Duality and Geometry in SVM Classifiers

2 authors:

Kristin P. Bennett
Rensselaer Polytechnic Institute
196 PUBLICATIONS **9,891** CITATIONS

SEE PROFILE

Erin J. Bredensteiner
University of Evansville
6 PUBLICATIONS **876** CITATIONS

SEE PROFILE

Duality and Geometry in SVM Classifiers

Kristin P. Bennett BENNEK@RPI.EDU
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Erin J. Bredensteiner EB6@EVANSVILLE.EDU
Department of Mathematics, University of Evansville, Evansville, IN 47722 USA

Abstract

We develop an intuitive geometric interpretation of the standard support vector machine (SVM) for classification of both linearly separable and inseparable data and provide a rigorous derivation of the concepts behind the geometry. For the separable case finding the maximum margin between the two sets is equivalent to finding the closest points in the smallest convex sets that contain each class (the convex hulls). We now extend this argument to the inseparable case by using a reduced convex hull reduced away from outliers. We prove that solving the reduced convex hull formulation is exactly equivalent to solving the standard inseparable SVM for appropriate choices of parameters. Some additional advantages of the new formulation are that the effect of the choice of parameters becomes geometrically clear and that the formulation may be solved by fast nearest point algorithms. By changing norms these arguments hold for both the standard 2-norm and 1-norm SVM.

1. Introduction

Support vector machines (SVM) are a very robust methodology for inference with minimal parameter choices. This should translate into the popular adaptation of SVM in many application domains by non-SVM experts. The popular success of prior methodologies like neural networks, genetic algorithms, and decision trees was enhanced by the intuitive motivation of these approaches, that in some sense enhanced the end users ability to develop applications independently and have a sense of confidence in the results. How do you sell a SVM to a consulting client, manager, etc? What quick description would allow an end user to grasp the fundamentals of SVM necessary for a successful application? There are three key ideas needed to understand SVM: maximizing margins, the dual formulation, and kernels. Most people intuitively grasp the idea that maximizing margins should help improve generalization. But changing from the primal to dual formulation is typically black magic for those uninitiated in duality theory. Duality is really the key concept frequently missing in the understanding of SVM.

In this paper we provide an intuitive geometric explanation of SVM for classification from the dual perspective along with a mathematically rigorous derivation of the ideas behind the geometry. We begin with an explanation of the geometry of SVM based on the idea of convex hulls. For the separable case, this geometric explanation has existed in various forms (Vapnik, 1996; Mangasarian, 1965; Keerthi et al., 1999; Bennett & Bredensteiner, in press). The new contribution is the adaptation of the convex hull argument for the inseparable case to the most commmonly used 2-norm and 1-norm soft margin SVM. The primal form resulting from this argument can be regarded as an especially elegant minor variant of the ν-SVM formulation (Schölkopf et al., 2000) or a soft margin form of the MSM method (Mangasarian, 1965). Related geometric ideas for the ν-SVM formulation were developed independently by Crisp and Burges (1999).

The primary contributions of this paper are:

- A simple intuitive explanation of SVM based on (reduced) convex hulls that allows nonexperts to grasp geometrically the main concepts of SVM.
- A new primal maximum (soft) margin SVM formulation that has as it's dual the problem of finding the nearest neighbors in the (reduced) convex hulls. Major benefits of this formulation are that the effects of the misclassification parameter choice are very clear and that it is amenable to solution with very fast closest points in poly-

tope algorithms (Keerthi et al., 1999) and a minor variant of sequential minimal optimization (SMO) (Platt, 1998).

- Proof of the equivalence, for appropriate choices of parameters, between the primal and dual forms of the reduced-convex-hull SVM to the primal and dual forms of the classic SVM.
- Extensions of the reduced convex hull arguments to the sparse 1-norm SVM formulation and a new infinity-norm SVM.

For compactness, we adopt matrix notation instead of the more typical summation notation. In particular, for a column vector x in the n-dimensional real space R^n, x_i denotes the i^{th} component of x. The notation $A \in R^{m \times n}$ will signify a real $m \times n$ matrix. For such a matrix, A_i will denote the i^{th} row. The transpose of x and A are denoted x' and A' respectively. The dot product of two vectors x and w is denoted by $x'w$. A vector of ones in a space of arbitrary dimension is denoted by e. The scalar 0 and a vector of zeros are both represented by 0. Thus, for $x \in R^m$, $x > 0$ implies that $x_i > 0$ for $i = 1, \ldots, m$. In general, for $x, y \in R^m$, $x > y$ implies that $x_i > y_i$ for $i = 1, \ldots, m$. Similarly, $x \geq y$ implies that $x_i \geq y_i$ for $i = 1, \ldots, m$. Several norms are used. The 1-norm of x, $\sum_{i=1}^{m} |x_i|$, is denoted by $\|x\|_1$. The 2-norm or Euclidean norm of x, $\sqrt{\sum_{i=1}^{m} x_i^2} = \sqrt{x'x}$, is denoted by $\|x\|$ and $\|x\|^2 = x'x$. The infinity-norm of x, $max_{i=1,\ldots,m}|x_i|$ is denoted by $\|x\|_\infty$.

2. Geometric Intuition: Separable Case

Assume that we are trying to construct a linear discriminant to separate two separable sets $\mathcal{A}$ and $\mathcal{B}$. Specifically, this linear discriminant is the plane $x'w = \gamma$, where w is the normal of the plane and $\frac{|\gamma|}{\|w\|}$ is the Euclidean distance of the plane from the origin. Let the coordinates of the points in $\mathcal{A}$ be given by the m rows of the $m \times n$ matrix A. Let the coordinates of the points in $\mathcal{B}$ be given by the k rows of the $k \times n$ matrix B. We say that the sets are **linearly separable** if w and γ exist such that: $Aw > e\gamma$ and $Bw < e\gamma$ where e is a vector of ones of appropriate dimension.

Figure 1 shows two such separable sets and two of the infinitely many possible planes that separate the sets with 100% accuracy. Which separable plane is preferable? With no other knowledge of the data, most people will prefer the solid line because it is further from each of the sets. In the case of the dotted line, small changes in the data will produce misclassification errors. So an intuitive idea would be to construct the plane that maximizes the minimum distance from the plane to each set. In fact we know this intuition coincides with the results in statistical learning theory (Vapnik, 1996) and is substantiated by results in Shawe-Taylor et al. (1998).

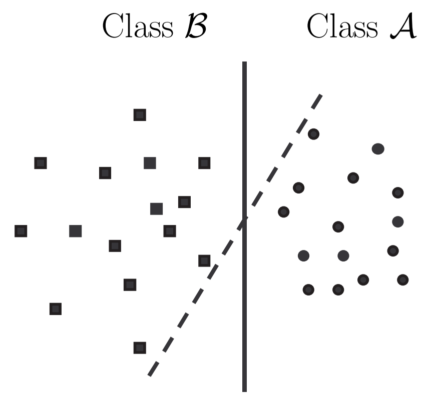

Figure 1. Which plane is best?

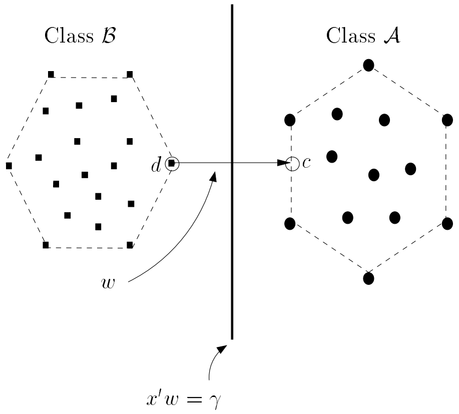

Figure 2. The two closest points of the convex hulls determine the separating plane.

One way to construct the plane as far as possible from both sets is to construct the smallest convex sets that contain all the data in each class (i.e. the convex hull) and find the closest points in those sets. Then, construct the line segment between the two points. The plane, orthogonal to the line segment, that bisects the line segment is chosen to be the separating plane. See, for example, Figure 2. The smallest convex set containing a set of points is called a convex hull. The convex hulls of $\mathcal{A}$ and $\mathcal{B}$ are shown with dashed lines. The convex hull consists of all points that can be written as a convex combination of the points in the orig-

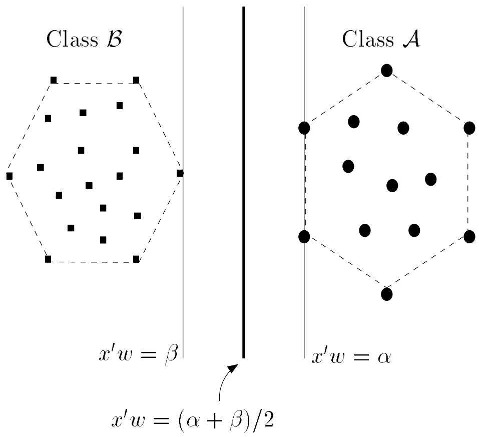

Figure 3. The primal problem maximizes the distance between two parallel supporting planes.

inal set. A convex combination of points is a positive weighted combination where the weights sum to one, e.g. a convex combination c of points in $\mathcal{A}$ is defined by $c' = u_1 A_1 + u_2 A_2 + \ldots + u_m A_m = u'A$ where $u \in R^m$, $u \geq 0$, and $\sum_{i=1}^{m} u_i = e'u = 1$ and a convex combination d of points in $\mathcal{B}$ is defined by $d' = v_1 B_1 + v_2 B_2 + \ldots + v_k B_k = v'B$ where $v \in R^k$, $v \geq 0$, and $e'v = 1$.

The problem of finding the two closest points in the convex hulls can be written as an optimization problem (C-Hull):

$$\begin{array}{ll} \min\limits_{u,v} & \frac{1}{2}\left\|A'u - B'v\right\|^2 \\ s.t. & e'u = 1 \quad e'v = 1 \quad u \geq 0 \quad v \geq 0 \end{array} \qquad (1)$$

The linear discriminant, $x'w = \gamma$, is constructed from the results of C-Hull (1). The normal w is exactly the vector between the two closest points in the convex hulls. Let $\bar{u}$ and $\bar{v}$ be an optimal solution of (1). The normal of the plane is the difference between the closest points, $c = A'\bar{u}$ and $d = B'\bar{v}$. Thus $w = c - d = A'\bar{u} - B'\bar{v}$. The threshold, γ, is the distance from the origin to the point halfway between the two closest points along the normal w: $\gamma = (\frac{c+d}{2})'w = \frac{(\bar{u}'Aw + \bar{v}'Bw)}{2}$.

There is an alternative approach to finding the best separating plane. Consider a set of parallel supporting planes as in Figure 3. These planes are positioned so that all the points in $\mathcal{A}$ satisfy $x'w \geq \alpha$ and at least one point in $\mathcal{A}$ lies on the plane $x'w = \alpha$. Similarly, all points in $\mathcal{B}$ satisfy $x'w \leq \beta$ and at least one point in $\mathcal{B}$ lies on the plane $x'w = \beta$. The optimal separating plane can be found by maximizing the distance between these two supporting hyperplanes. The distance between the two parallel supporting hyperplanes is $\frac{\alpha-\beta}{\|w\|}$. Therefore the distance between the two planes can be maximized by minimizing $\|w\|$ and maximizing $(\alpha - \beta)$.

The problem of maximizing the distance between the two supporting hyperplanes can be written as the following optimization problem (C-Margin):

$$\begin{array}{ll} \min\limits_{w,\alpha,\beta} & \frac{1}{2}\left\|w\right\|^2 - (\alpha - \beta) \\ s.t. & Aw - \alpha e \geq 0 \quad -Bw + \beta e \geq 0 \end{array} \qquad (2)$$

The final separating plane is the plane halfway between the two parallel planes: $x'\hat{w} = \frac{\hat{\alpha}+\hat{\beta}}{2}$. Note that the maximum distance between the supporting planes yields the distance between the two convex hulls. The two closest points for each convex hull must then lie on the supporting planes. The line segment between the two closest points in the convex hulls must be orthogonal to the supporting planes, otherwise a contradiction exists. Such a contradiction could be that either the two supporting planes are not as far apart as possible or these two points are not the closest points in the convex hulls. Therefore the solutions of both approaches are exactly the same. This is an example of duality. As stated later in Theorem 4.1, the dual of C-Margin (2) is C-Hull (1). See Bennett and Bredensteiner (in press) for the derivation. We can formulate and solve the problem in either space as is convenient for us. If there is no degeneracy, we will always get the same plane.

The primal C-Margin (2) and dual C-Hull (1) formulations provide a unifying framework for explaining other SVM formulations. By transforming C-Margin (2) into mathematically equivalent optimization problems, different SVM formulations are produced. If we set $\alpha - \beta = 2$ by defining $\alpha = \gamma + 1$ and $\beta = \gamma - 1$, then Problem (2) becomes the standard primal SVM 2-norm formulation (Vapnik, 1996)

$$\begin{array}{ll} \min\limits_{w,\gamma} & \frac{1}{2}\left\|w\right\|^2 \\ s.t. & Aw - (\gamma + 1)e \geq 0 \quad -Bw + (\gamma - 1)e \geq 0 \end{array} \qquad (3)$$

In fact, as stated in Theorem 4.2, the classic 2-norm SVM (3) and C-Margin (2) are mathematically equivalent on separable problems. They will produce the exact same separating plane or an equally good plane if multiple solutions exist (see Burges & Crisp, 1999).

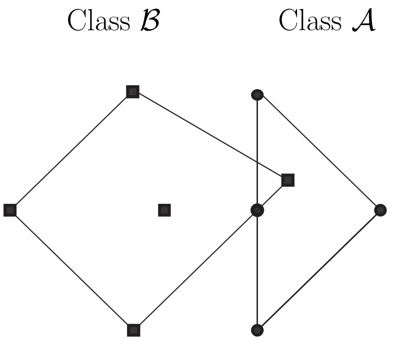

Figure 4. The convex hulls of inseparable sets intersect.

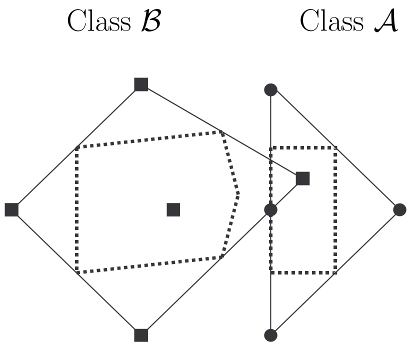

Figure 5. Convex hull and reduced convex hull with $K = 2$.

3. Geometric Intuition: Inseparable Case

For inseparable problems, the convex hulls of the two sets will intersect. Consider Figure 4. The difficult-to-classify points of one set will be in the convex hull of the other set. In a problem amenable to linear classification, most points of one class will not be in the convex hull of the other. If we could restrict the influence of outlying points then we could return to the usual convex hull problem. It is undesirable to let one point, particularly a difficult point, excessively influence the solution. Therefore, we want the solution to be based on a lot of points, not just a few bad ones. Say we want the solution to depend on at least K points. This can be done by contracting or reducing the convex hull by putting an upperbound on the multiplier in the convex combination for each point. The reduced convex hull is defined as follows.

Definition 3.1 (Reduced Convex Hull). *The set of all convex combinations $c = A'u$ of points in $\mathcal{A}$ where $e'u = 1$, $0 \leq u \leq De$, $D < 1$.*

Typically we choose $D = \frac{1}{K}$ and $K > 1$. Note that the reduced convex hull is nonempty as long as $K \leq m$ where m is the number of points in set $\mathcal{A}$.

We reduce our feasible set away from the boundaries of the convex hulls so that no extreme point or noisy point can excessively influence the solution. In Figure 5, the reduced convex hulls with $K = 2$ are given. Note that the reduced sets no longer intersect. Further examples of reduced convex hulls can be seen in Crisp and Burges (1999), who refer to our reduced convex hulls as "soft convex hulls". We believe that this is a misnomer because softening implies that the convex hulls are expanding but in fact they are being reduced. As we will see later, the concept of reducing the convex hulls to avoid error is the dual concept to enlarging margins by softening them to allow error. For sets with lots of redundant points, reducing the convex hull has little effect. But for a set with a single outlier the effect is quite marked. Note that for small D the reduced convex hulls no longer intersect. In general, we will need to choose K sufficiently large to ensure that the convex hulls do not intersect. We can now proceed as in the separable case using the reduced convex hulls instead. We will minimize the distance between the reduced convex hulls so that a few bad points will not dominate the solution.

The problem of finding two closest points in the reduced convex hulls can be written as an optimization problem (RC-Hull):

$$\begin{array}{ll} \min\limits_{u,v} & \frac{1}{2}\left\|A'u - B'v\right\|^2 \\ s.t. & e'u = 1 \quad e'v = 1 \quad 0 \leq u \leq De \quad 0 \leq v \leq De \end{array} \tag{4}$$

Immediately we can see the effect of our choice of parameter $D = \frac{1}{K}$. Note that each point can contribute no more than $\frac{1}{K}$ to the optimal solution. So the solution will be robust in some sense since it depends on at least $2K$ points. If K is too large or conversely D is too small the problem will be infeasible. So K must be smaller than the number of points in each set. Increasing D larger than 1 will produce no advantage over the solution where $D = 1$. If we have varying confidence in the points or if our classes are skewed in size we can choose different values of D for each point or class. The reader should consult (Schölkopf et al., 2000) for a more formal derivation of these and additional properties for the ν-SVM formulation which also has been shown to solve the closest points between the reduced convex hulls problem (Crisp & Burges, 1999). RC-Hull (4) is suitable for solution by nearest points in convex polytope algorithms; see (Keerthi et al., 1999).

If we add a soft margin error term to the separable C-Margin Problem (2), we get the following problem for the inseparable case (RC-Margin):

$$\begin{array}{ll} \min\limits_{w,\xi,\eta,\alpha,\beta} & D(e'\xi + e'\eta) + \frac{1}{2}\|w\|^2 - \alpha + \beta \\ s.t. & Aw - \alpha e + \xi \geq 0 \quad \xi \geq 0 \\ & -Bw + \beta e + \eta \geq 0 \quad \eta \geq 0 \end{array} \tag{5}$$

with $D = \frac{1}{K} > 0$. As we will prove in Theorem 4.3, the dual of RC-Margin (5) is exactly RC-Hull (4) which finds the closest points in the reduced convex hulls.

As in the linearly separable case, one transformation of this problem is to fix $\alpha - \beta$ by setting $\alpha = \gamma + 1$ and $\beta = \gamma - 1$. This results in the classic support vector machine approach (Vapnik, 1996):

$$\begin{array}{ll} \min\limits_{w,\xi,\eta,\gamma} & C(e'\xi + e'\eta) + \frac{1}{2}\|w\|^2 \\ s.t. & Aw - \gamma e + \xi \geq e \\ & -Bw + \gamma e + \eta \geq e \\ & \xi \geq 0 \quad \eta \geq 0 \end{array} \tag{6}$$

where $C > 0$ is a fixed constant. Note that the constant C is now different due to an implicit rescaling of the problem. As we will show in Theorem 4.4 the RC-Margin (5) and classic inseparable SVM (6) are equivalent for appropriate choices of C and D.

4. Equivalence to Classic Formulation

We now rigorously examine the claims of the previous section. We begin with the separable case. For both the separable and inseparable cases, the theorems establish that the dual of our SVM (soft) maximum margin formulation is exactly the (reduced) convex hull formulation and that our (reduced) convex hull based SVM formulations are equivalent to the classic SVM form for appropriate choices of parameters. The first theorem states that the problem of finding the two closest points in the convex hulls of two separable sets is the Wolfe dual (or equivalently Lagrangian dual) of the problem of finding the best separating plane.

Theorem 4.1 (Convex Hulls is Dual). *The Wolfe dual of C-Margin SVM (2) is the closest points of the convex hull problem C-Hull (1) or :*

$$\begin{array}{ll} \max\limits_{u,v} & -\frac{1}{2}\|A'u - B'v\|^2 \\ s.t. & e'u = e'v = 1, \; u \geq 0, \; v \geq 0 \end{array} \tag{7}$$

Proof of this theorem can be found in full detail in (Bennett & Bredensteiner, in press) or can easily be derived as a variant of the corresponding theorem for the inseparable case.

Problem C-Margin (2), the primal form of the dual C-Hull of finding the closest two points in the convex hulls, is equivalent to the classic inseparable 2-norm SVM (3) in Vapnik (1996). Specifically, every solution to one problem can be used to construct a corresponding solution to the other by simple scaling. The theorem assumes that the degenerate solution $w = 0$ is not optimal. This is equivalent to saying that the convex hulls do not intersect. For convex quadratic programs with linear constraints, a solution is optimal if and only if it (along with the corresponding Lagrangian multipliers) satisifies the Karush-Kuhn-Tucker (KKT) optimality conditions of primal feasibility, dual feasibility, and complementary slackness. We call a set of primal C-Margin and dual C-Hull solutions a KKT point. We can establish the equivalence of the C-Margin/C-Hull formulations with the classic inseparable SVM formulation by showing that a KKT point of one can be used to derive a KKT point of the other. The optimal separating plane of one solution will also be optimal for the other form, but the weights and threshold are scaled by a constant.

Theorem 4.2 (Equivalence of Separable Forms). *Assume C-Margin (2) has a solution with $\|\hat{w}\| > 0$. Then $(\bar{w}, \bar{\gamma}, \bar{u}, \bar{v})$ is a KKT point of the classic separable SVM (3) if and only if $(\hat{w}, \hat{\alpha}, \hat{\beta}, \hat{u}, \hat{v})$ is a KKT point of C-Margin (2) where $\delta = e'\bar{u} = \frac{2}{\hat{\alpha}-\hat{\beta}}$, $\hat{w} = \frac{\bar{w}}{\delta}$, $\hat{\alpha} = \frac{\bar{\gamma}+1}{\delta}$, $\hat{\beta} = \frac{\bar{\gamma}-1}{\delta}$, $\hat{u} = \frac{\bar{u}}{\delta}$, and $\hat{v} = \frac{\bar{v}}{\delta}$.*

Proof. Each KKT point of the classic separable SVM (3) satisfies:

$$\begin{array}{ll} A\bar{w} - (\bar{\gamma}+1)e \geq 0 & -B\bar{w} + (\bar{\gamma}-1)e \geq 0 \\ \bar{u}'(A\bar{w} - (\bar{\gamma}+1)e) = 0 & \bar{v}'(-B\bar{w} + (\bar{\gamma}-1)e) = 0 \\ \bar{w} = A'\bar{u} - B'\bar{v} & e'\bar{u} = e'\bar{v} \\ \bar{u} \geq 0 & \bar{v} \geq 0. \end{array} \tag{8}$$

Dividing each constraint by δ or δ^2 as appropriate yields a KKT point of the C-Margin SVM (2) satisfying:

$$\begin{array}{ll} A\hat{w} - \hat{\alpha}e \geq 0 & -B\hat{w} + \hat{\beta}e \geq 0 \\ \hat{u}'(A\hat{w} - \hat{\alpha}e) = 0 & \hat{v}'(-B\hat{w} + \hat{\beta}e) = 0 \\ \hat{w} = A'\hat{u} - B'\hat{v} & 1 = e'\hat{u} = e'\hat{v} \\ \hat{u} \geq 0 & \hat{v} \geq 0. \end{array} \tag{9}$$

Similarly, multiplying the KKT conditions (9) of C-Margin (2) by $\delta = \frac{2}{\hat{\alpha}-\hat{\beta}}$ or δ^2 yields the KKT conditions (8) of the standard separable SVM (3). We know $\hat{\alpha} - \hat{\beta} > 0$ because by strong duality the primal and dual objectives will be equal thus

$$\frac{1}{2}\|\hat{w}\|^2 - \hat{\alpha} + \hat{\beta} = -\frac{1}{2}\|A'u - B'v\|^2 < 0.$$

□

The theorems can be directly generalized to the inseparable case based on reduced convex hulls. The Wolfe dual (for example, see Mangasarian, 1969) of RC-Margin (5) is precisely the closest points in the reduced convex hull problem, RC-Hull (4).

Theorem 4.3 (Reduced Convex Hulls is Dual). *The Wolfe dual of the RC-Margin (5) is RC-Hull (4) or equivalently:*

$$\begin{array}{ll} \max\limits_{u,v} & -\frac{1}{2}\left\|A'u - B'v\right\|^2 \\ s.t. & e'u = e'v = 1,\ De \geq u \geq 0,\ De \geq v \geq 0 \end{array} \tag{10}$$

Proof. The dual problem maximizes the Lagrangian function of (5), $L(w, \alpha, \beta, \xi, \eta, u, v, r, s)$, subject to the constraints that the partial derivatives of the Lagrangian with respect to the primal variables are equal to zero (Mangasarian, 1969). Specifically, the dual of (5) is:

$$\begin{array}{ll} \max\limits_{w,\alpha,\beta,\xi,\eta,u,v,r,s} & L(w, \alpha, \beta, \xi, \eta, u, v, r, s) = \\ & \frac{1}{2}\left\|w\right\|^2 - \alpha + \beta + De'\xi + De'\eta \\ & \quad - u'(Aw - \alpha e + \xi) \\ & \quad - v'(-Bw + \beta e + \eta) - r'\xi - s'\eta \\ s.t. & \frac{\partial L}{\partial w} = w - A'u + B'v = 0 \\ & \frac{\partial L}{\partial \alpha} = -1 + e'u = 0,\quad u \geq 0 \\ & \frac{\partial L}{\partial \beta} = 1 - e'v = 0,\quad v \geq 0 \\ & \frac{\partial L}{\partial \xi} = De - u = r \geq 0 \\ & \frac{\partial L}{\partial \eta} = De - v = s \geq 0 \end{array} \tag{11}$$

where $\alpha, \beta \in R,\ w \in R^n,\ \xi, u, r \in R^m$, and $\eta, v, s \in R^k$. To simplify the problem, substitute in $w = (A'u - B'v)$, $r = De - u$ and $s = De - v$:

$$\begin{array}{ll} \max\limits_{\alpha,\beta,u,v} & \frac{1}{2}\left\|A'u - B'v\right\|^2 - (\alpha - \beta) + De'\xi + De'\eta \\ & \quad - u'A(A'u - B'v) + v'B(A'u - B'v) \\ & \quad + e'u\alpha - e'v\beta - u'\xi - v'\eta \\ & \quad - De'\xi - De'\eta + u'\xi + v'\eta \\ s.t. & e'u = e'v = 1,\ De \geq u \geq 0,\ De \geq v \geq 0 \end{array} \tag{12}$$

and then simplify to yield RC-Hull (10). □

Optimizing the reduced-convex-hull form of SVM with parameter D is equivalent to optimizing the classic 2-norm SVM (6) with parameter C. The parameters D and C are related by multiplication of a constant factor based on the size of the optimal margin. If the appropriate values of D and C are chosen, then once again a KKT point of one will be a KKT point of the other. A similar result for the ν-SVM formulation is given in Proposition 13 in Schölkopf et al. (2000).

Theorem 4.4 (Equivalence of Inseparable Forms). *Assume RC-Margin (5) has a solution with $\|\hat{w}\| > 0$. Then $(\bar{w}, \bar{\gamma}, \bar{\xi}, \bar{\eta}, \bar{u}, \bar{v})$ is a KKT point of the classic inseparable SVM (6) with parameter C if and only if $(\hat{w}, \hat{\alpha}, \hat{\beta}, \hat{\xi}, \hat{\eta}, \hat{u}, \hat{v})$ is a KKT point of RC-Margin (5) with parameter D where $\delta = e'\bar{u} = \frac{2}{\hat{\alpha}-\hat{\beta}}$, $\hat{w} = \frac{\bar{w}}{\delta}$, $\hat{\alpha} = \frac{\bar{\gamma}+1}{\delta}$, $\hat{\beta} = \frac{\bar{\gamma}-1}{\delta}$, $\hat{\xi} = \frac{\bar{\xi}}{\delta}$, $\hat{\eta} = \frac{\bar{\eta}}{\delta}$, $\hat{u} = \frac{\bar{u}}{\delta}$, $\hat{v} = \frac{\bar{v}}{\delta}$, and $D = \frac{C}{\delta}$.*

Proof. Each KKT point of the classic SVM (6) with parameter C satisfies:

$$\begin{array}{ll} A\bar{w} - (\bar{\gamma}+1)e + \bar{\xi} \geq 0 & \bar{w} = A'\bar{u} - B'\bar{v} \\ -B\bar{w} + (\bar{\gamma}-1)e + \bar{\eta} \geq 0 & e'\bar{u} = e'\bar{v} \\ \bar{\xi} \geq 0 & Ce \geq \bar{u} \geq 0 \\ \bar{\eta} \geq 0 & Ce \geq \bar{v} \geq 0 \\ \bar{u}'(A\bar{w} - (\bar{\gamma}+1)e + \bar{\xi}) = 0 & \bar{\xi}(Ce - \bar{u}) = 0 \\ \bar{v}'(-B\bar{w} + (\bar{\gamma}-1)e + \bar{\eta}) = 0 & \bar{\eta}(Ce - \bar{v}) = 0. \end{array} \tag{13}$$

Dividing each constraint by δ or δ^2 as appropriate yields a KKT point of the RC-Margin (5) with parameter D satisfying:

$$\begin{array}{ll} A\hat{w} - \hat{\alpha}e + \hat{\xi} \geq 0 & \hat{w} = A'\hat{u} - B'\hat{v} \\ -B\hat{w} + \hat{\beta}e + \hat{\eta} \geq 0 & 1 = e'\hat{u} = e'\hat{v} \\ \hat{\xi} \geq 0 & De \geq \hat{u} \geq 0 \\ \hat{\eta} \geq 0 & De \geq \hat{v} \geq 0 \\ \hat{u}'(A\hat{w} - \hat{\alpha}e + \hat{\xi}) = 0 & \hat{\xi}(De - \hat{u}) = 0 \\ \hat{v}'(-B\hat{w} + \hat{\beta}e + \hat{\eta}) = 0 & \hat{\eta}(De - \hat{v}) = 0. \end{array} \tag{14}$$

Similarly, multiplying the KKT conditions (14) of the RC-Margin SVM (5) with parameter D by $\delta = \frac{2}{\hat{\alpha}-\hat{\beta}}$ or δ^2 yields the KKT conditions (13) of the standard SVM (6) with parameter C. We know $\hat{\alpha} - \hat{\beta} > 0$ by equality of the primal and dual objectives

$$\begin{aligned} De'\hat{\xi} + De'\hat{\eta} + \frac{1}{2}\left\|\hat{w}\right\|^2 - \hat{\alpha} + \hat{\beta} \\ = -\frac{1}{2}\left\|A'u - B'v\right\|^2 < 0. \end{aligned}$$

□

This theorem proves that for appropriate parameter choice, the solution set of optimal parallel max-margin planes produced by the classic SVM with parameter C ($x'\bar{w} = \bar{\gamma} + 1$ and $x'\bar{w} = \bar{\gamma} - 1$) will also be optimal for the reduced-convex-hull problem with parameter D ($x'\hat{w} = \hat{\alpha}$ and $x'\hat{w} = \hat{\beta}$) using the relationship defined above and vice versa. But it is not true that the sets of final single separating planes produced by the two methods are identical. The plane bisecting the closest points in the reduced convex hulls i.e.

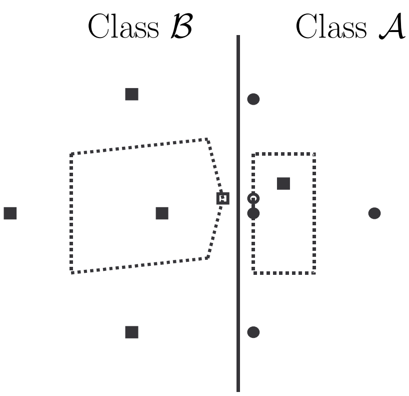

Figure 6. Optimal plane bisecting the closest points in the reduced convex hulls.

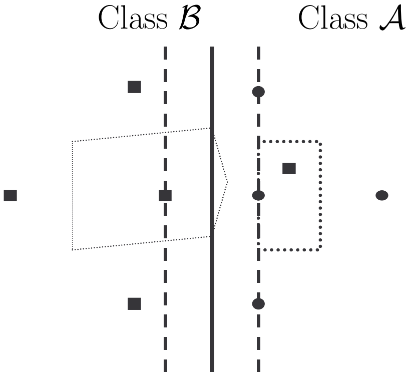

Figure 7. Optimal plane bisecting parallel maximum soft margin planes.

$x'\hat{w} = \frac{\hat{w}'(A'\hat{u}-B'\hat{v})}{2}$, is parallel to but not identical to the plane $x'\bar{w} = \frac{\bar{\alpha}+\bar{\beta}}{2}$ that would also be a solution of the original SVM problem once scaled. The thresholds differ. This is illustrated by Figures 6 and 7.

Figure 6 gives the solution found by the reduced-convex-hull SVM formulation which finds the two closest points in the reduced convex hull and as a heuristic selects the threshold halfway between the points. But there is nothing explicit about the choice of threshold in the reduced-convex-hull formulation RC-Hull. In Figure 6, the closest points in the reduced convex hull are represented by an open square and open circle. The solution found by the classic SVM is given in Figure 7. The classic SVM formulation assumes that the best plane bisects the two parallel margin planes. Note that the plane that bisects the closest points is nearer to Class $\mathcal{A}$. In some sense the plane is shifted toward the class in which we have more confidence. It is not a priori evident which assumption for the choice of threshold is best. This property was also noted with the ν-SVM formulation (Crisp & Burges, 1999).

Our reduced-convex-hull SVM formulation differs from the ν-SVM formulation in that there are distinct margin thresholds α and β for each class instead of a single variable for both. Extensions of the ν-SVM formulation using parametric models for the margins are suggested in Schölkopf et al. (2000). Similar analysis to the above can be performed for the ν-SVM. We refer readers to Crisp and Burges (1999) which uses a related but different argument for establishing the correspondence of ν-SVM with the reduced-convex-hull formulation. Assuming there exists a unique nonzero solution to the closest points in the reduced convex hull problem and appropriate parameter choices are made, the reduced-convex-hull, classic, and ν-SVM will all yield a plane with the same orientation, i.e. w is the same modulo a positive scaling factor. But they do not produce the exact same final planes because the assumptions used to construct the thresholds differ.

5. Alternative Norm Variations

We have shown that the classical 2-norm SVM formulation is equivalent to finding the closest points in the reduced convex hulls. This same explanation works for versions of SVM based on alternative norms. For example consider the case of finding the closest points in the reduced convex hulls as measured by the infinity-norm:

$$\begin{array}{ll} \min\limits_{u,v} & \|A'u - B'v\|_\infty \\ s.t. & e'u = e'v = 1,\ De \geq u \geq 0,\ De \geq v \geq 0 \end{array} \tag{15}$$

One method for converting the problem into a linear program (LP) produces:

$$\begin{array}{ll} \min\limits_{u,v,\rho} & \rho \\ s.t. & -\rho e \leq A'u - B'v \leq \rho e \\ & e'u = e'v = 1,\ De \geq u \geq 0,\ De \geq v \geq 0 \end{array} \tag{16}$$

The dual is

$$\begin{array}{ll} \max\limits_{w,\alpha,\beta,\xi,\eta} & \alpha - \beta - De'\xi - De'\eta \\ s.t. & Aw - \alpha e + \xi \geq 0, \quad \xi \geq 0 \\ & -Bw + \beta e + \eta \geq 0, \quad \eta \geq 0 \\ & \|w\|_1 = 1 \end{array} \tag{17}$$

For an appropriate choice of C, this is equivalent to solving the typical 1-norm SVM

$$\begin{array}{ll} \min\limits_{w,\gamma,\xi,\eta} & Ce'\xi + Ce'\eta + \|w\|_1 \\ s.t. & Aw - (\gamma+1)e + \xi \geq 0, \quad \xi \geq 0 \\ & -Bw + (\gamma-1)e + \eta \geq 0, \quad \eta \geq 0 \end{array} \tag{18}$$

Similarly finding the closest points of the reduced convex hulls using the 1-norm is equivalent to constructing a SVM regularized using an infinity-norm on w. Specifically, solving the problem

$$\begin{array}{ll}\min\limits_{u,v} & \|A'u - B'v\|_1 \\ s.t. & e'u = e'v = 1,\ De \geq u \geq 0,\ De \geq v \geq 0\end{array} \tag{19}$$

is equivalent to solving (for appropriate choices of D and C)

$$\begin{array}{ll}\min\limits_{w,\gamma,\xi,\eta} & Ce'\xi + Ce'\eta + \|w\|_\infty \\ s.t. & Aw - (\gamma+1)e + \xi \geq 0,\quad \xi \geq 0 \\ & -Bw + (\gamma-1)e + \eta \geq 0,\quad \eta \geq 0\end{array} \tag{20}$$

Limited space does not allow a full development of this argument.

6. Conclusion

The simple geometric argument of finding the closest points in the convex hulls or reduced convex hulls of two classes can be used to derive an intuitive geometric SVM formulation. Users can grasp visually the primary notions of SVM necessary for successful implementation without getting hung up on notions of duality. The reduced-convex-hull formulation forces the optimal solution to depend on more points depending on the parameter $D \in (0,1)$. If D is too large, the reduced convex hulls intersect, and the meaningless solution $w = 0$ results. If D is too small, the dual problem will be infeasible. We rigorously showed this formulation is exactly equivalent to the classic SVM formulation for appropriate choices of parameters. Assuming the parameters are well-defined, the solution sets of the problems are the same modulo a scaling factor dependent on the size of the margin. But the final choice of threshold will vary depending on the assumptions of the user. From an optimization perspective the reduced-convex-hull formulations may be preferable due to the interpretability of the misclassification parameter and the availability of fast nearest point in polytope algorithms (Keerthi et al., 1999). If the 1-norm or infinity-norm is used to measure the closest points in the reduced convex hull the analogous analysis can be performed showing that the primal problem corresponds to the SVM regularized with the infinity-norm or 1-norm of w respectively. Thus the reduced convex hull argument holds for 1-norm SVM linear programming approaches.

Acknowledgements

This material is based on research supported by Microsoft Research and NSF Grants 949427 and IIS-9979860.

References

Bennett, K. P., & Bredensteiner, E. J. (in press). Geometry in learning. In C. Gorini et al. (Eds.), *Geometry at work*. MAA Press. Also available as http://www.rpi.edu/∼bennek/geometry2.ps.

Burges, C. J. C., & Crisp, D. J. (1999). Uniqueness of the svm solution. *Proceedings of Neural Information Processing 12*. Cambridge, MA: MIT Press.

Crisp, D. J., & Burges, C. J. C. (1999). A geometric interpretation of ν-svm classifiers. *Proceedings of Neural Information Processing 12*. Cambridge, MA: MIT Press.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (1999). *A fast iterative nearest point algorithm for support vector machine classifier design* (Technical Report TR-ISL-99-03). Intelligent Systems Labs, Department of Computer Science and Automation, Indian Institute of Science, Bangalor, India.

Mangasarian, O. L. (1965). Linear and nonlinear separation of patterns by linear programming. *Operations Research*, *13*, 444–452.

Mangasarian, O. L. (1969). *Nonlinear programming.* New York: McGraw–Hill.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods - Support vector learning.* Cambridge, MA: MIT Press.

Schölkopf, B., Smola, A., Williamson, R., & Bartlett, P. (2000). New support vector algorithms. *Neural Computation*, *12*, 1083–1121.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., & Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, *44*, 1926–1940.

Vapnik, V. N. (1996). *The nature of statistical learning theory.* New York: Wiley.