Estimating the Support of a High-Dimensional Distribution

Bernhard Schölkopf[*], John C. Platt[‡],

John Shawe-Taylor[†], Alex J. Smola[§],

Robert C. Williamson[§]

[*] Microsoft Research Ltd, 1 Guildhall Street, Cambridge CB2 3NH, UK

[‡] Microsoft Research, 1 Microsoft Way, Redmond, WA, USA

[†] Royal Holloway, University of London, Egham, UK

[§] Department of Engineering, Australian National University,
Canberra 0200, Australia

bsc@microsoft.com, jplatt@microsoft.com, john@dcs.rhbnc.ac.uk
Alex.Smola@anu.edu.au, Bob.Williamson@anu.edu.au

27 November 1999

Technical Report
MSR-TR-99-87

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Abstract

Suppose you are given some dataset drawn from an underlying probability distribution P and you want to estimate a "simple" subset S of input space such that the probability that a test point drawn from P lies outside of S is bounded by some a priori specified ν between 0 and 1.

We propose a method to approach this problem by trying to estimate a function f which is positive on S and negative on the complement. The functional form of f is given by a kernel expansion in terms of a potentially small subset of the training data; it is regularized by controlling the length of the weight vector in an associated feature space. The expansion coefficients are found by solving a quadratic programming problem, which we do by carrying out sequential optimization over pairs of input patterns. We also provide a preliminary theoretical analysis of the statistical performance of our algorithm.

The algorithm is a natural extension of the support vector algorithm to the case of unlabelled data.

Keywords. Support Vector Machines, Kernel Methods, Density Estimation, Unsupervised Learning, Novelty Detection

1 Introduction

During recent years, a new set of kernel techniques for supervised learning has been developed (Vapnik, 1995; Schölkopf et al., 1999a). Specifically, support vector (SV) algorithms for pattern recognition, regression estimation and solution of inverse problems have received considerable attention.

There have been a few attempts to transfer the idea of using kernels to compute inner products in feature spaces to the domain of unsupervised learning. The problems in that domain are, however, less precisely specified. Generally, they can be characterized as estimating *functions* of the data which tell you something interesting about the underlying distributions. For instance, kernel PCA can be characterized as computing functions which on the training data produce unit variance outputs while having minimum norm in feature space (Schölkopf et al., 1999b). Another kernel-based unsupervised learning technique, regularized principal manifolds (Smola et al., 1999), computes functions which give a mapping onto a lower-dimensional manifold minimizing a regularized quantization error. Clustering algorithms are further examples of unsupervised learning techniques which can be kernelized (Schölkopf et al., 1999b).

An extreme point of view is that unsupervised learning is about estimating densities. Clearly, knowledge of the density of P would then allow us to solve whatever problem can be solved on the basis of the data.

The present work addresses an easier problem: it proposes an algorithm which computes a binary function which is supposed to capture regions in input space where the probability density lives (its support), i.e. a function such that most of the data will live in the region where the function is nonzero (Schölkopf et al., 1999). In doing so, it is in line with Vapnik's principle never to solve a problem which is more general than the one we actually need to solve. Moreover, it is applicable also in cases where the density of the data's distribution is not even well-defined, e.g. if there are singular components.

The article is organized as follows. After a review of some previous work in Sec. 2, we propose SV algorithms for the considered problem. Sec. 4 gives details on the implementation of the optimization procedure, followed by theoretical results characterizing the present approach. In Sec. 6, we apply the algorithm to artificial as well as real-world data. We conclude with a discussion.

2 Previous Work

Part of the motivation for the present work was a paper of Ben-David and Lindenbaum (1997). It turns out that there is a considerable amount of prior work in the statistical literature, and in this section we briefly summarise that. We do not attempt a detailed comparison of the proof techniques of the specific results achieved, but confine ourselves to placing the previous work in context.

In order to summarize the methods, it is convenient to introduce the following definition of a (multi-dimensional) quantile function (introduced by Einmal and Mason (1992)). Let $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ be i.i.d. random variables in a set $\mathcal{X}$ with distribution P. Let $\mathcal{C}$ be a class of measurable subsets of $\mathcal{X}$ and let λ be a real-valued function defined on $\mathcal{C}$. The *quantile function* with respect to $(P, \lambda, \mathcal{C})$ is

$$U(\alpha) = \inf\{\lambda(C) \colon P(C) \geq \alpha, C \in \mathcal{C}\} \quad 0 < \alpha \leq 1.$$

If P_ℓ is the empirical distribution ($P_\ell(C) = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathbf{1}_C(\mathbf{x}_i)$), the *empirical quantile function* is

$$U_\ell(\alpha) = \inf\{\lambda(C) \colon P_\ell(C) \geq \alpha, C \in \mathcal{C}\} \quad 0 < \alpha \leq 1.$$

We denote by $C(\alpha)$ and $C_\ell(\alpha)$ the (not necessarily unique) $C \in \mathcal{C}$ that attains the infimum (when it is achievable). The most common choice of λ is Lebesgue measure, in which case $C(\alpha)$ is the minimum volume $C \in \mathcal{C}$ that contains at least a fraction α of the probability mass. We will assume λ is Lebesgue measure from here on. Estimators of the form $C_\ell(\alpha)$ are called *minimum volume estimators*.

Estimating the support of a density. Observe that for $\mathcal{C}$ being all Borel measurable sets, $C(1)$ is the *support* of the density p corresponding to P, assuming it exists. (Note that $C(1)$ is well defined even when p does not exist.) For smaller classes $\mathcal{C}$, $C(1)$ is the minimum volume $C \in \mathcal{C}$ containing the support of p. The problem of estimating $C(1)$ appears to have first been studied by Geffroy (1964) who considered $\mathcal{X} = \mathbb{R}^2$ with piecewise constant estimators. There have been a number of works studying a natural nonparametric estimator of $C(1)$ (e.g. Chevalier (1976); Devroye and Wise (1980); Cuevas (1990); see (Gayraud, 1997) for further references). The nonparametric estimator is simply

$$\hat{C}_\ell = \bigcup_{i=1}^{\ell} B(\mathbf{x}_i, \epsilon_n) \tag{1}$$

where $B(\mathbf{x}, \epsilon)$ is the $l_2(\mathcal{X})$ ball of radius ϵ centered at $\mathbf{x}$ and $(\epsilon_n)_n$ is an appropriately chosen decreasing sequence. Devroye and Wise (1980) showed the asymptotic consistency of (1) with respect to the symmetric difference between $C(1)$ and $\hat{C}_\ell$. Cuevas (1990) did the same, but for Hausdorff distance. Cuevas and Fraiman (1997) studied the asymptotic consistency of a *plug-in* estimator of $C(1)$: $\hat{C}^{\mathrm{plug-in}} = \{\mathbf{x}\colon \hat{p}_\ell(\mathbf{x}) > 0\}$ where $\hat{p}_\ell$ is a kernel density estimator. In order to avoid problems with $\hat{C}^{\mathrm{plug-in}}$ they actually analyzed $\hat{C}_\beta^{\mathrm{plug-in}} := \{\mathbf{x}\colon \hat{p}_\ell(\mathbf{x}) > \beta_\ell\}$ where $(\beta_\ell)_\ell$ is an appropriately chosen sequence. Clearly for a given distribution, α is related to β, but this connection can not be readily exploited by this type of estimator.

The most recent work relating to the estimation of $C(1)$ is by Gayraud (1997) who has made an asymptotic minimax study of estimators of *functionals* of $C(1)$. Two examples are $\operatorname{vol} C(1)$ or the center of $C(1)$. (See also (Korostelev and Tsybakov, 1993, Chapter 8).)

Estimating high probability regions ($\alpha \neq 1$). Turning to the case where $\alpha < 1$, it seems the first work was reported by Sager (1977) and then Hartigan (1987) who considered $\mathcal{X} = \mathbb{R}^2$ with $\mathcal{C}$ being the class of closed convex sets in $\mathcal{X}$. (They actually considered density contour clusters; see below for a definition.) Nolan (1991) considered higher dimensions with $\mathcal{C}$ being the class of ellipsoids.

Tsybakov (1997) has studied an estimator based on piecewise polynomial approximation of $C(\alpha)$ and has shown it attains the asymptotically minimax rate for certain classes of densities p.

Polonik (1997) has studied the estimation of $C(\alpha)$ by $C_\ell(\alpha)$. He derived asymptotic rates of convergence in terms of various measures of richness of $\mathcal{C}$. He considers both VC classes and classes with a log ϵ-covering number with bracketing of order $O(\epsilon^{-r})$ for $r > 0$. He also summarizes a number of other previous works on minimum volume estimators which we have not mentioned here.

Polonik (1995b) has also studied the use of the "excess mass approach" (Müller, 1992) to construct an estimator of "generalized α-clusters" which are related to $C(\alpha)$.

Define the *excess mass over* $\mathcal{C}$ *at level* α as

$$E_{\mathcal{C}}(\alpha) = \sup\{H_\alpha(C) \colon C \in \mathcal{C}\}$$

where $H_\alpha(\cdot) = (P - \alpha\lambda)(\cdot)$ and again λ denotes Lebesgue measure. Any set $\Gamma_{\mathcal{C}}(\alpha) \in \mathcal{C}$ such that

$$E_{\mathcal{C}}(\alpha) = H_\alpha(\Gamma_{\mathcal{C}}(\alpha))$$

is called a *generalized* α*-cluster in* $\mathcal{C}$. Replace P by P_ℓ in these definitions to obtain their empirical counterparts $E_{\ell,\mathcal{C}}(\alpha)$ and $\Gamma_{\ell,\mathcal{C}}(\alpha)$. In other words, his estimator is

$$\Gamma_{\ell,\mathcal{C}}(\alpha) = \arg\max\{(P_\ell - \alpha\lambda)(C) : C \in \mathcal{C}\}$$

where the $\max$ is not necessarily unique. Now whilst $\Gamma_{\ell,\mathcal{C}}(\alpha)$ is clearly different to $C_\ell(\alpha)$, it is related to it in that it attempts to find small regions with as much excess mass (which is similar to finding small regions with a given amount of probability mass). Actually $\Gamma_{\ell,\mathcal{C}}(\alpha)$ is more closely related to the determination of *density contour clusters* at level α:

$$c_p(\alpha) := \{\mathbf{x} \colon p(\mathbf{x}) \geq \alpha\}.$$

Simultaneously, and independently, Ben-David and Lindenbaum (1997) studied the problem of estimating $c_p(\alpha)$. They too made use of VC classes but stated their results in a stronger form which is meaningful for finite sample sizes.

Finally we point out a curious connection between minimum volume sets of a distribution and its differential entropy in the case that $\mathcal{X}$ is one dimensional. Suppose X is a one dimensional random variable with density p. Let $S = C(1)$ be the support of p and define the *differential entropy* of X by

$$h(X) = -\int_S p(x)\log p(x)dx.$$

For $\epsilon > 0$ and $\ell \in \mathbb{N}$, define the *typical set* $A_\epsilon^{(\ell)}$ with respect to p by

$$A_\epsilon^{(\ell)} = \{(x_1,\dots,x_\ell) \in S^\ell \colon | -\tfrac{1}{\ell}\log p(x_1,\dots,x_\ell) - h(X)| \leq \epsilon\},$$

where $p(x_1,\dots,x_\ell) = \prod_{i=1}^{\ell} p(x_i)$.

If $(a_\ell)_\ell$ and $(b_\ell)_\ell$ are sequences, the notation $a_\ell \doteq b_\ell$ means $\lim_{\ell\to\infty} \frac{1}{\ell} \log \frac{a_\ell}{b_\ell} = 0$. (Cover and Thomas, 1991, p.227) show that for all $\epsilon, \delta < \frac{1}{2}$, then

$$\operatorname{vol} A_\epsilon^{(\ell)} \doteq \operatorname{vol} C_\ell(1-\delta) \doteq 2^{\ell h}.$$

They point out that this result "indicates that the volume of the smallest set that contains most of the probability is approximately $2^{\ell h}$. This is a ℓ-dimensional volume, so the corresponding side length is $(2^{\ell h})^{1/\ell} = 2^h$. This provides an interpretation of differential entropy."

Applications. A number of applications have been suggested for these techniques. They include problems in medical diagnosis (Tarassenko et al., 1995), marketing (Ben-David and Lindenbaum, 1997), condition monitoring of machines (Devroye and Wise, 1980), estimating manufacturing yields (Stoneking, 1999), econometrics and generalized nonlinear principal curves (Tsybakov, 1997; Korostelev and Tsybakov, 1993), regression and spectral analysis (Polonik, 1997), tests for multimodality and clustering (Polonik, 1995b) and others (Müller, 1992).

Polonik (1995a) has shown how one can use estimators of $C(\alpha)$ to construct density estimators. The point of doing this is that it allows one to encode a range of prior assumptions about the true density p that would be impossible to do within the traditional density estimation framework. He has shown asymptotic consistency and rates of convergence for densities belonging to VC-classes or with a known rate of growth of metric entropy with bracketing.

Relationship with the Present Work. The present paper describes an algorithm which finds regions close to $C(\alpha)$. Our class $\mathcal{C}$ is defined implicitly via a kernel k and the smoothness of the boundary of C can be controlled by the choice of k. We do not try and find *the* minimum volume such region. On the other hand, our algorithm has tractable computational complexity, even in several variables. Our theory, which uses very similar tools to those used by Polonik, gives results that we expect will be of more use in a finite sample size setting.

3 Algorithms

We first introduce terminology and notation conventions. We consider training data

$$\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathcal{X}, \tag{2}$$

where $\ell \in \mathbb{N}$ is the number of observations, and $\mathcal{X}$ is some set. For simplicity, we think of it as a compact subset of $\mathbb{R}^N$. Let Φ be a feature map $\mathcal{X} \to F$, i.e. a map into a dot product space F such that the dot product in the image of Φ can be computed by evaluating some simple kernel (Boser et al., 1992; Vapnik, 1995; Schölkopf et al., 1999a)

$$k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})), \tag{3}$$

such as the Gaussian kernel

$$k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/c}. \tag{4}$$

Indices i and j are understood to range over $1, \ldots, \ell$ (in compact notation: $i, j \in [\ell]$). Bold face greek letters denote ℓ-dimensional vectors whose components are labelled using normal face typeset.

In the remainder of this section, we shall develop an algorithm which returns a function f that takes the value $+1$ in a "small" region capturing most of the data points, and -1 elsewhere. Our strategy is to map the data into the feature space corresponding to the kernel, and to separate them from the origin with maximum margin. For a new point $\mathbf{x}$, the value $f(\mathbf{x})$ is determined by evaluating which side of the hyperplane it falls on, in feature space. Via the freedom to utilize different types of kernel functions, this simple geometric picture corresponds to a variety of nonlinear estimators in input space.

To separate the data set from the origin, we solve the following quadratic program:

$$\min_{w \in F, \boldsymbol{\xi} \in \mathbb{R}^\ell, \rho \in \mathbb{R}} \quad \tfrac{1}{2}\|w\|^2 + \tfrac{1}{\nu\ell}\textstyle\sum_i \xi_i - \rho \tag{5}$$

$$\text{subject to} \quad (w \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \;\; \xi_i \geq 0. \tag{6}$$

Here, $\nu \in (0, 1)$ is a parameter whose meaning will become clear later.

Since nonzero slack variables ξ_i are penalized in the objective function, we can expect that if w and ρ solve this problem, then the decision function

$$f(\mathbf{x}) = \operatorname{sgn}((w \cdot \Phi(\mathbf{x})) - \rho) \tag{7}$$

will be positive for most examples $\mathbf{x}_i$ contained in the training set, while the SV type regularization term $\|w\|$ will still be small. The actual trade-off between these two goals is controlled by ν.

Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian

$$L(w, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho - \sum_i \alpha_i((w \cdot \Phi(\mathbf{x}_i)) - \rho + \xi_i) - \sum_i \beta_i \xi_i, \tag{8}$$

and set the derivatives with respect to the primal variables $w, \boldsymbol{\xi}, \rho$ equal to zero, yielding

$$w = \sum_i \alpha_i \Phi(\mathbf{x}_i), \tag{9}$$

$$\alpha_i = \frac{1}{\nu\ell} - \beta_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \tag{10}$$

In (9), all patterns $\{\mathbf{x}_i : i \in [\ell],\ \alpha_i > 0\}$ are called Support Vectors. Together with (3), the SV expansion transforms the decision function (7) into a kernel expansion

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho\right). \tag{11}$$

Substituting (9) – (10) into L (8), and using (3), we obtain the dual problem:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \quad \text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \tag{12}$$

One can show that at the optimum, the two inequality constraints (6) become equalities if α_i and β_i are nonzero, i.e. if $0 < \alpha_i < 1/(\nu\ell)$. Therefore, we can recover ρ by exploiting that for any such α_i, the corresponding pattern $\mathbf{x}_i$ satisfies

$$\rho = (w \cdot \Phi(\mathbf{x}_i)) = \sum_j \alpha_j k(\mathbf{x}_j, \mathbf{x}_i). \tag{13}$$

Note that if ν approaches 0, the upper boundaries on the Lagrange multipliers tend to infinity, i.e. the second inequality constraint in (12) becomes void. The problem then resembles the corresponding *hard margin* algorithm, since the penalization of errors becomes infinite, as can be seen from the primal objective function (5). It is still a feasible problem, since we have placed no restriction on ρ, so ρ can become a large negative number in order to satisfy (6). If we had required $\rho \geq 0$ from the start, we would have ended up with the constraint $\sum_i \alpha_i \geq 1$ instead of the corresponding equality constraint in (12), and the multipliers α_i could have diverged.

To conclude this section, we note that one can also use *balls* to describe the data in feature space, close in spirit to the algorithms of Schölkopf et al. (1995), with hard boundaries, and Tax and Duin (1999), with "soft margins." Again, we try to put *most of* the data into a small ball by solving, for $\nu \in (0, 1)$,

$$\begin{aligned} &\min_{R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell, c \in F} && R^2 + \tfrac{1}{\nu\ell} \textstyle\sum_i \xi_i \\ &\text{subject to} && \|\Phi(\mathbf{x}_i) - c\|^2 \leq R^2 + \xi_i, \ \ \xi_i \geq 0 \quad \text{for } i \in [\ell]. \end{aligned} \tag{14}$$

This leads to the dual

$$\min_{\boldsymbol{\alpha}} \quad \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) \tag{15}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1 \tag{16}$$

and the solution

$$c = \sum_i \alpha_i \Phi(\mathbf{x}_i), \tag{17}$$

corresponding to a decision function of the form

$$f(\mathbf{x}) = \operatorname{sgn}\left(R^2 - \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + 2 \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - k(\mathbf{x}, \mathbf{x}) \right). \tag{18}$$

Similar to the above, R^2 is computed such that for any $\mathbf{x}_i$ with $0 < \alpha_i < 1/(\nu\ell)$ the argument of the sgn is zero.

For kernels $k(\mathbf{x}, \mathbf{y})$ which only depend on $\mathbf{x} - \mathbf{y}$, $k(\mathbf{x}, \mathbf{x})$ is constant. In this case, the equality constraint implies that the linear term in the dual target function is constant, and the problem (15–16) turns out to be equivalent to (12). It can be shown that the same holds true for the decision function, hence the two algorithms coincide in that case.

4 Optimization

The last section has formulated quadratic programs (QPs) for computing regions that capture a certain fraction of the data. These constrained optimization problems can be solved via an off-the-shelf QP package to compute the solution. They do, however, possess features that set them apart from generic QPs, most notably the simplicity of the constraints. In the present section, we describe an algorithm which takes advantage of these features and empirically scales better to large data set sizes than a standard QP solver with time complexity of order $O(\ell^3)$ (cf. Platt, 1999). The algorithm is a modified version of SMO (Sequential Minimal Optimization), an SV training algorithm originally proposed for classification (Platt, 1999), and subsequently adapted to regression estimation (Smola and Schölkopf, 1998).

The strategy of SMO is to break up the constrained minimization of (12) into the smallest optimization steps possible. Due to the constraint on the sum of the dual variables, it is impossible to modify individual variables separately without possibly violating the constraint. We therefore resort to optimizing over pairs of variables.

Elementary optimization step. For instance, consider optimizing over α_1 and α_2 with all other variables fixed. Using the shorthand $K_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$, (12) then reduces to

$$\min_{\alpha_1,\alpha_2} \frac{1}{2} \sum_{i,j=1}^{2} \alpha_i \alpha_j K_{ij} + \sum_{i=1}^{2} \alpha_i C_i + C, \tag{19}$$

with $C_i := \sum_{j=3}^{\ell} \alpha_j K_{ij}$ and $C := \sum_{i,j=3}^{\ell} \alpha_i \alpha_j K_{ij}$, subject to

$$0 \leq \alpha_1, \alpha_2 \leq \frac{1}{\nu\ell}, \quad \sum_{i=1}^{2} \alpha_i = \Delta, \tag{20}$$

where $\Delta := 1 - \sum_{i=3}^{\ell} \alpha_i$.

We discard C, which is independent of α_1 and α_2, and eliminate α_1 to obtain

$$\min_{\alpha_2} \frac{1}{2}(\Delta - \alpha_2)^2 K_{11} + (\Delta - \alpha_2)\alpha_2 K_{12} + \frac{1}{2}\alpha_2^2 K_{22} + (\Delta - \alpha_2)C_1 + \alpha_2 C_2, \tag{21}$$

with the derivative

$$-(\Delta - \alpha_2)K_{11} + (\Delta - 2\alpha_2)K_{12} + \alpha_2 K_{22} - C_1 + C_2. \tag{22}$$

Setting this to zero and solving for α_2, we get

$$\alpha_2 = \frac{\Delta(K_{11} - K_{12}) + C_1 - C_2}{K_{11} + K_{22} - 2K_{12}}. \tag{23}$$

Once α_2 is found, α_1 can be recovered from $\alpha_1 = \Delta - \alpha_2$. If the new point (α_1, α_2) is outside of $[0, 1/(\nu\ell)]$, the constrained optimum is found by projecting α_2 from (23) into he region allowed by the constraints, and the re-computing α_1.

The offset ρ is recomputed after every such step.

Additional insight can be obtained by rewriting the last equation in terms of the outputs of the kernel expansion on the examples $\mathbf{x}_1$ and $\mathbf{x}_2$ before the optimization step. Let α_1^*, α_2^* denote the values of their Lagrange parameter before the step. Then the corresponding outputs (cf. (11)) read

$$O_i := K_{1i}\alpha_1^* + K_{2i}\alpha_2^* + C_i. \tag{24}$$

Using the latter to eliminate the C_i, we end up with an update equation for α_2 which does not explicitly depend on α_1^*,

$$\alpha_2 = \alpha_2^* + \frac{O_1 - O_2}{K_{11} + K_{22} - 2K_{12}}, \tag{25}$$

which shows that the update is essentially the fraction of first and second derivative of the objective function along the direction of ν-constraint satisfaction.

Clearly, the same elementary optimization step can be applied to any pair of two variables, not just α_1 and α_2. We next briefly describe how to do the overall optimization.

Initialization of the algorithm. We start by setting a random fraction ν of all α_i to $1/(\nu\ell)$. If $\nu\ell$ is not an integer, then one of the examples is set to a value in $(0, 1/(\nu\ell))$ to ensure that $\sum_i \alpha_i = 1$. Moreover, we set the initial ρ to $\max\{O_i : i \in [\ell],\ \alpha_i > 0\}$.

Optimization algorithm. We then select a first variable for the elementary optimization step in one of the two following ways. Here, we use the shorthand SV_{nb} for the indices of variables which are not at bound, i.e. $SV_{nb} := \{i : i \in [\ell],\ 0 < \alpha_i < 1/(\nu\ell)\}$. At the end, these correspond to points that will sit exactly on the hyperplane, and that will therefore have a strong influence on its precise position.

(i) We scan over the entire data set[1] until we find a variable violating a KKT condition (Bertsekas, 1995, e.g.), i.e. a point such that $(O_i - \rho) \cdot \alpha_i > 0$ or $(\rho - O_i) \cdot (1/(\nu\ell) - \alpha_i) > 0$. Once we have found one, say α_i, we pick α_j according to

$$j = \arg\max{}_{n \in SV_{nb}} |O_i - O_n|. \tag{26}$$

(ii) Same as (i), but the scan is only performed over SV_{nb}.

In practice, one scan of type (i) is followed by multiple scans of type (ii), until there are no KKT violators in SV_{nb}, whereupon the optimization goes back to a single scan of type (i). If the type (i) scan finds no KKT violators, the optimization terminates.

In unusual circumstances, the choice heuristic (26) cannot make positive progress. Therefore, a hierarchy of other choice heuristics is applied to ensure positive progress. These other heuristics are the same as in the case of pattern recognition, cf. (Platt, 1999), and have been found to well in our experiments to be reported below.

In our experiments with SMO applied to distribution support estimation, we have always found it to converge. However, to ensure convergence even in rare pathological conditions, the algorithm can be modified slightly, cf. (Keerthi et al., 1999).

We end this session by stating a trick which is of importance in practical implementations. In practice, one has to use a nonzero accuracy tolerance such that two quantities are considered equal if they differ by less than that. In particular, comparisons of this type are used in determining whether a point lies on the margin. Since we want the final decision function to evaluate to 1 for points which lie *on* the margin, we need to subtract this constant from ρ at the end.

[1] This scan can be accelerated by not checking patterns which are on the correct side of the hyperplane by a large margin, using the method of Joachims (1999).

5 Theory

In this section, we analyse the algorithm theoretically, starting with the uniqueness of the hyperplane (Proposition 2). We then describe the connection to binary classification (Proposition 3), and show that the parameter ν characterizes the fractions of SVs and outliers (Proposition 4). Following that, we give a robustness result for the soft margin (Proposition 5) and finally we briefly state error bounds (Theorem 9).

In this section, we will use italic letters to denote the feature space images of the corresponding patterns in input space, i.e.

$$x_i := \Phi(\mathbf{x}_i). \tag{27}$$

Definition 1 *A data set*

$$x_1, \ldots, x_\ell \tag{28}$$

is called separable if there exists some $w \in F$ such that $(w \cdot x_i) > 0$ for $i \in [\ell]$.

Proposition 2 *If the data set (28) is separable, then there exists a unique* supporting hyperplane *with the properties that (1) it separates all data from the origin, and (2) its distance to the origin is maximal among all such hyperplanes. For any $\rho > 0$, it is given by*

$$\min_{w \in F} \frac{1}{2}\|w\|^2 \text{ subject to } (w \cdot x_i) \geq \rho, \;\; i \in [\ell]. \tag{29}$$

Proof Due to the separability, the convex hull of the data does not contain the origin. The existence and uniqueness of the hyperplane then follows from the supporting hyperplane theorem (e.g. Bertsekas, 1995).

Moreover, separability implies that there actually exists some $\rho > 0$ and $w \in F$ such that $(w \cdot x_i) \geq \rho$ for $i \in [\ell]$ (by rescaling w, this can be seen to work for arbitrarily large ρ). By the Cauchy-Schwartz inequality, the distance of the hyperplane $\{z \in F : (w \cdot z) = \rho\}$ to the origin is $\rho/\|w\|$. Therefore the optimal hyperplane is obtained by minimizing $\|w\|$ subject to these constraints, i.e. by the solution of (29). ∎

The following result elucidates the relationship between single-class classification and binary classification.

Proposition 3 *(i) Suppose* (w, ρ) *parametrizes the supporting hyperplane for the data (28). Then* $(w, 0)$ *parametrizes the optimal separating hyperplane (passing through the origin (Vapnik, 1995)) for the labelled data set*

$$\{(x_1, 1), \ldots, (x_\ell, 1), (-x_1, -1), \ldots, (-x_\ell, -1)\}. \tag{30}$$

(ii) Suppose $(w, 0)$ *parametrizes the optimal separating hyperplane passing through the origin for a labelled data set*

$$\{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}, \quad (y_i \in \{\pm 1\} \text{ for } i \in [\ell]). \tag{31}$$

Suppose, moreover, that w *is aligned such that* $(w \cdot x_i)$ *is positive whenever* $y_i = 1$*, and that* $\rho / \|w\|$ *is the margin of the optimal hyperplane. Then* (w, ρ) *parametrizes the supporting hyperplane for the unlabelled data set*

$$\{y_1 x_1, \ldots, y_\ell x_\ell\}. \tag{32}$$

Proof *Ad (i).* Observe that $(-w, \rho)$ parametrizes the supporting hyperplane for the data set reflected through the origin, and that it is parallel to the one given by (w, ρ). This provides an optimal separation of the two sets, with distance 2ρ, and a separating hyperplane $(w, 0)$.
Ad (ii). By assumption, we have $y_i(w \cdot x_i) \geq \rho$ (cf. Vapnik, 1995), hence $(w \cdot y_i x_i) \geq \rho$ for $i \in [\ell]$. ∎

Note that this relationship holds true also if we consider nonseparable problems. In that case, *margin errors* in binary classification (i.e. points which are either on the wrong side of the separating hyperplane or which fall inside the margin) translate into *outliers* in single-class classification, i.e. into points which fall on the wrong side of the hyperplane. Proposition 3 then holds, cum grano salis, for the training sets with margin errors and outliers, respectively, removed.

The utility of Proposition 3 lies in the fact that it allows us to recycle certain results proven for binary classification (Schölkopf et al., 1999c) for use in the single-class scenario. The following, explaining the significance of the parameter ν, is such a case.

Proposition 4 *Assume the solution of (6) satisfies* $\rho \neq 0$*. The following statements hold:*
(i) ν *is an upper bound on the fraction of outliers.*
(ii) ν *is a lower bound on the fraction of SVs.*

(iii) Suppose the data (28) were generated independently from a distribution $P(\mathbf{x})$ *which does not contain discrete components. Suppose, moreover, that the kernel is analytic and non-constant. With probability 1, asymptotically,* ν *equals both the fraction of SVs and the fraction of outliers.*

Parts (i) and (ii) follow directly from Proposition 3 and the fact that outliers are dealt with in exactly the same way as margin errors in the optimization problem for the binary classification case (Schölkopf et al., 1999c). The basic idea is that (10) imposes constraints on the fraction of patterns that can have $\alpha_i = 1/(\nu\ell)$, upper bounding the fraction of outliers, and on the fraction of patterns that must have $\alpha_i > 0$, the SVs. Alternatively, the result can be proven directly based on the primal objective function (5), as sketched presently: to this end, note that when changing ρ, the term $\sum_i \xi_i$ will change proportionally to the *number* of points that have a nonzero ξ_i (the outliers), plus, when changing ρ in the positive direction, the number of points which are just about to get a nonzero ρ, i.e. which sit *on* the hyperplane (the SVs). At the optimum of (5), we therefore have (i) and (ii).

Part (iii) can be proven by a uniform convergence argument showing that since the covering numbers of kernel expansions regularized by a norm in some feature space are well-behaved, the fraction of points which lie exactly on the hyperplane is asymptotically negligible (cf. Schölkopf et al., 1999c).

Proposition 5 (Resistance) *Local movements of outliers parallel to* w *do not change the hyperplane.*

Proof Suppose x_o is an outlier, i.e. $\xi_o > 0$, hence by the KKT conditions (e.g. Bertsekas, 1995) $\alpha_o = 1/(\nu\ell)$. Transforming it into $x'_o := x_o + \delta \cdot w$, where $|\delta| < \xi_o/\|w\|$, leads to a slack which is still nonzero, i.e. $\xi'_o > 0$, hence we still have $\alpha_o = 1/(\nu\ell)$. Therefore, $\boldsymbol{\alpha}' = \boldsymbol{\alpha}$ is still feasible, as is the primal solution $(w', \boldsymbol{\xi}', \rho')$. Here, we use $\xi'_i = (1 + \delta \cdot \alpha_o)\xi_i$ for $i \neq o$, $w' = w + \delta \cdot \alpha_o w$, and ρ' as computed from (13). Finally, the KKT conditions are still satisfied, as still $\alpha'_o = 1/(\nu\ell)$. Thus (Bertsekas, 1995, e.g.), $\boldsymbol{\alpha}$ is still the optimal solution. ∎

Note that although the hyperplane does not change, its parametrization in w and ρ does.

We now move on to the subject of generalization. Our goal is to bound the probability that a novel point drawn from the same underlying distribution lies outside of the estimated region by a certain margin. We start by introducing a common tool for measuring the capacity of a class $\mathcal{F}$ of functions that map $\mathcal{X}$ to $\mathbb{R}$.

Definition 6 *Let (X, d) be a pseudo-metric space,*[2] *let A be a subset of X and $\epsilon > 0$. A set $B \subseteq X$ is an* ϵ-cover *for A if, for every $a \in A$, there exists $b \in B$ such that $d(a, b) \leq \epsilon$. The* ϵ-covering number *of A, $\mathcal{N}_d(\epsilon, A)$, is the minimal cardinality of an ϵ-cover for A (if there is no such finite cover then it is defined to be ∞).*

The idea is that B should be finite but approximate all of A with respect to the pseudometric d. We will use the l_∞ distance over a finite sample $X = (x_1, \ldots, x_\ell)$ for the pseudo-metric in the space of functions,

$$d_X(f, g) = \max_{i \in [\ell]} |f(x_i) - g(x_i)|. \tag{33}$$

Let $\mathcal{N}(\epsilon, \mathcal{F}, \ell) = \sup X \in \mathcal{X}^\ell \mathcal{N}_{d_X}(\epsilon, \mathcal{F})$. Below, logarithms are to base 2.

Theorem 7 *Consider any distribution P on $\mathcal{X}$ and any $\theta \in \mathbb{R}$. Suppose $x_1, \ldots, x_\ell$ are generated i.i.d. from P. Then with probability $1 - \delta$ over such an ℓ-sample, if we find $f \in \mathcal{F}$ such that $f(x_i) \geq \theta + \gamma$ for all $i \in [\ell]$,*

$$P\{x : f(x) < \theta - \gamma\} \leq \tfrac{2}{\ell}(k + \log \tfrac{2\ell}{\delta}),$$

where $k = \lceil \log \mathcal{N}(\gamma, \mathcal{F}, 2\ell) \rceil$.

The basis of the proof is (Shawe-Taylor et al., 1998, Lemma 3.9).

We now consider the possibility that for a small number of points $f(x_i)$ fails to exceed $\theta + \gamma$. This corresponds to having a non-zero slack variable ξ_i in the algorithm, where we take $\theta + \gamma = \rho / \|w\|$ and use the class of linear functions in feature space in the application of the theorem. There are well-known bounds for the log covering numbers of this class. We first introduce notation for the size of the shortfall in $f(x)$.

Definition 8 *Let f be a real valued function on a space $\mathcal{X}$. Fix $\theta \in \mathbb{R}$. For $x \in \mathcal{X}$, define*

$$d(x, f, \gamma) = \max\{0, \theta + \gamma - f(x)\}.$$

Similarly for a training sequence X, we define

$$\mathcal{D}(X, f, \gamma) = \sum_{x \in X} d(x, f, \gamma).$$

Theorem 9 *Fix $\theta \in \mathbb{R}$. Consider a fixed but unknown probability distribution P on the input space $\mathcal{X}$ and a class of real valued functions $\mathcal{F}$ with range $[a, b]$. Then*

[2] i.e. with a distance function that differs from a metric in that it is only semidefinite

with probability $1-\delta$ *over randomly drawn training sequences* x *of size* ℓ*, for all* $\gamma > 0$ *and any* $f \in \mathcal{F}$*,*

$$P\{x\colon f(x) < \theta - \gamma \textit{ and } x \notin X\} \leq \tfrac{2}{\ell}(k + \log \tfrac{4\ell}{\delta}),$$

where $k = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2\ell) + \frac{64(b-a)\mathcal{D}(X,f,\gamma)}{\gamma^2} \log\left(\frac{e\ell\gamma}{8\mathcal{D}(X,f,\gamma)}\right) \log\left(\frac{32\ell(b-a)^2}{\gamma^2}\right)\right\rceil$.

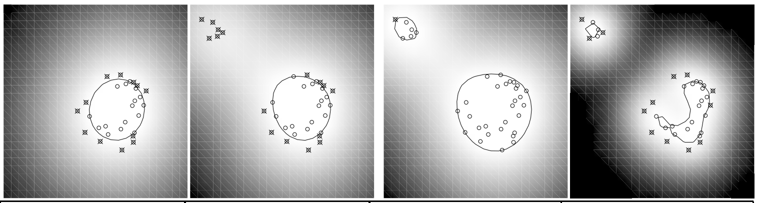
The proof is based on similar proofs for the classification case in (Shawe-Taylor and Cristianini, 1999, Theorem 3). The theorem bounds the probability of a new point falling in the region for which $f(x)$ has value less than $\theta - \gamma$, this being the complement of the estimate for the support of the distribution. In the algorithm described in this paper, one would use the hyperplane shifted by $2\gamma/\|w\|$ towards the origin to define the region. Note that there is no restriction placed on the class of functions though these functions could be probability density functions.

The choice of γ gives a trade-off between the size of the region over which the bound holds (increasing γ increases the size of the region) and the size of the probability with which it holds (increasing γ decreases the size of the log covering numbers).

The result shows that we can bound the probability of points falling outside the region of estimated support by a quantity involving the ratio of the log covering numbers (which can be bounded by the fat shattering dimension at scale proportional to γ) and the number of training examples, plus a factor involving the 1-norm of the slack variables.

The result is stronger than related results given by Ben-David and Lindenbaum (1997), since their bound involves the square root of the ratio of the Pollard dimension (the fat shattering dimension when γ tends to 0) and the number of training examples.

The above bounds are, nevertheless, not entirely satisfactory, and their inclusion here is much more as a sanity check than as a "closed-case" theory for the algorithm presented. Whilst most of the apparent technical gaps can be readily filled (for example determining the covering numbers for the class of functions induced by use of a particular kernel using methods as in Williamson et al. (1999)), there are still considerable gaps. These gaps do not invalidate the algorithm; they simply indicate an incomplete theory, one we hope to complete at some stage. The key difficulty is relating the margin achieved by the algorithm to the parameter γ. Unlike in the support vector machine case, there is no natural linkage imposed by the problem itself. Furthermore, whilst not immediately apparent, the results stated do not actually give guidance as to how to chose the kernel parameter, although they would if a connection between γ and the margin achieved were forced. The latter connection is not necessary, but it could be motivated by noting that it



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
frac. SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38
margin $\rho/\|w\|$	0.84	0.70	0.62	0.48

Figure 1: *First two pictures:* A single-class SVM applied to two toy problems; $\nu = c = 0.5$, domain: $[-1, 1]^2$. Note how in both cases, at least a fraction of ν of all examples is in the estimated region (cf. table). The large value of ν causes the additional data points in the upper left corner to have almost no influence on the decision function. For smaller values of ν, such as 0.1 *(third picture)*, the points cannot be ignored anymore. Alternatively, one can force the algorithm to take these 'outliers' (OLs) into account by changing the kernel width (4): in the *fourth picture*, using $c = 0.1, \nu = 0.5$, the data is effectively analyzed on a different length scale which leads the algorithm to consider the outliers as meaningful points.

seems plausible that if we obtain a very large margin of separation to the origin, we would be more likely to accept a large γ (with the associated risk of ending up with more false positives from the "unknown" class). Measuring γ relative to the margin would then lead to bounds which depend on the margin, and which justify our algorithm that tries to maximize the margin.

Equivalently, we could argue that we try to maximize the margin in order to have the freedom to use a large γ, leading to smaller values of the error bounds, while still not including the "unknown" class. Evidently, this argument implicitly makes prior assumptions about the unknown class, in particular that it is in some sense centered around the origin from which we try to separate the data. The algorithm could be modified to accomodate this case, but presently, we shall not go into further detail on that matter.

6 Experiments

We apply the method to artificial and real-world data. Figure 1 displays 2-D toy examples, and shows how the parameter settings influence the solution.

Figure 2 shows a plot of the outputs $(w \cdot \Phi(\mathbf{x}))$ on training and test sets of the US postal service database of handwritten digits. The database contains 9298 digit images of size $16 \times 16 = 256$; the last 2007 constitute the test set. We fed our

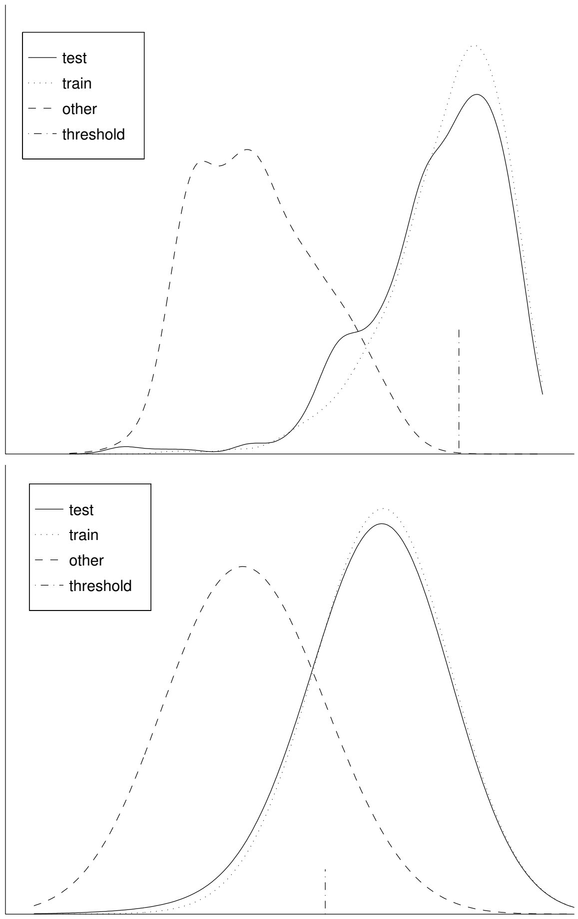


Figure 2: Experiments on the US postal service OCR dataset. Recognizer for digit 0; output histogram for the exemplars of 0 in the training/test set, and on test exemplars of other digits. The x-axis gives the output values, i.e. the argument of the sgn function in (11). For $\nu = 50\%$ *(top)*, we get 50% SVs and 49% outliers (consistent with Proposition 4), 44% true positive test examples, and zero false positives from the "other" class. For $\nu = 5\%$ *(bottom)*, we get 6% and 4% for SVs and outliers, respectively. In that case, the true positive rate is improved to 91%, while the false positive rate increases to 7%. The threshold ρ is marked in the graphs.

Note, finally, that the plots show a Parzen windows density estimate of the output histograms. In reality, many examples sit exactly at the threshold value (the non-bound SVs). Since this peak is smoothed out by the estimator, the fractions of outliers in the training set appear slightly larger than it should be.

algorithm, using a Gaussian kernel (4) of width $c = 0.5 \cdot 256$ (a common value for SVM classifiers on that data set, cf. Schölkopf et al. (1995)), with the training instances of digit 0 only. Testing was done on both digit 0 and on all other digits. As shown in figure 2, $\nu = 50\%$ leads to *zero* false positives (i.e. even though the learning machine has not seen any non-0-s during training, it correctly identifies all non-0-s as such), while still recognizing 44% of the digits 0 in the test set. Higher recognition rates can be achieved using smaller values of ν: for $\nu = 5\%$, we get 91% correct recognition of digits 0 in the test set, with a fairly moderate false positive rate of 7%.

Whilst leading to encouraging results, this experiment did not really address the actual task the algorithm was designed for. Therefore, we next focussed on a problem of novelty detection. Again, we utilized the USPS set; however, this time we trained the algorithm on the test set and used it to identify outliers — it is folklore in the community that the USPS test set (Fig. 3) contains a number of patterns which are hard or impossible to classify, due to segmentation errors or mislabelling (e.g. Vapnik, 1995). In the experiment, we augmented the input patterns by ten extra dimensions corresponding to the class labels of the digits. The rationale for this is that if we disregarded the labels, there would be no hope to identify mislabelled patterns as outliers. Vice versa, with the labels, the algorithm has the chance to identify both unusual patterns and usual patterns with unusual labels. Fig. 4 shows the 20 worst outliers for the USPS test set, respectively. Note that the algorithm indeed extracts patterns which are very hard to assign to their respective classes. In the experiment, we used the same kernel width as above, and a ν value of 5%.

In the last experiment, we tested the scaling behaviour of the proposed SMO solver which is used for training the learning machine (Fig. 5). It was found to depend on the value of ν utilized. For the small values of ν which are typically used in outlier detection tasks, the algorithm scales very well to larger data sets, with a dependency of training times on the sample size which is at most quadratic.

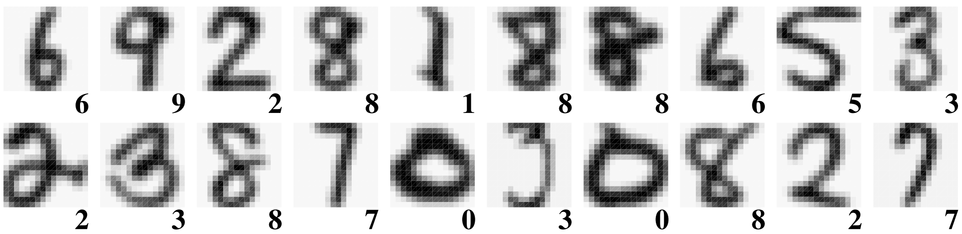


Figure 3: A subset of 20 examples randomly drawn from the USPS test set, with class labels.

ν	fraction of OLs	fraction of SVs	training time
1%	0.0%	10.0%	36
2%	0.0%	10.0%	39
3%	0.1%	10.0%	31
4%	0.6%	10.1%	40
5%	**1.4%**	**10.6%**	**36**
6%	1.8%	11.2%	33
7%	2.6%	11.5%	42
8%	4.1%	12.0%	53
9%	5.4%	12.9%	76
10%	6.2%	13.7%	65
20%	16.9%	22.6%	193
30%	27.5%	31.8%	269
40%	37.1%	41.7%	685
50%	47.4%	51.2%	1284
60%	58.2%	61.0%	1150
70%	68.3%	70.7%	1512
80%	78.5%	80.5%	2206
90%	89.4%	90.1%	2349

Table 1: Experimental results for various values of the outlier control constant ν. Note that ν bounds the fractions of outliers and support vectors from above and below, respectively (cf. Proposition 4). As we are not in the asymptotic regime, there is some slack in the bounds; nevertheless, ν can be used to control the above fractions. Note, moreover, that training times (CPU time in seconds on a Pentium II running at 450 MHz) increase as ν approaches 1. This is related to the fact that almost all Lagrange multipliers will be at the upper bound in that case (cf. Sec. 4). The system used in the outlier detection experiments is shown in bold face.

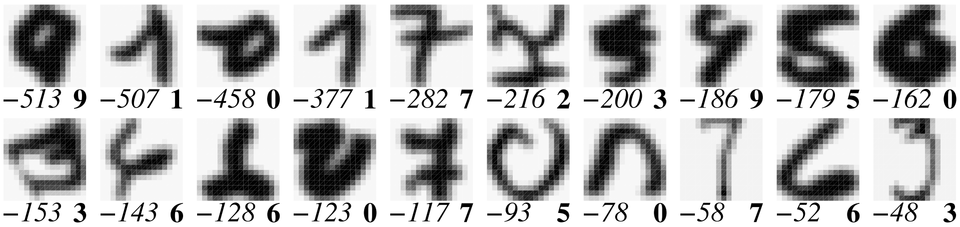


Figure 4: Outliers identified by the proposed algorithm, ranked by the negative output of the SVM (the argument of (11)). The outputs (for convenience in units of 10^{-5}) are written underneath each image in italics, the (alleged) class labels are given in bold face. Note that most of the examples are "difficult" in that they are either atypical or even mislabelled.

7 Discussion

One could view the present work as an attempt to provide a new algorithm which is in line with Vapnik's principle never to solve a problem which is more general than the one that one is actually interested in. E.g., in situations where one is only interested in detecting *novelty*, it is not always necessary to estimate a full density model of the data. Indeed, density estimation is more difficult than what we are doing, in several respects.

Mathematically speaking, a density will only exist if the underlying probability measure possesses an absolutely continuous distribution function. However, the general problem of estimating the measure for a large class of sets, say the sets measureable in Borel's sense, is not solvable (for a discussion, see e.g. Vapnik, 1998). Therefore we need to restrict ourselves to making a statement about the measure of *some* sets. Given a small class of sets, the simplest estimator which accomplishes this task is the empirical measure, which simply looks at how many training points fall into the region of interest. Our algorithm does the opposite. It starts with the number of training points that are supposed to fall into the region, and then estimates a region with the desired property. Often, there will be many such regions — the solution becomes unique only by applying a regularizer, which in our case enforces that the region be small in a feature space associated to the kernel.

Therefore, we must keep in mind that the measure of smallness in this sense depends on the kernel used, in a way that is no different to any other method that regularizes in a feature space. A similar problem, however, appears in density estimation already when done in input space. Let p denote a density on $\mathcal{X}$. If we

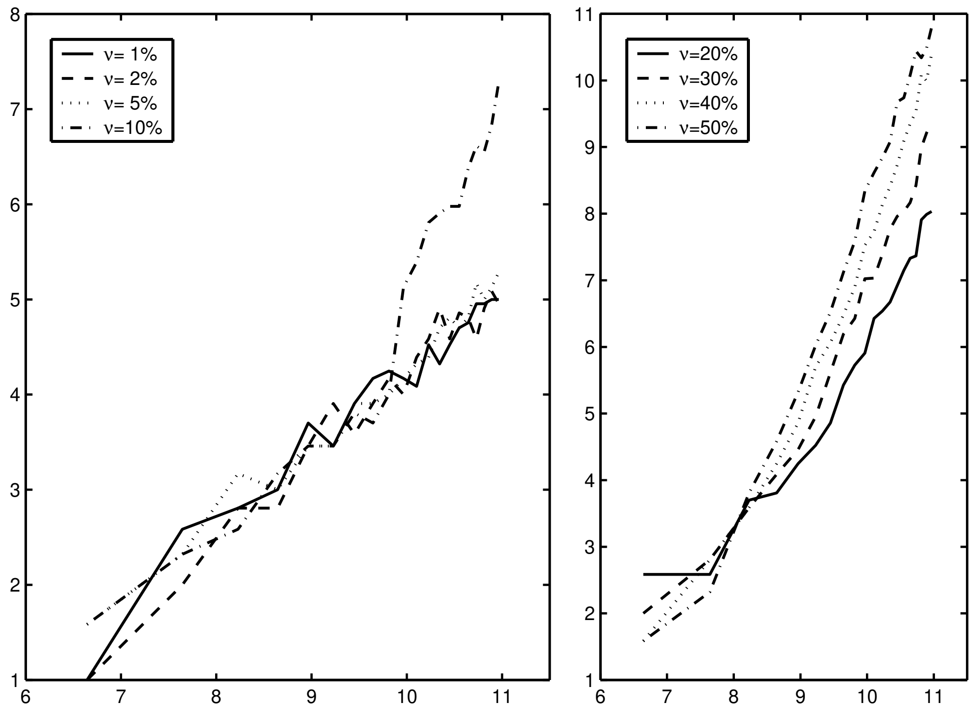


Figure 5: Training times vs. data set sizes (both axes depict logs at base 2; CPU time in seconds on a Pentium II running at 450 MHz). As in Table 1, it can be seen that larger values of ν generally lead to longer training times (note that the plots use different y-axis ranges). However, they also differ in their scaling with the sample size. For large values of ν, training times are roughly proportional to the sample size raised to the power of 2.5 (right plot). For values $\nu \leq 10\%$ (left plot), i.e. those typically used in outlier detection experiments (in Fig. 4, we used $\nu = 5\%$), the scaling exponent is below 2 (the exponents can be directly read off from the slope of the graphs, as they are plotted in log scale with equal axis spacing). Note that the scalings are better than the cubic one that one would expect when solving the optimization problem using all patterns at once, cf. Sec. 4. As in the other experiments, we used $c = 0.5 \cdot 256$, however we only trained on subsets of the USPS test set.

perform a (nonlinear) coordinate transformation in the input domain $\mathcal{X}$, then the density values will *change*; loosely speaking, what remains constant is $p(x) \cdot dx$, while dx is transformed, too. When directly estimating the probability *measure* of regions, we are not faced with this problem, as the regions automatically change accordingly.

An attractive property of the measure of smallness that we chose to use is that it can also be placed in the context of regularization theory, leading to an interpretation of the solution as maximally smooth in a sense which depends on the specific kernel used. More specifically, let us assume that k is Green's function of P^*P for an operator P mapping into some dot product space (Smola et al., 1998; Girosi, 1998), and take a look at the dual objective function that we minimize,

$$\begin{aligned}\sum_{i,j} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i,j} \alpha_i \alpha_j (k(\mathbf{x}_i, .) \cdot \delta_{\mathbf{x}_j}(.)) \\ &= \sum_{i,j} \alpha_i \alpha_j (k(\mathbf{x}_i, .) \cdot (P^*Pk)(\mathbf{x}_j, .)) \\ &= \sum_{i,j} \alpha_i \alpha_j ((Pk)(\mathbf{x}_i, .) \cdot (Pk)(\mathbf{x}_j, .)) \\ &= ((P \sum_i \alpha_i k)(\mathbf{x}_i, .) \cdot (P \sum_j \alpha_j k)(\mathbf{x}_j, .)) \\ &= \|Pf\|^2,\end{aligned}$$

using $f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x})$. The regularization operators of common kernels can be shown to correspond to derivative operators (Poggio and Girosi, 1990) — therefore, minimizing the dual objective function corresponds to maximizing the smoothness of the function f (which is, up to a thresholding operation, the function we estimate). This, in turn, is related to a prior $p(f) \sim e^{-\|Pf\|^2}$ on the function space.

Interestingly, as the minimization of the dual objective function also corresponds to a maximization of the margin in feature space, an equivalent interpretation is in terms of a prior on the distribution of the unknown other class (the "novel" class in a novelty detection problem) — trying to separate the data from the origin amounts to assuming that the novel examples lie around the origin.

The main inspiration for our approach stems from the earliest work of Vapnik and collaborators. In 1962, they proposed an algorithm for characterizing a set of unlabelled data points by separating it from the origin using a hyperplane (Vapnik and Lerner, 1963; Vapnik and Chervonenkis, 1974). However, they quickly moved on to two-class classification problems, both in terms of algorithms and in terms of the theoretical development of statistical learning theory which originated in those days.

From an algorithmic point of view, we can identify two shortcomings of the original approach which may have caused research in this direction to stop for more than three decades. Firstly, the original algorithm in (Vapnik and Chervonenkis, 1974) was limited to linear decision rules in input space, secondly, there was no way of dealing with outliers. In conjunction, these restrictions are indeed severe — a generic dataset need not be separable from the origin by a hyperplane in input space.

The two modifications that we have incorporated dispose of these shortcomings. First, the kernel trick allows for a much larger class of functions by nonlinearly mapping into a high-dimensional feature space, and thereby increases the chances of a separation from the origin being possible. In particular, using a Gaussian kernel (4), such a separation exists for any data set $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$: to see this, note that $k(\mathbf{x}_i, \mathbf{y}_j) > 0$ for all i, j, thus all dot products between mapped patterns are positive, implying that all patterns lie inside the same orthant. Moreover, since $k(\mathbf{x}_i, \mathbf{x}_i) = 1$ for all i, they all have unit length. Hence they are separable from the origin. The second modification directly allows for the possibility of outliers. We have incorporated this 'softness' of the decision rule using the ν-trick (Schölkopf et al., 1999c) and thus obtained a direct handle on the fraction of outliers.

We believe that our approach, proposing a concrete algorithm with well-behaved computational complexity (convex quadratic programming) for a problem that so far has mainly been studied from a theoretical point of view has abundant practical applications. To turn the algorithm into an easy-to-use black-box method for practicioners, questions like the selection of kernel parameters (such as the width of a Gaussian kernel) have to be tackled. It is our expectation that the theoretical results which we have briefly outlined in this paper will provide a solid foundation for this formidable task.

Acknowledgement. This work was supported by the ARC and the DFG (# Ja 379/9-1). Parts of it were done while BS and AS were with GMD FIRST, Berlin. Thanks to S. Ben-David, C. Bishop, N. Oliver, J. Platt, C. Schnörr, and M. Tipping for helpful discussions.

References

S. Ben-David and M. Lindenbaum. Learning distributions by their density levels: A paradigm for learning without a teacher. *Journal of Computer and System Sciences*, 55:171–182, 1997.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

J. Chevalier. Estimation du support et du contour du support d'une loi de probabilité. *Annales de l'Institut Henri Poincaré. Section B. Calcul des Probabilités et Statistique. Nouvelle Série*, 12(4):339–364, 1976.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

A. Cuevas. On pattern analysis in the non-convex case. *Kybernetes*, 19(6):26–33, 1990.

A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6):2300–2312, 1997.

L. Devroye and G.L. Wise. Detection of abnormal behaviour via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.

J.H.J. Einmal and D.M. Mason. Generalized quantile processes. *The Annals of Statistics*, 20(2):1062–1078, 1992.

G. Gayraud. Estimation of functional of density support. *Mathematical Methods of Statistics*, 6(1):26–46, 1997.

J. Geffroy. Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, 13:191–210, 1964.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

J.A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82:267–270, 1987.

T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 185 – 208. MIT Press, Cambridge, MA, 1999.

S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. Technical Report CD-99-14, Dept. of Mechanical and Production Engineering, Natl. Univ. Singapore, Singapore, 1999.

A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction*. Springer, New York, 1993.

D.W. Müller. The excess mass approach in statistics. Beiträge zur Statistik, Universität Heidelberg, 1992.

D. Nolan. The excess mass ellipsoid. *Journal of Multivariate Analysis*, 39:348–371, 1991.

J. Platt. Fast training of SVMs using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 185 – 208. MIT Press, Cambridge, MA, 1999.

T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), 1990.

W. Polonik. Density estimation under qualitative assumptions in higher dimensions. *Journal of Multivariate Analysis*, 55(1):61–81, 1995a.

W. Polonik. Measuring mass concentrations and estimating density contour clusters — an excess mass approach. *The Annals of Statistics*, 23(3):855–881, 1995b.

W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69:1–24, 1997.

T.W. Sager. An iterative method for estimating a multivariate mode and isopleth. *Journal of the American Statistical Association*, 74(366):329–339, 1977.

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.

B. Schölkopf, C. J. C. Burges, and A. J. Smola. *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA, 1999a.

B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA, 1999b. 327 – 352.

B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *To appear in: Neural Computation*, 1999c. Also: NeuroColt2-TR 1998-031.

B. Schölkopf, R. Williamson, A. Smola, and J. Shawe-Taylor. Single-class support vector machines. In J. Buhmann, W. Maass, H. Ritter, and N. Tishby, editors, *Unsupervised Learning*, Dagstuhl-Seminar-Report 235, pages 19 – 20, 1999.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory*, 44(5): 1926–1940, 1998.

J. Shawe-Taylor and N. Cristianini. Margin distribution bounds on generalization. In Paul Fischer and Hans Ulrich Simon, editors, *Computational Learning Theory, 4th European Conference, EuroCOLT'99*. Springer, 1999.

A. Smola and B. Schölkopf. A tutorial on support vector regression. NeuroColt2-TR 1998-030, http://svm.first.gmd.de, 1998.

A. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

A. Smola, R. C. Williamson, S. Mika, and B. Schölkopf. Regularized principal manifolds. In *Computational Learning Theory: 4th European Conference*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 214 – 229. Springer, 1999.

D. Stoneking. Improving the manufacturability of electronic designs. *IEEE Spectrum*, 36(6):70–76, 1999.

L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings Fourth IEE International Conference on Artificial Neural Networks*, pages 442 – 447, Cambridge, 1995.

D.M.J. Tax and R.P.W. Duin. Data domain description by support vectors. In M. Verleysen, editor, *Proceedings ESANN*, pages 251 – 256, Brussels, 1999. D Facto.

A.B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.

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

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979).

V. Vapnik and A. Lerner. Pattern recognition using generalized portraits. *Avtomatika i Telemekhanika*, 24:774 – 780, 1963.

R. Williamson, A. Smola, and B. Schölkopf. Entropy numbers, operators and support vector kernels. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 127 – 144. MIT Press, Cambridge, MA, 1999.