Theoretical Computer Science 529 (2014) 2–10

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Learning bounds via sample width for classifiers on finite metric spaces

Martin Anthony [a], Joel Ratsaby [b,*]

[a] Department of Mathematics, London School of Economics and Political Science, Houghton Street, London WC2A2AE, UK
[b] Electrical and Electronics Engineering Department, Ariel University of Samaria, Ariel 40700, Israel

ARTICLE INFO

Article history:
Received 21 June 2012
Received in revised form 5 February 2013
Accepted 11 July 2013
Communicated by J.N. Kok

Keywords:
Generalization error
Machine learning
Learning algorithms

ABSTRACT

In a recent paper [M. Anthony, J. Ratsaby, Maximal width learning of binary functions, Theoretical Computer Science 411 (2010) 138–147] the notion of *sample width* for binary classifiers mapping from the real line was introduced, and it was shown that the performance of such classifiers could be quantified in terms of this quantity. This paper considers how to generalize the notion of sample width so that we can apply it where the classifiers map from some finite metric space. By relating the learning problem to one involving the domination numbers of certain graphs, we obtain generalization error bounds that depend on the sample width and on certain measures of 'density' of the underlying metric space. We also discuss how to employ a greedy set-covering heuristic to bound generalization error.

1. Introduction

1.1. Overview

In [3], the notion of *sample width* for binary classifiers mapping from the real line was introduced, and in [4,5], related ideas were developed to explore the performance of hybrid classifiers based on unions of boxes and a nearest-neighbor paradigm. 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). Precise details are given below, but the idea is to define sample width to be at least γ if the classifier achieves the correct classifications on the sample and if, in addition, for each sample point, the minimum distance to a point of the domain having opposite classification is at least γ. We then relate the learning problem in this context to that of learning with a large margin. In order to obtain bounds on classifier accuracy, we consider the domination numbers of graphs associated with the underlying metric space and, using some previous 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. We also discuss how to employ the well-known greedy set-covering heuristic to bound generalization error.

1.2. The underlying metric space and the width of a classifier

Let $\mathcal{X} = [N] := \{1, 2, \ldots, N\}$ be a finite set on which is defined a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $D = [d(i, j)]$ be the corresponding 'distance matrix'. Then D is symmetric with (i, j)th element $d(i, j) \geqslant 0$, and with $d(i, j) = 0$ if and only if $i = j$.

* Corresponding author.
E-mail addresses: m.anthony@lse.ac.uk (M. Anthony), ratsaby@ariel.ac.il (J. Ratsaby).

0304-3975/$ – see front matter
http://dx.doi.org/10.1016/j.tcs.2013.07.004

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

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

We define the *diameter* of $\mathcal{X}$ as follows:

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

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

By a binary function on $\mathcal{X}$, we mean a mapping $h : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y} = \{-1, +1\}$. We will denote by $\mathcal{H}$ the class of all binary functions h on $\mathcal{X}$.

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*

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

We say that $h : \mathbb{R} \to \{-1, 1\}$ achieves sample margin at least γ on ξ 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 is then possible to quantify (in a probabilistic model of learning) the accuracy of learning in terms of the sample width. (More precisely, generalization error bounds are derived that involve the sample margin, within a version of the PAC model of learning. More detail about probabilistic modelling of learning is given in Section 2.)

In this paper we use an analogous notion of width to analyze 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 $h : \mathcal{X} \to \{-1, 1\}$ which are defined as follows:

$$S_-^h := \big\{x \in \mathcal{X} \colon h(x) = -1\big\}, \qquad S_+^h := \big\{x \in \mathcal{X} \colon h(x) = +1\big\}. \tag{1}$$

We will often omit the superscript h. In [4–6] we analyzed learning that was based on a class of real-valued functions defined as the difference between the distances of a point x from two non-overlapping subsets S_+, S_- of $\mathcal{X}$: of particular interest was the case in which S_+ and S_- are each unions of boxes (labeled 1 and -1, respectively), where the union $S_+ \cup S_-$ need not cover the domain. Here, we define the width in a slightly different way by starting with a given binary function h (rather than with two arbitrary non-overlapping sets). Given such a binary function h we define the *width* $w_h(x)$ of h at a point $x \in \mathcal{X}$ 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 $h(x) = -1$):

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

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:

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

This is commonly also referred to as the functional *margin* of h at 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.

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

$$\mathcal{F} := \big\{f_h(x) \colon h \in \mathcal{H}\big\}. \tag{2}$$

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

2. Measuring the accuracy of learning

2.1. Probabilistic modelling of learning

We work in the framework of the popular 'PAC' model of computational learning theory (see [23,10]). This model assumes that the labeled examples (x_i, y_i) in the training sample ξ have been generated randomly according to some fixed (but unknown) probability distribution P on $Z = \mathcal{X} \times \mathcal{Y}$. (This includes, as a special case, the situation in which each x_i is drawn according to a fixed distribution on $\mathcal{X}$ and is then labeled deterministically by $y_i = t(x_i)$ where t is some fixed function.) Thus, a sample ξ of length m can be regarded as being drawn randomly according to the product probability