Is the Maximal Margin Hyperplane Special In a Feature Space?

Bin Zhang
Software Technology Laboratory
HP Laboratories Palo Alto
HPL-2001-89
April 10th , 2001*

E-mail: bzhang@hpl.hp.com

machine learning, supervised learning, Support Vector Machines (SVM)

Recent developments in Support Vector Machines (SVM) generalize the Linear Support Vector Machines (L-SVM) to learn non-linear separating surfaces by applying a feature mapping first. We construct an example in this paper to show that *any* Separating Hyperplane (SH), f, in any feature space can be mapped to a maximal-margin SH (mmSH), f_o, in another feature space of the same dimension such that f and f_o give exactly the same separating surface in the original input space. Then the question is: which feature space's maximal margin hyperplane gives the best generalization guarantee?

* Internal Accession Date Only
Approved for External Publication

Is the Maximal Margin Hyperplane Special in a Feature Space?

Bin Zhang BZHANG@HPL.HP.COM
Hewlett-Packard Research Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304

Abstract

Recent developments in Support Vector Machines (SVM) generalize the Linear Support Vector Machines (L-SVM) to learn non-linear separating surfaces by applying a feature mapping first. We construct an example in this paper to show that *any* Separating Hyperplane (SH), *f*, in any feature space can be mapped to a maximal-margin SH (mmSH), f_0, in another feature space of the same dimension such that *f* and f_0 give exactly the same separating surface in the original input space. Then the question is: which feature space's maximal margin hyperplane gives the best generalization guarantee?

Keywords Support Vector Machines (SVM), Kernel Method, Classification, Binary Prediction, Supervised Learning

1. Introduction

Given a *training dataset* $\{(x_1, y_1),...,(x_l, y_l)\} \subset R^n \times \{\pm 1\}$, L-SVM learns a function (a classifier) $f : R^n \to R$ so that $sign(f(x))$ will predict the class label well for an unseen data *x*. Even a function which does well on every training data, i.e. $sign(f(x_i)) = y_i$ for all *i*, need not generalizes well to unseen data. Statistical learning theory or VC dimension theory (Vapnik & Chervonenkis, 1974; Vapnik 1979) set the foundation for bounding the generalization errors. Based on this theory, Vapnik developed the *structural risk minimization principle* and applied it to the *linearly separable* two-class learning problem above. He discovered that the maximal margin hyperplane gives the best generalization guarantee for learning a linear classifier of two classes, which gave rise to the L-SVM learning algorithm. L-SVM was generalized to finding non-linear separating surfaces by first embedding the *original input space* $X = R^n$ into a higher dimensional space, the so-called *feature space* $\Im = R^N$, and then applying L-SVM in the feature space.

1.1 Linear Support Vector Machines (L-SVM)

Linear Support Vector Machines developed by Vapnik (Vapnik 1998 and references there) for finding linear classifiers come with a theoretical guarantee on its generalization performance. L-SVM also resulted in a quadratic programming problem, which can be solved by efficient algorithms (Joachims 1997, Platt 1998). Its popularity soared and many success-stories occurred in the literature (Cristianini & Shawe-Taylor 2000 and the references there). We give a brief introduction in this section; for more details, see Vapnik's book.

A linear function $f(x) = w \cdot x + b : R^n \to R$ defines a *Separating Hyperplane* (*SH*)[1] if $sign(f(x_i)) = y_i$ for all training examples.

[1] The hyperplane is defined by *f(x)=0*. Without any confusion, we use notation *f* also for the hyperplane defined by it. From this definition, it is clear that none of the training examples is on a SH.

Among all separating hyperplanes, there exists a unique one with the maximum margin,

$$\rho = \max_{w,b} \min\{\| x - x_i \| : x \in X, w \cdot x + b = 0, i = 1,...,l\}.$$

Based on the structural risk minimization theory, this maximal margin SH minimizes the following error bound function (Vapnik 1998, page 430) on generalization:

$$R = \frac{D^2}{\rho^2} \tag{1}$$

The margin is reduced if the hyperplane is a) "tilted" or b) biased (see Figure 1), for which the error bounds derived by Vapnik are greater than the error bound for the maximal margin SH.

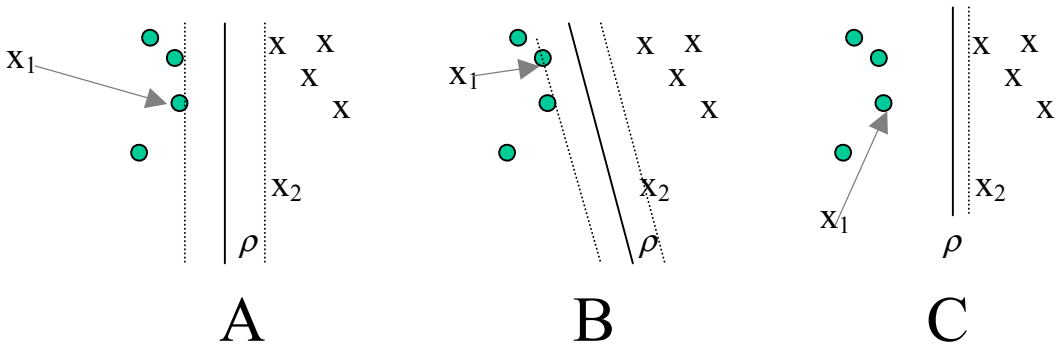

Figure 1. The maximal margin SH is in (A). The margin is reduced if the separating hyperplane is "titled" (B) or biased (C).

Scaling the function by a proper constant without changing the hyperplane, we can make f satisfy $\min_i(|\, w \cdot x_i + b \,|) = 1$, which is called the canonical form in Vapnik's book. The margin is calculated using canonical linear function satisfies inequality

$$\frac{w}{\| w \|} \cdot (x_1 - x_2) \geq \frac{2}{\| w \|} \tag{2}$$

where x_1, x_2 are the two training examples closest to the hyperplane from each side.

Vapnik proved that the maximal margin hyperplane is unique and the equality holds in (2) at the maximal margin hyperplane (page 402-404 of Vapnik 1998). Finding the maximal margin hyperplane becomes a constrained minimization problem (coefficient 1/2 in the objective function is introduced for technical convenience)

$$\begin{aligned} &\min \frac{1}{2} \| w \|^2 \quad with \\ &y_i (w \cdot x_i + b) \geq 1, \quad i = 1,...,l \end{aligned} \tag{3}$$

which is dealt with by solving its *dual problem*, derived from introducing Lagrange multipliers $\alpha_i \geq 0$ and a Lagrangian,

$$L(w,b,\alpha) = \frac{1}{2} \| w \|^2 + \sum_{i=1}^{l} \alpha_i [1 - y_i (w \cdot x_i + b)], \quad (4)$$

and taking partial derivative of the Lagrangian w.r.t the primal variables,

$$\frac{\partial}{\partial b} L(w,b,\alpha) = \sum_{i=1}^{l} y_i \alpha_i = 0, \quad (5)$$

$$\frac{\partial}{\partial w} L(w,b,\alpha) = w - \sum_{i=1}^{l} \alpha_i y_i x_i = 0 . \quad (6)$$

Equation (5) gives a linear constraint on the dual variables α_i and (6) gives the translation between the primal variables *w* and the dual variables, $w = \sum_{i=1}^{l} \alpha_i y_i x_i$. The training examples whose α_i is non-zero in (6) are called support vectors. Other examples have no impact on the solution. The primal variable *b* is calculated from the Karush-Kuhn-Tuker complementarity, $\alpha_i [1 - y_i (w \cdot x_i + b)] = 0$. Substituting (6) into the Lagrangian (4), we get the dual optimization problem

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad \text{subject to } \sum_{i=1}^{l} y_i \alpha_i = 0, \quad \alpha_i \geq 0, \quad i = 1, ..., l . \quad (7)$$

The hyperplane decision function is

$$D(x) = sign(\sum_{i=1}^{l} \alpha_i y_i (x_i \cdot x) + b) . \quad (8)$$

Rewriting (7) in matrix form, we have

$$\max W(\alpha) = \alpha^T * 1 - \frac{1}{2} \alpha^T Q \alpha \quad \text{subject to } \alpha^T y = 0, \qquad \alpha \geq 0 \quad (9)$$

where $Q = [y_i y_j (x_i \cdot x_j)]$, a *l*-by-*l* semi-positive symmetric matrix, called *kernel matrix*.

It is clear from Vipnik's original writings that the maximal margin hyperplane does not necessarily gives the smallest generalization error (even among all other SH's). It gives only the smallest value of the error *bound* function. More discussion on this issue can be found in Section 3.

1.2 Feature Space and "Non-linear" Support Vector Machines

To extend the power of L-SVM to learn non-linear separating surfaces, the original input space is (non-linearly) embedded into another space $\Im = R^N$, so called *feature space*, and then the L-SVM is applied in the feature space. The dimensionality of the feature space, *N*, is usually, but does not have to be, much larger than the dimensionality of the original space.

Let $\Phi : X \to \Im$ be an embedding. The training examples in the feature spaces are $\{(\Phi(x_i), y_i) \mid i = 1,...,l\}$ and the dual optimization problem takes exactly the same form as in (9) except that $Q = [y_i y_j (\Phi(x_i) \cdot \Phi(x_j))]$, the inner product in the feature space $\Im$ replaced the inner product in the original space. A computational short-cut is available to avoid the explicit construction of the feature space when the a so-called *kernel function*, $K(x, x') = (\Phi(x) \cdot \Phi(x'))$, can be directly computed in the original space.[2] The non-linear decision function is given by

$$D(x) = sign(\sum_{i=1}^{l} \alpha_i y_i K(x_i \cdot x) + b),$$

where the coefficients, $\{\alpha_i\}$, are the solution of the following constrained quadratic optimization,

$$\max W(\alpha) = \alpha^T * 1 - \frac{1}{2}\alpha^T Q \alpha \quad \text{subject to} \quad \alpha^T y = 0, \qquad \alpha \geq 0 \tag{9a}$$

where $Q = [y_i y_j K(x_i \cdot x_j)]$.

With the introduction of the feature space and the non-linear feature mapping, the meaning of the best generalization guarantee in a general feature space has not been carefully re-examined in Boser, Guyon & Vapnik (1992) and Cortes & Vapnik (1995).

We show in the next section that any SH in one feature space can be transformed into a maximal margin SH in another feature space of the same dimension. Therefore there is the question of which feature space's maximal margin hyperplane gives the best generalization performance guarantee? Given an arbitrary SH, *f*, in an arbitrary feature space, we show that there exists a maximal margin hyperplane (and its feature space) such that the best generalization performance guarantee anyone can possibly get on this maximal margin hyperplane is as bad as the generalization performance of *f*.

The rest of the paper is as follows. A topological homomorphism is constructed that maps a given SH into a maximal margin SH in Section 2. Discussions are in Section 3 and Conclusions in Section 4.

2. Any SH Is Equivalent to a Maximal Margin SH in Some Feature Space

For any SH f in a feature space $\Im$, a topological homomorphism[3] Ψ of the feature space exists to map f into a maximal margin SH f_0 without moving any of the training examples, only the area between the + and - examples shrinks or stretches. The topologically deformed space is another feature space with the same dimension as $\Im$. Both f and f_0 give the same separating surface in the original input space; therefore, they have the same generalization performance.

[2] Illustrative examples of feature spaces and kernels can be found in [CS00].
[3] A homomorphism is a continuous one-to-one onto mapping and the inverse of it is also continuous.

We provide a concrete procedure for the construction of Ψ in this section after the following thought experiment that gives the intuition behind our construction:

- The training examples are submerged into a rubber media and nailed to a rigid body behind (easier to imagine "behind" when dim($\Im$) < *3*).
- The SH is glued to the rubber media before moving it to the maximal margin SH.
- As we move the SH towards the maximal margin SH, the rubber media shrinks or stretches to keep the space "flat".

2.1 An One-dimensional Example

We use the lower case φ for the homomorphism in one-dimensional case so that it can be used in high dimension construction without conflicting notations.

In an one dimensional feature space, without loss of generality, we assume that there are only two training examples and they are at $x_1 = a$ and $x_2 = b$ (see Figure 2). Any point $f \in (a,b)$ is a separating "hyperplane". The maximal margin SH is $f_0 = (a+b)/2$.

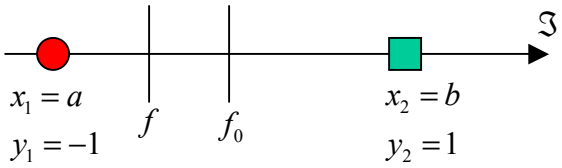

Figure 2. A linearly separable learning problem in one-dimensional feature space.

We construct a continuous strictly monotone increasing piece-wise linear $\varphi : R \to R$ that maps the given f to f_0 without moving the examples (and any point beyond the examples) as

$$\varphi(x;a,f,f_0,b) = \begin{cases} x & x \notin (a,b) \\ \dfrac{f_0 - a}{f - a}x + \dfrac{f_0 - f}{f - a} & x \in (a,f) \\ \dfrac{b - f_0}{b - f}x + \dfrac{f_0 - f}{b - f} & x \in (f,b) \end{cases}. \tag{10}$$

The same construction works for both $a < f < f_0 < b$ and $a < f_0 < f < b$. The construction does not require $f_0 = (a+b)/2$. φ is invertible. Its inverse is also continuous and can be constructed by the same procedure:

$$\varphi^{-1}(x;a,f,f_0,b) = \varphi(x;a,f_0,f,b) \tag{11}$$

The mapping φ we just defined has only a piece-wise continuous derivative. At *x= a, f, and b,* the derivatives are not defined. But constructing such a φ with a continuous derivative is

possible but more complex. We do not have to introduce this complexity because our discussion in Section 3 does not depend on it.

2.2 An Two-dimensional Example

Let $\Phi : X \to \Im$ be a feature mapping and $f_1 : \Im \to R$ a SH. We do not exclude $\Im = X$, the original space, and $\Phi = identity$. The 2-dimensional vectors in $\Im$ are written as (τ_1, τ_2). Since the maximal margin SH is invariant under orthogonal transformation and translation, without loss generality, we can always transform the maximal margin SH $f_0 : \Im \to R$ to $\tau_2 = 0$, the horizontal axis. One class of training examples have $\tau_2 \le -\rho$ and the other class have $\tau_2 \ge \rho > 0$, where ρ is the maximal geometric margin.

We construct a new feature mapping $\tilde{\Phi} : X \to \tilde{\Im}$ by "deforming" $\Im$ under a topological homomorphism $\Psi : \Im \to \Im$ without moving the training examples. More precisely, $\tilde{\Im} = \Im$, $\Phi_2 = \Psi \circ \Phi_1$, and $\Psi(\Phi(x_i)) = \Phi(x_i)$, for all *i*.

Case 1: If f_1 and f_0 are parallel, Ψ can be defined as a simple extension of φ in (10),

$$\Psi(\tau_1, \tau_2) = (\tau_1, \varphi(\tau_2 \mid -\rho, f_1, f_0, \rho)) .$$ Done.

Case 2: If f_1 and f_0 are not parallel, we can always translate everything along $\tau_1 - axis$ to make the intersection of f_1 and f_0 equal to (0,0). Representing everything in a polar coordinate system $(r, \theta) \in (-\infty, \infty) \times (-\pi/2, \pi/2)$, f_0 is given by $\theta = 0$ and f_1 by $\theta = \theta_1$ for some θ_1. Since the number of training examples is finite and none of them is in the interval $[0, \theta_1]$, there exists a sufficiently small $\delta > 0$, such that no training examples will fall in the interval $(-\delta, \theta_1 + \delta)$ [4]. Let $\theta_a = -\delta$ and $\theta_b = \theta_1 + \delta$. We have $0, \theta_1 \in (\theta_a, \theta_b)$. Applying φ in (10) to the angle coordinate, we define

$$\Psi(r, \theta) = (r, \varphi(\theta \mid \theta_a, \theta_1, 0, \theta_b)) . \quad (12)$$

$\Psi(r, \theta)$ is continuous and invertible because both *r* and $\varphi(\theta \mid \theta_a, \theta_1, 0, \theta_b)$ are continuous. The inverse of $\Psi(r, \theta)$ is also continuous and

$$\Psi^{-1}(r, \theta) = (r, \varphi^{-1}(\theta \mid \theta_a, \theta_1, 0, \theta_b)) . \quad (13)$$

Intuitively, $\Psi(r, \theta)$ is defined by shrinking the angle between θ_a and θ_1, and stretching the angle between θ_1 and θ_b until f_1 becomes horizontal (the maximal margin SH). The radius remains unchanged for all points in the feature space.

We point out that under the topological homomorphism, the maximal margin SH is still a SH after the deformation by our construction. This fact is used later.

---

[4] If $\theta_1 < 0$, $\theta_a < \theta_1 - \delta$ and $\theta_b = \delta$.

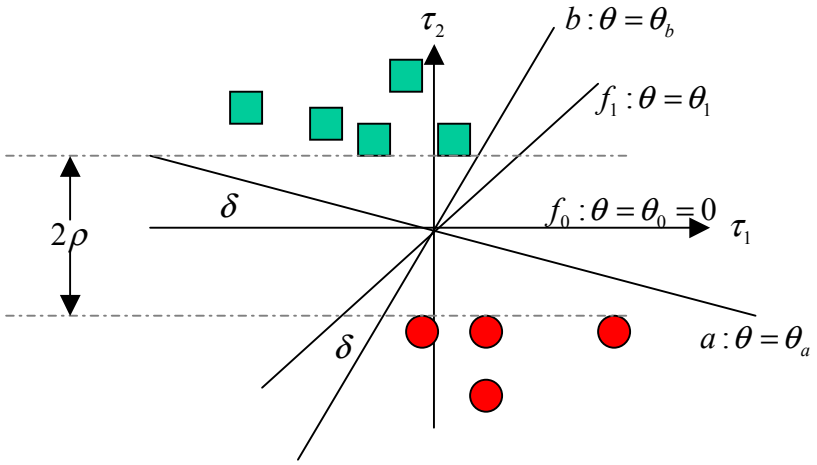

Figure 3. The two-dimensional example in a polar-coordinate system.

2.3 High Dimensional Case

If the two hyperplanes are parallel, the construction of Ψ is done along the norm vectors. Without loss of generality, we assume that all the norms are parallel to the τ_1-axis. Let $\hat{f}_0$ and $\hat{f}_1$ be the intersections of f_0 and f_1 with the τ_1-axis. Ψ is defined as,

$$\Psi(\tau_1,\tau_2,...,\tau_N)=(\varphi(\tau_1\mid \hat{f}_0-\rho,\hat{f}_1,\hat{f}_0,\hat{f}_0+\rho),\tau_2,...,\tau_N)\,.\ \text{Done.}$$

When the two hyperplanes are not parallel, the construction is reduced to a construction in $V_1=span(norm(f_1),norm(f_0))$, the two dimensional subspace of $\Im$ spanned by the two normal vectors of the SH f_1 and the maximal margin SH f_0. Let $V_2\subset\Im$ be the *N-2* dimensional subspace perpendicular to V_1. In the orthogonal decomposition $\Im=V_1\oplus V_2$, every vector in $\Im$ is uniquely written as $v=(v_1,v_2)$ where $v_1\in V_1$ and $v_2\in V_2$. Projecting all the training examples and the SH's onto the 2-dimensional subspace V_1 by the orthogonal projection P_1 and then following the procedure in Section 2.2, we construct the 2-dimensional homomorphism Ψ on V_1. $(\Psi\circ P_1,P_2):\Im\to\Im$ gives the topological homomorphism we need on the whole space $\Im$, where $P_2:\Im\to V_2$ is the orthogonal projection onto V_2.

Intuitively, we keep the intersection of the two hyperplanes and all the training examples fixed, and rotate $norm(f_1)$ towards $norm(f_0)$ as if they are in a rubber media.

3. Discussions

Starting with a feature space $\Phi:X\to\Im$ and a SH $f_1:\Im\to R$, we constructed a new feature mapping $\tilde{\Phi}=\Psi\circ\Phi:X\to\tilde{\Im}$ such that $f_1\circ\Psi^{-1}:\tilde{\Im}\to R$ is the maximal margin SH of the training examples $\tilde{\Phi}(x_i)=\Psi\circ\Phi(x_i)$ in $\tilde{\Im}$. Moreover, by our construction, the maximal margin SH f_0 is still a SH in $\tilde{\Im}$. Therefore, f_0 can be viewed as the image of a SH from $\tilde{\Im}$

under the topological homomorphism $\tilde{\Psi} = \Psi^{-1}$. The relationship between the triplets $(\Phi : X \to \Im, f_0, f_1)$ and $(\tilde{\Phi} : X \to \tilde{\Im}, \tilde{f}_0, \tilde{f}_1)$ is completely symmetric:

$$\begin{aligned} &(\tilde{\Phi} = \Psi \circ \Phi : X \to \tilde{\Im}, \tilde{f}_0 = f_1 \circ \Psi^{-1}, \tilde{f}_1 = f_0 \circ \Psi^{-1}) \\ &(\Phi = \tilde{\Psi} \circ \tilde{\Phi} : X \to \Im, f_0 = \tilde{f}_1 \circ \tilde{\Psi}^{-1}, f_1 = \tilde{f}_0 \circ \tilde{\Psi}^{-1}) \end{aligned} . \tag{14}$$

We could have constructed $\Phi : X \to \Im$ out of $\tilde{\Phi} : X \to \tilde{\Im}$ by exactly the same procedure (see (11)).

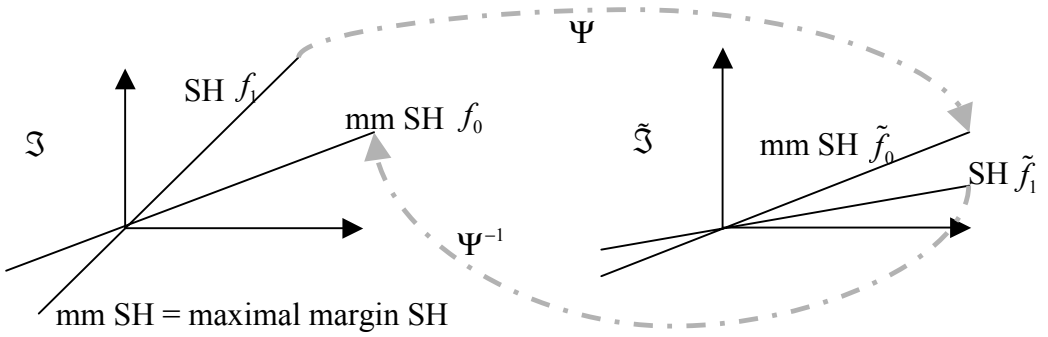

Figure 4. The relationship of the two feature spaces, the SH's and maximal margin SH's are completely symmetric.

3.1 Does the Maximal Margin Hyperplane Generalize Better than Any Other Separating Hyperplanes in the Same Feature Space?

From (14), we have $\tilde{f}_0 \circ \tilde{\Phi} = f_1 \circ \Psi^{-1} \circ \Psi \circ \Phi = f_1 \circ \Phi$; therefore, $\tilde{f}_0$ and f_1 give the same separating surface in the original input space and they must have the same generalization property. So are f_0 and $\tilde{f}_1$.

If the maximal margin SH always generalizes better than an arbitrary SH in the same feature space, we have the following two conclusions:

In $\Phi : X \to \Im$, f_0 generalizes better than f_1 and $\tilde{f}_0$.
In $\tilde{\Phi} : X \to \tilde{\Im}$, $\tilde{f}_0$ generalizes better than $\tilde{f}_1$ and f_0.

Based on both conclusions, f_0 and $\tilde{f}_0$ must be equally good. Since $\tilde{f}_0$ is mapped from an arbitrary SH in $\Phi : X \to \Im$, we conclude that all SH's in $\Phi : X \to \Im$ are equally good, a statement that is obviously false.

The maximal margin hyperplane can not always generalize better than other SH's in the same feature space. This statement also applies to the maximal margin hyperplane in the original input space. L-SVM does not give optimal generalization performance.

3.2 Any Separating Surface Given by a SH Is a Solution of (9a) Under Some Kernel Function

The separating surface defined by any SH f_1 in any feature space $\Phi: X \to \Im$, including the original input space, can also be given by a maximal margin hyperplane $\tilde{f}_0$ in $\tilde{\Phi} = \Psi \circ \Phi: X \to \tilde{\Im}$; therefore, it is the unique solution of (9a) under the kernel $K(x, x') = (\tilde{\Phi}(x), \tilde{\Phi}(x')) = (\Psi \circ \Phi(x), \Psi \circ \Phi(x'))$.

The constrained optimization in (9a) does not imply optimal generalization performance of the separating surface. Especially any SH in the original input space is the solution of (9a) under some kernel.

Since the discussion in Section 3.1 and 3.2 covers the original input space, the L-SVM does not give optimal generalization performance either. It gives only the best generalization guarantee as stated in Vapnik's original writings.

3.3 What Is the Best Generalization Guarantee?

With the L-SVM, the "optimality" of the maximal margin SH is derived by minimizing an error *bound function* instead of the error function itself. The relationship between the error bound and the actual error is not necessarily monotone increasing. A smaller value of the error bound function does not imply a smaller generalization error. The minimum of the error bound function is not necessarily the minimum of the error function. Figure 5 shows a hypothetical situation to illustrate a possible relationship; the horizontal axis represents the set of all SH's which is not necessarily one dimensional.

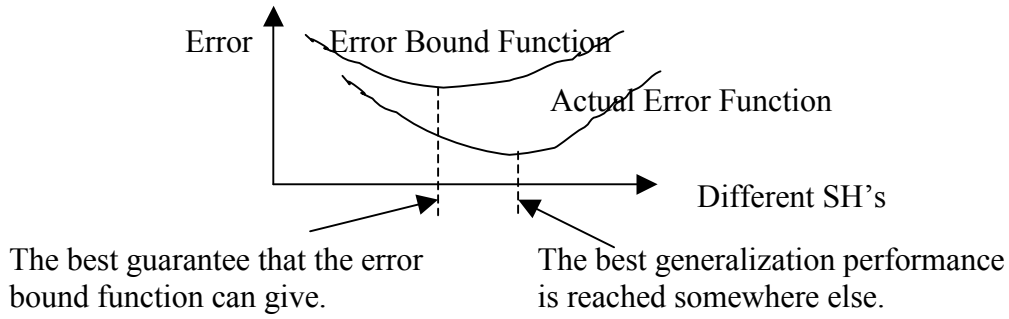

Figure 5. The minimum of the error bound function does not line-up with the minimum of the actual error function.

Maximal margin SH is not the optimal SH unless the error bound function has a monotone increasing relationship at least locally near its minimum with the actual error.

The amount of freedom granted by an arbitrary feature mapping allows us to "shift" the relationship between the error bound function and the actual error function in either direction – better aligned or worse aligned. We have shown that there exists a feature space $\tilde{\Phi}: X \to \tilde{\Im}$ in which the best generalization guarantee on its maximal margin hyperplane $\tilde{f}_0$ can not be better than the generalization performance of a randomly chosen SH f_1 in an arbitrary feature space $\Phi: X \to \Im$.

3.4 Overfitting

The following example (See Figure 5) shows that non-linear SVM may also overfit the data.

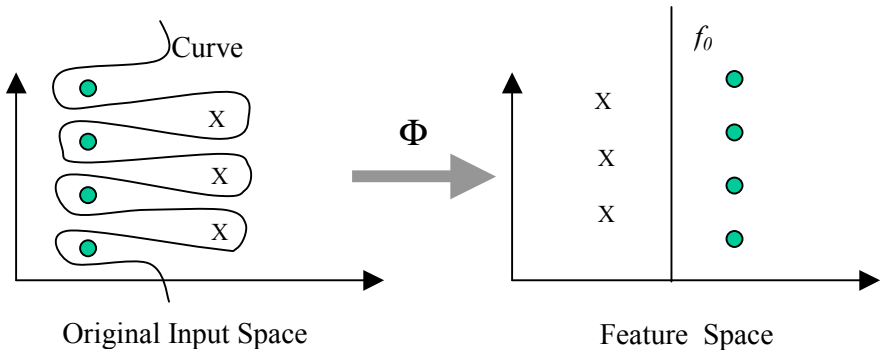

Figure 5. Non-linear SVM may overfit data just like other learning algorithms can.

The zigzag separating curve ("surface") together with the original input space and the training data is topologically equivalent to the maximal margin hyperplane in the feature space on the right with the new locations of the training examples. The topological homomorphism serves as the feature mapping. The best generalization guarantee given to the maximal margin hyperplane by any error bound function in the feature space can not be better than the actual generalization performance of the zigzag separating curve in the original input space.

3.5 Optimal Feature Mappings?

After the non-linear generalization of SVM, the optimization question becomes which separating surface (rather than hyperplane) generalizes best. The best separating surface is determined by both the feature mapping and the separating hyperplane in the feature space. The optimization goes back to the ill-posed problem of optimizing the pair $(\Phi : X \to \Im,\ f : \Im \to R)$ for the best generalization performance under the given training examples.

We have shown that any separating surface defined by a SH in a feature space can also be defined by a maximal margin hyperplane in another feature space. Therefore the optimal choice can be found by fixing the SH $f : \Im \to R$ to be the maximal margin SH and only optimizing over the feature mapping. Choosing the maximal margin SH merges the optimization over $f : \Im \to R$ into the optimization over $\Phi : X \to \Im$, which is the ill-posed problem faced by the practitioners. The old technique of regularization can be used on the feature mapping to turn it into a well defined optimization problem.

3.6 When The Original Input Space Is a Feature Space

In real-world problems, the training data is given in features that may have complicated correlations among them. For example the data collected along a manufacture line, satellite image data, etc. In such situations, we start with a feature space. The original input space does not occupy a special position among all feature spaces; it is just one of them. This argument reconfirms that L-SVM's best generalization guarantee could be as bad as the generalization performance of any (non-maximal margin) SH in any feature space. To prove this, all we have to

do is to deform the feature space to make the SH a maximal margin SH and call the new feature space the original input space.

3.7 "SVM Outperformed Most Other Systems in a Wide Variety of Applications"

The subtitle of this section is a citation from page 7 of Cristianini & Shawe-Taylor's book (2000). The issue we discuss in this paper is about the optimality of the maximal margin hyperplane in a general feature space, including the original input space. It does not conflict with the facts that SVM (including the non-linear SVM) has outperformed most other systems in a wide variety of applications. Even when the theoretical optimality of the maximal margin SH does not exist in a general feature space, the SVM algorithm could still be a good heuristic algorithm and have its own specialty that may make it outperform on certain applications.

4. Conclusions

The separating surface given by any separating hyperplane in any feature space can also be given by a maximal margin hyperplane in another feature space of the same dimension. Therefore, the maximal margin hyperplane does not have any special property to give the separating surface a better generalization performance. Without regularization on the feature mapping, the best performance guarantee given to the maximal margin hyperplane could be as bad as the generalization performance of a randomly chosen separating hyperplane in an arbitrary feature space.

Acknowledgements

I thank Dr. Meichun Hsu at HP Labs and Prof. Charles Elkan at UCSD for reading my manuscript and giving me valuable comments. The author of this paper is solely responsible for any mistakes.

References

Bennett, K.P. & Campbell, C. (2000), Support Vector Machines: Hype or Hallelujah? SIGKDD Explorations. Vol. 2, Issue 2, page 1.

Boser, B., Guyon, I. & Vapnik, V. (1992), A Training Algorithm for Optimal Margin Classifiers, In the ACM proceedings of the 5th International Workshop on Computational Learning Theory, p144-152.

Cortes, C. & Vapnik, V. (1995), Support-Vector Networks, Machine Learning 20, p273-297.

Cristianini, N. & Shawe-Taylor, J., An Introduction to Support Vector Machines and Other Kernel Based Learning Methods, Cambridge University Press, 2000.

Duda, R. & Hart, P. (1972), Pattern Classification and Scene Analysis, John Wiley & Sons.

Osuna, E., Freund, R. & Girosi, F. (1997), Training Support Vector Machines: An application to face detection, In Proc. Computer Vision and Pattern Recognition '97, p130-136.

Scholkopf, B., Burges, C. & Smola, A. (1998), Advances in Kernel Methods – Support Vector Learning, MIT Press.

Vapnik, V. (1998), Statistical Learning Theory, John Wiley & Sons.