Extension of the ν-SVM Range for Classification

Fernando Pérez-Cruz[†], Jason Weston[‡], Daniel J. L. Hermann[‡]
and Bernhard Schölkopf[‡]
[†]Dpto. Teoría de la Señal y Comunicaciones,
Universidad Carlos III de Madrid, Leganes, Spain
[‡]Max Plank Institute for Biological Cybernetics, Tübingen, Germany
e-mail: fernandop@ieee.org

2nd December 2002

Abstract

The ν-Suport Vector Machine for classification (ν-SVC) has been presented as a different formulation for solving SVMs in which the C parameter is transformed by a more meaningful parameter ν, that roughly represents the fraction of support vectors. The value of ν cannot always take all possible values between 0 and 1, which limits the range of possible soutions. Either, because the training set is non-separable in the feature space, or because the classes are unbalanced. In this paper, we will deal with both restrictions, presenting a new Extended ν-SVC, in which the value of ν can move from 0 to 1 in any circumstance. The modification to extend the range up to 1 is trivial, we only need to modify the cost associated to the margin errors to balance the classes. The modification to extend the range down to zero is far more complex. We will first need to revisit how maximum margin classifiers can be obtained for a separable training set, to enable us to construct "hard" margin classifiers for non-separable datasets. This can be achieved by finding the separation in which incorrectly classified samples have the smallest *negative* margin. This re-interpretation of the maximum margin classifier, when viewed as a soft margin formulation, will allow us to extend the range of ν-SVC to any number of support vectors. Experiments with real and synthetic data confirm the validity of the proposed approach.

1 Introduction

Support Vector Machines (SVMs) are state-of-the-art tools for linear and nonlinear knowledge discovery [14]. They were initially developed for linearly separable problems, known as the optimal hyperplane decision rule (OHDR) [18]. In a nutshell, the OHDR finds the classification boundary that linearly separates the given data and is furthest from the data. The OHDR is computed as the maximization of the minimum distance of the samples to the separating hyperplane. This minimum distance is known as the margin, and the OHDR is also known as the maximum margin classifier. The maximum margin classifiers were generalized to cover nonlinear problems, through the "kernel trick" [14, 2]; and non-separable problems, via slack variables that relax the conditions in the original formulation [19, 6].

The SVM, when solved for nonlinear problems, has to set the value of a weight parameter C which measures the trade off between the training errors and the maximization of the margin. This weight is hard to choose a priori and it is difficult to infer which result can be expected for a C value over any given problem. There is an alternative formulation, known as ν-SVM, in which the weight parameter is replaced by another more intuitive parameter ν. This parameter roughly represents the fraction of expected support vectors, therefore for any given $\nu \in (0, 1]$, we will know a priori how the classifier will be formed. Also, it allows to easily scan the whole range of possible solutions, because choosing ν between 0 and 1 will give all the possible outcomes.

The ν-SVM for classification (ν-SVC) has a limitation in terms of its usable range. The value of ν can be upper bounded by a value less than 1, if the classes are not balanced [7], and it can be lower bounded by a value greater than 0 for some data sets, if the VC dimension of the used classifier is not infinite [4]. These two limitations also exist in the formulation using the C parameter, although they are not explicit with this parameter, explaining why they had not been previously addressed as limitations. The first limitation can be easily avoided if one requires so, as we will show herein. The second is not readily overcome and the main body of this article is devoted to it.

We will propose a reinterpretation on how maximum margin hyper-planes are constructed. This reinterpretation will lead to a unified formulation for both separable and inseparable sets: a maximum positive margin solution, if the training data set is separable, and a minimum negative margin solution (to be described in the following sections), if the training samples are not. This unified formulation will be constructed using a ν-SVM type parameterization and, consequently, we will be able to control the number of SVs for the full range of possible values of ν. We will refer to it as Extended ν-SVM (Eν-SVM). Therefore, we will be able to select the SVM optimal solution from the whole range of possible solutions, i.e, all the solutions with any number of SVs.

We will start with a full description of the ν-SVM and its relationship with SVMs (using the C parameter, also known as C-SVM) in Section 2. Then in Section 3, we will focus on its two limitations, the upper and lower bounds over the value of ν. The first can be easily overcome by re-weighting the errors in the ν-SVM. The second will need a further study, we will define a negative margin classifier and how they have to be solved in Section 4. We will show that the negative margin classifiers can be expressed in a unified formulation similar to ν-SVM in Section 5, which we will refer to as Extended ν-SVM. In Section 6, We show by means of computer experiments the validity of the proposed approach using real and synthetic data. We will end with some concluding remarks in Section 7.

2 ν Support Vector Classifiers

The Support Vector Machine for binary classification (C-SVC) [20] finds the optimum of a quadratic constrained problem:

$$\min_{\mathbf{w},\xi_i,b} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \tag{1}$$

subject to

$$y_i(\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b) \geq 1 - \xi_i \qquad \forall i = 1,\ldots,n \tag{2}$$

$$\xi_i \geq 0 \qquad \forall i = 1,\ldots,n \tag{3}$$

where the data set $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_n, y_n)$ ($\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$) has been sampled independently and identically distributed (i.i.d.) from a joint probability density function $p(\mathbf{x}, y)$ that relates each vector $\mathbf{x}_i$ with its corresponding class label y_i. The nonlinear mapping $\boldsymbol{\phi}(\cdot)$ ($\mathbb{R}^d \xrightarrow{\boldsymbol{\phi}(\cdot)} \mathbb{R}^{\mathcal{H}}$) transforms the input data to a higher dimensional space, the feature space $\mathcal{H}$. The linear classifier $(\mathbf{w}, b)$ in the feature space is usually nonlinear in the input space, unless $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$.

In the above formulation, C is a parameter determining the trade-off between two conflicting goals: minimizing the training error, and maximizing the margin. Unfortunately, C is a rather unintuitive parameter, and we have no a priori way to select it. Therefore, a modification was proposed in [16], which replaces C by a parameter ν; the latter will turn out to control the number of margin errors and, consequently, the Support Vectors (SVs).

As a primal problem for this approach, termed the ν-SVM for classification (ν-SVC), we consider

$$\min_{\mathbf{w},\xi_i,\rho,b} \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \tag{4}$$

subject to (3) and

$$y_i(\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b) \geq \rho - \xi_i \qquad \forall i = 1, \ldots, n \tag{5}$$

$$\rho \geq 0 \tag{6}$$

A new non-negative variable ρ has been included in the objective functional and it has to minimized. However, it has been shown [7] that the constraint enforcing ρ to be positive, (6), is unnecessary and that the above optimization problem will always end with a ρ greater or equal than 0. Intuitively, if a solution with $\rho \leq 0$ is feasible, we can set $\mathbf{w} = 0$, $b = 0$ and $\xi_i = 0$, which will fulfill the constraint in (5), and will give the lowest possible values of the first and third terms of (4), $\rho = 0$ being the one that minimizes it most. Therefore, a negative value of ρ cannot reduce the value of (4). Also, it can be shown that the functional (4) cannot become negative (it can be readily seen from the dual of this problem), therefore the solution in which (4) is zero cannot be improved.

To explain the significance of ν, let us first define the term *margin error*: by this, we denote points with $\xi_i > 0$. These are points which are either training errors ($\xi_i > \rho$), or lie within the margin ($\xi_i \in (0, \rho]$). Formally, the fraction of margin errors is

$$R^{\rho}_{\text{emp}}[\mathbf{w}, b] := \frac{1}{n} \left| \{i | y_i(\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b) < \rho\} \right|. \tag{7}$$

The following proposition was stated and proven in [16] and it allows to understand the role of ν and what to expect once the solution has been reached.

Proposition 1 *Suppose we run ν-SVC with a given kernel on some data with the result that $\rho > 0$. Then*

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

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

(iii) Suppose the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ were generated i.i.d. from a distribution $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$, such that neither $p(\mathbf{x}, y = 1)$ nor $p(\mathbf{x}, y = -1)$ contains any discrete component. Suppose, moreover, that the kernel used is analytic and non-constant. With probability 1, asymptotically, ν equals both the fraction of SVs and the fraction of errors.

We would like to show with a toy example the solutions that one can expect when solving the ν-SVM, before explaining how it is actually solved. We show in Figure 1 the solution for various different ν for a two dimensional problem solved with a Gaussian

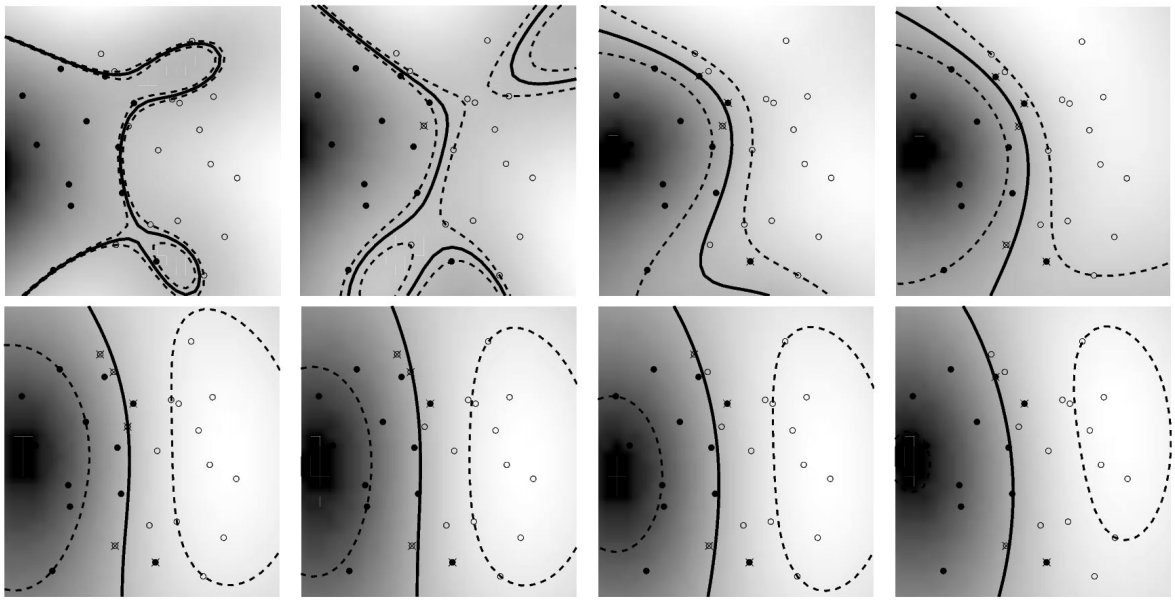

Figure 1: Toy problem (task: separate circles from disks) solved using ν-SV classification, with parameter values ranging from $\nu = 0.1$ (top left) to $\nu = 0.8$ (bottom right). The larger we make ν, the more points are allowed to lie inside the margin (depicted by dotted lines). Results are shown for a Gaussian kernel, $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2)$ [14].

kernel. The fraction of margin errors and support vectors, discussed in the previous proposition, can be seen in Table 1.

The ν-SVM for classification can include the linear restrictions in the objective functional using Lagrange multipliers, requiring one to minimize

$$L(\mathbf{w}, \xi_i, b, \rho, \alpha_i, \mu_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \\ - \sum_{i=1}^{n}(\alpha_i(y_i(\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b) - \rho + \xi_i) + \mu_i\xi_i), \quad (8)$$

with respect to $\mathbf{w}$, ξ_i, ρ and b and maximize it with respect to the Lagrange multipliers, $\alpha_i, \beta_i \geq 0$. We have not imposed the condition in (6), following [7]. The solution to this problem is given by the Karush-Kuhn-Tucker Theorem [9], that imposes the following

Table 1: Fractions of errors and SVs, along with the margins of class separation, for the toy example in Figure 1.
Note that ν upper bounds the fraction of errors and lower bounds the fraction of SVs, and that increasing ν, i.e., allowing more margin errors, increases the margin.

ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
fraction of errors	0.00	0.07	0.25	0.32	0.39	0.50	0.61	0.71
fraction of SVs	0.29	0.36	0.43	0.46	0.57	0.68	0.79	0.86
margin $\rho/\|\mathbf{w}\|$	0.005	0.018	0.115	0.156	0.364	0.419	0.461	0.546

conditions: (5), (3) and

$$\frac{\partial L_p}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i) = \mathbf{0} \tag{9}$$

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{10}$$

$$\frac{\partial L_p}{\partial \rho} = \sum_{i=1}^{n} \alpha_i - \nu = 0 \tag{11}$$

$$\frac{\partial L_p}{\partial \xi_i} = \frac{1}{n} - \alpha_i - \mu_i = 0 \qquad \forall i = 1, \ldots, n \tag{12}$$

$$\alpha_i, \mu_i \geq 0 \qquad \forall i = 1, \ldots, n \tag{13}$$

$$\alpha_i \{ y_i (\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b) - \rho + \xi_i \} = 0 \qquad \forall i = 1, \ldots, n \tag{14}$$

$$\mu_i \xi_i = 0 \qquad \forall i = 1, \ldots, n \tag{15}$$

which are known as the KKT conditions.

The ν-SVC, like the C-SVC, gives the solution as a linear combination of the samples in the feature space (9), called the SV expansion. The α_i that are non-zero correspond to a constraint (5) which is precisely met.

The regular way of solving Support Vector Machines is by substituting (9) to (12) into L in (8), leaving us with the following quadratic optimization problem for ν-SV classification:

$$\max_{\alpha_i} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{16}$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{n}, \tag{17}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{18}$$

$$\sum_{i=1}^{n} \alpha_i = \nu. \tag{19}$$

The $k(\cdot,\cdot)$ represents a dot product (kernel) of two variables in the feature space, $k(\cdot,\cdot) = \boldsymbol{\phi}^T(\cdot)\boldsymbol{\phi}(\cdot)$. The conditions for any function to be a kernel in a Hilbert space is given by the Mercer theorem [3].

The resulting decision function can be expressed as a linear combination of kernels:

$$\begin{aligned} f(x) = \operatorname{sgn}\, \left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + b\right) = \\ = \operatorname{sgn}\, \left(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{\phi}^T(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}) + b\right) = \operatorname{sgn}\, \left(\sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b\right). \end{aligned} \tag{20}$$

Therefore, we do not need to specify the whole nonlinear mapping only its kernel.

To compute the threshold b and the margin parameter ρ, we consider two sets $S_\pm$, containing SVs $\mathbf{x}_i$ with $0 < \alpha_i < 1/n$ and $y_i = \pm 1$, respectively. We choose $s = \min(|S_+|, |S_-|)$, and limit the larger $S_\pm$ set to contain s elements. Then, due to the KKT conditions, (5) becomes an equality with $\xi_i = 0$ for all the samples in $S_\pm$. Hence, in terms of kernels,

$$b = -\frac{1}{2s} \sum_{\mathbf{x} \in S_+ \cup S_-} \sum_{j=1}^{n} \alpha_j y_j k(\mathbf{x}, \mathbf{x}_j), \tag{21}$$

$$\rho = \frac{1}{2s}\Big(\sum_{\mathbf{x} \in S_+} \sum_{j=1}^{n} \alpha_j y_j k(\mathbf{x}, \mathbf{x}_j) - \sum_{\mathbf{x} \in S_-} \sum_{j=1}^{n} \alpha_j y_j k(\mathbf{x}, \mathbf{x}_j)\Big). \tag{22}$$

Note that for the decision function, only b is actually required.

In the case that either S_+ or S_- are the empty set, they will be, respectively, formed by a one element set:

$$S_+ = \operatorname*{argmax}_{\mathbf{x}_i | \alpha_i \neq 0, y_i = 1} \left\{\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\right\}$$

and

$$S_- = \operatorname*{argmin}_{\mathbf{x}_i | \alpha_i \neq 0, y_i = -1} \left\{\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\right\}$$

as detailed in [8] for the C-SVM.

A connection to standard SV classification, and a somewhat surprising interpretation of the regularization parameter C, is described by the following result:

Proposition 2 (Connection ν-SVC — C-SVC [16]) *If ν-SV classification leads to $\rho > 0$, then C-SV classification, with C set a priori to $1/\rho$, leads to the same decision function.*

The proof of this proposition can be found in [16]. This proposition ensures that any ν providing a non-trivial solution[1] with $\rho \neq 0$, a C value can be obtained for the C-SVC formulation that will lead to the same solution obtained with such ν, up to a scaling factor, which is the value of C. For further details on the connection between ν-SVMs and C-SVMs see [7, 1].

3 Limitation in the range of ν

A complete account of the relation between the C-SVM and ν-SVM has been given in [4], where they have shown that the value of ν can not always take the full range from 0 to 1. They have stated and proven the following theorem in which the maximum and minimum value of ν are bounded:

Theorem 1 *We can define*

$$\nu_* = \lim_{C \to \infty} \frac{1}{nC} \sum_{i=1}^{n} \alpha_i^C \tag{23}$$

and

$$\nu^* = \lim_{C \to 0} \frac{1}{nC} \sum_{i=1}^{n} \alpha_i^C \tag{24}$$

where α_i^C are the Lagrange multipliers associated with the constraints in (2) *in the C-SVM and $\nu_* > 0$ and $\nu^* \leq 1$. For any $\nu > \nu^*$* (16) *is infeasible and for any $0 < \nu \leq \nu_*$* (16) *is feasible with zero optimal objective value (the trivial solution). For $\nu_* < \nu \leq \nu^*$* (16) *is feasible and its solution is equal) to the solution of the dual of* (1)*, up to a scaling factor ($\alpha_i^C = Cn\alpha_i^\nu$).*

We will not enter in the demonstration of the theorem, which is detailed in [4], but we will give some intuitions of the results provided by the theorem. The minimal value of ν for which (4) is nonzero (greater than 0) was discussed earlier in this section, once ρ becomes zero (with $\mathbf{w} = 0$ and $b = 0$) there is no incentive in the objective functional (4) to make ρ go negative. Therefore, it can be seen that this will only happen if the kernel matrix $\mathbf{H}$ ($(\mathbf{H})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$) is singular (not full-rank), because otherwise (16) will only be zero iff $\alpha_i = 0\ \forall i$. The solutions of the ν-SVM in which ν lies between ν_*

[1] We understand by a trivial solution a value of ν that forces $\rho = 0$ and, consequently, $\mathbf{w} = \mathbf{0}$ and $b = 0$.

and ν^* are feasible and meaningful, and its relationship with C-SVM is the one stated by Proposition 2.

Finally, the infeasibility of (16) for ν greater than ν^* is easily understood when examining restrictions (18) and (19). These two restrictions enforce that the sum of the α_i of class $+1$ has to be equal to $\nu/2$ and equal to the sum of the α_i of class -1. Therefore, as the maximum contribution of each sample to this sum is $1/n$ (see KKT condition (12)), then the maximum value ν can take is $2\min(n^+, n^-)/n$, where n^+ and n^- are, respectively, the number of sample in class $+1$ and -1. If the classes are not balanced ν has to be less than 1, because $2\min(n^+, n^-) < n$.

In this contribution, we are interested in extending the range of ν-SVM for classification. We will first show that extending the range up to 1 is readily obtained and we will dedicate the rest of the paper to detail how the range can be extended for a value of ν less than ν_* without ending in the trivial solution ($\rho = 0$, $\mathbf{w} = 0$ and $b = 0$).

To extended the value of ν up to 1 ($\nu = 1$), we will need that $\sum_{i=1|y_i=1}^{n} \alpha_i = \sum_{i=1|y_i=-1}^{n} \alpha_i = 1/2$, which can be obtained if we set, respectively, $\alpha_i = 1/(2n^+)$ and $\alpha_i = 1/(2n^-)$, if the samples belongs to class $+1$ or class -1. As the maximum value α_i can take is the multiplicative factor for ξ_i in (4), we can then extend the range by considering a different penalty factor for the ξ_i of positive and negative samples, leading to the modification of (4) by:

$$\min_{\mathbf{w},\xi_i,\rho,b} \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{2n^+}\sum_{i=1|y_i=1}^{n} \xi_i + \frac{1}{2n^-}\sum_{i=1|y_i=-1}^{n} \xi_i \tag{25}$$

This modification is similar to the one proposed in [13, 11] for solving the C-SVM with a different cost for positive and negative classes in unbalanced problems, to obtain a good balanced error performance.

Proposition 1 still holds and it will also hold for each class independently. Using (25) the value of ν will be an upper bound for the fraction margin errors of class $+1$ (class -1) and will be a lower bound for the fraction of support vectors of class $+1$ (class -1), which did not hold for the previous formulation.

4 Negative Margin Minimization

We will now address the problem of reducing the value of ν below ν_* without being led to the trivial solution. To do so, we will need to revist the regular SVM solution for linearly separable data sets and try to find a different solution for non-separable sets.

The SVM enforces a solution in which for positive class samples, $\mathbf{x}_i^T\mathbf{w} + b \geq \rho$, and for negative class samples, $\mathbf{x}_i^T\mathbf{w} + b \leq -\rho$, where $\mathbf{w}$ and b define a linear classifier

with $\rho > 0$. Among the solutions that fulfill these requirements the SVM picks the one in which the samples are the furthest apart from the classification boundary (the maximum margin solution) [20], as shown in Figure 2a. To construct the maximum margin solution, the SVM fixes ρ to 1 and minimizes $\|\mathbf{w}\|^2$. For non-separable problems the SVM includes slack variables in the previous constraints and minimizes the one-norm of the slack variables (to approximately minimize the number of training errors), leading to the optimization of (1) subject to (2) and (3).

A typical solution of a non-separable problem is shown in Figure 2b, in which there are 17 SVs. This solution presents the least number of SVs, and no other value of C will reduce it, therefore the minimum value of ν will be between $14/60 < \nu \leq 17/60$. But, analyzing the obtained solution for the separable problem in Figure 2a, one could expect the solution for a non-separable problem to be the one shown in Figure 2c, in which the solution is obtained with the extreme vectors as in Figure 2a, instead of the one shown in Figure 2b. In Figure 2a we have an exclusion zone (no training sample is allowed to have a machine output, $\mathbf{x}_i^T\mathbf{w} + b$, between $-\rho$ and $+\rho$) and in Figure 2c we have an intersection zone, into which samples from both classes can be placed without becoming SVs. This intersection zone can be implemented through the use of the following constraint:

$$\begin{aligned} \mathbf{x}_i^T\mathbf{w} + b &\geq -1 \qquad \text{if } \; y_i = +1 \\ \mathbf{x}_i^T\mathbf{w} + b &\leq +1 \qquad \text{if } \; y_i = -1 \end{aligned}$$

To obtain the maximum margin solution in Figure 2a, we maximize the exclusion zone ($\min \|\mathbf{w}\|^2$) forcing the samples to be as far apart from the classification boundary as possible. To obtain the solution in Figure 2c, we would like the intersection zone to be as small as possible to reduce the number of samples that lie inside it (i.e to minimize the number of potential errors). To reduce the intersection zone, we will have to minimize $1/\|\mathbf{w}\|$ ($\min -\|\mathbf{w}\|^2$).

The maximum of the positive margin and the minimum of the negative margin can be unified in a single optimization problem by looking back at how the maximum margin is constructed. In Figure 2a the solution can be obtained by solving:

$$\max_{\rho,\mathbf{w},b} \frac{\rho}{\|\mathbf{w}\|}$$

subject to

$$y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq \rho$$

As there is a multiplicative factor between $\rho/\|\mathbf{w}\|$, there are infinitely many different solutions that only differ by a scaling factor. To resolve this, one can fix ρ and maximize $1/\|\mathbf{w}\|$ (or minimize $\|\mathbf{w}\|^2$ as the SVM does). Note that this only works for positive

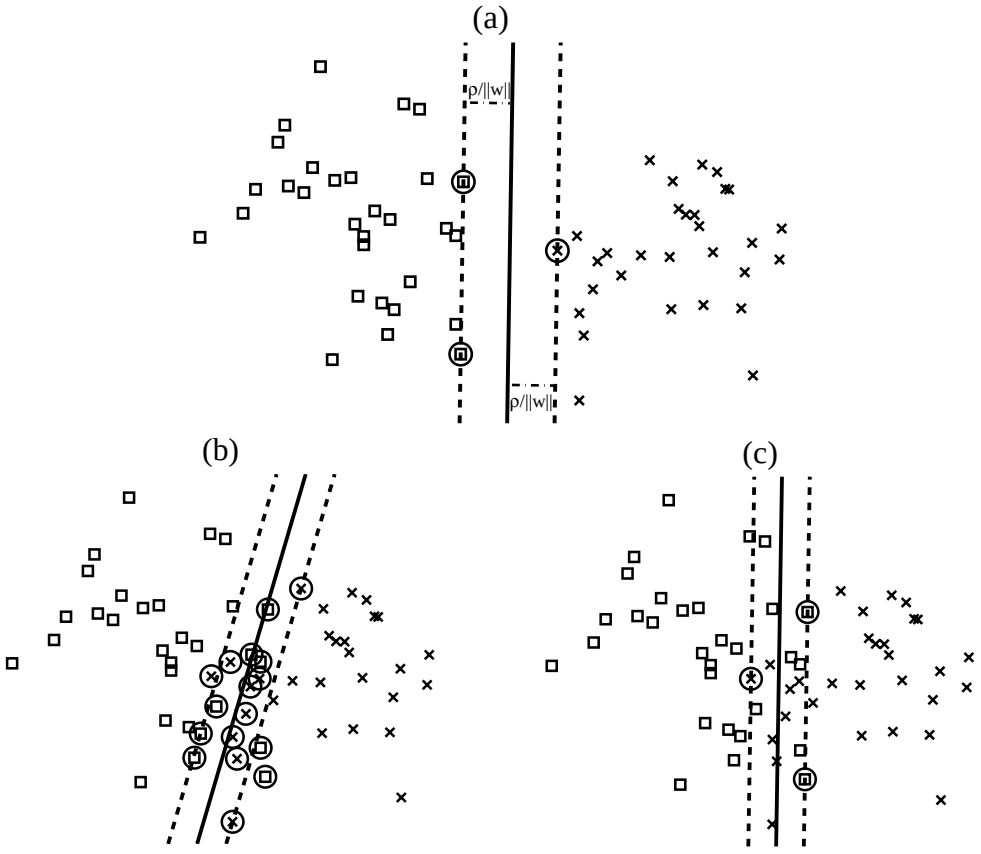

Figure 2: The solid lines represent the classification boundary, the dashed lines the $\pm\rho/\|\mathbf{w}\|$ margins. Class $+1$ is shown by crosses, Class -1 by squares, and the SVs are ringed. In (a) we show the maximum margin solution for a linearly separable data set. In (b) we show the SVM solution for a non-linearly separable data set. In (c) we show the solution with the negative margin classifier in which the solution is constructed by the extreme vectors as in (a).

ρ, since for negative ρ, we would have to minimize $1/\|\mathbf{w}\|$. Alternatively one can fix $\|\mathbf{w}\|$, which accounts for a non-convex constrain, and maximize ρ. If this is the case and the problem is separable we will end up with the maximum margin solution and a positive ρ, for non-separable problems we will end up with a negative ρ and with the least possible intersection zone.

5 Extended ν-SVM

In the previous section, we have motivated how the hard maximum margin can be modified to deal with separable and non-separable training data sets. In this section, we will formulate the new learning problem with a ν-SVM like formulation (soft margin)

that will allow us to control ρ in an intuitive way. We solve the problem:

$$\min_{\rho,\mathbf{w},b,\xi_i} -n\nu\rho + \sum_{i=1}^{n} \xi_i \tag{26}$$

subject to:

$$y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq \rho - \xi_i \qquad \forall i = 1,\ldots,n \tag{27}$$

$$\xi_i \geq 0 \qquad \forall i = 1,\ldots,n \tag{28}$$

$$\frac{1}{2}\|\mathbf{w}\|^2 = 1 \tag{29}$$

The objective function is linear but there is a non-convex constraint (29); therefore we can expect local minima in its solution[2], which have to be dealt with either using several initializations or starting at a controlled point. Note that in this case ν can not run up to 1 unless the classes are balanced. The modification needed is the same one proposed at the end of Section 3, but we have not included it to make the development of the Eν-SVM clearer.

We can solve this problem directly for linear classifiers by linearizing the quadratic constraint in (29). We select a starting point labeled as $\widetilde{\mathbf{w}}$ and we then replace (29) by: $\widetilde{\mathbf{w}}^T\mathbf{w} = 2$. We thus get a linear problem that can be solved using any linear programming tool such as `linprog` from MATLAB®. Once we have computed the solution, we obtain a new $\widetilde{\mathbf{w}}$ and continue iterating until there is no further modification in either $\mathbf{w}$, b or ρ. To construct the new $\widetilde{\mathbf{w}}$ we do not directly use the value of $\mathbf{w}$ due to the linearizing step. We will construct it using a convex combination between $\widetilde{\mathbf{w}}$ and $\mathbf{w}$: $\widetilde{\mathbf{w}} = \gamma\widetilde{\mathbf{w}} + (1-\gamma)\mathbf{w}$. We have found experimentally that $\gamma = \frac{9}{10}$ is a good compromise value. We have written down an algorithmic implementation of the proposed approach in Table 2. The initial value of $\widetilde{\mathbf{w}}$ can be a random guess (not very convenient) or a solution to the ν-SVC with a ν above ν_*. The benefits of using this starting point is that we will be looking for a solution close to the best solution provided by the ν-SVC, which should do as a good starting point for avoiding local minima (or at least a bad local minima).

5.1 Kernelization in the dual

The above learning machine can be used to solve nonlinear problems using kernels. In order to construct a nonlinear classifier, we need to map non-linearly the vectors

[2]If the two data sets are not linearly separable, the problem of finding two half-spaces with parallel boundary and minimal overlap, containing the respective classes, is not convex. So this is not a weakness of the chosen aproach, but an important aspect of the considered problem.

Table 2: Eν-SVC algoritmic implementation.

0. Initialize $\widetilde{\mathbf{w}}$.
1. Solve the linear problem:

$$\min_{\rho,\mathbf{w},b,\xi_i} -n\nu\rho + \sum_{i=1}^{n} \xi_i$$

subject to:

$$y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq \rho - \xi_i,$$
$$\xi_i \geq 0 \text{ and } \widetilde{\mathbf{w}}^T\mathbf{w} = 2.$$

2. Compute $\widetilde{\mathbf{w}} = \gamma\widetilde{\mathbf{w}} + (1-\gamma)\mathbf{w}$.
3. If $\widetilde{\mathbf{w}} = \mathbf{w}$ end, otherwise go to Step 1.

$\mathbf{x}_i$ to the feature space, through a nonlinear transformation $\boldsymbol{\phi}(\cdot)$. To solve (26) with a possibly unknown $\boldsymbol{\phi}(\cdot)$, we will introduce it in the constraints in (27)-(29) using Lagrange multipliers. This requires us to minimize

$$L_p = -n\nu\rho + \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(y_i(\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b) - \rho + \xi_i\right) - \sum \mu_i\xi_i + \lambda\left(\frac{1}{2}\|\mathbf{w}^2\| - 1\right) \quad (30)$$

with respect to $\rho, \mathbf{w}, b, \xi_i$ and maximize it with respect to the Lagrange multipliers, α_i, μ_i and λ. The solution to this problem is given by the Karush-Kuhn-Tucker (KKT)

theorem [9], which imposes the following conditions over (30):

$$\frac{\partial L_p}{\partial \rho} = -n\nu - \sum_{i=1}^{n} \alpha_i = 0 \tag{31}$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = \lambda \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i) = \mathbf{0} \tag{32}$$

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{33}$$

$$\frac{\partial L_p}{\partial \xi_i} = 1 - \alpha_i - \mu_i = 0 \qquad \forall i = 1, \ldots, n \tag{34}$$

$$\alpha_i, \mu_i \geq 0 \qquad \forall i = 1, \ldots, n \tag{35}$$

$$\alpha_i \{ y_i (\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w} + b) - \rho + \xi_i \} = 0 \qquad \forall i = 1, \ldots, n \tag{36}$$

$$\mu_i \xi_i = 0 \qquad \forall i = 1, \ldots, n \tag{37}$$

$$\lambda \left(\frac{1}{2} \|\mathbf{w}\|^2 - 1 \right) = 0 \tag{38}$$

The dual formulation, which is the usual way of solving SVMs, cannot be used for solving the Eν-SVC, because it is not a convex problem and the dual formulation only holds for convex problems. But in our problem if $\lambda > 0$, the constrain in (29) can be transformed to $\frac{1}{2}\|\mathbf{w}\|^2 \leq 1$, which makes the problem convex. Therefore for positive λ, we can obtain the dual formulation by substituting (31), (32), (33), (34) into (30), in which one needs to maximize with respect to α_i and λ:

$$L_D = -\frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \boldsymbol{\phi}^T(\mathbf{x}_i) \boldsymbol{\phi}(\mathbf{x}_j) - \lambda \tag{39}$$

subject to (31), (33) and $0 \leq \alpha_i \leq 1$. This problem cannot be easily solved because λ depends on α. This dependence can be obtain using (29) and (32):

$$\lambda = + \sqrt{\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)} \tag{40}$$

To solve (39), we fix λ for example equal to 1. If in the solution the optimal objective value is nonzero one can compute λ as shown by (40). If it is zero, it will mean that the solution of the Eν-SVC will be obtained with a negative λ and the dual cannot be used. If we replace this new value in the functional the solution will be the same. Therefore, if λ is positive we only need to solve:

$$L_D = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \boldsymbol{\phi}^T(\mathbf{x}_i) \boldsymbol{\phi}(\mathbf{x}_j)$$

which is the same functional used for the ν-SVM. Therefore if λ is positive the solution for the ν-SVC and the Eν-SVC is the same one up to a scaling factor.

For negative values of λ, however, if one fixes λ and tries to optimize (39), will obtain $\sum_i \sum_j y_i y_j \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) = 0$ and one therefore one will not get a feasible solution. This is something already known, because for negative λ the constraint in (29) is non-convex and the dual problem can not be stated. Therefore, if the dual ends with a nonzero λ the solution of the Eν-SVC and ν-SVC are equal, up to a scaling factor, and if $\lambda = 0$ the solution of the Eν-SVC will have to be obtained by other means.

5.2 Kernelization in the primal

We need a different approach to solve the Eν-SVC using kernels and without explicitly computing λ. We will use a kernel PCA (KPCA) decomposition of the input data using the selected kernel. KPCA [15, 14] computes the principal components of the mapped data and it will give at most n (the number of training samples) principal components. Once we have performed KPCA, we will map the data onto these components, so we will have a finite dimensional representation of each training vector in the feature space. Consequently, we can use the KPCA representation of the input data to train a linear classifier using (26), which will become a nonlinear machine in the original representation of the data (input space). A similar procedure has been used for incorporating invariances in SVMs [5].

When we have solved the linear problem with the KPCA representation, we can obtain the values of α_i and λ using the KKT conditions and the obtained solution ($\mathbf{w}$, b, ρ and ξ_i), which we use to distinguish between solutions that can be obtained with the classic ν-SVM or SVM ($\lambda > 0$) and those solutions which are not feasible with them ($\lambda < 0$).

We believe that this learning algorithm can be used together with the new trend in the machine learning community in which the kernel matrix is learnt instead [10]. Once the learning matrix has been optimized for a given problem, if it is not full-rank, all the values of ν might not be feasible and the Eν-SVC will provide a wider range of possible solutions to be evaluated.

The theorem in [16] in which it is stated that ν is an upper bound in the fraction of bounded SVs and a lower bound in the fraction of SVs also holds for this learning algorithm.

We conclude this section by noting that a this approach is related to one used for boosting [12] (related to 1-norm SVMs), but in which ρ was constrained positive, although it is not a necessary constraint for solving the problem.

6 Experiments

We have shown in the previous section that the ν-SVM can be modified to be able to extend the range of ν value down to zero, by allowing ρ to take negative values. We first show a simple 2D example, in which we can picture the solution given by the Extended ν-SVM. Class $+1$ samples are drawn equally from two normal distributions of means [2 0] and [0 2] and Class -1 from a normal distibution of zero mean. Both classes are equally likely and their variance matrices are the identity. In Figure 3, we show the obtained solution using a linear classifier for different values of ν. Of the ones shown, the solution in Figure 3a ($\nu = 0.69$) is the only one that can be achieved by the classic ν-SVM and SVM described in [14]. The solution depicted in Figure 3b ($\nu = 0.51$) actually presents a ρ which is still positive, but the associated value of λ is negative, so this solution cannot be obtained with the classic ν-SVM. In Figure 3c ($\nu = 0.33$), we show a solution for $\rho \simeq 0$ in which the solution is constructed by all and only incorrectly classified samples. Finally, in Figure 3d ($\nu = 0.01$), we show the hard negative margin solution in which all the slack variables are equal to zero.

We have also solved this problem with two types of nonlinear kernels for $\nu = 0.51$, which is a value unreachable for the classic ν-SVM. In Figure 4a we show the obtained solution with an inhomogeneous polynomial kernel of second degree and in Figure 4b, an RBF kernel with standard deviation $\sigma = 8$. The probability of error of the Bayesian classifier is 0.200 and the achieved solution for the polynomial and RBF kernels are, respectively, 0.205 and 0.217, which are the best possible results for the whole range of ν.

We performed experiments on a real application using a dataset from the UCI Machine Learning Repository (http://www.ics.uci.edu/~ mlearn). We have chosen the Liver-disorder database, because it is a noisy dataset with a large classification error, which significantly limits the minimum value of ν for the classic ν-SVM. We preprocessed the dataset so every feature has zero mean and unit standard deviation. Then, we solved the problem using a linear kernel and measured success using 5-fold cross-validation over different possible values from the whole range of ν.

In Figure 5, we have plotted the mean training and test error for the 5 splits of the data. In this problem as the classification error is so high the allowed values of ν in the classic ν-SVM formulation are restricted to the two highest ($\nu = 0.76$ and $\nu = 0.81$). Therefore the best solution for the classic ν-SVM is obtained for $\nu = 0.76$, achieving a probability of error of 0.325. The Extended ν-SVM is able to obtain the solution for any ν value so we can find that the best solution is found for $\nu = 0.41$ with a probability of error of 0.293, a reduction of more than three percentage points.

We have reported in Table 3 the values of ρ, λ and the fraction of SVs for the

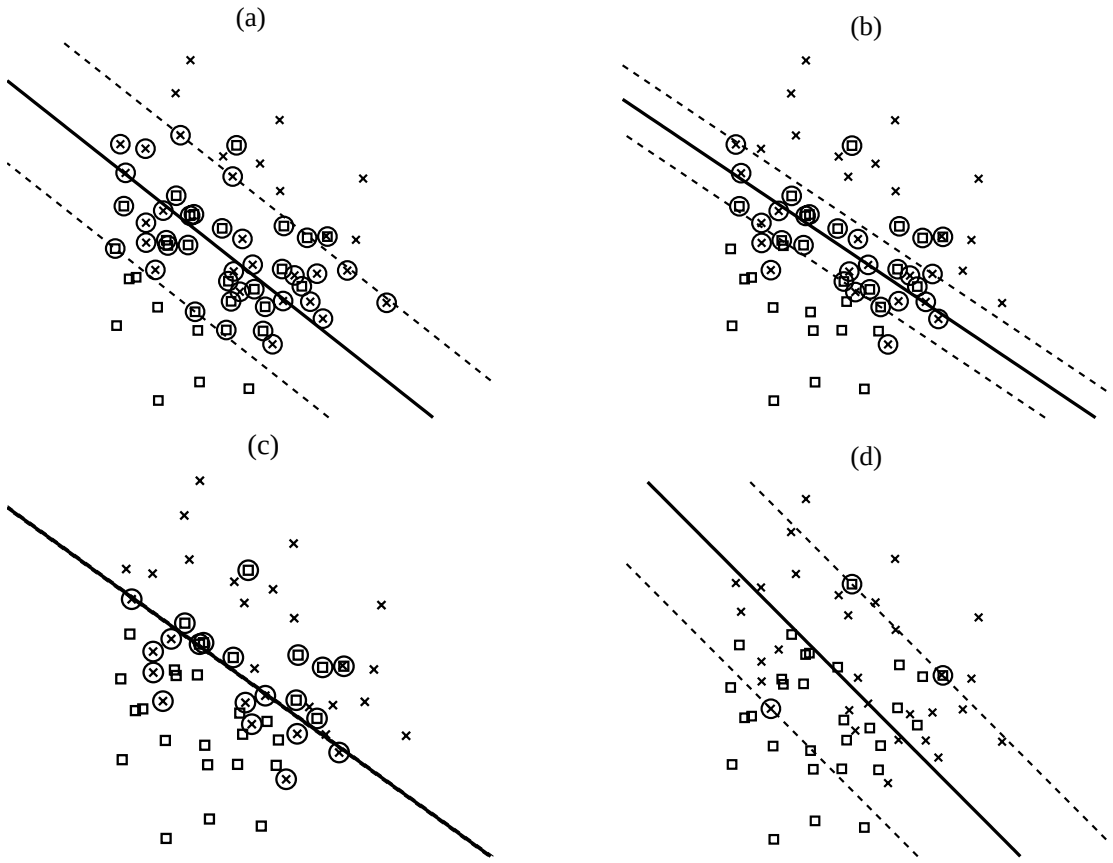

Figure 3: The solid lines represent the classification boundary and the dashed lines the $\pm\rho$ margins. Class $+1$ is shown by crosses, Class -1 by squares, and the SVs are ringed. In (a) we show the solution for $\nu = 0.69$ ($\rho = 1.59$ and $\lambda = 1.78$). In (b) we show the solution for $\nu = 0.51$ ($\rho = 0.69$ and $\lambda = -4.60$). In (c) we show the solution for $\nu = 0.33$ ($\rho = -0.04$ and $\lambda = -7.01$). And, in (d) we show the solution for $\nu = 0.01$ ($\rho = -1.79$ and $\lambda = -0.54$).

tested values of ν. In it, we can see that for $\nu \leq 0.71$ the values of λ are negative and therefore this solution cannot be achieved by the classic ν-SVM nor by the regular SVM (recall that the two types of SVMs can be shown to provide identical solutions for suitable parameter values [14]). It can be also pointed out that the value of ρ remains positive for ν ranging from 0.71 to 0.31. In this case we will have a positive margin but we will want to minimize it, because the number of SVs correctly classified do not outnumber the incorrectly classified ones, but there are still training samples that are correctly classified and are SVs. For values of $\nu \leq 0.26$ the solution will be constructed exclusively using incorrectly classified samples and, consequently, the margin ρ will be negative.

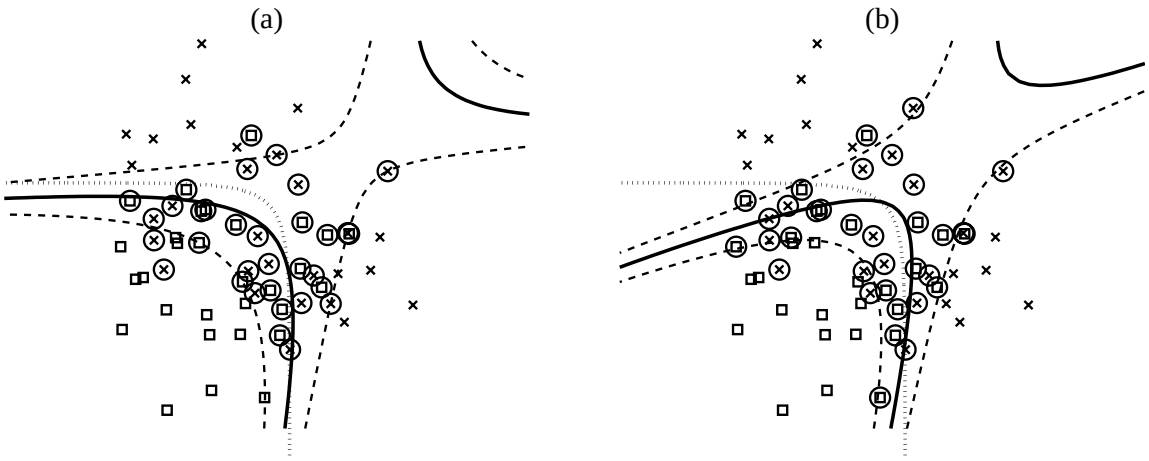

Figure 4: We show the results for polynomial and RBF kernels for $\nu = 0.51$, respectively in (a) and (b). The solid lines represent the classification boundary and the dashed lines the $\pm\rho$ margins. Class $+1$ is shown by crosses, Class -1 by squares, and the SVs are ringed. The dotted lines represent the Bayes classifier.

ν	0.01	0.16	0.26	0.31	0.36	0.41	0.56	0.71	0.76	0.81
ρ	-1.12	-0.27	-0.07	0.02	0.09	0.15	0.34	0.54	0.69	1.09
λ	-0.78	-13.3	-15.7	-15.8	-15.5	-14.6	-9.59	-0.63	3.59	9.46
fSVs	.025	.167	.265	.312	.361	.420	.569	.716	.767	.811

Table 3: The values of ρ, λ, and the fraction of support vectors (fSVs) mean values for the 5 split of the data.

7 Conclusions and Further Work

In this paper, we have reinterpreted how the maximum margin can be constructed for separable training data sets, and so we were able to obtain an alternative SVM solution for non-separable data sets. Moreover, we have developed a formulation like the ν-SVM, Extended ν-SVM, in which the former is contained as a special case (for $\lambda > 0$). This extended ν-SVM allows us to construct the machine with any number of SVs and if the best possible solution lies in the zone in which $\lambda < 0$ we can obtain a solution that can be better than the best solution achieved by the classic SVM. And it is possible that the best solution can be obtained for a $\lambda < 0$ because it has been recently proven that the optimal ν is equal to twice the Bayes error [17].

There are various possible extensions of the work presented. The training procedure has to be improved so it is easier to solve when using kernels. And, there are also possible theoretical extensions, to prove bounds and convergence rates as well as to

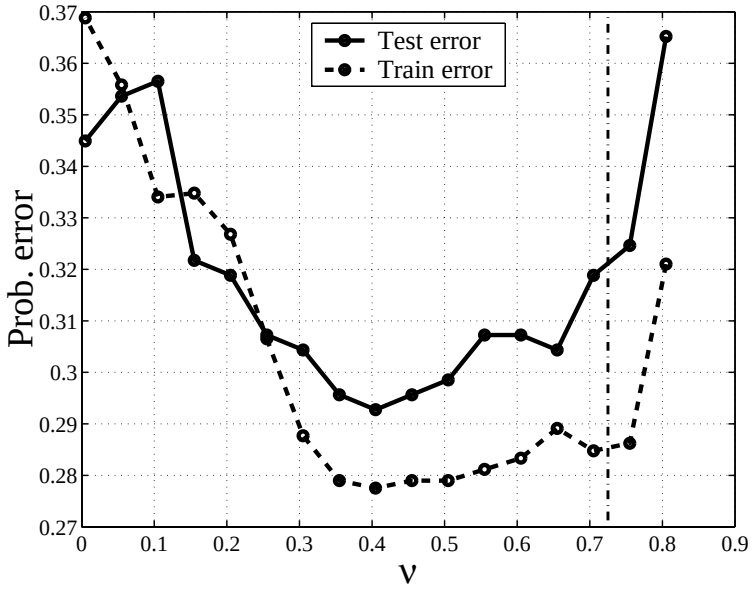

Figure 5: The training and test error are represented, respectively, by the dashed and solid lines for the liver disorder database. The dash-dotted line separates the results of the classic ν-SVM (to the right) and those that can be obtained only with the Extended ν-SVM.

determine the role of the Lagrange multiplier λ.

Acknowledgements

We would like to express our gratitude to Chih-Jen Lin for its helpful comments for making the paper more readable and avoiding the introduction of major mistakes.

References

[1] K. Bennett and E. Bredensteiner. Duality and geometry in svm classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 57–64, San Francisco, CA, December 2000.

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

[3] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Knowledge Discovery and Data Mining*, 2(2):121–167, 1998.

[4] C.-C. Chang and C.-J. Lin. Training ν-support vector classifiers: Theory and algorithms. *Neural Computation*, 13(9):2119–2147, 2001.

[5] O. Chapelle and B. Scholkopf. Incorporating invariances in non-linear SVMs. In *Advances in Neural Information Processing Systems 14*, Cambridge, MA, December 2001. M.I.T. Press.

[6] C. Cortes and V. N. Vapnik. Support Vector Networks. *Machine Learning*, 20:273–297, 1995.

[7] D.J. Crisp and C.J.C. Burges. A geometric interpretation of nu-SVM classifiers. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA, November 1999. M.I.T. Press.

[8] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machine.* Cambridge University Press, 1999.

[9] R. Fletcher. *Practical Methods of Optimization.* Wiley, Chichester, 2^{nd} edition, 1987.

[10] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semi-definite programming. In *Proceedings of the nineteenth International Conference on Machine Learning*, pages 57–64, Sidney, Australia, 2002.

[11] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

[12] G. Rätsch, B. Schölkopf, A. J. Smola, S. Mika, T. Onoda, and K.-R. Müller. Robust ensemble learning. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 207–220. M.I.T. Press, Cambridge, (MA), 2000.

[13] M. Schmidt and H. Gish. Speaker identification via support vector classifiers. In *Proceedings ICASSP'96*, pages 105–108, Atlanta, GA, 1996.

[14] B. Schölkopf and A. Smola. *Learning with kernels.* M.I.T. Press, 2001.

[15] B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(15):1299–1319, 1998.

[16] B. Schölkopf, A. Smola, R. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, May 2000.

[17] I. Steinwart. On the optimal choice for the nu-support vector machines. Technical report, Driedrich-Schiller-Universitat, 2002.

[18] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer–Verlag, 1982.

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

[20] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.