massachusetts institute of technology — artificial intelligence laboratory

b

T. Poggio, S. Mukherjee, R. Rifkin, A. Rakhlin and A. Verri

AI Memo 2001-011 July 2001
CBCL Memo 198

© 2001 massachusetts institute of technology, cambridge, ma 02139 usa — www.ai.mit.edu

Abstract

In this note we characterize the role of b, which is the constant in the standard form of the solution provided by the Support Vector Machine technique $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$.

This report describes research done within the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

This research was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032

Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., and The Whitaker Foundation.

1 Introduction

Support Vector Machines (SVMs), originally introduced by Vapnik [13] in the context of Statistical Learning Theory, can be shown to be a special case of a regularization approach to the ill-posed problem of regression or classification from sparse and finite data [5, 15]. This derivation makes it clear that SVMs and a large body of different learning and approximation techniques, known as Regularization Networks (RNs) [9], can be obtained from the same general principles. The aim of this short note[1] is to clarify the role of b, the constant term in the solution to the learning problem obtained in the standard derivation of SVMs due to Vapnik [13] and found also in some RNs. These issues might have some impact for a) exploring the theoretical connection between SVMs, RNs, and other techniques (see [7], for example, for some of the difficulties arising for the connection between SVMs for regression and a special version of Basis Pursuit Denoising) and b) for developing efficient algorithms for SVMs (such as the Kernel Adatron [6]).
This note is organized as follows: we motivate this study in section 2, present our analysis in section 3, and then list remarks and conclusions. Throughout the paper we assume some familiarity with both SVMs and RNs as presented in [5] (we review the main mathematical concepts and definition in section 5).

2 Motivation

Given ℓ training pairs $(\mathbf{x}, y_1), ..., (\mathbf{x}_\ell, y_\ell)$, a regularized solution to the learning problem is found by minimizing the following functional for fixed λ

$$I[f] = \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_K^2 \tag{1}$$

corresponding to the minimization of the empirical loss – the first term – under capacity control – the second term. The choice of the loss function V determines the learning scheme. Minimization of $I[f]$ corresponds to standard Regularization Networks for $V(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$, to SVM Classification for $V(y_i, f(\mathbf{x}_i)) = |1 - y_i f(\mathbf{x}_i)|_+$ and to SVM Regression for $V(y_i, f(\mathbf{x}_i)) = |y_i - f(\mathbf{x}_i)|_\epsilon$, where $|x|_+ = x$ if $x > 0$ and zero otherwise and $|\cdot|_\epsilon$, called ϵ-insensitive loss, is defined as:

$$|x|_\epsilon \equiv \begin{cases} 0 & \text{if } |x| < \epsilon \\ |x| - \epsilon & \text{otherwise.} \end{cases}$$

The term $\|f\|_K$ is the norm of the function f in the Reproducing Kernel Hilbert Space (RKHS) induced by a positive definite kernel K (for definitions and properties of kernels see the Appendix) [1, 15]. It can be shown that the minimizer of $I[f]$ for both RN [15, 9] and SVM [14, 7, 5, 16] belongs to the RKHS induced by K and can be written as $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ for some coefficients α_i which depend on the ℓ examples and on λ. In the original formulation of SVMs due to Vapnik, the minimizer is actually written as

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{2}$$

with $-1/(2\lambda\ell) \leq \alpha_i \leq 1/(2\lambda\ell)$,

[1] Besides entering the Guinness Book of World Records for the shortest title.

$$\sum_{i=1}^{\ell} \alpha_i = 0, \tag{3}$$

The offset parameter b is estimated (like the coefficients α_i) from the ℓ examples. From the dual formulation of the optimization problem of SVMs (see [13], for example), it can be seen that the equality constraint 3 is induced[2] by the form of the solution assumed in equation 2.
The case of standard RNs (quadratic V) is well known. If the kernel is a positive definite function (i.e., in the case of Gaussian Radial Basis Functions) one chooses the solution in the linear span of the RKHS written as $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ without any constraint on the α_i. If the kernel is a conditionally positive definite function of order 1 (i.e., in the case of piecewise linear splines) a constant term, b, is added to the solution which becomes $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$, subject to the same equality constraint on the α. As shown in [5] the solution $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$ with the constant b (and the equality constraint for the α) is allowed even in the case of a positive definite kernel K. This choice – rather than the usual one without b – effectively corresponds to a different (semi)norm and to a different RKHS [3].
This paper is devoted to answering the following questions: When should b be used? Is there a choice of using or not using b? What does the choice mean? Are the answers different for RNs and SVMs? Is the hypothesis of K positive definite really necessary for SVMs?

2.1 Mercer's theorem: pd and cpd kernels

The key tool in our analysis is the result published by Mercer in 1909. Mercer's theorem [3] states that a symmetric, *pd* kernel $K : X \times X \to \mathbb{R}$, with X a bounded domain (see the Appendix for definitions of *pd* and *cpd* kernels), can always be written as

$$K(\mathbf{x}, \mathbf{x}') = \sum_{q=1}^{N} \mu_q \phi_q(\mathbf{x}) \phi_q(\mathbf{x}') \tag{4}$$

with the ϕ_q being orthonormal eigenfunctions of the integral equation

$$\int_X K(\mathbf{x}, \mathbf{x}') \phi(\mathbf{x}) d\mathbf{x} = \mu \phi(\mathbf{x}'). \tag{5}$$

The μ_q are strictly positive numbers (decreasing to 0 if N is infinite); N is either infinite or finite; in the latter case the kernel is called *degenerate*. The RKHS induced by K is the Hilbert space of the functions spanned by the ϕ

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_q \phi_q(\mathbf{x}),$$

equipped with the scalar product $< f, g >= \sum_{i=1}^{N} \frac{c_q d_q}{\mu_q}$ and with finite norm in the RKHS

$$\|f\|_K^2 = \sum_{q=1}^{N} \frac{c_q^2}{\mu_q}.$$

[2] This property of b was not discussed in [13].

[3] Equation 4.10 in [5] is incorrect and should be replaced with $(K' + \lambda I)\alpha + \mathbf{1}b = (K + \lambda I)\alpha + \mathbf{1}b = \mathbf{y}$. The key equation 4.12 and the conclusions are however correct.

It is important to emphasize that the expansion 4 – which is not valid for a general symmetric kernel – remains valid when the kernel has *a finite number k of negative eigenvalues*, eg when it is a *cpd* kernel of order k. In the latter case the kernel can be made positive definite by subtracting the terms $\mu_q \phi_q(\mathbf{x})\phi_q(\mathbf{x}')$ belonging to negative eigenvalues[3]. Also notice that a positive definite kernel remains positive definite if one of the terms in the expansion 5 is eliminated.

3 Analysis

3.1 Conditionally positive definite kernel

For a *cpd* kernel (of order 1) the interpretation of equation 1 is the following. Formally we replace the *cpd* kernel K by the *pd* kernel $K' = K + a$ where $a = -\mu_0 \phi_1 \phi_1$ is a positive constant (we assume in this paper that $\phi_1(\mathbf{x})\phi_1(\mathbf{x}')$ corresponds to a constant), to eliminate the negative term in K. The stabilizer term in $I[f]$ is then interpreted as $\|f\|_K^2 = \sum_{q=2}^N \frac{c_q^2}{\mu_q}$, that is as $\|f\|_{K'}^2 = \|P_{K'} f\|_{K'}^2$ where $P_{K'}$ projects f into the RKHS induced by K' (see [15]). It is then natural to consider, as the space of functions in which the minimizer of $I[f]$ is sought, functions of the form $f(\mathbf{x}) = \sum_{q=2}^N c_q \phi_q(\mathbf{x}) + b$. The latter are functions in the RKHS induced by the *pd* kernel K' completed with the constants (which are not in the RKHS). In the rest of the paper we use $\|f\|_{K'}^2$ to mean the norm of the projection of f onto the space induced by K' [1]. Consider now the functional $I[f]$ as a function of the coefficients c_q and b. Taking derivatives of I with respect to the coefficients and setting them equal to zero (following the *pd* case in [7]) yields $-\sum_{i=1}^\ell V'(y_i - f(\mathbf{x}_i))\phi_q(\mathbf{x}_i) + \frac{\lambda c_q}{\mu_q} = 0, -\sum_{i=1}^\ell V'(y_i - f(\mathbf{x}_i)) = 0$. The previous equations can be rewritten defining $\alpha_i = \frac{1}{\lambda} V'(y_i - f(\mathbf{x}_i))$ as

$$c_q = \mu_q \sum_{i=1}^\ell \alpha_i \phi_q(\mathbf{x}_i)$$

with

$$\sum_{i=1}^\ell \alpha_i = 0. \tag{6}$$

Therefore the minimizer of $I[f]$ is

$$f(\mathbf{x}) = \sum_{q=2}^N c_q \phi_q(\mathbf{x}) + b = \sum_{i=1}^\ell \alpha_i \sum_{q=2}^N \mu_q \phi_q(\mathbf{x}_i)\phi_q(\mathbf{x}) + b = \sum_{i=1}^\ell \alpha_i K'(\mathbf{x}, \mathbf{x}_i) + b = \sum_{i=1}^\ell \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$$

where we have used equation 4 and, in the last step, equation 6. By interpreting V' in a generalized sense, the proof is valid also for the non-differentiable V used by SVM (regression and classification), as shown explicitly in appendix B.2 of [7].
The construction given above can be derived more directly without the Mercer expansion but with an additional assumption. Given a *cpd* kernel K, a *pd* kernel K' can be obtained through Property 5.3 in the Appendix, yielding $f(\mathbf{x}) = \sum_{i=1}^\ell \alpha_i K'(\mathbf{x}, \mathbf{x}_i) + b$. *Assuming* 6, we obtain $f(\mathbf{x}) = \sum_{i=1}^\ell \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$, with K *cpd*.

Thus conditionally positive definite kernels can be used not only for standard RNs but also for SVMs[4]. In both cases the term b is needed in the solution.

3.2 Positive definite kernel

It is clear that if the kernel is *pd*, b is not needed and there is no equality constraint on the α. In the case of SVM this induces a different kind of machine from the one originally proposed by Vapnik. It is also correct – as Vapnik originally proposed – to use solutions with b (that is solutions of the form $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$) since a positive definite kernel K is also *cpd*. In this case, the equality constraint on the α induced by b "kills" the constant term in the kernel expansion.
Thus when K is *pd* one can choose between two types of solution:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

or

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i' K(\mathbf{x}, \mathbf{x}_i) + b.$$

The first solution corresponds to the minimizer of $I[f] = \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_K^2$ and the second solution to the minimizer of $I[f] = \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{K'}^2$, where $K' = K - a$ is cpd (for an appropriate constant a).
In the RN case with quadratic V the minimization of the two different I yields the linear equations (see [5]): $(K + \lambda I)\alpha = \mathbf{y}$ and $(K + \lambda I)\alpha' + \mathbf{1}b = \mathbf{y}, \mathbf{1}\alpha = 0$. Thus in the standard RN case it is possible to compute one solution from the other one using $\Delta\alpha = (K + \lambda I)^{-1}\mathbf{1}b$, which is effectively equivalent to recomputing the solution. In the SVM case minimization yields two different QP problems; the relation between α and (α', b) cannot be given in closed form.

3.3 Remarks

1. *b or not b?* In the case of a *cpd* kernel the choice of a minimizer of the form $f(\mathbf{x}) = \sum_{q=2}^{N} c_q \phi_q(\mathbf{x}) + b$ is not the only possibility. In the finite dimensional case for instance one may choose simply $f(\mathbf{x}) = \sum_{q=2}^{N} c_q \phi_q(\mathbf{x})$ and *only* consider f of this form in the minimization of $I[f]$ in 1. In the infinite dimensional case, however, the first choice is necessary for a natural space of functions in the minimization of $I[f]$.
2. *Finite and infinite dimensional kernels* With a given and fixed degenerate – eg finite dimensional – kernel it is in general impossible to approximate arbitrarily well a continuous function on a bounded interval. This is the case, for instance, with the fixed degree polynomial kernels often used in SVMs. The infinite dimensional case is typical for RN, such as radial basis functions and tensor product splines. In the classical regression setup with a quadratic V in equation 1 the goal is to approximate a continuous function f from scattered data. It is well known that for certain *pd* kernels, such as the Gaussian, the RKHS (on a bounded domain) induced by K is dense in the space of continuous functions on the same domain [8]. Thus it is sufficient to look for the

[4]This extension was correctly suggested in [12] together with other confusing or wrong statements (such as equation (46)).

minimizer in the form of $\sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$. If however the kernel K is *cpd*, as is usually the case for splines, it is necessary to add a polynomial [15]—footnoteThe eigenfunctions ϕ associated with the positive eigenvalues of the *cpd* kernel are not complete in $L2$ on X, X bounded, without an additional polynomial.. Thus for kernels which are cpd of order 1, it is necessary to look for a minimizer of I of the form $\sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$. Notice that the RKHS norm penalizes the constant in the *pd* K but does not penalize b when a *cpd* kernel is used.
3. *Unbounded domain* The Mercer expansion of a *pd* kernel requires the domain to be bounded. Certain very useful kernels, however, such as the Gaussian and thin plate splines, are functions over an unbounded domain. It is possible to define a RKHS restricted to a bounded domain for such kernels (see for instance [4, 7]), by appropriately restricting the space of functions f [1]. We start with a set of functions defined on the unbounded domain E, forming a Hilbert space and with a reproducing kernel K. This is possible, since the definition of a RKHS does not depend on Mercer theorem. As an example of RKHSs defined on an unbounded domain $E = I\!R^n$ consider the integral equation (5). The spectrum of the eigenvalues of is now continuous. Consider translational invariant kernels, that is kernels of the type $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x} - \mathbf{x}')$. Following [7], if $\tilde{f}(\boldsymbol{\omega})$ denotes the Fourier Transform of $f(\mathbf{x})$, a RKHS can be defined as the space of functions for which the norm $\|f\|_K = \int_{-\infty}^{+\infty} \frac{|\tilde{f}(\boldsymbol{\omega})|^2}{\tilde{K}(\boldsymbol{\omega})} d\boldsymbol{\omega}$ exists finite. The well-posedness of the above definition rests upon Bochner's theorem, which ensures that the positivity of the Fourier Transform $\tilde{K}$ of a function K is a necessary and sufficient condition for the positive definiteness of K. Consider now the bounded domain E_1 contained in E. K for points in E_1 is still a positive definite matrix. This means that K corresponds to a class of functions F_1 over the domain E_1. A theorem by Aronszajn [1] shows that K restricted to E_1 is the reproducing kernel of the class F_1 of all restrictions of functions of F to the subset E_1. For the RKHS F_1 equation 4 is valid (with $N = \infty$) (consider the example of the Fourier series discussed in [7]).

4 Conclusion

In summary, *pd* kernels are not a necessary requirement for SVMs. *Cpd kernels* are sufficient. With order 1 *cpd* kernels, the b term – and the associated constraint equation 6 – is a natural choice and is "de facto" required for infinite dimensional kernels to allow for a natural interpretation of f in equation 1. If the kernel is *pd* the normal choice of the solution is without b; however, it is possible to use a positive definite K *and* the constant b. The latter choice is effectively the choice of a different kernel and a different feature space relative to the initial K used in the standard solution without b: the constant feature "disappears" from the RKHS norm and therefore is not "penalized". This choice is often quite reasonable, since in many regression problems shifts of f by a constant should not be penalized. This is especially true for the binary classification framework originally considered by Vapnik. Only the sign of the function f found by SVM is used for classification. A constant b plays therefore the role of a threshold; using a solution of the form $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$ corresponds to the very reasonable assumption that there is no privileged value – such as 0 – for the classification threshold.
We conclude by mentioning an interesting topic for future research, arising from the freedom of choosing an arbitrary polynomial, for instance, to be added to $\sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ with K positive definite. The polynomial could be chosen to satisfy certain invariances dictated by the problem, to avoid penalization in the stabilizer. Of course, for high dimensional spaces the degree of the polynomial that may be reasonably considered is not higher than linear.

5 Appendix: positive definite and conditionally positive definite kernels

We review here some basic facts on kernels. More on this subject for the case of $K(x,y) = K(x-y)$ can be found in [11, 2, 10].
Let X be some set, for example a subset of $\mathbb{R}^d$ or $\mathbb{R}^d$ itself. A *kernel* is a symmetric function $K : X \times X \to \mathbb{R}$. We have the following two definitions.

Definition 5.1
A kernel $K(\mathbf{t},\mathbf{s})$ *is* positive definite (pd) *if* $\sum_{i,j=1}^n c_i c_j K(\mathbf{t}_i,\mathbf{t}_j) \geq 0$ *for any natural number* n *and choice of* $\mathbf{t}_1,...,\mathbf{t}_n \in X$ *and* $c_1,...,c_n \in \mathbb{R}$.

An equivalent definition could be given in terms of positive semidefiniteness of the matrix $K_{ij} = K(\mathbf{t}_i,\mathbf{t}_j)$.

Definition 5.2
A kernel $K(\mathbf{t},\mathbf{s})$ *is* conditionally positive definite (cpd) *of order 1 if* $\sum_{i,j=1}^n c_i c_j K(\mathbf{t}_i,\mathbf{t}_j) \geq 0$ *for any natural number* n *and choice of* $\mathbf{t}_1,...,\mathbf{t}_n \in X$ *and* $c_1,...,c_n \in \mathbb{R}$ *subject to the constraint* $\sum_{i=1}^n c_i = 0$.

In the literature, $-K$, the negative of a *cpd* kernel of order 1 K, is also known as a *negative definite* kernel [2]. Although *cpd* kernels of order higher than 1 can be defined [10], in what follows we refer to *cpd* kernels of order 1 simply as *cpd* kernels.
While every *pd* kernel is also *cpd*, the converse is not true. For example, the kernel $K(\mathbf{t},\mathbf{s}) = -\|\mathbf{t}-\mathbf{s}\|^2$ is *cpd* but not *pd*. Indeed we have

$$-\sum_{i,j=1}^n c_i c_j \|\mathbf{t}_i - \mathbf{t}_j\|^2 = -\sum_{i,j=1}^n c_i c_j \|\mathbf{t}_i\|^2 - \sum_{i,j=1}^n c_i c_j \|\mathbf{t}_j\|^2 + 2\sum_{i,j=1}^n c_i c_j \mathbf{t}_i \cdot \mathbf{t}_j.$$

Since the c_i sum to zero, the first two terms in the r.h.s. vanish and the conditionally positive definiteness follows from the positive definiteness of the inner product. We see that this kernel is not *pd* by setting all the c_i equal to 1.
Let us now define $K_a = K + a$ for some real number a. We have two simple properties whose proofs are trivial.

Property 5.1
If K *is* cpd, K_a *is* cpd *for every* a.

Property 5.2
If K *is* pd *and* $a \geq 0$, K_a *is also* pd.

If K is *pd* but a is negative, K_a may or may not be *pd*. For example, while $K(\mathbf{t},\mathbf{s}) = (1+\mathbf{t}\cdot\mathbf{s})^2$ is positive definite, the kernel $K_a(\mathbf{t},\mathbf{s}) = (1+\mathbf{t}\cdot\mathbf{s})^2 + a$ is *pd* for all $a \geq -1$ but only *cpd* for all a. Similarly, while $(\mathbf{t}\cdot\mathbf{s})$ is *pd*, $(\mathbf{t}\cdot\mathbf{s}) - a$ with $a > 0$ is only *cpd*. In what follows we need the following result whose proof can be found, for example, in [2].

Property 5.3
Let any $\mathbf{t}_0 \in X$. *Then* $K'(\mathbf{s},\mathbf{t}) = K(\mathbf{s},\mathbf{t}) - K(\mathbf{t},\mathbf{t}_0) - K(\mathbf{s},\mathbf{t}_0) + K(\mathbf{t}_0,\mathbf{t}_0)$. *The kernel* K' *is* pd *if and only if the kernel* K *is* cpd.

Acknowledgements

We wish to thank Federico Girosi for many useful discussions and a legacy of insights.

References

[1] N. Aronszajn "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 686, pp. 337-404, 1950.

[2] C. Berg, J.P.R. Christensen, and P. Ressel *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions* Springer, 1984.

[3] R. Courant and D. Hilbert *Methods of Mathematical Physics, Vol 2* Interscience, 1962.

[4] F. Cucker and S. Smale On the Mathematical Foundations of Learning, *Bulletin of AMS, in press*

[5] T. Evgeniou, M. Pontil, and T. Poggio "Regularization Networks and Support Vector Machines," *Advances in Computational Mathematics*, vol. 13, pp. 1–50, 2000.

[6] T. Friess, N. Cristianini, and C. Campbell. "The Kernel-Adatron: A fast and simple learning procedure for Support Vector Machines," *Proceedings of the 15th International Conference in Machine Learning*, pages 188-196, 1998.

[7] F. Girosi "An equivalence between sparse approximation and support vector machines," *Neural Computation*, vol. 10, pp. 1455–1480, 1998.

[8] F. Girosi and T. Poggio "Networks and the Best Approximation Property," *Biological Cybernetics*, vol. 63, pp. 169–176, 1990.

[9] F. Girosi, M. Jones, and T. Poggio "Regularization theory and neural network architectures," *Neural Computation,* vol. 7, pp. 219–269, 1995.

[10] C.A. Micchelli "Interpolation of scattered data: distance matrices and conditionally positive definite functions," *Constructive Approximation*, vol. 2, pp. 11–22, 1986.

[11] I.J. Schoenberg "Metric spaces and completely monotone functions" *Ann. of Math.*,vol. 39, pp.811–841, 1938.

[12] Alex J. Smola and Bernhard Schölkopf and Klaus-Robert Müller "The connection between Regularization Operators and Support Vector Kernels" *Neural Networks*, vol.11, num. 4, pp. 637–649", 1998.

[13] V.N. Vapnik The Nature of Statistical Learning Theory Springer, 1995.

[14] V.N. Vapnik Statistical Learning Theory Wiley, 1998.

[15] G. Wahba *Spline models for observational data* SIAM, 1990.

[16] Lin, Y., Lee, Y., and Wahba, G. *Support Vector Machines for Classification in Nonstandard Situations* TR 1016, March 2000. To appear, Machine Learning