Support Vector Machine Classification via Parameterless Robust Linear Programming *

O. L. Mangasarian†

Abstract

We show that the problem of minimizing the sum of arbitrary-norm *real* distances to misclassified points, from a pair of parallel bounding planes of a classification problem, divided by the margin (distance) between the two bounding planes, leads to a simple parameterless linear program. This constitutes a linear support vector machine (SVM) that simultaneously minimizes empirical error of misclassified points while maximizing the margin between the bounding planes.Nonlinear kernel SVMs can be similarly represented by a parameterless linear program in a typically higher dimensional feature space.

1 Introduction

Support vector machines (SVMs) [11, 5] constitute the method of choice for classification problems. Constructing an SVM classifier typically entails the solution of either a quadratic or a linear program that almost always involves at least one parameter. Determining the size of this parameter is time consuming and its size significantly affects the correctness of the classifier obtained. Furthermore, distances to misclassified points from a bounding plane are not represented correctly as distances in these formulations. We address both these issues in this work, by taking as the objective of our optimization problem a parameterless *ratio* of the sum of arbitrary-norm *real* distances from the bounding planes to points lying on the wrong sides of these planes, divided by the margin between these two planes, also measured by the same arbitrary norm. This simple method of minimizing the error in data fitting while maximizing the margin between the separating planes, leads to a simple linear programming formulation first considered in [6] and later in [10, 2], even though none of these formulations explicitly took into account the margin between the separating planes of the problem. We will show in this work that the linear programming

*Data Mining Institute Technical Report 03-01, March 2003. Research supported by NSF GrantCR-0138308.

† *olvi@cs.wisc.edu.* Computer Sciences Department, University of Wisconsin, Madison, WI 53706, and Department of Mathematics, University of California at San Diego, La Jolla, CA 92093.

models of [10, 2] in fact can be considered to be effective and probably the simplest models for a support vector machine.

In Section 2 we define our classification problem and give standard SVM formulations involving a linear or a quadratic program with a parameter in the objective function. In Section 3 we derive our parameterless linear formulation with distances measured using any desired norm defined on the input space and derive some properties of this formulation. In Section 4 we formulate nonlinear kernel classifiers as parameterless linear programs. Section 5 concludes the paper.

A word about our notation and background material. All vectors will be column vectors unless transposed to a row vector by a prime superscript $'$. The scalar product of two vectors x and y in the n-dimensional real space R^n will be denoted by $x'y$. For $x \in R^n$ and $p \in [1, \infty)$, the norm $\|x\|_p$ will denote the p-norm: $(\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ and $\|x\|_\infty$ will denote $\max_{1\leq i\leq n} |x_i|$. For $x \in R^n$, $(x_+)_i = \max\ \{0, x_i\}$, $i = 1, \ldots, n$. For an $m \times n$ matrix A, A_i will denote the ith row of A, $A_{\cdot j}$ will denote the jth column of A and A_{ij} will denote the element in row i and column j. The identity matrix in a real space of arbitrary dimension will be denoted by I, while a column vector of ones of arbitrary dimension will be denoted by e. For a general norm $\|\cdot\|$ on R^n, the dual norm $\|\cdot\|'$ on R^n is defined as

$$\|x\|' := \max_{\|y\|=1} x'y, \tag{1}$$

from which follows the generalized Cauchy-Schwarz inequality,

$$\pm x'y \leq |x'y| \leq \|x\|'\|y\|. \tag{2}$$

For $p, q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$, the p-norm and q-norm are dual norms by the classical Hölder inequality [1]. A norm $\|\cdot\|$ on R^n is said to be monotonic (or absolute) if either of the following equivalent conditions hold:

$$\begin{array}{rcl} x, y \in R^n,\ |x| \leq |y| & \Longrightarrow & \|x\| \leq \|y\|, \\ \|\ |x|\ \| & = & \|x\|\ \forall x \in R^n. \end{array} \tag{3}$$

For $p \in [1, \infty]$, the p-norm is monotonic. For $A \in R^{m\times n}$ and $B \in R^{n\times k}$, a *kernel* $K(A, B)$ maps $R^{m\times n} \times R^{n\times k}$ into $R^{m\times k}$. In particular, if x and y are column vectors in R^n then, $K(x', y)$ is a real number, $K(x', A')$ is a row vector in R^m and $K(A, A')$ is an $m \times m$ matrix. The base of the natural logarithm will be denoted by ε. A frequently used kernel in nonlinear classification is the Gaussian kernel [12, 4, 9] whose ijth element, $i = 1\ldots, m,\ j = 1\ldots, k$, is given by: $(K(A, B))_{ij} = \varepsilon^{-\mu\|A_i{}' - B_{\cdot j}\|^2}$, where $A \in R^{m\times n}$, $B \in R^{n\times k}$ and μ is a positive constant.

2 The Classification Problem

We consider the problem of classifying m points in the n-dimensional real space R^n, represented by the $m \times n$ matrix A, according to membership of each point A_i in the classes +1 or -1 as specified by a given $m \times m$ diagonal matrix D with ones or minus ones along its diagonal. For this problem the standard support vector machine with a linear kernel AA' [12, 4] is given by the following quadratic program for some $\nu > 0$:

$$\begin{array}{rrcl} \displaystyle\min_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \frac{1}{2}\|w\|_2^2 & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e, \\ & y & \geq & 0. \end{array} \tag{4}$$

As depicted in Figure 1, w is the normal to the bounding planes:

$$\begin{array}{rcl} x'w & = & \gamma + 1, \\ x'w & = & \gamma - 1, \end{array} \tag{5}$$

and γ determines their location relative to the origin. The first plane above bounds the class $A+$ points (circles) and the second plane bounds the class $A-$ points (asterisks) when the two classes are strictly linearly separable, that is when the slack variable $y = 0$. The linear separating surface is the plane

$$x'w = \gamma, \tag{6}$$

midway between the bounding planes (5). If the classes are linearly inseparable (as is the case depicted in Figure 1), then the two planes bound the two classes with a "soft margin" determined by a nonnegative slack variable y, that is:

$$\begin{array}{cccccccl} x'w & - & \gamma & + & y_i & \geq & +1, & \text{for } x' = A_i \text{ and } D_{ii} = +1, \\ x'w & - & \gamma & - & y_i & \leq & -1, & \text{for } x' = A_i \text{ and } D_{ii} = -1. \end{array} \tag{7}$$

The 1-norm of the slack variable y is minimized with parameter weight ν in (4) relative to the quadratic term in (4). This latter term which is twice the reciprocal of the square of the 2-norm distance $\frac{2}{\|w\|_2}$ between the two bounding planes of (5) in the n-dimensional space of $w \in R^n$, maximizes that distance, called the "margin". We immediately note three shortcomings of this standard SVM formulation:

(i) At a solution of (4), the error y is given by:

$$y = (e - D(Aw - e\gamma))_+, \tag{8}$$

which is a surrogate for the distance of misclassified points to their bounding planes (5), that is:

$$y = \frac{(e - D(Aw - e\gamma))_+}{\|w\|_2}. \tag{9}$$

A component of this error distance is depicted in Figure 1, for the dual $\|\cdot\|'$ of a general norm $\|\cdot\|$ on R^n.

(ii) The margin is represented by its square in (4), whereas the misclassification error surrogates are not squared.

(iii) The parameter ν typically requires extensive tuning experiments before its optimal value is arrived at, which is typically critical in determining correctness of the classifier.

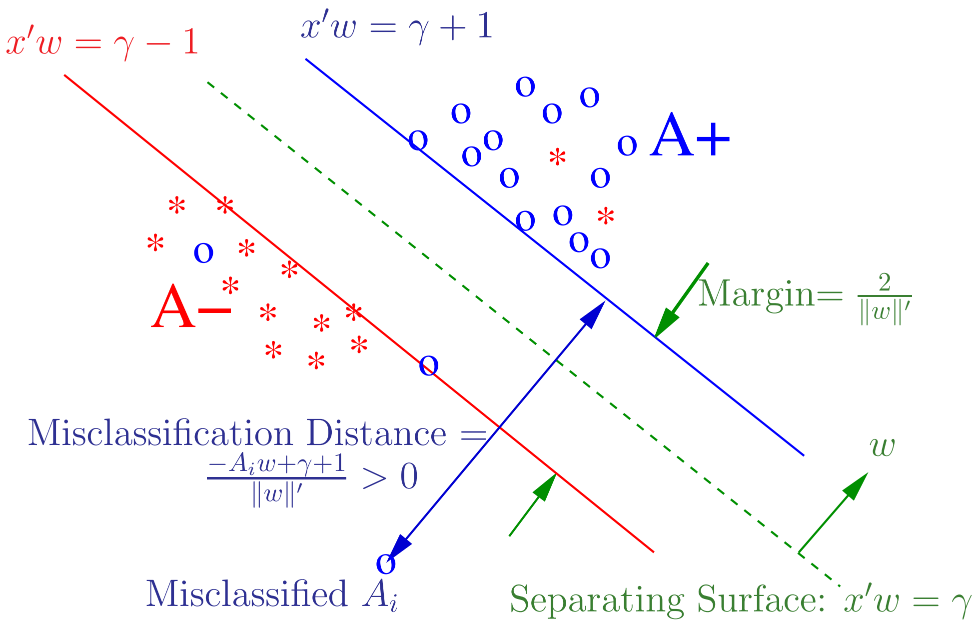

Figure 1: **The bounding planes (5) with margin $\frac{2}{\|w\|'}$, and the plane (6) approximately separating $A+$, the points (circles) represented by rows of A with $D_{ii} = +1$, from $A-$, the points (asterisks) represented by rows of A with $D_{ii} = -1$.**

One way to overcome these shortcomings is to use instead of the standard quadratic programming formulation (4) a linear programming formulation by replacing the 2-norm in (4) by a 1-norm as follows [3, 9]:

$$
\begin{array}{lrcl}
\displaystyle\min_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \|w\|_1 & & \\
\text{s.t.} & D(Aw - e\gamma) + y & \geq & e, \\
 & y & \geq & 0.
\end{array}
\tag{10}
$$

This reformulation in effect maximizes the margin, the distance between the two bounding planes of Figure 1, using a different norm, the ∞-norm, and results with a margin in terms of the 1-norm, $\frac{2}{\|w\|_1}$, instead of $\frac{2}{\|w\|_2}$ [8]. More generally, the margin between the bounding planes (5) measured using a p-norm, is given by $\frac{2}{\|w\|_q}$ [8], where $\frac{1}{p} + \frac{1}{q} = 1$, that is the p-norm and q-norm are dual norms for $1 \leq p,\, q \leq \infty$. However this linear programming formulation still does not rectify the shortcomings (i) and (iii) above. We propose instead the following formulation that takes care of all these shortcomings. We minimize the ratio of the sum of *real* distances of misclassified points to their bounding plane divided by the margin between the bounding planes, with all distances measured using the same *arbitrary monotonic* norm $\|\cdot\|$ on R^n. We require a monotonic norm in order to ensure that larger violations have larger distances from their bounding plane. This leads to the following fractional mathematical program:

$$
\min_{(w,\gamma)\in R^{n+1}} \frac{\frac{e'(e-D(Aw-e\gamma))_+}{\|w\|'}}{\frac{2}{\|w\|'}},
\tag{11}
$$

which is equivalent to the following linear program:

$$
\begin{array}{lrcl}
\displaystyle\min_{(w,\gamma,y)\in R^{n+1+m}} & e'y & & \\
\text{s.t.} & D(Aw - e\gamma) + y & \geq & e, \\
 & y & \geq & 0,
\end{array}
\tag{12}
$$

with the somewhat surprising fact, that this mathematical program is invariant with respect to the norm employed. That is, all norms lead to the same linear program.

A linear programming formulation for the linearly separable case was first introduced in [6], for the inseparable case in [10], and in more general form than (12) in [2].

We turn now to properties of the proposed linear programming formulation (12).

3 SVM as a Parameterless Linear Program and its Properties

We show in this section, by techniques similar to those of [2], that our parameterless robust linear programming SVM (12) has a trivial solution $w = 0$ if the arithmetic mean of points of $A+$ weighted by the number of points in $A+$, equals the arithmetic mean of points of $A-$ weighted by the number of points in $A-$, that is:

$$e'DA = m_1 \cdot (\frac{e'A+}{m_1}) - m_2 \cdot (\frac{e'A-}{m_2}) = 0. \tag{13}$$

Here, with a slight abuse of notation, $A+$ and $A-$ also denote respectively the submatrices of rows of A that are in classes $A+$ and $A-$. We note however that when (13) is satisfied, the trivial solution is not unique except in certain circumstances as outlined below.

Proposition 3.1 Existence of the Trivial Solution $w = 0$

(i) *For $m_1 \geq m_2$, the robust linear program SVM (12) has a trivial solution $w = 0$ if and only if:*

$$\sum_{i\in I+} u_i A_i = \sum_{j\in I-} A_j, \text{ for some } u_i \in [0,1],\ i \in I+,\ \sum_{i\in I+} u_i = m_2, \tag{14}$$

where $I+$ is the index set with cardinality m_1 of the rows of A that are in the class $A+$ and $I-$ is the index set with cardinality m_2 of the rows of A that are in the class $A-$. It follows in such a case that the minimum of the linear program (12) is given by:

$$e'y = 2m_2. \tag{15}$$

(ii) *Equality of the weighted means (13) implies (14).*

Proof

(i) Note first that the linear program is solvable because it is feasible and the objective function is bounded below by zero. Also, the assumption that $m_1 \geq m_2$ does not cause any loss of generality because the two classes can be interchanged if that were not the case. Furthermore, $(w = 0, \gamma, y)$ solves the linear program (12) if and only if:

$$y = (e + De\gamma)_+,\ e'y = e'u, \tag{16}$$

where u solves the dual of the linear program (12):

$$\begin{array}{lrcl} \max\limits_{(u)\in R^m} & e'u & & \\ \text{s.t.} & A'Du & = & 0, \\ & -e'Du & = & 0, \\ & 0 \leq u & \leq & e. \end{array} \tag{17}$$

It follows from (16) that:

$$\begin{array}{lcl} 1 + D_{ii}\gamma \leq 0 & \Longrightarrow & y_i = 0, \\ 1 + D_{ii}\gamma > 0 & \Longrightarrow & y_i = 1 + D_{ii}\gamma > 0. \end{array} \tag{18}$$

Under the assumption that $m_1 \geq m_2$, it follows that there are only two cases to be considered which satisfy (16) and (17), all other cases leading to a contradiction in the choice of γ.

Case I We have in this case that:

$$\begin{array}{rll} 1 + D_{ii}\gamma \leq 0, \ y_i = 0, & \forall i \in I+, & \gamma \leq -1, \\ 1 + D_{jj}\gamma > 0, \ y_j = 1 + D_{jj}\gamma > 0, & \forall j \in I-, & \gamma < 1, \\ u_j = 1, & \forall j \in I-, & \\ \sum_{i \in I+} u_i A_i - \sum_{j \in I-} A_j = 0, & & \\ 0 \leq u_i \leq 1, & \forall i \in I+, & \\ e'y = e'u = 2m_2. & & \end{array} \tag{19}$$

These conditions are equivalent to conditions (14) and (15) holding.

Case II We have in this case that:

$$\begin{array}{rll} 1 + D_{ii}\gamma > 0, \ y_i = 1 + D_{ii}\gamma > 0, & \forall i \in I+, & \gamma > -1, \\ 1 + D_{jj}\gamma > 0, \ y_j = 1 + D_{jj}\gamma > 0, & \forall j \in I-, & \gamma < 1, \\ u_i = 1, & \forall i \in I+, & \\ u_j = 1, & \forall j \in I-, & \\ \sum_{i \in I+} A_i - \sum_{j \in I-} A_j = 0, & & \\ e'y = e'u = m_1 + m_2 = 2m_1 = 2m_2. & & \end{array} \tag{20}$$

These conditions are equivalent to (14) and (15) if we set $u_i = 1$, $i \in I-$, in (14).

(ii) To see that equality of the weighted means (13) implies (14), just set $u_i = \frac{m_2}{m_1}$, $i \in I+$ in Case I above, while in Case II, (14) is directly implied by (13). $\square$

We give conditions that ensure that the linear program (12) has a nontrivial solution $w \neq 0$ in addition to the trivial solution.

Proposition 3.2 Existence of Nontrivial Solution to (12) *Let $m_1 \geq m_2$ and let the conditions (14) that ensure the existence of a trivial solution $w = 0$ hold. There exists another solution to the linear program (12) with $w \neq 0$ if and only if for some $h \in R^n$ the following system has* **no** *solution (r, ξ):*

$$\begin{array}{rcl} A'Dr & = & h, \\ e'Dr & = & 0, \\ -r + e\xi & \geq & 0, \\ e'r - 2m_2\xi & \geq & 0, \\ (r, \xi) & \geq & 0. \end{array} \tag{21}$$

Proof By Proposition 3.1 the linear program (12) has a solution with a trivial $w = 0$ if and only conditions (14) hold. Hence the linear program (12) has a nontrivial solution $w \neq 0$, if and only if:

$$\begin{array}{rcl} e'y & \leq & 2m_2, \\ D(Aw - e\gamma) + y & \geq & e, \\ y & \geq & 0, \end{array} \tag{22}$$

has a solution $(w \neq 0, \gamma, y)$. This is equivalent to

$$\begin{array}{rcl} -e'y + 2m_2\zeta & \geq & 0, \\ DAw - De\gamma + y - e\zeta & \geq & 0, \\ y & \geq & 0, \\ \zeta & > & 0, \\ -h'w & > & 0, \end{array} \tag{23}$$

has solution (w, γ, y, ζ) for some $h \in R^n$. By Motzkin's Theorem of the alternative [7], this is equivalent to:

$$\begin{array}{rcl} A'Dr - h\sigma & = & 0, \\ -e'Dr & = & 0, \\ -e\xi + r + s & = & 0, \\ 2m_2\xi - e'r + \rho & = & 0, \\ (\xi, r, s) & \geq & 0, \\ 0 \neq (\rho, \sigma) & \geq & 0, \end{array} \tag{24}$$

having no solution $(\xi, r, s, \rho, \sigma)$ for some $h \in R^n$. The case that (24) has no solution such that $(\sigma = 0, \rho > 0)$ follows from the assumption that the linear program (12) has a solution with $w = 0$, that is:

$$-De\gamma + y \geq e,\ y \geq 0,\ e'y \leq 2m_2, \tag{25}$$

which together with (24) having a solution with $(\sigma = 0, \rho > 0)$ leads to the contradiction:

$$0 \leq -r'De\gamma + r'y - r'e = e'y\xi - s'y - 2m_2\xi - \rho \leq 2m_2\xi - s'y - 2m_2\xi - \rho \leq -\rho < 0. \tag{26}$$

Hence, $(\sigma > 0,\ \rho \geq 0)$ in (24). Dividing all terms of (24) by σ and letting $\frac{r}{\sigma} \longrightarrow r$ and $\frac{\xi}{\sigma} \longrightarrow \xi$, reduces (24) not having a solution for some $h \in R^n$ to (21) not having a solution for some $h \in R^n$, which is the desired result.$\Box$

We consider now two examples that demonstrate the results given above.

Example 3.3 Exclusive-Or (XOR) *This classical example in R^2 for which the data is not linearly separable is given by:*

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix},\ De = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}. \tag{27}$$

For this example equality of the weighted means (13) holds, and hence by Proposition 3.1 the linear program (12) has a solution with $w = 0$, $\gamma \in [-1, 1]$ and $e'y = 4$. However, for $h \in R^2$ and $h_1 > h_2$, the system (21) can be shown (after some algebra) to have no solution (r, ξ). Hence by Proposition 3.2, the linear program (12) has another solution with $w \neq 0$. In fact such a solution is given by $w = [-2 \ \ -2]'$, $\gamma = -1$, $y = [0\ 0\ 0\ 4]'$, with a dual optimal solution of $u = e$. This results in the separating surface:

$$-2x_1 - 2x_2 = -1, \tag{28}$$

that misclassifies one point: $[1 \ \ 1]'$. This is the best that a linear classifier can do for this problem.

This example shows that the proposed robust parameterless linear programming SVM (12) is capable of handling such standard "troublemaker" examples for linear classifiers as the XOR example. However there are some synthetic examples that do not satisfy the criteria of Proposition 3.2 for having nonzero w for a solution. One such example is given in [2] which we cite now.

Example 3.4 Unique Trivial Solution *This is a one dimensional example with:*

$$A = \begin{bmatrix} 1 \\ 2 \\ -1 \\ 0 \\ 4 \end{bmatrix}, \ De = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}. \tag{29}$$

For this example condition (14) is satisfied by $u_1 = 1,\ u_2 = 0,\ u_3 = 1$, and hence the linear programming SVM (12) has the trivial solution $w = 0$. However, it can be checked for this example, that the system (21) has a solution for any $h \in R$ and hence the linear program (12) does not have a a solution with a nontrivial w. To overcome this difficulty for this somewhat unusual example the linear program (12) objective can be weighted by the reciprocal of the number of points in each class resulting in an average error distance for each class instead of the sum of the error distances. In fact this leads to the robust linear program of [2] which we state explicitly now:

$$\begin{array}{ll} \displaystyle\min_{(w,\gamma,y)\in R^{n+1+m}} & \displaystyle\frac{1}{m_1}\sum_{i\in I+} y_i + \frac{1}{m_2}\sum_{i\in I-} y_i \\ \text{s.t.} & \begin{array}{rcl} D(Aw - e\gamma) + y & \geq & e, \\ y & \geq & 0, \end{array} \end{array} \tag{30}$$

which has a solution with nonzero w for this example. In fact, even though (30) was not considered to be a support vector machine in [2], it is in fact an SVM using the same fractional justification that we have utilized to derive linear program (12).

We turn now to our nonlinear parameterless kernel formulation.

4 Nonlinear SVM Classifiers

To generate a parameterless kernel-based nonlinear classifier we proceed in a manner similar to that for the linear classifier, except that instead of measuring errors by distances in the input space, we measure distances in a higher dimensional space feature space [12, 11, 5] by utilizing a dual representation [9] as follows.

We start with the linear 1-norm SVM (10) and use the dual representation $w = A'Du$ [9] which leads to the following linear program:

$$\begin{array}{rrcl} \min\limits_{(u,\gamma,y)\in R^{m+1+m}} & \nu e'y + \|A'Du\|_1 & & \\ \text{s.t.} & D(AA'Du - e\gamma) + y & \geq & e, \\ & y & \geq & 0. \end{array} \tag{31}$$

If we now replace AA' by any nonlinear kernel $K(A, A')$ as defined in the Introduction and replace $\|A'Du\|_1$ by $\|u\|_1$ in the objective function we obtain a standard linear programming nonlinear kernel SVM:

$$\begin{array}{rrcl} \min\limits_{(u,\gamma,y)\in R^{m+1+m}} & \nu e'y + \|u\|_1 & & \\ \text{s.t.} & D(K(A, A')Du - e\gamma) + y & \geq & e, \\ & y & \geq & 0. \end{array} \tag{32}$$

Again this problem has a parameter ν in the objective function which balances the surrogate distance error

$$e'y = e'(e - D(K(A, A')Du - e\gamma))_+, \tag{33}$$

against the margin $\frac{2}{\|u\|_1}$ in the m-dimensional feature space represented by the linear bounding planes:

$$\begin{array}{rcl} z'u & = & \gamma + 1, \\ z'u & = & \gamma - 1, \end{array} \tag{34}$$

where:

$$z = K(x', A')D. \tag{35}$$

A parameterless linear program ensues if we minimize the ratio of the sum of the real distances of misclassified points to their bounding planes planes divided by the margin between the bounding planes in the m-dimensional space. We have then the fractional mathematical program:

$$\min_{(u,\gamma)\in R^{m+1}} \frac{\frac{e'(e-D(K(A,A')Du-e))_+}{\|u\|'}}{\frac{2}{\|u\|'}}. \tag{36}$$

This is equivalent to the following linear program:

$$\begin{array}{rrcl} \min\limits_{(u,\gamma,y)\in R^{m+1+m}} & e'y & & \\ \text{s.t.} & D(K(A, A')Du - e\gamma) + y & \geq & e, \\ & y & \geq & 0. \end{array} \tag{37}$$

This is the desired parameterless linear program for a nonlinear kernel SVM classifier.

5 Summary & Conclusion

We have reduced both linear and nonlinear kernel classification to a parameterless linear programming problem which is independent of the norm used to measure the margin between the bounding planes and the distance of misclassified points to their bounding planes. Past work [2] has made use of these formulations but without pointing out their support vector machine aspect of implicitly maximizing the margin between bounding planes. It is hoped that this connection will be investigated more thoroughly and computationally with the possible outcome of generating of one of the simplest representations of a support vector machine.

References

[1] E. F. Beckenbach and R. Bellman. *Inequalities.* Springer–Verlag, Berlin, 1961.

[2] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[3] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)*, pages 82–90, San Francisco, California, 1998. Morgan Kaufmann. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps.

[4] V. Cherkassky and F. Mulier. *Learning from Data - Concepts, Theory and Methods.* John Wiley & Sons, New York, 1998.

[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge, MA, 2000.

[6] O. L. Mangasarian. Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13:444–452, 1965.

[7] O. L. Mangasarian. *Nonlinear Programming.* SIAM, Philadelphia, PA, 1994.

[8] O. L. Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24:15–23, 1999. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07r.ps.

[9] O. L. Mangasarian. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135–146, Cambridge, MA, 2000. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps.

[10] F. W. Smith. Pattern classifier design by linear programming. *IEEE Transactions on Computers*, C-17:367–372, 1968.

[11] A. Smola and B. Schölkopf. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[12] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, second edition, 2000.