Global Geometry of SVM Classifiers*

Dengyong Zhou
AI Lab, Institute of Automation,
Chinese Academy of Scinces
Beijing 100080, China
dengyong.zhou@mail.ia.ac.cn

Baihua Xiao
AI Lab, Institute of Automation,
Chinese Academy of Scinces
Beijing 100080, China
baihua.xiao@mail.ia.ac.cn

Huibin Zhou
Department of Mathematics,
Cornell University
Ithaca, NY 14853-4201, USA
hbzhou@math.cornell.edu

Ruwei Dai
AI Lab, Institute of Automation,
Chinese Academy of Scinces
Beijing 100080, China
ruwei.dai@mail.ia.ac.cn

Abstract

We construct an alternative geometry framework for Support Vector Machine (SVM) classifiers. Within this framework, separating hyperplanes, dual descriptions and solutions of SVM classifiers all are constructed clearly by a pure geometry fashion. Now all kinds of SVM formulations and their dual descriptions including the arbitrary-norm cases are only different expressions of the underlying common geometry essentials. Compared with the optimization theory in SVM classifiers, we don't need redundant confused computations any more. Instead, every step in our theory is guided by elegant geometry intuitions. Our framework can make people understand SVM in a totally visual fashion. In addition, it is also helpful to expose the correlations between SVM and other learning algorithms.

1 Introduction

The basic ideas of SVM are very intuitive [9, 10]. If the data is linearly separable, the strategy of SVM is to separate the data with the maximal margin hyperplane. While, if the data isn't linearly separable, slack variables are introduced to allow the margin constraints to be violated and the data is separated with the so-called soft maximal margin. These strategies are implemented by reducing them into convex optimization problems, that is minimizing convex functions under linear inequality constraints. Within the framework of constrained optimization theory, these optimization problems are then converted to their alternative dual forms, which are easier to be solved than the primal problems. So the dual transform is an important concept for understanding the mechanism of SVM. But in fact it is

*Technical Report in AI Lab, Institute of Automation, Chinese Academy of Sciences. Submitted to NIPS 2002. June 1, 2002.

almost back magic for many people to change from the primal to dual presentations [1].

Recent researches have shown that there exist nice geometry interpretations for the dual formulations in SVM that can make people grasp visually this key idea [1, 2, 4]. In geometry, for the separable case finding the maximum margin between two classes is equivalent to finding the nearest neighbors in the convex hulls of each class; for the inseparable case finding the soft maximum margin between two classes is equivalent to finding the closet points in the so-called reduced convex hulls of each class. But these geometry intuitions are proved within the framework of optimization theory. So the proofs are still back magic for many people. It seems that the geometry explanations emerge suddenly from a series of computations.

In this paper, we construct an alternative geometry framework for SVM classifiers motivated by the related researches [1, 2, 4, 7]. Within this framework, separating hyperplanes, dual descriptions and solutions of SVM classifiers all are constructed clearly by a pure geometry fashion. Now all kinds of SVM formulations including the arbitrary-norm cases and their dual descriptions are only different expressions of the underlying common geometry essentials. Compared with the optimization theory in SVM classifiers, we don't need redundant confused computations any more. Instead, every step in our theory is guided by elegant geometry intuitions. So our framework can make people understand SVM in a totally visual fashion. In addition, it is also helpful to expose the correlations between SVM and other learning algorithms, such as boosting [8].

This paper is organized as follows. Section 2 introduces some preliminary mathematical notions. Section 3 discusses arbitrary-norm separating hyperplanes. Section 4 discusses the dual problems in arbitrary-norm SVM classifiers. Finally, section 5 investigates solutions of 2-norm SVM classifiers. In this case, the solution has the best geometry intuition.

2 Preliminary

In this section, we introduce some mathematical notions. For more details, you can refer to the book [5, 6].

A normed linear space is a vector space $\mathcal{X}$ on which there is defined a real-valued function which maps each element x in $\mathcal{X}$ into a real number $\|x\|$. The norm satisfies the following axioms:
1. $\|x\| \geq 0$ for all $x \in \mathcal{X}$,$\|x\| = 0$ if only if $x = \theta$.
2. $\|x + y\| \leq \|x\| + \|y\|$ for each x, $y \in \mathcal{X}$.
3. $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all scalars α and each $x \in \mathcal{X}$.

If A, B are subsets of a normed vector space, their distance is $d(A, B) = \inf \|x - y\|$, $x \in A$, $x \in B$.

A Linear functional f on a normed space is bounded if there is a constant λ such that $|f(x)| \leq \lambda\|x\|$ for all $x \in X$. The space of all bounded linear functionals on $\mathcal{X}$ is called the normed dual of $\mathcal{X}$ and is denoted $\mathcal{X}^*$. The norm of $f \in \mathcal{X}^*$ is $\|f\| = \inf\{\lambda : |f(x)| \leq \lambda\|x\|$, for all $x \in \mathcal{X}\}$. Generally, we let x^* denote an element in $\mathcal{X}^*$, and employ the notion $\langle x, x^*\rangle$ for the value of the functional x^* at a point $x \in \mathcal{X}$. By the norm definition we have $\langle x, x^*\rangle \leq \|x\|\|x^*\|$.

The translation of a subspace is said to be a linear manifolds. A hyperplane H in a linear vector space $\mathcal{X}$ is a maximal proper manifolds, that is, a linear manifolds H such that $H \neq \mathcal{X}$, and if V is any linear manifolds containing H, then either $V = \mathcal{X}$ or $V = H$. Given a hyperplane H, there is a linear functional x^* on $\mathcal{X}$ and a constant c such that $H = \{x : \langle x, x^* \rangle = c\}$. Conversely, if x^* is a nonzero linear functional on $\mathcal{X}$, the set $\{x : \langle x, x^* \rangle = c\}$ is a hyperplane in $\mathcal{X}$.

A real inner space is a real linear vector space $\mathcal{X}$ together with an inner product, which is a map from $\mathcal{X} \times \mathcal{X}$ to R and denoted by $\langle x, y \rangle$ where $x, y \in \mathcal{X}$. The inner product satisfies the following axioms:
1. $\langle x, y \rangle = \langle y, x \rangle$.
2. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.
3. $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$.
4. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = \theta$.

We say x is perpendicular to y when $\langle x, y \rangle = 0$. $\sqrt{\langle x, y \rangle}$ is denoted by $\|x\|$ because it is indeed a norm (2-norm). For all x, y in an inner product space, $|\langle x, y \rangle| \leq \|x\| \|y\|$. This is called Cauchy-Schwarz inequality. Equality holds if and only if $x = \lambda y$ or $y = \theta$. A complete real inner space is called a real Hilbert space. If x^* is an element in the dual space $\mathcal{H}^*$ of a Hilbert space $\mathcal{H}$, there exists a unique vector $x \in \mathcal{H}$ such that for all $y \in \mathcal{H}$, $\langle y, x^* \rangle = \langle y, x \rangle$. Furthermore, we have $\|x^*\| = \|x\|$.

3 Hyperplane

In this section, we discuss arbitrary-norm separating hyperplanes.

Theorem 3.1 Let x_0 be a point in a real normed vector space $\mathcal{X}$, and H a hyperplane $\{x : \langle x, x^* \rangle = c\}$, $x^* \in \mathcal{X}^*$. Then the distance from x_0 to H is

$$d = \inf_{x \in H} \|x_0 - x\| = \frac{|\langle x_0, x^* \rangle - c|}{\|x^*\|}.$$

Especially, $d = |c|$ when $\|x^*\| = 1$ and $x = \theta$.
If the infimum on the left is achieved by some point $x_1 \in H$ and if $\langle x_0, x^* \rangle > c$, then x^* is aligned with $x_0 - x_1$, i.e. $\langle x_0 - x_1, x^* \rangle = \|x_0 - x_1\| \|x^*\|$.

Proof. For simplicity, let $x = \theta$ science the general case can then be deduced by translations. Thus we must show that

$$d = \inf_{y \in H} \|y\| = \frac{|c|}{\|x^*\|}.$$

Let $H = z + M$ where M is maximal proper subspace of $\mathcal{X}$. Elements of $\mathcal{X}$ are uniquely representable in the form $x = \alpha z + m$, with $m \in M$. Let $\langle z, x^* \rangle = c$, so that $H = \{x : \langle x, x^* \rangle = c\}$. We have

$$\begin{aligned} \|x^*\| &= \sup_{x \in X} \frac{|\langle x, x^* \rangle|}{\|x\|} = \sup \frac{|\alpha c|}{\|\alpha z + m\|} \\ &= \sup \frac{|\alpha c|}{|\alpha| \|z + \frac{m}{\alpha}\|} = \frac{|c|}{\inf \|z + \frac{m}{\alpha}\|}. \end{aligned}$$

Note $d = \inf \|z + \frac{m}{\alpha}\|$, so $d = |c| / \|x^*\|$.
If the infimum on the left is achieved by some point $x_1 \in H$, and if $\langle x_0, x^* \rangle > c$, then, by the above result, we have

$$\begin{aligned} d &= \|x_0 - x_1\| = \frac{\langle x_0, x^* \rangle - c}{\|x^*\|} \\ &= \frac{\langle x_0, x^* \rangle - \langle x_1, x^* \rangle}{\|x^*\|} = \frac{\langle x_0 - x_1, x^* \rangle}{\|x^*\|}. \end{aligned}$$

So $\langle x_0 - x_1, x^* \rangle = \|x_0 - x_1\| \|x^*\|$.
This proves our theorem.

Remark 1 In Hilbert space $\mathcal{H}$, for any $x^* \in \mathcal{H}^*$, there exists a unique vector $w \in \mathcal{H}$ such that $\|w\| = \|x^*\|$ and $\langle x, w\rangle = \langle x, x^*\rangle$ for all $x \in \mathcal{H}$. In the above theorem, by substituting w for x^*, we have $\langle x_0 - x_1, w\rangle = \|x_0 - x_1\|\|w\|$. Equality holds if and only if $x_0 - x_1 = \mu w$, $\mu > 0$. So for any $m \in M$, we have $\langle x_0 - x_1, m\rangle = \langle \mu w, m\rangle = \mu\langle w, m\rangle = 0$. This means $x_0 - x_1$ is perpendicular to M. This situation is sketched in Figure 3.1.

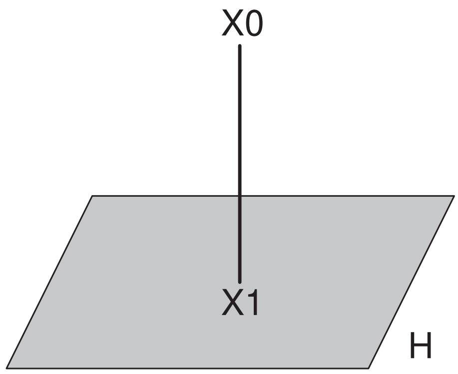

Figure 3.1 Minimum distance: from a point to a hyperplane

Remark 2 In an arbitrary normed space, there does not have a unique solution possibly. We consider a simple two-dimensional example[]. Let $\mathcal{X}$ be the space of pairs of real numbers $x = (\xi_1, \xi_2)$ with $\|x\| = \max_{i=1,\ 2} |\xi_i|$. Let H be the subspace $H = \{x : \xi_2 = 0\}$, and consider the fixed point $x = (2, 1)$. The distance from x to H is obviously 1, but any vector m in the subset $\{x : 1 \leq \xi_1 \leq 3,\ \xi_2 = 0\}$ of H satisfies $\|x - m\| = 1$.

Remark 3 The set of all the nearest neighbors in H of x_0 is convex. Let x_1 and x_2 be two nearest neighbors, and $x = (1-\lambda)x_1 + \lambda x_2$, $0 \leq \lambda \leq 1$. Obviously $x \in H$, so $\|x_0 - x\| \geq d$. In addition, by the triangle inequality we have

$$\begin{aligned}
\|x_0 - x\| &= \|x_0 - [(1-\lambda)x_1 + \lambda x_2]\| \\
&= \|(1-\lambda)(x_0 - x_1) + \lambda(x_0 - x_2)\| \\
&\leq (1-\lambda)\|x_0 - x_1\| + \lambda\|x_0 - x_2\| \\
&= (1-\lambda)d + \lambda d = d.
\end{aligned}$$

So $\|x_0 - x\| = d$, i.e., x is one of the nearest neighbor of x_0. This proves our statement.

Example Let p be a real number of $1 < p < \infty$.The space l_p consists of all sequences of scalars $\{\xi_1, \xi_2, ...\}$ for which

$$\sum_{i=1}^{\infty} |\xi_i|^p < \infty.$$

The norm of an element $x = \{\xi_i\}$ in l_p is defined as

$$\|x\|_p = (\sum_{i=1}^{\infty} |\xi_i|^p)^{\frac{1}{p}}.$$

Let q be the positive number (necessarily > 1) such that

$$\frac{1}{p} + \frac{1}{q} = 1,$$

and call q the dual exponent of p. The dual space of l_p is l_q[5, 6].
The space l_∞ consists of bounded sequences. The norm of an element $x = \{\xi_i\}$ is defined as

$$\|x\| = \sup_i |\xi_i|.$$

The dual space of l_∞ is l_1[5, 6].
Let $x_0 \in l_p (1 \leq p < \infty, \text{if } p = 1, \text{we take } q = \infty)$ be any point which is not on the hyperplane $H = \{x : \langle x, w \rangle = 0\}$ where $w \in l_q$. Then the distance from x_0 to the hyperplane H is

$$d = \inf_{x \in H} \|x_0 - x\|_p = \frac{|\langle x_0, w \rangle|}{\|w\|_q}.$$

You also can get this result by a series of computations based on Karush-Kuhn-Tucker optimality criterion[7, 8].

Theorem 3.2 Let $\mathcal{X}$ be a real normed vector space and $\mathcal{X}^*$ is its dual space. Given the two parallel hyperplanes $H_1 = \{x : \langle x, x^* \rangle = c_1\}$ and $H_2 = \{x : \langle x, x^* \rangle = c_2\}$ where $x^* \in \mathcal{X}^*$, $c_1, c_2 \in I\!R$ and $c_1 \neq c_2$, then the distance between H_1 and H_2 is

$$d = \inf_{\substack{x \in H_1 \\ y \in H_2}} \|x - y\| = \frac{|c_1 - c_2|}{\|x^*\|}.$$

If the infimum on the left is achieved by $x_0 \in H_1, y_0 \in H_2$ and if $c_1 > c_2$, then x^* is aligned with $x_0 - y_0$, i.e. $\langle x_0 - y_0, x^* \rangle = \|x_0 - y_0\| \|x^*\|$.

Proof. Let $H = \{z : z = x_1 - x_2, x_1 \in H_1, x_2 \in H_2\}$. Then $d = \inf\{\|z\| : z \in H\}$. Note $H = \{z : |\langle z, x^* \rangle = c_1 - c_2\}$, i.e. H is still a hyperplane. By theorem 3.1 we have $d = |c_1 - c_2| / \|x^*\|$.
If the infimum on the left is achieved by $x_0 \in H_1$ and $y_0 \in H_2$, i.e., $\langle x_0, x^* \rangle = c_1$ and $\langle x_0, x^* \rangle = c_2$, and if $c_1 > c_2$, then

$$\begin{aligned} d &= \|x_0 - y_0\| = \frac{c_1 - c_2}{\|x^*\|} \\ &= \frac{\langle x_0, x^* \rangle - \langle y_0, x^* \rangle}{\|x^*\|} = \frac{\langle x_0 - y_0, x^* \rangle}{\|x^*\|}. \end{aligned}$$

Note $d = \|x_0 - y_0\|$, we have $\langle x_0 - y_0, x^* \rangle = \|x_0 - y_0\| \|x^*\|$.
This proves our theorem.

Example 1 Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in I\!R^n \times \{-1, 1\}$ be our training sample. Let $A = \{x_i : y_i = 1, 1 \leq i < m, m \leq n\}$ and $B = \{x_i : y_i = -1, m \leq i \leq n\}$. There exists a $w \in I\!R^n$ (the dual space of $I\!R^n$ is still $I\!R^n$)and a number b such that $\langle x, w \rangle \geq b + 1$ for all $x \in A$ and $\langle x, w \rangle \leq b - 1$ for all $x \in B$. Then $H_1 = \{x : \langle x, w \rangle = b + 1\}$and $H_1 = \{x : \langle x, w \rangle = -b - 1\}$ are the two support hyperplanes. By this theorem the distance between H_1 and H_2 is $|(-b + 1) - (-b - 1)| / \|w\|_2 = 2 / \|w\|_2$. So maximizing the distance of the the two support hyperplanes is to solve the following optimization problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|_2^2 \\ \text{subject to} \quad & y_i(\langle x_i, w \rangle + b) \geq 1 \quad i = 1, 2, \ldots, n. \end{aligned}$$

This is the standard primal SVM 2-norm formulation[9, 10].

Example 2 Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$ be our training sample where $\mathcal{X}$ is a l_p space. Let $A = \{x_i : y_i = 1, 1 \leq i < m, m \leq n\}$ and $B = \{x_i : y_i = -1, m \leq i \leq n\}$. There exists a $w \in X^*$, $\|w\|_q = 1$ (note X^* is a l_q space) and a number $\rho > 0$ such that $\langle x, w\rangle \geq \rho$ for all $x \in A$ and $\langle x, w\rangle \leq -\rho$ for all $x \in B$. Let $H_1 = \{x : \langle x, w\rangle = \rho\}$ and $H_1 = \{x : \langle x, w\rangle = -\rho\}$. By this theorem the distance between H_1 and H_2 is $|\rho - (-\rho)|/\|w\|_q = 2\rho$. So maximizing the distance of these two separating hyperplanes is to solve the following optimization problem:

$$\begin{array}{ll} \max\limits_{\rho \in I\!\!R_+} & \rho \\ \text{subject to} & y_i(\langle x_i, w\rangle) \geq \rho \quad i = 1, 2, \ldots, n \\ & \|w\|_q = 1. \end{array}$$

You also can get this result by a series of computations based on Karush-Kuhn-Tucker optimality criterion[7, 8].

4 Duality

In this section we discuss the dual problems in arbitrary-norm SVM classifiers.

Let K be a convex set in a real normed vector space $\mathcal{X}$. The functional $h_K(x^*) = \sup_{x \in K} \langle x, x^*\rangle$ defined on $\mathcal{X}^*$ is called the upper support functional of K and the hyperplane $H = \{x : \langle x, x^*\rangle = h_K(x^*)\}$ a upper support hyperplane for K. Correspondingly, the functional $l_K(x^*) = \inf_{x \in K} \langle x, x^*\rangle$ is called the lower support functional and $H = \{x : \langle x, x^*\rangle = l_K(x^*)\}$ a lower support hyperplane.

Theorem 4.1 (Minimum Norm Duality) Let x_0 be a point in a real normed vector space $\mathcal{X}$ and K is a convex set and let d denote its distance from the convex set K; then

$$d = \inf_{x \in K} \|x_0 - x\| = \max_{x^* \in \mathcal{X}^*} \frac{[\langle x_0, x^*\rangle - h_K(x^*)]}{\|x^*\|}$$

where the maximum on the right is achieved by some $x_1^* \in \mathcal{X}^*$.
If the infimum on the left is achieved by some point $x_1 \in K$, then x_1^* is aligned with $x_0 - x_1$, i.e. $\langle x_0 - x_1, x_1^*\rangle = \|x_0 - x_1\|\|x_1^*\|$.

Proof. For simplicity we take $x_0 = \theta$ since the general case can then be deduced by translation. Thus we must show that

$$d = \inf_{x \in K} \|x\| = \max_{x^* \in \mathcal{X}^*} \frac{-h_K(x^*)}{\|x^*\|}.$$

We first show that for any $x^* \in \mathcal{X}^*$, we have $d \geq -h_K(x^*)/\|x^*\|$. For this we may only consider $\{x^* : h(x^*) < 0\}$. Obviously, if $h_K(x^*) < 0$, then the hyperplane $H = \{x : \langle x, x^*\rangle = h_K(x^*)\}$ separates K and θ.
Let $S(r) = \{x : \|x\| \leq r\}$. For any $x^* \in \mathcal{X}^*$ with $h_K(x^*) < 0$, let r_0 be the supremum of the r for which the hyperplane $\{x : \langle x, x^*\rangle = h_K(x^*)\}$ separates K and $S(r)$. Obviously, we have $h_K(x^*) = \inf_{x \in X^*} \langle x, x^*\rangle = -r_0\|x^*\| \geq -d\|x^*\|$ (note that $r_0 \leq d$).
On the other hand, since K contains no interior points of $S(d)$, there is a hyperplane separating $S(d)$ and K. Therefore, their is a $x_1^* \in \mathcal{X}^*$, such that $-h_K(x_1^*) = d\|x_1^*\|$. Thus the first part of this theorem is proved.
To prove the second part on alignment, suppose that $x_1 \in K$, $\|x_1\| = d$. Then $\langle x_1, x_1^*\rangle \leq h_K(x_1^*) = -d\|x_1^*\|$. However, $-\langle x_1, x_1^*\rangle \leq \|x_1\|\|x_1^*\| = d\|x_1^*\|$. Thus $-\langle x_1, x_1^*\rangle = \|x_1\|\|x_1^*\|$ and x_1^* is aligned with $-x_1$.

Remark 1 This theorem is adopted from the book[6] with minor variations. For the convenience of understanding the following materials, the proof of this theorem is also contained.

Remark 2 Some complements on the separation theorem used in the above proof. Let A and B be the subsets in a real normed vector space $\mathcal{X}$. We say they are linearly separable if there is a $x^* \in \mathcal{X}^*$ such that

$$\sup_{x \in A} \langle x, x^* \rangle \leq \inf_{x \in B} \langle x, x^* \rangle.$$

In other words, there exists a $x^* \in \mathcal{X}^*$ and a number $\gamma \in I\!R$ such that $\langle x, x^* \rangle \geq \gamma$ for all $x \in A$ and $\langle x, x^* \rangle \leq \gamma$ for all $x \in B$. When these inequalities hold strictly, we say they are strictly linearly separable. If A and B are convex, and A has interior points and B contains no interior point of A, then A and B are linearly separable[5, 6].

Remark 3 This theorem is very intuitive in geometry. In fact, it means that the minimum distance from a point to a convex set K is equal to the maximum of the distance from the point to support hyperplanes separating from the point and the convex set K. This situation is sketched in Figure 4.1.

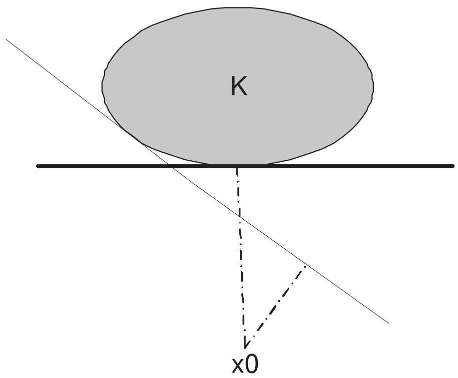

Figure 4.1 Maximal margin duality: from a point to a convex

Theorem 4.2 Let K_1 and K_2 are the complete, disjoint and convex sets in a real normed vector space $\mathcal{X}$ and let d denote the distance between K_1 and K_2; then

$$d = \inf_{\substack{y \in K_1 \\ z \in K_2}} \|y - z\| = \max_{x^* \in X^*} \frac{l_{K_1}(x^*) - h_{K_2}(x^*)}{\|x^*\|}$$

where the maximum on the right is achieved by some $x_0^* \in \mathcal{X}^*$.
If the infimum on the left is achieved by $y_0 \in K_1$ and $z_0 \in K_2$, then x_0^* is aligned with $y_0 - z_0$, i.e. $\langle y_0 - z_0, x_0^* \rangle = \|y_0 - z_0\| \|x_0^*\|$.

Proof. Let $K = \{x : x = z - y,\ y \in K_1,\ z \in K_2\}$. Obviously, K is still a convex set. Furthermore, $K_1 \cap K_2 = \phi$ since K_1 and K_2 are disjoint, so θ is not the element of K. By theorem 4.1, we have

$$d = \inf_{x \in K} \|x\| = \max_{x^* \in X^*} \frac{-h_K(x^*)}{\|x^*\|}.$$

Note that

$$\inf_{x \in K} \|x\| = \inf_{\substack{y \in K_1 \\ z \in K_2}} \|y - z\|,$$

and

$$-h_K(x^*) = l_{K_1}(x^*) - h_{K_2}(x^*).$$

Thus we have proved the first part of this theorem.
To prove the second part, suppose that $y_0 \in K_1$ and $z_0 \in K_2$, $\|y_0 - z_0\| = d$. Let $x_0 = z_0 - y_0$. By theorem 4.1, we have $-\langle z_0 - y_0, x_0^* \rangle = -\langle x_0, x_0^* \rangle = \|x_0\|\|x_0^*\| = \|z_0 - y_0\|\|x_0^*\|$. So x_0^* is aligned with $y_0 - z_0$.

Remark This theorem is very intuitive in geometry. In fact, it means that the minimum distance between the two convex sets K_1 and K_2 is equal to the maximum of the distance between a pair of parallel support hyperplanes separating K_1 and K_2 one from the other. This situation is sketched in Figure 4.2.

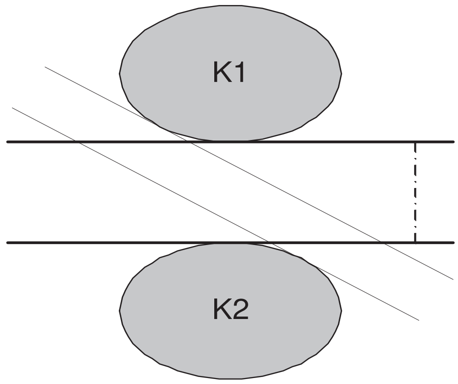

Figure 4.2 Maximal margin duality: from a convex to another one

Example 1 Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in I\!R^n \times \{-1, 1\}$ be our training sample. Let $A = \{x_i : y_i = 1\}$ and $B = \{x_i : y_i = -1\}$. Let $co(A)$ and $co(B)$ are the convex hulls of A and B respectively. Suppose that $w \in I\!R^n$, $\langle x, w \rangle \geq \alpha$ for all $x \in A$ and $\langle x, w \rangle \leq \beta$ for all $x \in B$. Then $l_{co(A)}(w) = \alpha$, and $h_{co(B)}(w) = \beta$. The problem of maximizing $[l_{co(A)}(w) - h_{co(B)}(w)]/\|w\|_2$ can be written as the the following optimization problem:

$$\begin{array}{ll} \displaystyle\min_{w,\alpha,\beta} & \displaystyle\frac{1}{2}\|w\|_2^2 - (\alpha - \beta) \\ \text{subject to} & \langle x_i, w \rangle \geq \alpha \text{ for any } x_i \in A \\ & \langle x_i, w \rangle \leq \beta \text{ for any } x_i \in B. \end{array}$$

By theorem 4.2 its dual problem is

$$\begin{array}{ll} \displaystyle\min_{u,v} & \displaystyle\frac{1}{2}\|\Sigma_{x_i \in A} u_i x_i - \Sigma_{x_i \in B} v_i x_i\| \\ \text{subject to} & \Sigma u_i = 1, \ \Sigma v_i = 1 \\ & u_i \geq 0, \ v_i \geq 0. \end{array}$$

This example shows that for the separable case finding the maximum margin between two classes is equivalent to finding the nearest neighbors in the convex hulls of each class.

Example 2 Now suppose that A and B in the above example are inseparable. Let $0 < \mu < 1$. Let $R(A, \ \mu) = \{x : x = \Sigma u_i x_i, \ \Sigma u_i = 1, \ 0 \leq u_i \leq \mu, \ x_i \in A\}$ and $R(B, \ \mu) = \{x : x = \Sigma v_i x_i, \ \Sigma v_i = 1, \ 0 \leq v_i \leq \mu, \ x_i \in B\}$. $R(A, \ \mu)$ and $R(B, \ \mu)$ are called the reduced convex hulls of A and B respectively. Suppose that $R(A, \ \mu)$ and $R(B, \ \mu)$ are separable now. So there exists $w \in I\!R^n$ such that $\langle x, w \rangle \geq \alpha$ for all $x \in R(A, \ \mu)$ and $\langle x, w \rangle \leq \beta$ for all $x \in R(B, \ \mu)$. Then we have

$$\langle x_i, w \rangle \geq \alpha - \xi_i, \text{ for any } x_i \in A$$

where $\xi_i = 0$ if $x \in R(A,\ \mu)$ or else $\xi_i = \alpha - \langle x_i, w\rangle$. Analogously,

$$\langle x_i, w\rangle \leq \beta + \eta_i,\ \text{ for any } x_i \in B$$

where $\xi_i = 0$ if $x \in R(B,\ \mu)$ or else $\eta_i = \langle x_i, w\rangle - \beta$. Obviously, $\xi_i \geq 0,\ \eta_i \geq 0$. So we have

$$\begin{array}{l} \langle x_i, w\rangle \geq \alpha - \mu\Sigma_{x_i\in A}\xi_i,,\ \text{ for any } x_i \in R(A,\ \mu) \\ \langle x_i, w\rangle \leq \beta + \mu\Sigma_{x_i\in B}\eta_i,\ \text{ for any } x_i \in R(B,\ \mu). \end{array}$$

This means

$$\begin{array}{r} l_{R(A,\mu)}(w) = \alpha - \mu\Sigma_{x_i\in A}\xi_i \\ h_{R(B,\mu)}(w) = \beta + \mu\Sigma_{x_i\in B}\eta_i. \end{array}$$

So maximizing

$$[l_{R(A,\mu)}(w) - h_{R(B,\mu)}(w)]/\|w\|_2$$

is equivalent to maximizing

$$[(\alpha - \mu\Sigma_{x_i\in A}\xi_i) - (\beta + \mu\Sigma_{x_i\in B}\eta_i)]/\|w\|_2.$$

The latter can be be written as the the following optimization problem:

$$\begin{array}{cl} \min\limits_{w,\alpha,\beta} & \mu(\Sigma\xi_i + \Sigma\eta_i) + \frac{1}{2}\|w\|_2^2 - (\alpha - \beta) \\ \text{subject to} & \langle x_i, w\rangle \geq \alpha - \xi_i,\ \text{ for any } x_i \in A \\ & \langle x_i, w\rangle \leq \beta + \eta_i,\ \text{ for any } x_i \in B. \end{array}$$

By theorem 4.2 the former can be be written as the the following optimization problem:

$$\begin{array}{cl} \min\limits_{u,v} & \frac{1}{2}\|\Sigma_{x_i\in A}u_i x_i - \Sigma_{x_i\in B}v_i x_i\| \\ \text{subject to} & \Sigma u_i = 1,\ \Sigma v_i = 1 \\ & 0 \leq u_i \leq \mu,\ 0 \leq v_i \leq \mu. \end{array}$$

This example shows that for the inseparable case finding the maximum soft margin between two classes is equivalent to finding the closet points in the reduced convex hulls of each class.

Remark The arguments in the above two examples are firstly proposed in Bennett's paper[1]. But she got them by a series of computations based on Karush-Kuhn-Tucker optimality criterion.

5 Solution

In this section we discuss the solution of SVM classifiers in Hilbert spaces.

Theorem 5.1 Let K_1 and K_2 are the complete, disjoint and convex sets in a real Hilbert space $\mathcal{H}$. Let $d(K_1, K_2) = \|x_1 - x_2\|$ where $x_1 \in K_1,\ x_2 \in K_2$. Let $w = x_1 - x_2, \langle x_1, w\rangle = c_1, \langle x_2, w\rangle = c_2, c_1 > c_2$. Let $H_1 = \{x : \langle x, w\rangle = c_1\}$ and $H_2 = \{x : \langle x, w\rangle = c_1\}$; then: (1) the distance between H_1 and H_2 is just the distance between K_1 and K_2; (2) H_1 is a lower support hyperplane of K_1 and H_2 is a upper support hyperplane of K_2.

Proof. We firstly show the first part. By theorem 3.2, we have

$$\begin{aligned} d(H_1, H_2) &= \frac{c_1 - c_2}{\|w\|} = \frac{\langle x_1, w\rangle - \langle x_2, w\rangle}{\|w\|} \\ &= \frac{\langle x_1 - x_2, w\rangle}{\|w\|} = \frac{\langle w, w\rangle}{\|w\|} \\ &= \|w\| = d(K_1, K_2). \end{aligned}$$

To prove the second part, suppose that there were a point $x \in K_1$ for which $\langle x, w\rangle < c_1$. Let $u = \lambda x + (1-\lambda)x_1, \lambda \in I\!R$, then

$$\begin{aligned} & \|u - x_2\|^2 \\ = & \langle \lambda x + (1-\lambda)x_1 - x_2, \lambda x + (1-\lambda)x_1 - x_2\rangle \\ = & \langle \lambda(x - x_1) + (x_1 - x_2), \lambda(x - x_1) + (x_1 - x_2)\rangle \\ = & \|x_1 - x_2\|^2 + \lambda^2\|x - x_1\|^2 + 2\lambda\langle x - x_1, x_1 - x_2\rangle. \end{aligned}$$

Let

$$\lambda_0 = 2\frac{\langle x_1 - x, w\rangle}{\|x - x_1\|^2}.$$

Since $\langle x_1, w\rangle = c_1$ and $\langle x, w\rangle < c_1$, we have $\lambda_0 > 0$. Let $\lambda_1 = \min\{\lambda_0, 1\}$. Then, for any $\lambda \in (0, \lambda_1)$, we have $\|u - x_2\| < \|x_1 - x_2\|$. This contradicts the condition that $d(K_1, K_2) = \|x_1 - x_2\|$, since, by the convexity of K_1, we have $u \in K_1$. Hence $\langle x, w\rangle \geq c_1$ for all $x \in K_1$. Analogously, we have $\langle x, w\rangle \leq c_2$ for all $x \in K_2$.
This proves our theorem.

Remark In geometry, this theorem means that if $x_1 \in K_1$ and $x_1 \in K_2$ are a pair of points that are closest to each other, then the hyperplane bisecting, and orthogonal to, the line segment between x_1 and x_2 is a separating hyperplane for K_1 and K_2 with the maximal margin. This situation is sketched in Figure 5.1.

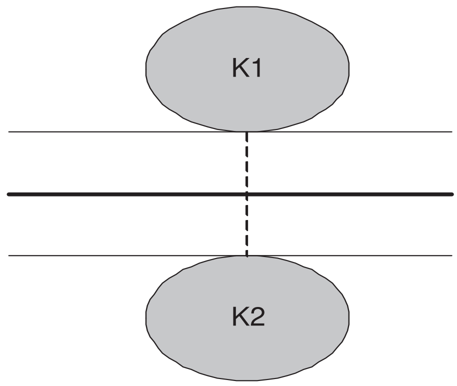

Figure 5.1 The maximal margin separating hyperplane

Theorem 5.2 Let K_1 and K_2 are the complete, disjoint and convex sets in a real Hilbert space $\mathcal{H}$. Then there is one and only one separating hyperplane for K_1 and K_2 with the maximal margin.

Proof. By theorem 5.1, there exists one separating hyperplane for the convex sets K_1 and K_2 with the maximal margin. Now we only need to show the uniqueness of such hyperplane. Suppose that $x_1 \in K_1,\ x_2 \in K_2,\ w \in \mathcal{H}$ satisfy

$$d(K_1, K_2) = \|x_1 - x_2\| = \frac{l_{K_1}(w) - h_{K_2}(w)}{\|w\|}.$$

Let H_1 and H_2 be the lower and upper support hyperplanes of the convex sets K_1 and K_2 respectively decided by w, i.e. $H_1 = \{x : \langle x, w\rangle \ = l_{K_1}(w)\}$ and $H_2 = \{x : \langle x, w\rangle \ = h_{K_2}(w)\}$. By theorem 4.2 we have $\langle x_1 - x_2, w\rangle = \|x_1 - x_2\|\|w\|$. Let $w_0 = x_1 - x_2$. By Cauchy-Schwarz inequality, there exists $\alpha \in I\!R$ such that $w = \alpha w_0$. Then $H_1 = \{x : \langle x, \alpha w_0\rangle \ = l_{K_1}(\alpha w_0)\} = \{x : \langle x, w_0\rangle \ = l_{K_1}(w_0)\}$. Analogously, $H_2 = \{x : \langle x, w_0\rangle \ = h_{K_2}(w_0)\}$. These mean H_1 and H_2 are decided

only by w_0.
This proves our theorem.

Remark Note that w_0 doesn't depend on the choice of the nearest neighbors in K_1 and K_2. In fact, suppose that $x_1,\ y_1 \in K_1,\ x_2,\ y_2 \in K_2$ satisfy

$$d(K_1, K_2) = \|x_1 - x_2\| = \|y_1 - y_2\|,$$

then we have $x_1 - x_2 = y_1 - y_2$. Let $C = \{z : z = x - y, x \in A, y \in B\}$. To prove our statement, we only need to prove that there exists only one element in c which has the least norm. Let d denote the least norm. Suppose that there are $z_1, z_2 \in C$ that satisfy$\|z_1\| = \|z_2\| = d$. Then

$$\begin{aligned} \|z_1 - z_2\|^2 &= \langle z_1 - z_2, z_1 - z_2 \rangle \\ &= 2\|z_1\|^2 + 2\|z_2\|^2 - \|z_1 + z_2\|^2 \\ &= 2d^2 + 2d^2 - 2^2\|\frac{z_1 + z_2}{2}\|. \end{aligned}$$

Since $(z_1 + z_2)/2 \in C$, we have

$$\|\frac{z_1 + z_2}{2}\| \geq d.$$

This means $\|z_1 - z_2\| \leq 0$. But we always have $\|z_1 - z_2\| \geq 0$. So $\|z_1 - z_2\| = 0$, i.e. $z_1 = z_2$. This proves our statement.

6 Conclusion

We construct an alternative geometry framework for SVM classifiers. Within this framework, separating hyperplanes, dual descriptions and solutions of SVM classifiers all are constructed clearly by a pure geometry fashion. Now all kinds of SVM formulations and their dual descriptions including the arbitrary-norm cases are only different expressions of the underlying common geometry essentials. Compared with the optimization theory in SVM classifiers, we don't need redundant confused computations any more. Instead, every step in our theory is guided by elegant geometry intuitions. So our framework can make people understand SVM in a totally visual fashion. In addition, it is also helpful to expose the correlations between SVM and other learning algorithms.

Acknowledgments

Dengyong Zhou wishes to thank Chris Ding of Lawrence Berkeley National Laboratory, Lei Guo and Yanxia Zhang of Chinese Academy of Sciences, and Sophy Zhu and Jing Han of National University of Science and Technology of China for helpful discussions and comments.

References

[1] K. P. Bennett & E. Bredensteiner. (2000). Duality, and Geometry in SVM Classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 57-64. Morgan Kaufmann.

[2] J. Bi & K. P. Bennett. (2002). Duality, Geometry, and Support Vector Regression. To appear in *Advances in Neural Information Processing Systems.*

[3] C. Burges & D. Crisp. (2000). Uniqueness of the SVM solution. In *Advances in Neural Information Processing Systems 12*, pages 223-229, Cambridge, MA, MIT Press.

[4] D. Crisp & C. Burges. (2000). A geometric interpretation of ν-SVM classifiers. In *Advances in Neural Information Processing Systems 12*, pages 244-251, Cambridge, MA, MIT Press.

[5] S. Lang. (1993). *Real and Functional Analysis* (Third Edition). Springer-Verlag, New York.

[6] D. G. Luenberger. (1976). *Optimization by Vector Space Methods.* John Wiley & Sons, New York.

[7] O. L. Mangasarian. (1999). Arbitrary-Norm Separating Plane. *Operations Research Letters* 24, pages 15-23.

[8] G. Rätsch, B. Schölkopf, S. Mika & KR. Müller. (2000). SVM and Boosting: One Class. Technical Report 119, GMD FIRST, Berlin.

[9] N. Cristianini & J. Shawe-Taylor. (2000). *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge, UK.

[10] V. Vapnik. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.