Quantifying accuracy of learning via sample width

Martin Anthony
Department of Mathematics
London School of Economics
Houghton Street
London WC2A 2AE, UK
Email: m.anthony@lse.ac.uk

Joel Ratsaby
Electrical and Electronics Engineering Department
Ariel University of Samaria
Ariel 40700, Israel
Email: ratsaby@ariel.ac.il

***Abstract*—In a recent paper, the authors introduced the notion of *sample width* for binary classifiers defined on the set of real numbers. It was shown that the performance of such classifiers could be quantified in terms of this sample width. This paper considers how to adapt the idea of sample width so that it can be applied in cases where the classifiers are defined on some finite metric space. We discuss how to employ a greedy set-covering heuristic to bound generalization error. Then, by relating the learning problem to one involving certain graph-theoretic parameters, we obtain generalization error bounds that depend on the sample width and on measures of 'density' of the underlying metric space.**

I. Introduction

By a (binary) classifier (or function) on a set X, we mean a function mapping from X to $\{-1, 1\}$. A classifier indicates to which of two classes objects from X belong and, in supervised machine learning, it is arrived at on the basis of a *sample*, a set of objects from X together with their classifications (-1 or 1). In [3], the notion of *sample width* for binary classifiers mapping from the real line $X = \mathbb{R}$ was introduced. In this paper, we consider how a similar approach might be taken to the situation in which classifiers map from some finite metric space (which would not generally have the linear structure of the real line). The definition of sample width is given below, but it is possible to indicate the basic idea at this stage: we define sample width to be at least γ if the classifier achieves the correct classifications on the sample and, furthermore, for each sample point, the minimum distance to a point of the domain having opposite classification is at least γ.

A key issue that arises in machine learning is that of *generalization error*: given that a classifier has been produced by some learning algorithm on the basis of a (random) sample of a certain size, how can we quantify the accuracy of that classifier, where by its accuracy we mean its likely performance in classifying objects from X correctly? In this paper, we seek answers to this question that involve not just the sample size, but the sample width. By relating the question to that of quantifying generalization error in 'large-margin' classification, we obtain generalization error bounds involving 'covering numbers' of the underlying metric space. We then discuss how to employ the well-known greedy set-covering heuristic to bound these covering numbers, and hence the generalization error. We next show that we can obtain bounds on generalization error by considering the domination numbers of certain graphs associated with the underlying metric space. Using some combinatorial results bounding domination number in terms of graph parameters, including number of edges and minimum degree, we obtain generalization error bounds that depend on measures of density of the underlying metric space.

II. Measuring the accuracy of learning

We work in a version of the popular 'PAC' framework of computational learning theory (see [23], [9]). This model assumes that the sample $\mathbf{s}$ consists of an ordered set (x_i, y_i) of labeled examples, where $x_i \in X$ and $y_i \in Y = \{-1, 1\}$, and that each (x_i, y_i) in the training sample $\mathbf{s}$ have been generated randomly according to some fixed (but unknown) probability distribution P on $Z = X \times \{-1, 1\}$. (This includes, as a special case, the situation in which each x_i is drawn according to a fixed distribution on X and is then labeled deterministically by $y_i = t(x_i)$ where t is some fixed function.) Thus, a sample $\mathbf{s}$ of length m can be thought of as being drawn randomly according to the product probability distribution P^m. In general, suppose that H is a set of functions from X to $\{-1, 1\}$. An appropriate measure of how well $h \in H$ would perform on further randomly drawn points is its *error*, $\mathrm{er}_P(h)$, the probability that $h(X) \neq Y$ for random (X, Y).

Given any function $h \in H$, we can measure how well h matches the training sample through its *sample error*

$$\mathrm{er}_{\mathbf{s}}(h) = \frac{1}{m}|\{i : h(x_i) \neq y_i\}|$$

(the proportion of points in the sample incorrectly classified by h). Much classical work in learning theory (see [9], [23], for instance) related the error of a classifier h to its sample error. A typical result would state that, for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $h \in H$ we have $\mathrm{er}_P(h) < \mathrm{er}_{\mathbf{s}}(h) + \epsilon(m, \delta)$, where $\epsilon(m, \delta)$ (known as a *generalization error bound*) is decreasing in m and δ. Such results can be derived using uniform convergence theorems from probability theory [24], [19], [12], in which case $\epsilon(m, \delta)$ would typically involve a quantity known as the growth function of the set of classifiers [24], [9], [23], [2]. More recently, emphasis has been placed on 'learning with a large margin'. (See, for instance [22], [2], [1], [21].) The rationale behind margin-based generalization error bounds is that if a classifier can be thought of as a geometrical separator between points, and if it has managed to achieve a 'wide' separation between the points of different classification, then this indicates that it is a good classifier, and it is possible that a better generalization error bound can be obtained. Margin-based results apply when the classifiers are derived from real-valued function by 'thresholding' (taking their sign). Although the classifiers we consider here are not of this type, we can deploy margin-based learning theory by working with the real-valued functions related to the classifiers.

III. The width of a classifier

We now discuss the case where the underlying set of objects X forms a finite metric space. Let $X = [N] := \{1, 2, \ldots, N\}$ be a finite set on which is defined a metric $d : X \times X \to \mathbb{R}$. So, $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $y = x$; and $d(x, y) = d(y, x)$. Furthermore, d satisfies the triangle inequality:

$$d(a, c) \leq d(a, b) + d(b, c). \tag{1}$$

Let $D = [d(i, j)]$ be the corresponding 'distance matrix'. D is symmetric with (i, j)th element $d(i, j) \geq 0$, and with $d(i, j) = 0$ if and only if $i = j$.

For a subset S of X, define the distance from $x \in X$ to S as follows:

$$\text{dist}\,(x, S) := \min_{y \in S} d(x, y).$$

We define the *diameter* of X to be

$$\text{diam}_D(X) := \max_{x,y \in X} d(x, y) = \|D\|_\infty$$

where $\|D\|_\infty$ is the max-norm for matrix D.

We will denote by $\mathcal{H}$ the class of all binary functions h on X.

The paper [3] introduced the notion of the width of a binary function at a point in the domain, in the case where the domain was the real line $\mathbb{R}$. Consider a set of points $\{x_1, x_2, \ldots, x_m\}$ from $\mathbb{R}$, which, together with their true classifications $y_i \in \{-1, 1\}$, yield a *training sample*

$$\mathbf{s} = ((x_j, y_j))_{j=1}^m = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)) .$$

We say that $h : \mathbb{R} \to \{-1, 1\}$ achieves sample margin at least γ on $\mathbf{s}$ if $h(x_i) = y_i$ for each i (so that h correctly classifies the sample) and, furthermore, h is constant on each of the intervals $(x_i - \gamma, x_i + \gamma)$. It was then possible to obtain generalization error bounds in terms of the sample width. In this paper we use an analogous notion of width to analyse classifiers defined on a finite metric space. We now define the notion of width that naturally suits this space.

Let us denote by S_-^h and S_+^h the sets corresponding to the function $h : X \to \{-1, 1\}$ which are defined as follows:

$$S_-^h := \{x \in X : h(x) = -1\}, \quad S_+^h := \{x \in X : h(x) = +1\}. \tag{2}$$

We will often omit the superscript h. We define the *width* $w_h(x)$ of h at a point $x \in X$ to be the following distance (where $\bar{h}(x)$ is the sign opposite to that of $h(x)$, meaning $-$ if $h(x) = 1$ and $+$ if $h(x) = -1$):

$$w_h(x) := \text{dist}\left(x, S_{\bar{h}(x)}\right).$$

In other words, it is the distance from x to the set of points that are labeled the opposite of $h(x)$. The term 'width' is appropriate since the functional value is just the geometric distance between x and the set $S_{\bar{h}(x)}$.

Let us define the signed width function, or *margin function*, f_h, as follows:

$$f_h(x) := h(x) w_h(x).$$

Note that the absolute value of $f_h(x)$ is, intuitively, a measure of how 'definitive' or 'confident' is the classification of x by h: the higher the value of $f_h(x)$ the greater the confidence in the classification of x. Note also that the error $\text{er}_P(h)$ of h can also be expressed in terms of the margin function f_h:

$$\text{er}_P(h) = P(h(X) \neq Y) = P\left(Y h(X) < 0\right) = P\left(Y f_h(X) < 0\right). \tag{3}$$

We define the class $\mathcal{F}$ of margin functions as

$$\mathcal{F} := \{f_h(x) : h \in \mathcal{H}\}. \tag{4}$$

Note that f_h is a mapping from X to the interval $[-\text{diam}_D(X), \text{diam}_D(X)]$. Henceforth, we will use $\gamma > 0$ to denote a *learning margin parameter* whose value is in the range $(0, \text{diam}_D(X)]$.

IV. Margin-based generalization bounds and covering numbers

A. Margin-based bounds

For a positive margin parameter $\gamma > 0$ and a training sample $\mathbf{s}$, the *empirical* (sample) γ-margin error is defined as

$$\hat{P}_{\mathbf{s}}(Y f_h(X) < \gamma) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}\left(y_j f_h(x_j) < \gamma\right).$$

(Here, $\mathbb{I}(A)$ is the indicator function of the set, or event, A.)

Our aim is to show that the generalization misclassification error $P(Y f_h(X) < 0)$ is not much greater than

$\hat{P}_{\mathbf{s}}(Yf_h(X) < \gamma)$. Such a result, which would constitute a large-margin generalization bound for the class of margin functions, will in this context be a generalization bound that involves the sample margin (since the class $\mathcal{F}$ of margin functions is defined in terms of the sample margins).

Explicitly, we aim for bounds of the form: for all $\delta \in (0,1)$, with probability at least $1-\delta$, for all $h \in H$ and for all $\gamma \in (0, \operatorname{diam}_D(X)]$, we have

$$\operatorname{er}_P(h) = P(h(X) \neq Y) < \hat{P}_{\mathbf{s}}(Yf_h(X) < \gamma) + \epsilon(m,\delta).$$

This will imply that if the learner finds a hypothesis which, for a large value of γ, has a small γ-margin error, then that hypothesis is likely to have small error. What this indicates, then, is that if a hypothesis has a large width on most points of a sample, then it will be likely to have small error, exactly the type of result we seek.

B. Covering numbers

To use techniques from margin-based learning, we consider *covering numbers*. We will discuss different types of covering numbers, so we introduce the idea in some generality to start with.

Suppose (A,d) is a metric space and that $\alpha > 0$. Then an α-cover of A (with respect to d) is a finite subset C of A such that, for every $a \in A$, there is some $c \in C$ such that $d(a,c) \leq \alpha$. If such a cover exists, then the mimimum cardinality of such a cover is the *covering number* $\mathcal{N}(A,\alpha,d)$.

Suppose now that F is a set of functions from a domain X to some bounded subset Y of $\mathbb{R}$. For a finite subset S of X, the $l_\infty(S)$-norm is defined by $\|f\|_{l_\infty(S)} = \max_{x\in S}|f(x)|$. For $\gamma > 0$, a γ-cover of F with respect to $l_\infty(S)$ is a subset $\hat{F}$ of F with the property that for each $f \in F$ there exists $\hat{f} \in \hat{F}$ with the property that for all $x \in S$, $|f(x) - \hat{f}(x)| \leq \gamma$. The *covering number* $\mathcal{N}(F,\gamma,l_\infty(S))$ is the smallest cardinality of a covering for F with respect to $l_\infty(S)$. In other words, and to place this in the context of the general definition just given, $\mathcal{N}(F,\gamma,l_\infty(S))$ equals $\mathcal{N}(F,\gamma,d_\infty(S))$ where $d_\infty(S)$ is the metric induced by the norm $l_\infty(S)$. The *uniform covering number* $\mathcal{N}_\infty(F,\gamma,m)$ is the maximum of $\mathcal{N}(F,\gamma,l_\infty(S))$, over all S with $S \subseteq X$ and $|S| = m$.

C. A generalization result

We will make use of the following result. (Most standard bounds, such as those in [7], [2], do not have a factor of 3 in front of the empirical margin error, but involve ϵ^2 rather than ϵ in the negative exponential. This type of bound is therefore potentially more useful when the empirical margin error is small.) The proof is omitted here but can be found in the full version of this paper [5].

Theorem 4.1: Suppose that F is a set of real-valued functions defined on a domain X and that P is any probability measure on $Z = X \times \{-1,1\}$. Let $\delta \in (0,1)$ and $B > 0$, and let m be a positive integer. Then, with P^m probability at least $1-\delta$, a training sample of length m will be such that: *for all* $f \in F$, and *for all* $\gamma \in (0,B]$, the error $P(Yf(X) < 0)$ is no more than

$$3\,\hat{P}_{\mathbf{s}}(Yf(X) < \gamma) + \frac{4}{m}\left(\ln \mathcal{N}_\infty(F,\gamma/4,2m) + \ln\left(\frac{8B}{\gamma\delta}\right)\right).$$

Note that, in Theorem 4.1, γ is not specified in advance, so γ can be chosen, in practice, after learning, and could, for instance, be taken to be as large as possible subject to having the empirical γ-margin error equal to 0.

V. Covering the class $\mathcal{F}$

To use the generalization result just presented, we need to bound the covering number of $\mathcal{F}$. Our approach is to construct and bound the size of a covering with respect to the sup-norm on X. (This is the norm given by $\|f\|_\infty = \sup_{x\in X}|f(x)|$.) Any such covering clearly also serves as a covering with respect to $l_\infty(S)$, for any S, since if $\|f - \hat{f}\|_\infty \leq \gamma$ then, by definition of the sup-norm, $\sup_{x\in X}|f(x) - \hat{f}(x)| \leq \gamma$ and, hence, for all $x \in X$ (and, therefore, for all $x \in S$ where S is some subset of X), $|f(x) - \hat{f}(x)| \leq \gamma$.

We first show that the margin (or signed width) functions are 'smooth', in that they satisfy a sort of Lipschitz condition.

A. $\mathcal{F}$ is smooth

We prove that the class $\mathcal{F}$ satisfies a Lipschitz condition, as follows:

Theorem 5.1: For every $f_h \in \mathcal{F}$,

$$|f_h(x) - f_h(x')| \leq 2d(x,x') \tag{5}$$

uniformly for any x, $x' \in X$.

Proof: Consider two points x, $x' \in X$. We consider bounding the difference $|f_h(x) - f_h(x')|$ from above. There are two cases to consider: $h(x)$ and $h(x')$ equal, or different.

Case I, in which $h(x) \neq h(x')$. Without loss of generality, assume that $h(x) = +1$, $h(x') = -1$. Then $S_{\overline{h}(x)} = S_-$ and $S_{\overline{h}(x')} = S_+$. We have

$$\operatorname{dist}(x, S_-) = \min_{z\in S_-} d(x,z) \leq d(x,x'),$$

since $x' \in S_-$. Similarly,

$$\operatorname{dist}(x', S_+) = \min_{z\in S_+} d(x',z) \leq d(x',x),$$

since $x \in S_+$. Hence,

$$\begin{aligned}
|f_h(x) - f_h(x')| &= |h(x)\mathrm{dist}(x, S_-) - h(x')\mathrm{dist}(x', S_+)| \\
&= |\mathrm{dist}(x, S_-) + \mathrm{dist}(x', S_+)| \\
&\leq d(x, x') + d(x', x) \\
&= 2d(x, x'),
\end{aligned}$$

since $d(x, x') = d(x', x)$ by symmetry of the metric.

Case II, in which $h(x) = h(x')$. Without loss of generality, assume that $h(x) = h(x') = +1$. Then $S_{\overline{h(x)}} = S_{\overline{h(x')}} = S_-$. We have,

$$\begin{aligned}
|f_h(x) - f_h(x')| &= |h(x)\mathrm{dist}(x, S_-) - h(x')\mathrm{dist}(x', S_-)| \\
&= |\mathrm{dist}(x, S_-) - \mathrm{dist}(x', S_-)| \\
&= \left| \min_{z \in S_-} d(x, z) - \min_{z \in S_-} d(x', z) \right|. \qquad (6)
\end{aligned}$$

Denote by s, s' the closest points in S_- to x, x', respectively. Then

$$\left| \min_{z \in S_-} d(x, z) - \min_{z \in S_-} d(x', z) \right| = |d(x, s) - d(x', s')| \, (7)$$

Assume that

$$d(x, s) \geq d(x', s') \qquad (8)$$

so that (7) equals $d(x, s) - d(x', s')$. We have

$$d(x, s) \leq d(x, s') \leq d(x, x') + d(x', s') \qquad (9)$$

where the last inequality follows from the fact that D satisfies the triangle inequality (1).

So combining (6), (7), (8) and (9) gives the following upper bound,

$$\begin{aligned}
|f_h(x) - f_h(x')| &\leq d(x, x') + d(x', s') - d(x', s') \\
&= d(x, x').
\end{aligned}$$

In the other case where the inequality (8) is reversed we also obtain this bound. ■

Next we use this 'smoothness' to obtain a cover for $\mathcal{F}$.

B. Covering $\mathcal{F}$

Let the subset $C_\gamma \subseteq X$ be a *minimal* size γ-cover for X with respect to the metric d. So, for every $x \in X$ there is some $\hat{x} \in C_\gamma$ such that $d(x, \hat{x}) \leq \gamma$. Denote by N_γ the cardinality of C_γ.

Let Λ_γ be the set of numbers of the form $\lambda_i = i\gamma$ as i runs through the integers from $-\left\lceil \frac{\mathrm{diam}_D(X)}{\gamma} \right\rceil$ to $\left\lceil \frac{\mathrm{diam}_D(X)}{\gamma} \right\rceil$ and define the class $\hat{F}$ to be all functions $\hat{f} : C_\gamma \to \Lambda_\gamma$. Clearly, a function $\hat{f}$ can be thought of simply as an N_γ-dimensional vector whose components are restricted to the elements of the set Λ_γ. Hence $\hat{F}$ is of a finite size equal to $|\Lambda_\gamma|^{N_\gamma}$. For any $\hat{f} \in \hat{F}$ define the extension $\hat{f}_{ext} : X \to [-1, 1]$ of $\hat{f}$ to the whole domain X as follows: given $\hat{f}$ (which is well defined on the points $\hat{x}_i$ of the cover) then for every point x in the ball $B_\gamma(\hat{x}_i) = \{x \in X : d(x, \hat{x}_i) \leq \gamma\}$, we let $\hat{f}_{ext}(x) = \hat{f}(\hat{x}_i)$, for all $\hat{x}_i \in C_\gamma$ (where, if, for a point x there is more than one point $\hat{x}_i$ such that $x \in B_\gamma(\hat{x}_i)$, we arbitrarily pick one of the points $\hat{x}_i$ in order to assign the value of $\hat{f}_{ext}(x)$). There is a one-to-one correspondence between $\hat{f}$ and $\hat{f}_{ext}$. Hence the set $\hat{F}_{ext} = \left\{ \hat{f}_{ext} : \hat{f} \in \hat{F} \right\}$ is of cardinality equal to $|\Lambda_\gamma|^{N_\gamma}$.

We claim that for any $f \in \mathcal{F}$ there exists an $\hat{f}_{ext}$ such that $\sup_{x \in X} |f(x) - \hat{f}_{ext}(x)| \leq 3\gamma$. To see that, first for every point $\hat{x}_i \in C_\gamma$ consider the value $f(\hat{x}_i)$ and find a corresponding value in Λ_γ, call it $\hat{f}(\hat{x}_i)$, such that $|f(\hat{x}_i) - \hat{f}(\hat{x}_i)| \leq \gamma$. (That there exists such a value follows by design of Λ_γ). By the above definition of extension, it follows that for all points $x \in B_\gamma(\hat{x}_i)$ we have $\hat{f}_{ext}(x) = \hat{f}(\hat{x}_i)$. Now, from (5) we have for all $f \in \mathcal{F}$,

$$\sup_{x \in B_\gamma(\hat{x}_i)} |f(x) - f(\hat{x}_i)| \leq 2d(x, \hat{x}_i) \leq 2\gamma. \qquad (10)$$

Hence for any $f \in \mathcal{F}$ there exists a function $\hat{f} \in \hat{F}$ with a corresponding $\hat{f}_{ext} \in \hat{F}_{ext}$ such that given an $x \in X$ there exists $\hat{x}_i \in C_\gamma$ such that $|f(x) - \hat{f}_{ext}(x)| = |f(x) - \hat{f}_{ext}(\hat{x}_i)|$. The right hand side can be expressed as

$$\begin{aligned}
|f(x) - \hat{f}_{ext}(\hat{x}_i)| &= |f(x) - \hat{f}(\hat{x}_i)| \\
&= |f(x) - f(\hat{x}_i) + f(\hat{x}_i) - \hat{f}(\hat{x}_i)| \\
&\leq |f(x) - f(\hat{x}_i)| + |f(\hat{x}_i) - \hat{f}(\hat{x}_i)| \\
&\leq 2\gamma + \gamma \qquad (11) \\
&= 3\gamma.
\end{aligned}$$

where (11) follows from (10) and by definition of the grid Λ_γ.

Hence the set $\hat{F}_{ext}$ forms a 3γ-covering of the class $\mathcal{F}$ in the sup-norm over X. Thus we have the following covering number bound (holding uniformly for all m).

Theorem 5.2: With the above notation,

$$\begin{aligned}
\mathcal{N}_\infty(\mathcal{F}, 3\gamma, m) &\leq |\Lambda_\gamma|^{N_\gamma} \qquad (12) \\
&= \left(2 \left\lceil \frac{\mathrm{diam}_D(X)}{\gamma} \right\rceil + 1 \right)^{N_\gamma}. \qquad (13)
\end{aligned}$$

VI. A generalization error bound involving covering numbers of X

Our central result, which follows from Theorem 4.1 and Theorem 5.2, is as follows.

Theorem 6.1: Suppose that X is a finite metric space of diameter $\mathrm{diam}_D(X)$. Suppose P is any probability measure on $Z = X \times \{-1, 1\}$. Let $\delta \in (0, 1)$. For a function

$h : X \to \{-1, 1\}$, let f_h be the corresponding margin (or signed width) function, given by

$$f_h(x) = h(x) w_h(x) = h(x) \text{dist}\left(x, S_{\bar{h}(x)}\right).$$

Then, for any positive integer m, the following holds with P^m-probability at least $1 - \delta$, for a training sample $\mathbf{s} \in Z^m$:

– for any function $h : X \to \{-1, 1\}$,

– for any $\gamma \in (0, \text{diam}_D(X)]$, $P(h(X) \neq Y)$ is at most

$$3\,\hat{P}_{\mathbf{s}}(Y f_h(X) < \gamma) + \epsilon,$$

where

$$\epsilon = \frac{4}{m}\left(N_{\gamma/12} \ln\left(\frac{27\text{diam}_D(X)}{\gamma}\right) + \ln\left(\frac{8\,\text{diam}_D(X)}{\gamma\delta}\right)\right).$$

Here, for any given $\alpha > 0$, $N_\alpha = \mathcal{N}(X, \alpha, d)$ is the α-covering number of X with respect to the metric d on X.

Proof: This follows directly from Theorem 4.1 and Theorem 5.2, together with the observation that, for γ in $(0, \text{diam}_{\text{D}}(X)]$,

$$\begin{aligned}
\mathcal{N}_\infty(\mathcal{F}, \gamma/4, 2m) &\leq \left(2\left\lceil\frac{12\,\text{diam}_D(X)}{\gamma}\right\rceil + 1\right)^{N_{\gamma/12}} \\
&\leq \left(2\left(\frac{12\,\text{diam}_D(X)}{\gamma} + 1\right) + 1\right)^{N_{\gamma/12}} \\
&\leq \left(\frac{27\,\text{diam}_D(X)}{\gamma}\right)^{N_{\gamma/12}}.
\end{aligned}$$

■

In order to use this result, we therefore would need to be able to bound N_γ, and this is the focus of the remainder of the paper.

VII. Using a greedy algorithm to estimate the covering number

We have seen that N_γ, the covering number of X at scale γ, plays a crucial role in our analysis. We now explain how it is possible to obtain a bound on N_γ by using the familiar greedy heuristic for set covering.

We start from the given distance matrix D. Given a fixed margin parameter value $\gamma > 0$ let us define the $N \times N$ $\{0, 1\}$-matrix

$$A_\gamma := [a(i, j)] \tag{14}$$

as follows:

$$a(i, j) := \begin{cases} 1 & \text{if} \quad d(i, j) \leq \gamma \\ 0 & \text{otherwise.} \end{cases}$$

The jth column $a^{(j)}$ of A_γ represents an incidence (binary) vector of a set, or a ball $B_\gamma(j)$, which consists of all the points $i \in X$ that are a distance at most γ from the point j.

We can view A_γ as an adjacency matrix of a graph $G_\gamma = (X, E_\gamma)$, where E_γ is the set of edges corresponding to all adjacent pairs of vertices according to A_γ: there is an edge between any two vertices i, j such that $d(i, j) \leq \gamma$. We note in passing that G_γ can be viewed as an extension (to general metric space) of the notion of a unit disk-graph [11], [17] which is defined in the Euclidean plane.

The problem of finding a minimum γ-cover C_γ for X can be phrased as a classical *set-cover problem* as follows: find a minimal cardinality collection of sets $C_\gamma := \{B_\gamma(j_l) : j_l \in X, 1 \leq l \leq N_\gamma\}$ whose union satisfies $\bigcup_l B_\gamma(j_l) = X$. It is well known [15], [10] that this can be formulated as a linear integer programming problem, as follows: Let the vector $v \in \{0, 1\}^N$ have the following interpretation: $v_i = 1$ if the set $B_\gamma(i)$ is in the cover C_γ and $v_i = 0$ otherwise. Denote by $\mathbf{1}$ the N-dimensional vector of all 1's. Then we wish to find a solution $v \in \{0, 1\}^N$ that minimizes the norm

$$\|v\|_1 = \sum_{j=1}^{N} v_j$$

under the constraints

$$A_\gamma v \geq \mathbf{1}, \quad v \in \{0, 1\}^N.$$

The constraint $A_\gamma v \geq \mathbf{1}$, which is

$$\sum_{j=1}^{N} a(i, j) v_j \geq 1, \quad \text{for every } 1 \leq i \leq N,$$

simply expresses the fact that for every $i \in X$, there must be at least one set $B_\gamma(j)$ that contains it.

It is well known that this problem is NP-complete. However, there is a simple efficient deterministic greedy algorithm (see [10]) which yields a solution — that is, a set cover — of size which is no larger than $(1 + \ln N)$ times the size of the minimal cover. Denote by $\hat{C}_\gamma$ this almost-minimal γ-cover of X and denote by $\hat{N}_\gamma$ its cardinality. Then $\hat{N}_\gamma$ can be used to approximate N_γ up to a $(1 + \ln N)$ accuracy factor:

$$N_\gamma \leq \hat{N}_\gamma \leq N_\gamma(1 + \ln N).$$

VIII. Bounding the covering number in terms of the domination number of a related graph

Next, we relate the problem of bounding N_γ to a graph-theoretical question about some related graphs.

Given a graph $G = (V, E)$ with order (number of vertices) N, let $A(G)$ be its adjacency matrix. Denote by $\deg(x)$ the degree of vertex $x \in V$ and by $\Delta_{min}(G)$, $\Delta_{max}(G)$ the minimum and maximum degrees over all vertices of G.

We now define a quantity we call density, which depends only on X and the distance matrix D.

Definition 8.1: Let $x \in X$. The *γ-density* induced by the distance matrix D at x, denoted $\rho_\gamma(x)$, is the number of points $y \in X$ such that $d(x, y) \leq \gamma$.

The more points in the ball $B_\gamma(x)$, the higher the density value $\rho_\gamma(x)$. Clearly, the degree of x in G_γ satisfies

$$\deg(x) = \rho_\gamma(x). \tag{15}$$

A *dominating* set of vertices $U \subseteq V(G)$ is a set such that for every vertex $v \in V(G) \setminus U$ there exists a vertex $u \in U$ such that u and v are adjacent. The *domination number* $\eta(G)$ is the size of the smallest dominating set of G. (It is usually denoted $\gamma(G)$, but we are using γ to denote widths and margins.) A useful and easy observation is that any dominating set of G_γ is also a γ-cover of X with respect to the distance matrix D (or underlying metric d). For, suppose $U = \{u_1, \ldots, u_k\}$ is a dominating set. Any $u \in U$ is evidently covered by U: there exists an element of U (namely, u itself) whose distance from u is 0 and hence is no more than γ. Furthermore, for $v \in V(G) \backslash U$, since U is a dominating set, there is some $u \in U$ such that u and v are adjacent in G_γ which, by definition of the graph, means that $d(v, u) \leq \gamma$. Hence U indeed serves as a γ-cover of X. This is, in particular, true also for the minimal dominating set of size is $\eta(G_\gamma)$. It follows that the covering number N_γ of X is bounded from above by the domination number of $G = (X, E_\gamma)$. That is,

$$N_\gamma \leq \eta(G_\gamma). \tag{16}$$

There are a number of graph theory results which provide upper bounds for the domination number of a graph in terms of various other graph-theoretic parameters. For instance (though we will not use these here), the domination number can be related to the *algebraic connectivity*, the second-smallest eigenvalue of the Laplacian of the graph [16], and it can also [20] be related to the girth of the graph, the length of the shortest cycle. Other bounds, such as those in [18], [14], involve the order, maximal or minimal degree, or diameter of a graph. We now mention some results which will enable us to bound the covering numbers in terms of a measures of density of the underlying metric space X. First, we have the following result (see [8], [25]):

$$\eta(G) \leq N + 1 - \sqrt{1 + 2\,\text{size}(G)}$$

where size(G) is the number of edges of G, equal to half the sum of the degrees $\sum_{i \in X} \deg(i)$. For G_γ we have $2\,\text{size}(G_\gamma) = \sum_{x \in X} \rho_\gamma(x)$. Let us make the following definition in order to involve quantities explicitly dependent on the metric on X.

Definition 8.2: The average density of X at scale γ (which depends only on the matrix D of distances) is

$$\overline{\rho}_\gamma(D) := \frac{1}{N} \sum_{x \in X} \rho_\gamma(x).$$

Applying this to G_γ, we therefore have

$$N_\gamma \leq \eta(G_\gamma) \leq N + 1 - \sqrt{1 + N\overline{\rho}_\gamma(D)} \tag{17}$$

Any bound on domination number in terms of the number of edges can, in a similar way, be translated into a covering number bound that depends on the average density. Equally, bounds involving the minimum or maximum degrees yield covering number bounds involving minimum or maximum densities. For instance, a bound from [18] upper-bounds $\eta(G)$ by

$$\left\lfloor \frac{1}{N-1} \left(N - \Delta_{max}(G) - 1\right)\left(N - \Delta_{min}(G) - 2\right) \right\rfloor + 2.$$

Letting

$$\rho_{min,\gamma}(D) \quad = \quad \min_{x \in X} \rho_\gamma(x)$$

and

$$\rho_{max,\gamma}(D) \quad = \quad \max_{x \in X} \rho_\gamma(x)$$

then gives as an upper bound on N_γ the quantity $\eta(G_\gamma)$, which is at most:

$$\left\lfloor \frac{1}{N-1} \left(N - \rho_{max,\gamma}(D) - 1\right)\left(N - \rho_{min,\gamma}(D) - 2\right) \right\rfloor + 2. \tag{18}$$

If G_γ has no isolated vertices (which means that each element of X is within distance γ of some other element) then, by a result of [6] (mentioned in [14]),

$$N_\gamma \leq \eta(G_\gamma) \leq N \left(\frac{1 + \ln\left(1 + \rho_{min,\gamma}\right)}{1 + \rho_{min,\gamma}} \right). \tag{19}$$

Note that from (19), the bound on N_γ can be made, for instance, as low as $O(\ln N)$ if D satisfies $\rho_{min,\gamma}(D) = \alpha N$ for $0 < \alpha < 1$.

In [14], it is shown that if G_γ has no cycles of length 4 and if $\rho_{min,\gamma} \geq 2$ then

$$N_\gamma \leq \eta(G_\gamma) \leq \frac{3}{7} \left(N - \frac{\left(3\rho_{min,\gamma} + 1\right)\left(\rho_{min,\gamma} - 2\right)}{6} \right).$$

The paper [14] also mentions some bounds that involve the diameter of the graph (Theorem 4.1-4.8).

We remark that, for a given γ, it is relatively straightforward to determine the average, maximum, and minimum degrees of G_γ by working from its incidence matrix A_γ, which itself is easily computable from the matrix D of metric distances in X.

IX. Conclusions

In this paper, we have considered the generalization error in learning binary functions defined on a finite metric space. Our approach has been to develop bounds that depend on 'sample width', a notion analogous to sample margin when real-valued functions are being used for classification. However, there is no requirement that the classifiers analysed here are derived from real-valued functions. Nor must they belong to some specified, limited, 'hypothesis class'. They can be *any* binary functions on the metric space. We have derived a fairly general bound that depends on the covering numbers of the metric space and we have related this, in turn, through some graph-theroretical considerations, to the 'density' of the metric space. We have also indicated that the covering numbers of the metric space (and hence the generalization error bounds) can be approximated by using a greedy heuristic. The results suggest that if, in learning, a classifier is found that has a large 'sample width' and if the covering numbers of the metric space are small, then good generalization is obtained. An approach based on classical methods involving VC-dimension would not be as useful, since the set of all possible binary functions on a metric space of cardinality N would have VC-dimension equal to N.

Acknowledgment

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

References

[1] M. Anthony and P. L. Bartlett. Function learning from interpolation. *Combinatorics, Probability, and Computing*, 9:213–225, 2000.

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

[3] M. Anthony and J. Ratsaby. Maximal width learning of binary functions. *Theoretical Computer Science*, 411:138–147, 2010.

[4] M. Anthony and J. Ratsaby. Robust cutpoints in the logical analysis of numerical data. *Discrete Applied Mathematics*, 160:355–364, 2012.

[5] M. Anthony and J. Ratsaby. Learning on Finite Metric Spaces. RUTCOR Research Report RRR 19-2012, Rutgers University, June 2012. (Submitted for journal publication.)

[6] V.I. Arnautov. Estimation of the exterior stability number of a graph by means of the minimal degree of the vertices. *Prikl. Mat. i Programmirovanic Vyp.*, 126 11: 3–8, 1974. (In Russian.)

[7] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[8] C. Berge. *Graphes et Hypergraphes*. Dunod, Paris, 1970.

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[10] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):pp. 233–235, 1979.

[11] B. Clark, C. Colbourn, and D. Johnson. Unit disk graphs. *Discrete Mathematics*, 86(1-3):165 – 177, 1990.

[12] R. M. Dudley (1999). *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, 63, Cambridge University Press, Cambridge, UK.

[13] U. Grenander. *Abstract Inference*. Wiley, 1981.

[14] B. Kupper and L. Volkmann. Upper bounds on the domination number of a graph in terms of order, diameter and minimum degree. *Australasian Journal of Combinatorics*, 35:133–140, 2006.

[15] L. Lovász. On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13(4):383 – 390, 1975.

[16] M. Lu, H. Liu and F. Tian. Lower bounds of the Laplacian spectrum of graphs based on diameter. *Linear Algebra and its Applications*, 402(0): 390–396, 2005.

[17] M. V. Marathe, H. Breu, H. B. Hunt, S. S. Ravi, and D. J. Rosenkrantz. Simple heuristics for unit disk graphs. *Networks*, 25(2):59–68, 1995.

[18] D. Marcu. An upper bound on the domination number of a graph. *Math. Scand.*, 59:41–44, 1986.

[19] D. Pollard (1984). *Convergence of Stochastic Processes*. Springer-Verlag.

[20] D. Rautenbach. A note on domination, girth and minimum degree. *Discrete Applied Mathematics*, 308:2325–2329, 2008.

[21] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5), 1996: 1926–1940.

[22] A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans. *Advances in Large-Margin Classifiers (Neural Information Processing)*. MIT Press, 2000.

[23] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[24] V.N. Vapnik and A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**(2): 264–280, 1971.

[25] V.G. Vizing. An estimate of the external stability number of a graph. *Dokl. Akad. Nauk. SSSR* 164: 729–731, 1965.