On the VC dimension of bounded margin classifiers

Don Hush  
Computer Research Group, CIC-3  
Los Alamos National Laboratory  
Los Alamos, NM, 87545  
dhush@lanl.gov

Clint Scovel  
Computer Research Group, CIC-3  
Los Alamos National Laboratory  
Los Alamos, NM, 87545  
jcs@lanl.gov

Dedicated to Ané.

Abstract

Existing proofs of Vapnik’s result on the VC dimension of bounded margin classifiers rely on the assumption that the minimum margin over all dichotomies of $k \leq n + 1$ points contained in a sphere in $\mathbb{R}^n$ can be maximized by placing these points on a regular simplex whose vertices lie on the surface of the sphere (See [8], page 324 or [9], page 353). Although this assumption has intuitive appeal, it has not been proven correct (cf. Burges [2], page 30). This paper provides such a proof.

1 Introduction

Vapnik’s support vector machines (SVMs) [8, 9] represent a powerful class of Machine Learning methods. These methods use a form of structural risk minimization where the complexity of the classifier is controlled via the margin, which is defined as follows.

Definition 1 Let $X = \mathbb{R}^n$ be the n-dimensional Euclidian space, and let $H$ be the family of linear classifiers $c(x) = \text{sign}(h(x))$ where $h(x)$ is an affine function. Further, let $H_\rho$ be the set of linear classifiers that dichotomize $X$ using hyperplanes of thickness $\rho$. More formally, define $H_\rho$ to be classifiers of the form

$$c_\rho(x) = c(x), \quad D(x|h = 0) > \frac{\rho}{2}$$

where $D(x|h = 0)$ is the distance from $x$ to the hyperplane $h = 0$. (Note that $c_\rho(x)$ is not defined for $\{x : D(x|h = 0) \leq \frac{\rho}{2}\}$.) The margin of classifiers in $H_\rho$ is defined to be $\rho$. Finally, let $H_{\rho^*}$ be the set of linear classifiers with thickness greater than or equal to $\rho$, that is $H_{\rho^*} = \cup_{\phi \geq \rho} H_\phi$. 
The SVM method produces classifiers of maximal margin that correctly classify a fixed size training set. The following theorem, due to Vapnik [8, 9], provides the essential link between margin and the generalization error bound for SVMs.

**Theorem 1** (Vapnik, 1982) Let \( X_r = \{x_1, x_2, \ldots, x_k\} \subset X \) denote a set of points contained within a sphere of radius \( r \). The VC dimension of \( H_{r^+} \) restricted to \( X_r \) satisfies

\[
VCdim(H_{r^+}) \leq \min\left(\left\lfloor \frac{4r^2}{\rho^2} \right\rfloor \right., n + 1.
\]

To prove this result it is sufficient to determine the largest set \( X_r \) that can be shattered by the set of hyperplanes of thickness \( \rho \). The upper bound is obviously \( n + 1 \) (the number shattered when \( \rho = 0 \)). Existing proofs of the (potentially) tighter bound, \( \left\lfloor \frac{4r^2}{\rho^2} \right\rfloor + 1 \) rely on the almost obvious assumption that the minimum margin over all dichotomies of \( k \leq n + 1 \) points in \( \mathbb{R}^n \) can be maximized by placing these points on a regular simplex whose vertices lie on the surface of the sphere.

See [8], page 324 or [9], page 353. Although this assumption has intuitive appeal, it has not been proven correct (cf. Burges [2], page 30). The purpose of this paper is to provide such a proof.

In closely related work, we note that Shawe-Taylor et. al. [7] prove a bound on the level fat shattering dimension of the set of linear classifiers as a corollary to Theorem 1. On the other hand, Gurvits [3] provides a bound that amounts to Vapnik’s theorem with a loose constant. Bartlett and Shawe-Taylor [1] use Gurvits’ idea to improve bounds on the level fat shattering dimension of homogeneous linear classifiers. With no constant term. We can prove Vapnik’s theorem for \( k \) even using a modification of the technique used by Gurvits [3] and Bartlett and Shawe-Taylor [1](instead of averaging over all subsets we do so over the subsets of size \( \frac{k}{2} \)). However, a complete proof using these ideas is still open.

We begin by establishing some useful facts with the following lemmas.

### 2 Preparation

Let \( x = (x_1, x_2, \ldots, x_k) \) denote a vector of \( k \) points in \( \mathbb{R}^n \). Define \( r(x) \) to be the radius of the smallest ball in \( \mathbb{R}^n \) that contains all \( k \) points.

**Lemma 1** Suppose that \( r(x) \leq 1 \). Then in the center of mass frame

\[
\sum |x_i|^2 \leq k.
\]

**Proof:** By definition,

\[
r(x)^2 = \min_{x^*} \max_i |x_i - x^*|^2 \leq 1
\]

However,

\[
r(x)^2 = \min_{x^*} \max_i |x_i - x^*|^2 = \min_{x^*} \max_i \sum_{\lambda} \lambda_i |x_i - x^*|^2
\]
where $\sum_i \lambda_i = 1$ and $\lambda_i \geq 0$ (cf. [6]). This is a min-max problem with a function which is convex in $x^*$ and concave in $\lambda$ and so (von Neumann[4])

$$\min_{x^*} \max_{\lambda} \sum_{i} \lambda_i |x_i - x^*|^2 = \max_{x^*} \min_{\lambda} \sum_{i} \lambda_i |x_i - x^*|^2.$$ 

The minimum is obtained at $x^* = \sum \lambda_i x_i$, and if we fix an origin, this minimum has a value of $\sum \lambda_i |x_i|^2 - \sum \lambda_i x_i^2$. Consequently,

$$\sum_{i} \lambda_i |x_i|^2 - \sum \lambda_i x_i^2 \leq 1$$

for any $\lambda$. Letting $\lambda_i = \frac{1}{k}$, $i = 1, 2, \ldots, k$ and moving the origin to the center of mass, kills the second term and gives the result.

$$\square$$

We now compute some important measurements on the regular simplex.

**Lemma 2** Let $t$ denote the regular k-simplex with vertices on the unit sphere in $\mathbb{R}^{k-1}$ and let $1 \leq s \leq k$. Let $\rho(s)$ denote the distance from the convex hull of any $s$ vertices to the convex hull of the remainder. Let $d_{ij}$ denote the distance from vertex $i$ to vertex $j$. Then

$$\rho(s)^2 = \frac{k^2}{(k-1)s(k-s)}$$

$$d_{ij}^2 = |t_i - t_j|^2 = \frac{k}{k-1}, \ i \neq j.$$ 

**Proof:** Represent the regular k-simplex in $\mathbb{R}^k$ by the $k$ basis vectors

$$t_i = \sqrt{\frac{k}{k-1}}(0, 0, \ldots, 1, 0, \ldots, 0), i = 1, 2, \ldots, k$$

which are zero in all except for the $i$-th component. The distance from each of these vectors to the centroid of the simplex $\sqrt{\frac{1}{k(k-1)}}(1, 1, \ldots, 1, 1)$ is 1 so that in the $\mathbb{R}^{k-1}$ affine subspace spanned by these points they represent $k$ points on a regular simplex on the unit sphere. Direct calculation finishes the proof.

$$\square$$

### 3 Statement and Proof of the Theorem

We now state and prove the main theorem.

**Theorem 2** Let $x = (x_1, x_2, \ldots, x_k)$ denote a vector of $k$ points in $\mathbb{R}^n$. Define $r(x)$ to be the radius of the smallest ball in $\mathbb{R}^n$ that contains all $k$ points. Let $s$ denote a proper subset of the $k$ integers $\{1, 2, \ldots, k-1, k\}$ and let $x_s$ denote the set of points corresponding to the subset $s$. Let $\rho(x_s)$ be the distance between the convex hull of $x_s$ and the convex hull of its complement $x_{sc}$.

Then the value

$$\max_{x : r(x) \leq r} \min_{s} \rho(x_s)^2$$

is obtained when $x$ is a regular simplex with vertices on the sphere of radius $r$. 

Proof:

We first note that since $k$ points span at most $k - 1$ dimensions we can restrict to $n = k - 1$. It is also clear that $\max_{x: r(x) \leq 1} \min_{s} \rho(x_s)^2$ is quadratic in $r$, so we need to prove that the value

$$ \max_{x: r(x) \leq 1} \min_{s} \rho(x_s)^2 $$

is obtained when $x$ is a regular simplex with vertices on the unit sphere. Define

$$ h(k) = \max_{x: r(x) \leq 1} \min_{s} \rho(x_s)^2 $$

where $x$ is constrained so that $r(x) \leq 1$ and $s$ varies over all the proper subsets of the $k$ points. This is a maximin game with payoff function $\rho(x_s)^2$ and lower value $h(k) = \max_{x} \min_{s} \rho(x_s)^2$ (the upper value $v(k) = \min_{x} \max_{s} \rho(x_s)^2$ always satisfies $h(k) \leq v(k)$).

Our plan of attack is as follows. We extend to a game with payoff function $f(x, y)$ with the same lower value. Then we explicitly construct a saddle point $(x_0, y_0)$ to this extended game with $x_0$ a regular $k$-simplex, where a saddle point $(x_0, y_0)$ satisfies $f(x, y_0) \leq f(x_0, y) \leq f(x_0, y)$ for all $x$ and $y$. By von Neumann’s Theorem (von Neumann and Morgenstern [5] pg 95),

$$ h(k) = f(x_0, y_0). $$

This proves the theorem.

To make $x_s$ a vector we define $x_s = (x_{i_1}, x_{i_2}, ..., x_{i_{|s|}})$, where $i_j$ are all in $s$ and they are monotonic $i_1 < i_2 < ... < i_{|s|}$. Observe that $\rho(x_s)^2$ itself is a minimization

$$ \rho(x_s)^2 = \min_{p^s, q^s} \left| \sum_{i \in s} p_i x_i - \sum_{j \in s^c} q_j x_j \right|^2 $$

where $p^s$ are vectors of length $|s|$, with $p_i \geq 0, i = 1, .., |s|$ and $\sum_{i=1,|s|} p_i = 1$ and likewise for $q^s$ except that it is of length $|s^c| = k - |s|$. Therefore we first rewrite the max-min game as a max-min game with payoff function $F(x, z) = |\sum_{i \in s} p_i x_i - \sum_{j \in s^c} q_j x_j|^2$ where $z = (s, p^s, q^s)$. We extend again by observing that

$$ \min_{(s, p^s, q^s)} \left| \sum_{i \in s} p_i x_i - \sum_{j \in s^c} q_j x_j \right|^2 = \min_{(\lambda, p, q)} \sum_{s} \lambda_s \left| \sum_{i \in s} p_i x_i - \sum_{j \in s^c} q_j x_j \right|^2, $$

where $\lambda$ varies over all probability distributions over the set of proper subsets and where $p = \prod_{s} \{p^s\}$ and $q = \prod_{s} \{q^s\}$ are the product variables. This forms a new min-max game with the same lower value as the original with payoff function $f(x, y) = \sum_{s} \lambda_s \left| \sum_{i \in s} p_i x_i - \sum_{j \in s^c} q_j x_j \right|^2$ where $y = (\lambda, p, q)$. Consequently the chain of extensions can be written

$$ h(k) = \max_{x} \min_{s} \rho(x_s)^2 = \max_{x} \min_{z} F(x, z) = \min_{x} \max_{y} f(x, y) $$

Lemma 3 The function

$$ f(x, y) = \sum_{s} \lambda_s \left| \sum_{i \in s} p_i x_i - \sum_{j \in s^c} q_j x_j \right|^2 $$
has a saddle point at \((t, y^*)\) where \(t\) is the regular simplex on the unit sphere and \(y^* = (1_{[k/2]}, P, Q)\) where \(1_{[k/2]}\) is the probability whose mass lies uniformly distributed over the set of subsets \(s\) such that \(|s| = \left[\frac{k}{2}\right]\) and \(P^* = \frac{1}{|s|} (1, 1, \ldots, 1, 1)\) and \(Q^* = \frac{1}{k-|s|} (1, 1, \ldots, 1, 1)\)

Proof: Recall the definition of a saddle at \((t, y^*)\):

\[
f(x, y^*) \leq f(t, y^*) \leq f(t, y)
\]

for all \(x\) and \(y\). We prove these inequalities one at a time.

Proof of \(f(t, y^*) \leq f(t, y)\):

The simplex is special in that

\[
\left| \sum_{i \in s} p_i^x t_i - \sum_{j \in s'} q_j^x t_j \right|^2 = \sum_{i \in s} (p_i^x)^2 + \sum_{j \in s'} (q_j^x)^2
\]

which has its minimum value \(\frac{k^2}{(k-1)|s||k-|s|}|\) at \(p^x = P^*\) and \(q^x = Q^*\).

Consequently,

\[
f(t, (\lambda, P, Q)) \leq f(t, (\lambda, p, q))
\]

Since the function \(\frac{k^2}{(k-1)|s||k-|s|}|\) is constant on the strata of subsets of size \(|s|\),

\[
f(t, (\lambda, P, Q)) = \frac{k^2}{k-1} \sum_s \frac{1}{|s|(k-|s|)} \lambda_s = \frac{k^2}{k-1} \sum_s \lambda_s \frac{1}{|s|(k-|s|)}.
\]

Since \(\frac{1}{|s|(k-|s|)}\) is minimal at \(|s| = \left[\frac{k}{2}\right]\), \(f(t, (\lambda, P, Q))\) is then minimized by placing all the mass of \(\lambda\) entirely on \(|s| = \left[\frac{k}{2}\right]\). Consequently,

\[
f(t, (1_{[k/2]}, P, Q)) \leq f(t, (\lambda, P, Q))
\]

and therefore

\[
f(t, y^*) \leq f(t, y).
\]

Proof of \(f(x, y^*) \leq f(t, y)\):

By definition

\[
f(x, y^*) = \frac{1}{|s|} \sum_{|s| = \left[\frac{k}{2}\right]} \frac{1}{|s|} \sum_{i \in s} x_i - \frac{1}{k - |s|} \sum_{i \in s'} x_i^2,
\]

but if we choose the origin to be at the center of mass so that \(0 = \sum x_i\) this becomes a positive multiple of

\[
\sum_{|s| = \left[\frac{k}{2}\right]} \sum_{i \in s} x_i^2.
\]
Reverse the order of summation and expand so that

\[ \sum_{x \in \mathcal{X}} \left| \sum_{i \in s} x_i \right|^2 = \sum_{i,j \in s, |s|=\lceil \frac{k}{2} \rceil} x_i \cdot x_j. \]

The interior sum \( \sum_{x \in \mathcal{X}, |s|=\lceil \frac{k}{2} \rceil} x_i \cdot x_j \) is \( x_i \cdot x_j \) times the number of subsets which contain both \( i \) and \( j \). When \( i = j \) it is \( \binom{k-1}{\lceil \frac{k}{2} \rceil - 2} \) but when \( i \neq j \), \( \binom{k-2}{\lceil \frac{k}{2} \rceil - 2} \). Consequently,

\[
\sum_{i,j \in s, |s|=\lceil \frac{k}{2} \rceil} x_i \cdot x_j = \left( \binom{k}{\lceil \frac{k}{2} \rceil - 1} - 1 \right) \sum_i |x_i|^2 + \left( \binom{k}{\lceil \frac{k}{2} \rceil - 2} - \binom{k-1}{\lceil \frac{k}{2} \rceil - 2} \right) \sum_{i,j} x_i \cdot x_j
\]

but since \( 0 = \sum x_i \) and \( \binom{k-1}{\lceil \frac{k}{2} \rceil - 1} > \binom{k-2}{\lceil \frac{k}{2} \rceil - 2} \) the second term vanishes and we are left with a positive multiple of

\[ \sum_i |x_i|^2 \]

From Lemma 1, we know that

\[ \sum_i |x_i|^2 \leq k \]

and for the simplex \( t \)

\[ \sum_i |t_i|^2 = k. \]

Therefore,

\[ f(x, y^*) \leq f(t, y^*). \]

The proof of Lemma 3 and therefore of Theorem 2 is finished.

\[ \square \]

Acknowledgments

This work was partially supported by the Los Alamos DOE Program in Applied Mathematical Sciences.

References


