xi^* = \mathbf{0}$ is an optimal solution to $\mathcal{R}$. Therefore, we restrict our attention to cases were $\epsilon < 1$. Loosely stated, our main result is that for ϵ sufficiently close to one, the optimal hyperplane and threshold for the support vector machine classification problem with regularization parameter C_c are equal to $\frac{1}{1-\epsilon}$ times the optimal hyperplane and threshold for the support vector machine regression problem with regularization parameter $C_r = (1-\epsilon)C_c$. We now proceed to formally derive this result.

We make the following variable substitution:

$$\eta_i = \begin{cases} \xi_i & \text{if } y_i = 1 \\ \xi_i^* & \text{if } y_i = -1. \end{cases} \quad , \quad \eta_i^* = \begin{cases} \xi_i^* & \text{if } y_i = 1 \\ \xi_i & \text{if } y_i = -1. \end{cases} \tag{2}$$

Combining this substitution with our knowledge that $y_i \in \{-1, 1\}$ yields the following modification of $\mathcal{R}$:

$$\begin{array}{llll} (\mathcal{R}') & \min\limits_{\mathbf{w}, b, \eta, \eta^*} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\ell}(\eta_i + \eta_i^*) & \\ & & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \epsilon - \eta_i & i = 1, \ldots, l \\ & & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \leq 1 + \epsilon + \eta_i^* & i = 1, \ldots, l \\ & & \eta, \eta^* \geq 0 & \end{array}$$

Continuing, we divide both sides of each constraint by $1-\epsilon$, and make the variable substitutions $w' = \frac{w}{1-\epsilon}$, $b' = \frac{b}{1-\epsilon}$, $\eta' = \frac{\eta}{1-\epsilon}$, $\eta^{*'} = \frac{\eta^*}{1-\epsilon}$:

$$\begin{array}{llll} (\mathcal{R}'') & \min\limits_{\mathbf{w}', b', \eta', \eta^{*'}} & \frac{1}{2}\|\mathbf{w}'\|^2 + \frac{C}{1-\epsilon}(\sum_{i=1}^{\ell}(\eta_i' + \eta_i^{*'})) & \\ & & y_i(\mathbf{w}'\cdot\mathbf{x}_i + b) \geq 1 - \eta_i' & i = 1, \ldots, l \\ & & y_i(\mathbf{w}'\cdot\mathbf{x}_i + b) \leq \frac{1+\epsilon}{1-\epsilon} + \eta_i^{*'} & i = 1, \ldots, l \\ & & \eta', \eta^{*'} \geq 0 & \end{array}$$

Looking at formulation $\mathcal{R}''$, one suspects that as ϵ grows close to 1, the second set of constraints will be "automatically" satisfied with $\eta^* = 0$. We confirm this suspicion by forming the Lagrangian dual:

$$\begin{array}{lll} (\mathcal{RD}'') & \min\limits_{\beta} & \frac{1}{2}\sum_{i,j=1}^{l}(\beta_i - \beta_i^*)D_{ij}(\beta_i - \beta_i^*) - \sum_i \beta_i + \frac{1+\epsilon}{1-\epsilon}\sum_i \beta_i^* \\ & & \sum_i y_i\beta_i = \sum_i y_i\beta_i^* \\ & & \beta_i, \beta_i^* \geq 0 \\ & & \beta_i, \beta_i^* \leq \frac{C}{1-\epsilon} \end{array}$$

where D is the symmetric positive semidefinite matrix defined by the equation $D_{ij} \equiv y_i y_j \mathbf{x}_i \mathbf{x}_j$. For all ϵ sufficiently close to one, the η_i^* will all be zero: to see this, note that $\eta = \mathbf{0}, \eta^* = \mathbf{0}$ is a feasible solution to $\mathcal{RD}''$ with cost zero, and if any η_i^* is positive, for ϵ sufficiently close to one, the value of the solution will be positive. Therefore, assuming that ϵ is sufficiently large, we may eliminate the η^* terms from $\mathcal{R}''$ and and the β^* terms from $\mathcal{D}''$. But removing these terms from $\mathcal{R}''$ leaves us with a quadratic program essentially identical to the dual of formulation $\mathcal{C}$:

$$
\begin{array}{lll}
(\mathcal{CD}'') & \min\limits_{\beta} & \frac{1}{2}\sum_{i,j=1}^{l}\beta_i D_{ij}\beta_j - \sum_i \beta_i \\
& & \begin{array}{rl} \sum_i y_i\beta_i & = 0 \\ \beta_i & \geq 0 \\ \beta_i & \leq \frac{C}{1-\epsilon} \end{array}
\end{array}
$$

Going back through the dual, we recover a slightly modified version of $\mathcal{C}$:

$$
\begin{array}{lll}
(\mathcal{C}') & \min\limits_{\mathbf{w},b,\xi} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{1-\epsilon}\sum_{i=1}^{\ell}\xi_i \\
& & \begin{array}{rll} y_i(\mathbf{w}\cdot\mathbf{x}_i + b) & \geq 1-\xi_i & \quad i = 1,\ldots,l \\ \xi & \geq 0 & \end{array}
\end{array}
$$

Starting from the classification problem instead of the regression problem, we have proved the following theorem:

Theorem 2.1 *Suppose the classification problem $\mathcal{C}$ is solved with regularization parameter C, and the optimal solution is found to be $(\mathbf{w}, b)$. Then, there exists a value $a \in (0, 1)$ such that $\forall\epsilon \in [a, 1)$, if problem $\mathcal{R}$ is solved with regularization parameter $(1-\epsilon)C$, the optimal solution will be $(1-\epsilon)(\mathbf{w}, b)$.*

Several points regarding this theorem are in order:

- **The η substitution.** This substitution has an intuitive interpretation. In formulation $\mathcal{R}$, a variable ξ_i is non-zero if and only if y_i lies *above* the ϵ-tube, and the corresponding ξ_i^* is non-zero if and only if y_i lies *below* the ϵ-tube. This is independent of whether y_i is 1 or -1. After the η substitution, η_i is non-zero if $y_i = 1$ and y_i lives above the ϵ-tube, or if $y_i = -1$ and y_i lives below the ϵ-tube. A similar interpretation holds for the η_i^*. Intuitively, the η_i correspond to error points which lie on the *same* side of the tube as their sign, and the η_i^* correspond to error points which lie on the *opposite* side. We might guess that as ϵ goes to one, only the former type of error will remain: the theorem provides a constructive proof of this conjecture.

- **Support Vectors.** Examination of the formulations, and their KKT conditions, shows that there is a one-to-one correspondence between support vectors of $\mathcal{C}$ and support vectors of $\mathcal{R}$ under the conditions of correspondence. Points which are not support vectors in $\mathcal{C}$ and therefore lie *outside* the margin and are correctly classified will lie strictly *inside* the ϵ-tube in $\mathcal{R}$. Points which lie on the margin in $\mathcal{C}$ will lie on the boundaries of the ϵ-tube in $\mathcal{R}$, and are support vectors for both problems. Finally, points which lie inside the margin or are incorrectly classified in $\mathcal{C}$ will lie strictly outside the ϵ-tube, *above the tube for points with $y = 1$, below the tube for points with $y = -1$*, and are support vectors for both problems.

- **Computation of a.** Using the KKT conditions associated with problem $(\mathcal{R}'')$, we can determine the value of a which satisfies the theorem. To do so, simply solve problem $\mathcal{C}'$, and choose a to be the smallest value such that when the constraints $(\mathbf{w}'\cdot\mathbf{x}_i + b) \leq \frac{1+\epsilon}{1-\epsilon} + \eta_i^{*'}, i = 1,\ldots,l$ are added, they are satisfied by the optimal solution to $\mathcal{C}'$. In particular, if we define m to be the maximal value of $y_i(\mathbf{w}\cdot\mathbf{x}_i + b)$, then $a = \frac{m-1}{m+1}$ will satisfy the theorem. Observe that as $w := \|\mathbf{w}\|$ gets larger (i.e., the separating hyperplane gets steeper), or as the *correctly classified* $\mathbf{x}_i$ get relatively (in units of the margin w^{-1}) farther away from the hyperplane we expect a to increase. More precisely it is easy to see that $m \leq wD$, with D the diameter of the smallest hypersphere containing all the points. Then $a \geq \frac{wD-1}{wD+1}$, which is an increasing function of w. Finally observe that incorrectly classified points will have $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) < 0$, and therefore they cannot affect m or a.

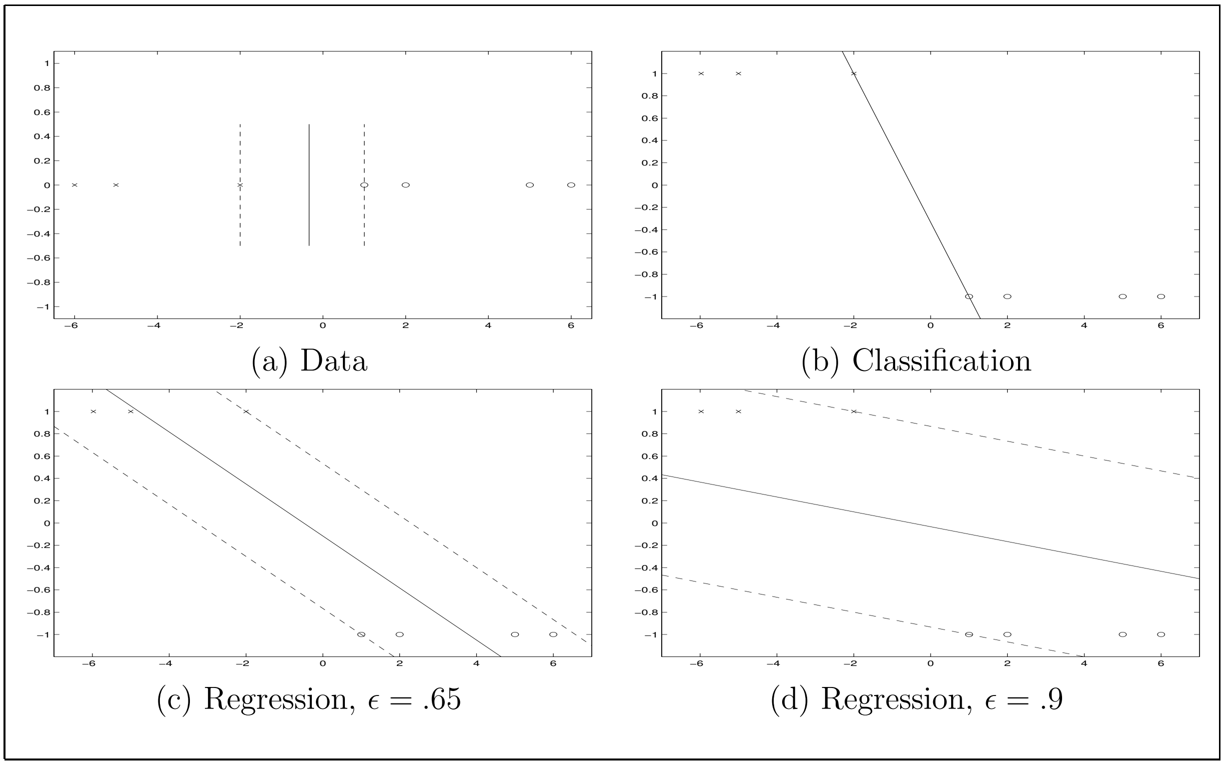
- **The ξ^2 case.** We may perform a similar analysis when the slack variables are penalized quadratically rather than linearly. The analysis proceeds nearly identically. In the transition from formulation $\mathcal{R}'$ to $\mathcal{R}''$, an extra factor of $(1-\epsilon)$ falls out, so the objective function in $\mathcal{R}''$ is simply $\frac{1}{2}\|\mathbf{w}'\|^2 + C(\sum_{i=1}^{\ell}((\eta_i')^2 + (\eta_i^{*'})^2))$. The theorem then states that for sufficiently large ϵ, if ($\mathbf{w}$, b) solves $\mathcal{C}$ with regularization parameter C, $(1-\epsilon)(\mathbf{w}, b)$ solves $\mathcal{R}$, also with regularization parameter C.

3 Examples

In this section, we present two simple one-dimensional examples that help to illustrate the theorem. These examples were both performed penalizing the ξ_i linearly.
In the first example, the data are linearly separable. Figure 1a shows the data points, and Figure 1b shows the separating hyperplane found by performing support vector classification with $C = 5$ on this data set. Note that in the classification problem, the data lie in one dimension, with the y-values being "labels". The hyperplane drawn shows the value of $\mathbf{w} \cdot \mathbf{x} + b$ as a function of $\mathbf{x}$. The computed value of a is approximately .63. Figure 1c shows the ϵ-tube computed for the regression problem with $\epsilon = .65$, and Figure 1d shows the same for $\epsilon = .9$. Note that every data point is correctly classified in the classification problem, and that every data point lies inside the ϵ-tube in the regression problems, for the values of ϵ chosen.
In the second example, the data are not linearly separable. Figure 2a shows the data points. Figure 2b shows the separating hyperplane found by performing classification with $C = 5$. The computed value of a is approximately .08. Figures 2c and d show the regression tubes for $\epsilon = .1$ and $\epsilon = .5$, respectively. Note that the points that lie at the edge of the margin for classification, $\mathbf{x} = -5$ and $\mathbf{x} = 6$ lie on the edge of the ϵ-tube in the regression problems, and that points that lie inside the margin, or are misclassified, lie outside the ϵ-tube. The point $\mathbf{x} = -6$, which is the only point that is strictly outside the margin in the classification problem, lies *inside* the ϵ-tubes. The image provides insight as to why a is much smaller in this problem than in the linearly separable example: in the linearly separable case, any ϵ-tube must be shallow and wide enough to contain all the points.

4 Conclusions and Future Work

In this note we have shown how SVMR can be related to SVMC. Our main result can be summarized as follows: if ϵ is sufficiently close to one, the optimal hyperplane and threshold for the SVMC problem with regularization parameter C_c are equal to $\frac{1}{1-\epsilon}$ times the optimal hyperplane and threshold for SVMR with regularization parameter $C_r = (1-\epsilon)C_c$. A direct consequence of this result is that SVMC can be regarded as a special case of SVMR. An important problem which will be study of future work is whether this result can help place SVMC and SVMR in the same common framework of structural risk minimization.



(a) Data (b) Classification

(c) Regression, $\epsilon = .65$ (d) Regression, $\epsilon = .9$

Figure 1: Separable data.

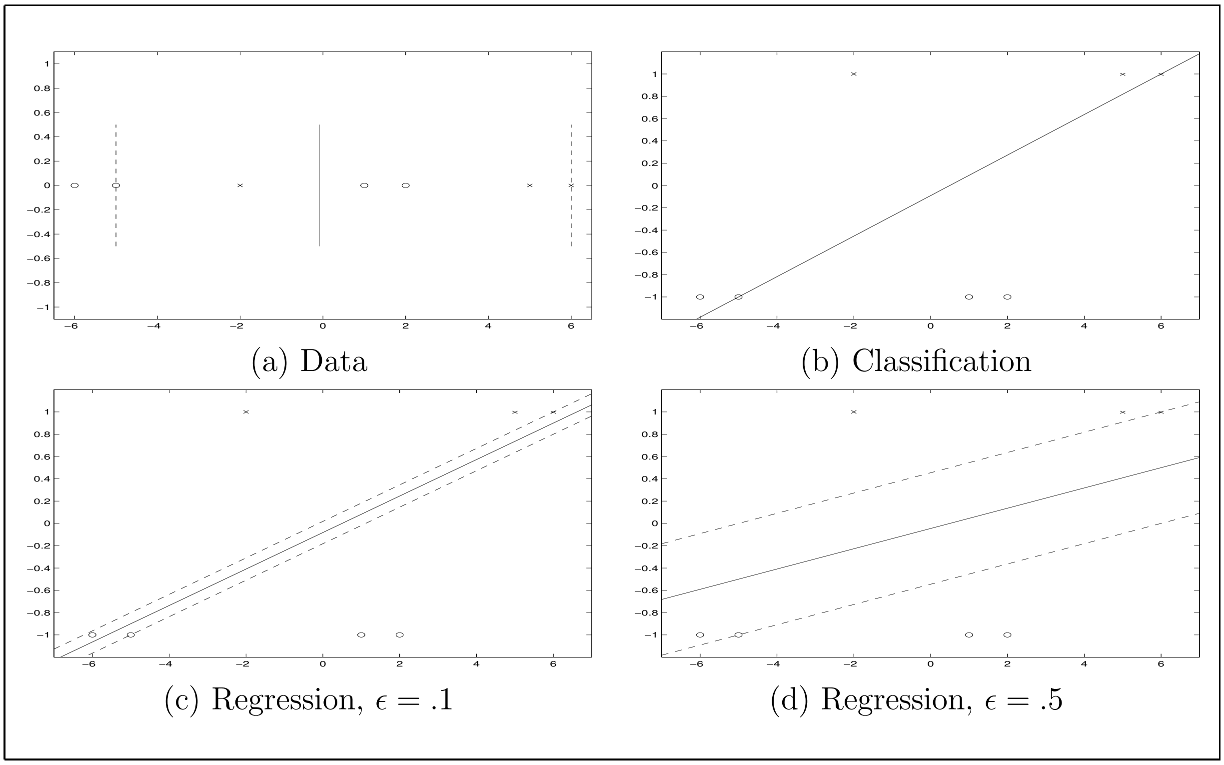
5 Acknowledgments

We wish to thank Alessandro Verri, Tomaso Poggio, Sayan Mukherjee and Vladimir Vapnik for useful discussions.

References

[1] C. Burges. A tutorial on support vector machines for pattern recognition. In *Data Mining and Knowledge Discovery*, volume 2, pages 1 – 43. Kluwer Academic Publishers, Boston, 1998.

[2] V. Vapnik. *Statistical Learning Theory.* John Wiley and sons, New York, 1998.



(a) Data

(b) Classification

(c) Regression, $\epsilon = .1$

(d) Regression, $\epsilon = .5$

Figure 2: Non-separable data.