Available online at www.sciencedirect.com
ScienceDirect

Theoretical Computer Science 363 (2006) 2–10

Theoretical Computer Science

www.elsevier.com/locate/tcs

Complexity of hyperconcepts

Joel Ratsaby

Ben Gurion University, P.O.B. 653, Beer-Sheva 84105, Israel

Abstract

In machine-learning, maximizing the sample margin can reduce the learning generalization error. Samples on which the target function has a large margin (γ) convey more information since they yield more accurate hypotheses. Let X be a finite domain and $\mathbb{S}$ denote the set of all samples $S \subseteq X$ of fixed cardinality m. Let $\mathcal{H}$ be a class of hypotheses h on X. A *hyperconcept* h' is defined as an indicator function for a set $A \subseteq \mathbb{S}$ of all samples on which the corresponding hypothesis h has a margin of at least γ. An estimate on the complexity of the class $\mathcal{H}'$ of hyperconcepts h' is obtained with explicit dependence on γ, the pseudo-dimension of $\mathcal{H}$ and m.

Keywords: Sample-dependent error-bounds; Large-margin samples; Learning complexity; Pseudo-dimension

1. Introduction

Over the course of the last decade the field of learning theory has benefited from a rich body of work and the inception of a few key mathematical quantities which concisely capture and describe the complexity of accurate learning. Among those that characterize the problem of learning pattern classification is the Vapnik–Chervonenkis (VC) dimension. Based on the VC-dimension, upper bounds on the worst-case deviation of the learner's error from the optimal error may be obtained. The bounds are a function of the sample size m and a complexity term which is dependent on the VC-dimension of the function class. They are independent of the actual training sample and hence become useful, i.e., have a value between 0 and 1, only for very large sample sizes (see [11]). Thus in this respect they are weak. More recently, sample-dependent bounds have been shown to be stronger [9,2,3,8]. They hold when the learner finds a hypothesis which maximizes the margin on a training sample over a class of functions.

As large-margin samples may yield better hypotheses (having lower error estimates) they convey more information about the target. Intuitively, this should result in fewer possible hypotheses. Since each hypothesis is associated with a particular set of training samples based on which the learner may be able to infer it to within a fixed given accuracy then we expect fewer possible sets of samples with an increasing amount of information, i.e., with an increase in the margin.

As part of a more comprehensive on-going work to formalize the worth of information for learning [7], we take here a combinatorial approach whereby we aim to estimate the cardinality of the class of sets of large-margin samples (each

E-mail address: ratsaby@bgu.ac.il
URL: http://www.bgu.ac.il/∼ ratsaby/

0304-3975/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2006.06.013

set is associated with a hypothesis). Its logarithm is taken as the (descriptional) complexity of any such sample. Our main result (Theorem 3) provides an estimate of this quantity.

We start with some notations and definitions.

2. Basic notations and definitions

Let X be a domain. For $a \in \mathbb{R}$, define $\operatorname{sgn}(a) = +1$ if $a \geqslant 0$ and -1 if $a < 0$. The following definitions can be found for instance in [1]. For a class $\mathcal{A}$ of real-valued functions on X the *Pseudo-dimension*, denoted as $\dim_p(\mathcal{A})$, is defined as the maximum cardinality m of a sample $S = \{x_{i_1}, \ldots, x_{i_m}\} \subset X$ such that there exists a *translate vector* $r = [r_1, r_2, \ldots, r_m] \in \mathbb{R}^m$ where for each vector $v \in \{-1, 1\}^m$ there is a function $a_v(\cdot) \in \mathcal{A}$ that satisfies $\operatorname{sgn}(a_v(x_{i_j}) - r_j) = v_j$ for $1 \leqslant j \leqslant m$. The sample S is said to be *shattered* by $\mathcal{A}$. In the special case where $\mathcal{A}$ consists of mappings from $X \to \{-1, 1\}$ and $r_j = 0$, $1 \leqslant j \leqslant m$, then $\dim_p(\mathcal{A})$ is called the *Vapnik–Chervonenkis dimension* of $\mathcal{A}$ and is denoted by $VC(\mathcal{A})$.

For any $\gamma > 0$ the *γ-dimension* of $\mathcal{A}$, denoted as $\operatorname{fat}_\gamma(\mathcal{A})$, is defined as the maximum cardinality m of a sample $S = \{x_{i_1}, \ldots, x_{i_m}\} \subset X$ such that there exists $r = [r_1, r_2, \ldots, r_m] \in \mathbb{R}^m$ where for each vector $v \in \{-1, 1\}^m$ there is a function $a_v(\cdot) \in \mathcal{A}$ that satisfies $a_v(x_{i_j}) \geqslant r_j + \gamma$ if $v_j = 1$ and $a_v(x_{i_j}) \leqslant r_j - \gamma$ if $v_j = -1$, for $1 \leqslant j \leqslant m$. The sample S is said to be *γ-shattered* by $\mathcal{A}$.

For $B > 0$ let $\mathcal{H}$ denote a class of real-valued functions h from X into $[0, B]$. Let $\mathbb{S} = X^m$ consist of all sets $S \subset X$ with cardinality $|S| = m$. For any $S = \{x_{i_1}, \ldots, x_{i_m}\}$ define the *margin* of h on S to be

$$d_S(h) = \min_{\{x_{i_j} : 1 \leqslant j \leqslant m\}} |h(x_{i_j}) - B/2|.$$

Let $A \subseteq X$ and denote by $\mathbb{I}(A)$ the indicator function which takes the value 1 for any $x \in A$ and zero otherwise. For brevity, we sometimes write $\mathbb{I}(|h(x) - B/2| - c)$ instead of $\mathbb{I}(\{x \in X : |h(x) - B/2| - c \geqslant 0\})$. Henceforth, X is assumed to be finite and well ordered.

3. Overview of the problem of learning

The generalized probably-approximately-correct (PAC) model of learning from examples [5] can be used to represent learning classification. Under this model, data used for training and testing are generated independently and identically (i.i.d) according to an unknown but fixed probability distribution P which is defined over the input/output pairing $(x, y) \in X \times \{-1, 1\}$ where the two possible classification categories are labeled as -1 and 1. In general, y need not be a function of x however for the purpose of our work here we assume $y = \operatorname{sgn}(t(x))$ where $t \in \mathcal{H}$ is some unknown *target* function in a class $\mathcal{H}$ of real-valued functions on X. To measure the misclassification error by any h in $\mathcal{H}$ the natural choice is the empirical error which is based on a randomly drawn sample $S_m = \{(x_{i_j}, y_{i_j})\}_{j=1}^m$ according to P. It is defined as

$$L_m(h) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}(\operatorname{sgn}(h(x_{i_j})) \neq y_{i_j}).$$

The (true) misclassification error of h is $L(h) = P((x, y) : \operatorname{sgn}(h(x)) \neq y)$ and is unknown to the learner as P is not known. According to this model, the process of learning amounts to finding a hypothesis $\hat{h} \in \mathcal{H}$ which minimizes the empirical error over $\mathcal{H}$, i.e.,

$$L_m(\hat{h}) = \min_{h \in \mathcal{H}} L_m(h).$$

The minimal true-misclassification error is clearly zero, i.e., $L(t) = 0$. But in general the error of learning, $L(\hat{h})$, may be greater than zero. One of the main aims of the research in this field is to understand which settings, i.e., different learning algorithms and hypothesis classes $\mathcal{H}$, yield lower error $L(\hat{h})$. Theoretical estimates of $L(\hat{h})$ exist for various scenarios (see for instance [1]), the simplest being the pure PAC setting. This is described in the following result which appears as Theorem 4.1 in [4].

Theorem 1. *Let $\mathcal{H}$ be a class of functions from X to $\{-1, 1\}$ with $VC(\mathcal{H}) = d$. For any probability distribution P on $X \times \{-1, 1\}$, let $S = \{(x_{i_j}, y_{i_j})\}_{j=1}^{m}$ be drawn according to P. Based on S, suppose $\hat{h} \in \mathcal{H}$ satisfies $L_m(\hat{h}) = 0$. Then for any $\varepsilon, \delta > 0$, with probability $1 - \delta$,*

$$L(\hat{h}) \leqslant \frac{2}{m}\left(d \log \frac{2em}{d} + \log \frac{2}{\delta}\right)$$

provided $m \geqslant \max\{d, 2/\varepsilon\}$.

Consider now the scenario where the learner obtains a sample S of cardinality m on which it is believed[1] that the target t has a large margin, i.e., $d_S(t) \geqslant \gamma$. Support vector machines and other kernel-based learning methods which use the principle of maximizing the margin [9,4] can be represented in this way. In general, the γ-dimension $\mathrm{fat}_\gamma(\mathcal{H})$ of a class $\mathcal{H}$ decreases with an increase in γ so, as the following result suggests, the error of the classifier is significantly reduced with a larger margin (this result appears as Corollary 4.14 in [4]).

Theorem 2. *Let $\mathcal{H}$ consist of real-valued functions from X to $[0, 1]$ and fix a $\gamma > 0$. Consider doing classification by thresholding functions h at $1/2$, i.e., $\mathrm{sgn}(h(x) - 1/2)$, and denote by $L_m(h)$, $L(h)$ the corresponding empirical and true-misclassification errors, respectively. For any probability distribution P on $X \times \{-1, 1\}$, let $S = \{(x_{i_j}, y_{i_j})\}_{j=1}^{m}$ be drawn according to P. Based on S suppose that $\hat{h} \in \mathcal{H}$ satisfies $L_m(\hat{h}) = 0$ and $d_S(\hat{h}) > \gamma$. Then*

$$L(\hat{h}) \leqslant \frac{2}{m}\left(\mathrm{fat}_{\gamma/8}(\mathcal{H}) \log \frac{8em}{\gamma\, \mathrm{fat}_{\gamma/8}(\mathcal{H})} \log \frac{32m}{\gamma^2} + \log \frac{4}{\gamma}\right) \equiv \varepsilon$$

provided $m \geqslant \max\{2/\varepsilon, \mathrm{fat}_{\gamma/8}(\mathcal{H})\}$.

This result is an example of a sample-dependent bound since in reality the value of γ is the margin achieved by $\hat{h}$ which obviously depends on the data. It is apparent that large-margin samples are worth more for learning since the bound decreases with increasing γ.

Our interest now is to estimate the complexity of large-margin samples. Since they convey more information about the target, i.e., yield hypotheses that have lower error-bounds with increasing γ, then as was discussed earlier in Section 1 one expects their complexity to decrease as γ increases.

In order to characterize this quantitatively, starting in the next section our approach is to consider all samples $S \subset X$ of size m on which a hypothesis h in $\mathcal{H}$ has a margin of at least $\gamma > 0$. Fixing any such h as a target t, we estimate the complexity, i.e., the number of bits sufficient to index the subset of samples S on which t has a margin of at least γ.

4. Hyperconcepts

Let the space $\mathbb{S}$ consist of all samples $S \subset X$ of size m. On $\mathbb{S}$ consider sets of the form

$$A_{\beta,h} = \{S \in \mathbb{S} : d_S(h) \geqslant \beta\}, \quad \beta > 0.$$

We refer to indicator functions on $\mathbb{S}$ which take the form

$$h'_{\beta,h}(S) = \mathbb{I}\left(A_{\beta,h}\right)$$

as *hyperconcepts* and we sometimes write just h'.

For any fixed margin parameter $\gamma > 0$ define the *hyperclass*

$$\mathcal{H}'_\gamma = \left\{h'_{\gamma,h} : h \in \mathcal{H}\right\}. \tag{1}$$

In words, $\mathcal{H}'_\gamma$ consists of all sets of samples $S \subseteq X$ of cardinality m on which the corresponding hypotheses h have a margin of at least γ.

[1] This belief is usually phrased as an assumption of being lucky to have such a target.

Considering γ, m and $\mathcal{H}$ as given and fixed but allowing the possible training sample $S \in \mathbb{S}$ to vary then the quantity $\log|\mathcal{H}'_\gamma|$ represents the description length of any hyperconcept $h' \in \mathcal{H}'_\gamma$ and is thus a measure of the richness or complexity of the hyperclass $\mathcal{H}'_\gamma$. It is the description length of a set t' (corresponding to the target t) which consists of γ-good samples S each of which may be chosen to learn t and yield $\hat{h}$ whose error $L(\hat{h})$ is bounded as in Theorem 2 (provided that S is drawn by the underlying probability distribution P^m).

As mentioned earlier, we expect this complexity to decrease with increasing γ since larger-margin samples produce better error-bounds which implicitly means having more information about t. So we expect fewer possible hypotheses and hence fewer possible sets of samples that induce them via learning.

The following main result gives an estimate of the dependence of this complexity on $\mathcal{H}$ and γ.

Theorem 3. *For any $1 \leqslant m \leqslant |X|$ and $\gamma > 0$, consider a class $\mathcal{H}$ of real-valued functions from X to the bounded interval $[0, B]$ and let $1 \leqslant \mathrm{fat}_{\gamma/2}(\mathcal{H}) \leqslant |X| - m$. Then*

$$\log\left|\mathcal{H}'_\gamma\right| \leqslant \mathrm{fat}_{\gamma/16}(\mathcal{H}) \log\left(\frac{8eB(|X|-m)}{\gamma\,\mathrm{fat}_{\gamma/16}(\mathcal{H})}\right) \log\left(\frac{16B^2(|X|-m)}{\gamma^2}\right) + 2.$$

The next section contains the technical work of the proof of this theorem.

5. Proof of Theorem 3

Viewing $\mathcal{H}'_\gamma$ as a class of Boolean sets on $\mathbb{S}$ then, in general, it may obtain a limited number of dichotomies of any *hypersample*[2] $\zeta_N = \{S^{(1)}, \ldots, S^{(N)}\}$ with $S^{(j)} \in \mathbb{S}$, $1 \leqslant j \leqslant N$ and $1 \leqslant N \leqslant |\mathbb{S}|$. The growth function $\Pi_{\mathcal{H}'_\gamma}(N)$, introduced by [10], bounds this number and is defined as:

$$\Pi_{\mathcal{H}'_\gamma}(N) \equiv \max_{\zeta_N \subseteq \mathbb{S}} \Pi_{\mathcal{H}'_\gamma}(\zeta_N) \equiv \max_{\zeta_N \subseteq \mathbb{S}} \left|\left\{[h'(S^{(1)}), \ldots, h'(S^{(N)})] : h' \in \mathcal{H}'_\gamma\right\}\right|. \tag{2}$$

Viewing $\mathbb{S}$ as a (maximal) finite hypersample, then $\left|\mathcal{H}'_\gamma\right| \leqslant \Pi_{\mathcal{H}'_\gamma}(|\mathbb{S}|)$. This allows to introduce a dependence on the pseudo-dimension of $\mathcal{H}$. The approach is to upper bound the growth function $\Pi_{\mathcal{H}'_\gamma}(N)$ and then evaluate it at $N = |\mathbb{S}|$.

Let N be a positive integer and consider any hypersample $\zeta_N = \left\{S^{(1)}, \ldots, S^{(N)}\right\} \subseteq \mathbb{S}$. Denote by $S_i^{(j)}$ the ith element of the sample $S^{(j)}$ based on the ordering of the elements of $S^{(j)}$ (which is induced by the ordering of the elements in X). Then

$$\begin{aligned}\Pi_{\mathcal{H}'_\gamma}(\zeta_N) &= \left|\left\{\left[\mathbb{I}\left(\min_{x\in S^{(1)}} |h(x) - B/2| - \gamma\right), \ldots, \mathbb{I}\left(\min_{x\in S^{(N)}} |h(x) - B/2| - \gamma\right)\right] : h \in \mathcal{H}\right\}\right| \\ &= \left|\left\{\left[\prod_{j=1}^{m} \mathbb{I}\left(|h(S_j^{(1)}) - B/2| - \gamma\right), \ldots, \prod_{j=1}^{m} \mathbb{I}\left(|h(S_j^{(N)}) - B/2| - \gamma\right)\right] : h \in \mathcal{H}\right\}\right|\end{aligned} \tag{3}$$

since the minimum of m functions exceeds γ only if all functions exceed it. Order the elements in each set of ζ_N by the underlying ordering on X. Then put the sets in lexical ordering starting with the first up to the mth element, so for instance, if $N = 3$, $m = 4$ and

$$\zeta_3 = \{\ \{x_2, x_8, x_9, x_{10}\}, \{x_2, x_5, x_8, x_9\}, \{x_3, x_8, x_{10}, x_{13}\}\}$$

the ordered version is

$$\{\{x_2, x_5, x_8, x_9\}, \{x_2, x_8, x_9, x_{10}\}, \{x_3, x_8, x_{10}, x_{13}\}\}.$$

For any point $x \in X$ let

$$\theta_h^\gamma(x) \equiv \mathbb{I}\left(|h(x) - B/2| - \gamma\right)$$

[2] We will allow the notation ζ_N to also represent more general hypersamples consisting of samples which are not necessarily of the same cardinality m.

and denote it more simply by $\theta_h(x)$. For any sample $S^{(i)}$ of cardinality $|S^{(i)}|\geqslant 1$ let

$$e_{S^{(i)}}(h) = \prod_{j=1}^{|S^{(i)}|} \theta_h(S_j^{(i)}).$$

For any hypersample ζ_N let

$$v_{\zeta_N}(h) \equiv \left[e_{S^{(1)}}(h), \ldots, e_{S^{(N)}}(h)\right],$$

where for brevity we sometimes write $v(h)$. Let

$$V_{\mathcal{H}}(\zeta_N) = \left\{v_{\zeta_N}(h) : h \in \mathcal{H}\right\}$$

or simply $V(\zeta_N)$. Then from (3) we have

$$\Pi_{\mathcal{H}'_\gamma}(\zeta_N) = |V_{\mathcal{H}}(\zeta_N)|\,. \tag{4}$$

For a positive integer N^*, define a mapping $Q : \mathbb{S}^N \to \mathbb{S}^{N^*}$ as follows:

Procedure Q. *Given a hypersample $\zeta_N \subseteq \mathbb{S}$. Construct a new hypersample ζ_{N^*} as follows: let $Y = X \setminus S^{(1)}$ and let the elements in Y be ordered according to their ordering on X (we will refer to them as $y_1, y_2, \ldots$). Let $S^{*(1)} = S^{(1)}$. For $2\leqslant i\leqslant |X| - m + 1$, let $S^{*(i)} = \{y_{i-1}\}$.*

Note that for $m > 1$, ζ_{N^*} is not contained in $\mathbb{S}$ since some of its elements are singletons of size less than m. Henceforth denote by $N^* \equiv |X| - m + 1$. We have the following:

Claim 1. *For any $1\leqslant N\leqslant |\mathbb{S}|$, and any $\zeta_N \subseteq \mathbb{S}$,*

$$|V_{\mathcal{H}}(\zeta_N)|\leqslant |V_{\mathcal{H}}(Q(\zeta_N))|.$$

Proof. Let $\zeta_{\tilde{N}} \equiv Q(\zeta_N)$. Note that by definition of Procedure Q, it follows that $\zeta_{\tilde{N}}$ is a hypersample having $\tilde{N} = N^*$ non-overlapping sets. They consist of a single set $\tilde{S}^{(1)}$ of cardinality m and sets $\tilde{S}^{(i)}$, $2\leqslant i\leqslant \tilde{N}$, each of which contains a single element of X. Their union $\bigcup_{i=1}^{\tilde{N}} \tilde{S}^{(i)} = X$. This holds for any ζ_N and $1\leqslant N\leqslant |\mathbb{S}|$.

Consider the sets $V_{\mathcal{H}}(\zeta_N)$, $V_{\mathcal{H}}(\zeta_{\tilde{N}})$ and denote them simply by V and $\tilde{V}$. For any $v \in V$ consider the following subset of $\mathcal{H}$:

$$B(v) = \{h \in \mathcal{H} : v(h) = v\}\,.$$

We consider two types of $v \in V$: the first does *not* have the following property: there exist functions $h_\alpha, h_\beta \in B(v)$ with $\theta^\gamma_{h_\alpha}(x) \neq \theta^\gamma_{h_\beta}(x)$ for at least one element $x \in X$. Denote by $\theta^\gamma_h \equiv [\theta^\gamma_h(x_1), \ldots, \theta^\gamma_h(x_{|X|})]$. Then in this case all $h \in B(v)$ have the same $\theta^\gamma_h = \theta$, where $\theta \in \{0, 1\}^{|X|}$. This implies that

$$e_{\tilde{S}^{(1)}}(h) = e_{S^{(1)}}(h) = v_1,$$

while for $2\leqslant j\leqslant \tilde{N}$ we have

$$e_{\tilde{S}^{(j)}}(h) = \theta_{k(j)},$$

where, using the fact that $\tilde{S}^{(j)}$ is a singleton, we alternatively refer to $\tilde{S}^{(j)}$ by the element $x_{k(j)} \in X$ and $\theta_{k(j)}$ denotes the $k(j)$th component of θ. Hence it follows that

$$|V_{B(v)}(\zeta_{\tilde{N}})| = |V_{B(v)}(\zeta_N)|.$$

Let the second type of v satisfy the complement condition, namely, there exist functions $h_\alpha, h_\beta \in B(v)$ with $\theta^\gamma_{h_\alpha}(x) \neq \theta^\gamma_{h_\beta}(x)$ for at least one point $x \in X$. If such x is an element of $S^{(1)}$ then the first part of the argument above holds and we still have

$$|V_{B(v)}(\zeta_{\tilde{N}})| = |V_{B(v)}(\zeta_N)|.$$

If, however, there is also such an x in $X \backslash S^{(1)}$ then since the sets $\tilde{S}^{(i)}$, $2 \leqslant i \leqslant \tilde{N}$, are singletons then there exists some $2 \leqslant k \leqslant \tilde{N}$

$$e_{\tilde{S}^{(k)}}(h_\alpha) \neq e_{\tilde{S}^{(k)}}(h_\beta).$$

Hence for this second type of v we have

$$|V_{B(v)}(\zeta_{\tilde{N}})| \geqslant |V_{B(v)}(\zeta_N)|. \tag{5}$$

Combining the above, then (5) holds for any $v \in V$.

Now, consider any two distinct $v_\alpha, v_\beta \in V$. Clearly, $B(v_\alpha) \cap B(v_\beta) = \emptyset$ since every h has only one unique $v(h)$. Moreover, for any $h_a \in B(v_\alpha)$ and $h_b \in B(v_\beta)$ we have $\tilde{v}(h_a) \neq \tilde{v}(h_b)$ for the following reason: there must exist some set $S^{(i)}$ with a point $x \in S^{(i)}$ such that $h_a(x) \neq h_b(x)$ (since $v_\alpha \neq v_\beta$). If $i = 1$ then they must differ on $\tilde{S}^{(1)}$, i.e., $e_{\tilde{S}^{(1)}}(h_\alpha) \neq e_{\tilde{S}^{(1)}}(h_\beta)$. If $i \neq 1$, then such a point x is in $X \backslash S^{(1)}$ hence in some $\tilde{S}^{(j)}$ and therefore $e_{\tilde{S}^{(j)}}(h_\alpha) \neq e_{\tilde{S}^{(j)}}(h_\beta)$, where $j \in \{2, \ldots, \tilde{N}\}$. Hence no two distinct v_α, v_β map to the same $\tilde{v}$. We therefore have

$$|V_{\mathcal{H}}(\zeta_N)| = \sum_{v \in V} |V_{B(v)}(\zeta_N)| \leqslant \sum_{v \in V} |V_{B(v)}(\zeta_{\tilde{N}})| = |V_{\mathcal{H}}(\zeta_{\tilde{N}})|, \tag{6}$$

where (6) follows from (5) which proves the claim. □

Note that by construction of Procedure Q, the dimensionality of the vectors in the set $V_{\mathcal{H}}(Q(\zeta_N))$ is N^*, i.e., $|X| - m + 1$, regardless of the cardinality N of ζ_N. In particular, the bound of Claim 1 holds for the hypersample which consists of N^* maximally overlapping sets S.

Let us denote by ζ_{N^*} any hypersample obtained by Procedure Q, namely,

$$\zeta_{N^*} \equiv \left\{ S^{*(1)}, S^{*(2)}, \ldots, S^{*(N^*)} \right\}$$

with any set $S^{*(1)} \subset X$ of cardinality m and

$$S^{*(k)} = \{x_{i_k}\} \quad \text{where } x_{i_k} \in X \setminus S^{*(1)}, \ \ k = 2, \ldots, N^*.$$

Hence we may now write

$$\max_{1 \leqslant N \leqslant |\mathbb{S}|} \max_{\zeta_N \subseteq \mathbb{S}} \Pi_{\mathcal{H}'_\gamma}(\zeta_N) \leqslant \max_{1 \leqslant N \leqslant |\mathbb{S}|} \max_{\zeta_N \subseteq \mathbb{S}} |V_{\mathcal{H}}(Q(\zeta_N))| \tag{7}$$

$$\leqslant \max_{\zeta_{N^*}} |V_{\mathcal{H}}(\zeta_{N^*})|, \tag{8}$$

where (7) follows from (3), (4) and Claim 1 while (8) follows by definition of ζ_{N^*}. Now,

$$|V_{\mathcal{H}}(\zeta_{N^*})| = |\{[e_{S^{*(1)}}(h), \ldots, e_{S^{*(N^*)}}(h)] : h \in \mathcal{H}\}| \leqslant 2|\{[e_{S^{*(2)}}(h), \ldots, e_{S^{*(N^*)}}(h)] : h \in \mathcal{H}\}|, \tag{9}$$

where (9) follows trivially since $e_{S^{*(1)}}(h)$ is binary. So from (8) we have

$$\max_{1 \leqslant N \leqslant |\mathbb{S}|} \max_{\zeta_N \subseteq \mathbb{S}} \Pi_{\mathcal{H}'_\gamma}(\zeta_N) \leqslant 2 \max_{\zeta_{N^*}} |\{[e_{S^{*(2)}}(h), \ldots, e_{S^{*(N^*)}}(h)] : h \in \mathcal{H}\}|$$

$$\leqslant 2 \max_{x_1, \ldots, x_{N^*-1}} |\{[\theta^\gamma_h(x_1), \ldots, \theta^\gamma_h(x_{N^*-1})] : h \in \mathcal{H}\}|, \tag{10}$$

where $x_1, \ldots, x_{N^*-1}$ run over any $N^* - 1$ points in X. Fix any subset $X_{N^*-1} = \{x_1, \ldots, x_{N^*-1}\} \subseteq X$. We henceforth denote

$$C_\gamma(X_{N^*-1}) \equiv \left| \left\{ \left[\theta^\gamma_h(x_1), \ldots, \theta^\gamma_h(x_{N^*-1}) \right] : h \in \mathcal{H} \right\} \right|. \tag{11}$$

We proceed to bound $C_\gamma(X_{N^*-1})$.

For any real number u define a quantization function

$$Q_\gamma(u) = \gamma \left\lfloor \frac{u}{\gamma} \right\rfloor.$$

For a function $h \in \mathcal{H}$ the function $Q_\gamma(h(x))$ maps from X into the finite subset $Z_\gamma = \{0, \gamma, 2\gamma, \ldots, \lfloor B/\gamma \rfloor \gamma\}$ of $[0, B]$. We denote by $Q_\gamma(\mathcal{H})$ the function class $\{Q_\gamma(h) : h \in \mathcal{H}\}$. Consider the restriction of $Q_\gamma(\mathcal{H})$ on X_{N^*-1} and denote it by $\mathcal{A}_\mathcal{H}$ with functions $\alpha_h \in \mathcal{A}_\mathcal{H}$ mapping X_{N^*-1} into Z_γ.

For any subset $Y \subset X$ let $\mathbb{R}^Y$ denote the space of real-valued functions on Y. For any $f \in \mathbb{R}^Y$ let the $l_\infty(Y)$-norm of f be defined as $\|f\|_Y = \max_{x \in Y} |f(x)|$. For any class F of functions on Y let $\mathcal{M}(\varepsilon, F, l_\infty(Y))$ be the packing number, i.e., the size of the maximal ε-separated set in F with respect to the $l_\infty(Y)$-norm. Let the uniform packing number be defined as

$$\mathcal{M}(\varepsilon, F, n) = \max\{\mathcal{M}(\varepsilon, F, l_\infty(Y)) : Y, |Y| = n\}.$$

Every pair of distinct elements $\alpha_1, \alpha_2 \in A_\mathcal{H}$ satisfy $\|\alpha_1 - \alpha_2\|_Y \geqslant \gamma$ since they must be different on at least one point $x \in Y$ and on x their values $\alpha_1(x)$ and $\alpha_2(x)$ are restricted to the set Z_γ. Hence $\mathcal{A}_\mathcal{H}$ is itself a maximal γ-separated set and therefore

$$|\mathcal{A}_\mathcal{H}| = \mathcal{M}(\gamma, \mathcal{A}_\mathcal{H}, l_\infty(X_{N^*-1})). \tag{12}$$

Note also that

$$C_\gamma(X_{N^*-1}) \leqslant |A_\mathcal{H}| \tag{13}$$

since given any $h \in \mathcal{H}$, for $x \in X_{N^*-1}$ with $\theta_h^\gamma(x) = 1$ then $\alpha_h(x)$ in general can take one of several possible values in Z_γ. Hence it now suffices to bound $\mathcal{M}(\gamma, \mathcal{A}_\mathcal{H}, N^* - 1)$.

Denoting by $d_\mathcal{A} = \text{fat}_\gamma(\mathcal{A}_\mathcal{H})$ then for $n \geqslant d_\mathcal{A}$ we have from Theorems 12.1 and 12.8 in [1] that

$$\mathcal{M}(8\gamma, \mathcal{A}_\mathcal{H}, n) < 2 \left(\frac{nB^2}{4\gamma^2} \right)^{d_\mathcal{A} \log(eBn/(\gamma d_\mathcal{A}))}. \tag{14}$$

This result holds for a general class of real-valued functions [3] (not necessarily with a discrete finite range as for $\mathcal{A}_\mathcal{H}$).

Now, if $\gamma < 2\varepsilon$ then

$$\text{fat}_\varepsilon(Q_\gamma(\mathcal{H})) \leqslant \text{fat}_{\varepsilon-\gamma/2}(\mathcal{H}) \tag{15}$$

since suppose $Q_\gamma(\mathcal{H})$ ε-shatters a set $\{x_{i_1}, \ldots, x_{i_k}\} \subseteq X$ with translate vector $r = [r_1, \ldots, r_k] \in \mathbb{R}^k$. Then for all $v \in \{-1, 1\}^k$ there is a function $f_v \in \mathcal{H}$ with

$$\begin{aligned} &Q_\gamma(f_v(x_{i_j})) - r_j \geqslant \varepsilon \quad \text{if } v_j = 1, \\ &Q_\gamma(f_v(x_{i_j})) - r_j \leqslant -\varepsilon \text{ if } v_j = -1, \end{aligned}$$

$1 \leqslant j \leqslant k$. It follows that

$$\begin{aligned} &f_v(x_{i_j}) - r_j \geqslant \varepsilon \qquad \text{if } v_j = 1, \\ &f_v(x_{i_j}) - r_j < -\varepsilon + \gamma \text{ if } v_j = -1 \end{aligned}$$

or equivalently

$$\begin{aligned} &f_v(x_{i_j}) - r_j - \gamma/2 \geqslant \varepsilon - \gamma/2 \quad \text{if } v_j = 1 \\ &f_v(x_{i_j}) - r_j - \gamma/2 < -\varepsilon + \gamma/2 \text{ if } v_j = -1, \end{aligned}$$

$1 \leqslant j \leqslant k$, so $\mathcal{H}$ $(\varepsilon - \gamma/2)$-shatters $\{x_{i_1}, \ldots, x_{i_k}\}$ with a translate vector $r = [r_1 + \gamma/2, \ldots, r_k + \gamma/2]$.

[3] Related to this, for a class H of binary-valued functions with a suitable definition of a functional margin, Ratsaby [6] estimated the cardinality of $F_\gamma \subset H$, where F_γ consists of all $h \in H$ having a margin $d_S(h) > \gamma$ on a set $S \subseteq X$.

Hence continuing from (15) and letting $\varepsilon = \gamma$ then

$$d_{\mathcal{A}} = \mathrm{fat}_{\gamma}(Q_{\gamma}(\mathcal{H})) \leqslant \mathrm{fat}_{\gamma/2}(\mathcal{H}).$$

So (14) becomes

$$\mathcal{M}(8\gamma, \mathcal{A}_{\mathcal{H}}, n) < 2\left(\frac{nB^2}{4\gamma^2}\right)^{\mathrm{fat}_{\gamma/2}(\mathcal{H})\log(eBn/(\gamma \mathrm{fat}_{\gamma/2}(\mathcal{H})))}. \tag{16}$$

Letting $n = N^* - 1$, from (12), (13) and (16) it follows that if $N^* - 1 \geqslant \mathrm{fat}_{\gamma/2}(\mathcal{H})$ then

$$C_{\gamma}(X_{N^*-1}) \leqslant 2\left(\frac{16(N^*-1)B^2}{\gamma^2}\right)^{\mathrm{fat}_{\gamma/16}(\mathcal{H})\log(8eB(N^*-1)/(\gamma \mathrm{fat}_{\gamma/16}(\mathcal{H})))}. \tag{17}$$

Together with (2), (10), (11) and recalling that $N^* - 1 = |X| - m$, we therefore have

$$\max_{1 \leqslant N \leqslant |\mathbb{S}|} \Pi_{\mathcal{H}'_{\gamma}}(N) \leqslant 4\left(\frac{16(|X|-m)B^2}{\gamma^2}\right)^{\mathrm{fat}_{\gamma/16}(\mathcal{H})\log(8eB(|X|-m)/(\gamma \mathrm{fat}_{\gamma/16}(\mathcal{H})))}.$$

Now, as noted earlier, $\left|\mathcal{H}'_{\gamma}\right| \leqslant \Pi_{\mathcal{H}'_{\gamma}}(|\mathbb{S}|)$ hence we have

$$\log\left|\mathcal{H}'_{\gamma}\right| \leqslant \mathrm{fat}_{\gamma/16}(\mathcal{H}) \log\left(\frac{8eB(|X|-m)}{\gamma\, \mathrm{fat}_{\gamma/16}(\mathcal{H})}\right) \log\left(\frac{16B^2(|X|-m)}{\gamma^2}\right) + 2.$$

This proves Theorem 3. □

Note that the function $\mathrm{fat}_{\gamma}()$ is non-increasing with γ hence since the first factor in the bound dominates the first log factor then for all $\gamma \geqslant c$, $\log\left|\mathcal{H}'_{\gamma}\right|$ is non-increasing with increasing γ and decreasing with increasing m, for some constant $c > 0$. When the margin parameter value γ goes to 0 the cardinality of $|\mathcal{H}'_{\gamma}|$ decreases to 1 so the above bound can clearly be improved in this case.

6. Conclusions

Recent results in learning theory indicate that larger-margin samples may yield improved learning rates and hence implicitly convey more information about the target. In this paper we posed the question of determining the dependence of the complexity of such good samples on the margin parameter. We introduced a new notion of a class $\mathcal{H}'_{\gamma}$ of hyperconcepts which are indicator functions of sets of all large-margin samples for hypotheses in a corresponding class $\mathcal{H}$. Based on the estimate of the cardinality of $\mathcal{H}'_{\gamma}$, we conclude that with more information, i.e., with a higher margin value γ, there are fewer possible sets of samples that can induce good (low-error) hypotheses. An open question is to obtain a tighter bound for small margin value.

References

[1] M. Anthony, P.L. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, Cambridge, 1999.

[2] A. Antos, B. Kgl, T. Linder, G. Lugosi, Data-dependent margin-based generalization bounds for classification, J. Mach. Learning Res. 3 (2002) 73–98.

[3] P.L. Bartlett, S. Boucheron, G. Lugosi, Model selection and error estimation, Mach. Learning 48 (2002) 85–113.

[4] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and other Kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000.

[5] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, Inform. and Comput. 100 (1) (1992) 78–150.

[6] J. Ratsaby, Density of smooth boolean functions. Proceedings of International Mathematical Conference—Topics in Mathematical Analysis and Graph Theory (MAGT'06), September 2006.

[7] J. Ratsaby, On the combinatorial representation of information, in: D.Z. Chen, D.T. Lee (Eds.), The Twelth Annual International Computing and Combinatorics Conference (COCOON'06), LNCS, Vol. 4112, Springer, 2006, pp. 479–488.

[8] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inform. Theory 44 (1998) 1926–1940.

[9] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[10] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (1971) 264–280.

[11] ⟨http://www.neurocolt.com/bounds2002.html⟩, NeuroCOLT Workshop on Generalization Bounds Less than 0.5, Windsor, 29 April–2 May, 2002.