Large width nearest prototype classification on general distance spaces

Martin Anthony[1] and Joel Ratsaby[2]

[1]Department of Mathematics, London School of Economics and Political Science, Houghton Street, London WC2A2AE, U.K.

[2]Department of Electrical and Electronics Engineering, Ariel University of Samaria, ARIEL 40700, ISRAEL

Abstract

In this paper we consider the problem of learning nearest-prototype classifiers in any finite distance space; that is, in any finite set equipped with a distance function. An important advantage of a distance space over a metric space is that the triangle inequality need not be satisfied, which makes our results potentially very useful in practice. We consider a family of binary classifiers for learning nearest-prototype classification on distance spaces, building on the concept of large-width learning which we introduced and studied in earlier works. Nearest-prototype is a more general version of the ubiquitous nearest-neighbor classifier: a prototype may or may not be a sample point. One advantage in the approach taken in this paper is that the error bounds depend on a 'width' parameter, which can be sample-dependent and thereby yield a tighter bound.

1 Introduction

Learning Vector Quantization (LVQ) and its various extensions introduced by Kohonen [21] are used successfully in many machine learning tools and applications. Learning pattern classification by LVQ is based on adapting a fixed set of labeled prototypes in Euclidean space and using the resulting set of prototypes in a nearest-prototype rule (winner-take-all) to classify any point in the input space. As [20] mentions, LVQ fails if Euclidean representation is not well-suited for the data; and there have been extensions of LVQ to try to allow different metrics [20, 24] and take advantage of samples for which a more confident (or a large margin) classification can be obtained. Generalization error bounds with dependence on this sample margin are stated in [20, 24] for learning over Euclidean spaces and, as is usually the case for large-margin learning [1], the error bounds are tighter than ones with no sample-margin dependence. The results of such work are important as they explain why LVQ works well in practice in Euclidean metric spaces.

There are learning domains in which it is difficult to formalize quantitative features that are encoded as numerical variables which together constitute a vector space (usually Euclidean) to discriminate between objects that belong to different classes [22]. Learning over such domains requires a qualitative approach which tries to describe the structure of objects in a way that is similar to how humans do, for instance, in terms of morphological elements of objects. Objects are then represented not by numerical vectors but by other means such as strings of symbols which can be compared using a dissimilarity (or distance) function. This approach is much more flexible than the one based on numerical features since there are many existing distance functions [18] and new ones can be defined easily for any kind of objects, for instance, bioinformatic sequences, graphs, images, etc., and they do not have to satisfy the requirements of a metric. However, most learning algorithms, in particular neural networks which have been very succesfull recently, require a Euclidean, or more generally vector spaces, that are represented by numerical features. Such problem domains are potential applications of prototype-based learning over non-Euclidean, or more generally, non-metric spaces.

In [3] we studied learning binary classification with nearest-prototype classifiers over metric spaces and obtained sample-dependent error bounds. In the current paper we consider learning binary classification on finite distance spaces; that is, finite sets equipped with a distance function (often called 'dissimilarity measure' [18]) where the classifiers (which we call nearest-prototype classifiers) are generalizations of the well known nearest neighbor classifier [14]. An important advantage of a distance space over a metric space is that the triangle inequality need not be satisfied, which makes our results potentially very useful in practice. Our definition of distance function is quite loose in that it does *not* need to satisfy any of the non-negativity, symmetry or reflexivity properties of a proper distance function [18]. We still call it a distance because, as far as we can expect in applying our learning results, any useful space has at least the non-negativity property and so we will assume in the paper that the distance function satisfies the non-negativity property.

We consider a family of binary classifiers for learning nearest-prototype classification on distance spaces, building on the concept of large-width learning which was introduced in [4] and expanded in various classification settings [5–10]. The advantage in this approach is that the error bounds depend on the 'width' parameter which can be sample-dependent and thereby yield a tighter error bound.

We define a width function which measures the difference between the distance from a test point x (to be classified) to its nearest negative prototype and the distance to its nearest positive prototype. The classifier's decision is defined as the sign of this difference. The set of prototypes from which these two nearest ones are obtained is very general in that it can be *any* set of points in the distance space. In particular, it can be a subset of the sample and can be determined via *any* algorithm. The error bounds that we state in the current paper apply regardless of the algorithm that is used to determine these prototypes. The fact that we deal with a distance space, rather than a metric space, means that the triangle inequality need not be satisfied. The error bound depends on quantities that can be evaluated directly from a matrix which consists of the half-space functions over the distance space. This matrix is a function of the distance matrix of the space and hence it can be computed efficiently using massively parallel processing techniques, for instance, as in [11, 12].

Also, as mentioned above, our use of the concept of distance is loose, so that, for instance,

not only that the triangle inequality need not be satisfied, but also none of the three standard properties of a distance need to be satisfied either.

2 Setup

2.1 Nearest-prototype classifiers

For a positive integer n, let $[n] := \{1, 2, \ldots, n\}$. We consider a finite set $\mathcal{X} := \{x_1, \ldots, x_N\}$ with a binary set $\mathcal{Y} = \{-1, 1\}$ of possible classifications. Let d, a 'distance function', be a function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. Let us assume that d is normalized such that

$$\operatorname{diam}(\mathcal{X}) := \max_{1 \le i,j \le N} d(x_i, x_j) = 1.$$

A *prototype* $p_i \in \mathcal{X}$ is a point in the distance space that has an associated label $\sigma_i \in \mathcal{Y}$. We denote by

$$p_i^+,\ p_j^-$$

a prototype with label $\sigma_i = 1$ and a prototype with label $\sigma_j = -1$, respectively. When the label of a prototype is not explicitly mentioned we write p_i. Given a fixed integer $n \ge 2$, we consider learning the family of classifiers that are defined by the nearest-neighbor rule defined by a set of n prototypes. We refer to any such classifier by $h_{R,\sigma}$ and it is defined by an ordered set of prototypes $R = \{p_i\}_{i=1}^n$ and their corresponding label vector $\sigma^T := [\sigma_1, \ldots, \sigma_n]$, $\sigma_i \in \mathcal{Y}$, $1 \le i \le n$. The order of the set R can be defined based on any ordering of $\mathcal{X}$ and it allows us to fix σ in advance and consider all classifiers obtained by all ordered sets R. We define $h_{R,\sigma}$ to be a *nearest-prototype classifier* as follows: let

$$N_+ = N_+(\sigma) := \{i \in [n] :\ \sigma_i = 1\},\ N_- = N_-(\sigma) := \{i \in [n] :\ \sigma_i = -1\}.$$

Then, given x,

$$h_{R,\sigma}(x) = \begin{cases} -1 & \text{if } \operatorname{argmin}_{1 \le i \le n} d(x, p_i) \in N_- \\ 1 & \text{otherwise.} \end{cases} \tag{2.1}$$

Let us denote by ξ a sample of m labeled examples

$$\xi := \{(X_i, Y_i)\}_{i=1}^m. \tag{2.2}$$

In the current paper, a prototype p_i may be any point in $\mathcal{X}$; in particular, one that depends on the sample ξ directly or via some learning algorithm. For instance, if $(R, \sigma) = \xi$, then $h_{R,\sigma}$ is the well known nearest-neighbor classifier [14]. If the labeled prototype set is only a subset of the sample, $(R, \sigma) \subseteq \xi$, then $h_{R,\sigma}$ belongs to a family of the so-called 'edited nearest-neighbor' classifiers [17]. The prototype set (R, σ) may not necessarily be a subset of the sample ξ, but could be derived from it via some adaptive procedure such as the LVQ algorithm [21], in which case $h_{R,\sigma}$ would be the LVQ classifier. In the current paper, any such classifier is referred to as a nearest-prototype classifier and, as mentioned above, is denoted by $h_{R,\sigma}$.

2.2 Probabilistic modeling of learning

We work in the framework of the popular 'PAC' model of computational learning theory (see [13, 27]). 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 (2.2) of length m can be regarded as being drawn randomly according to the product probability distribution P^m. In general, suppose that H is a set of functions from $\mathcal{X}$ 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) which can be expressed as

$$\mathrm{er}_P(h) = P(h(X) \neq Y) = P\left(Yh(X) < 0\right). \tag{2.3}$$

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

$$\mathrm{er}_\xi(h) = \frac{1}{m}|\{i : h(X_i) \neq Y_i\}|$$

(the proportion of points in the sample incorrectly classified by h). Much classical work in learning theory (see [13, 27], 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 $\mathrm{er}_P(h) < \mathrm{er}_\xi(h) + \epsilon(m, \delta)$, 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 [19, 23, 28], in which case $\epsilon(m, \delta)$ would typically involve a quantity known as the growth function of the set of classifiers [1, 13, 27, 28]. More recently, emphasis has been placed on 'learning with a large margin'. (See, for instance [1, 2, 25, 26].) The rationale behind margin-based, or width-based generalization error bounds is that if a classifier 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). A more direct approach which does not require real-valued functions as a basis for classification margin, uses the concept of width (introduced in [4]) and studied in various settings in [5–10].

2.3 Width and error of a classifier

We define the *width* of $h_{R,\sigma}$ at a point $x \in \mathcal{X}$ as follows:

$$w_{h_{R,\sigma}}(x) := \min_{1 \leq i \leq n:\, \sigma_i \neq h(x)} d\left(x, p_i\right) - \min_{1 \leq i \leq n:\, \sigma_i = h(x)} d\left(x, p_i\right). \tag{2.4}$$

In words, the width of h at x is the difference between the distance to the nearest-unlike-prototype of x and the distance to the nearest-prototype to x where unlike means of a different sign than x. In [10] we consider binary classifiers that are based on a pair of oppositely labeled prototypes and use this definition of width, which, in this case, becomes simply $d(x, p_-) - d(x, p_+)$.

The signed width (or margin) function corresponding to (2.4) is defined as

$$f_{R,\sigma}(x) := f_{h_{R,\sigma}}(x) = h_{R,\sigma}(x) w_{h_{R,\sigma}}(x). \tag{2.5}$$

Note that this definition means that for x equidistant from two oppositely labeled prototypes p, $q \in R$ that are each the closest to x from all other prototypes of the same label, the value of the margin $f_{R,\sigma}(x)$ at this x is zero. This definition is intuitive and actually makes the analysis simpler compared to an alternative definition of width [7].

This definition of width is an application of a more general definition of width, introduced in [5, 6], which takes the form $f(x) = d(x, S_-) - d(x, S_+)$, where S_- and S_+ are any disjoint subsets of the input space that are labeled -1 and 1, respectively. (In [7–9], a slightly different definition of width was used where the union of the disjoint sets S_- and S_+ equals the input space.)

In [3] we considered binary classifiers which are also based on prototypes where the decision is not based on the nearest-prototype but is based on the combined influence of several prototypes based on certain regions of influence. The present notion of width was not explicitly utilized there.

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

$$\hat{P}_m(Y f_{R,\sigma}(X) \le \gamma) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}\left(Y_j f_{R,\sigma}(X_j) \le \gamma\right).$$

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

Define the function

$$\operatorname{sgn}(a) := \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a \le 0. \end{cases}$$

For the purpose of bounding the generalization error it is convenient to express the classification $h(X)$ in terms of the signed width as follows,

$$h_{R,\sigma}(X) = \operatorname{sgn}\left(f_{R,\sigma}(X)\right).$$

Therefore the generalization error $\mathrm{er}_P(h_{R,\sigma})$ can be bounded as follows

$$\begin{aligned} \mathrm{er}_P(h_{R,\sigma}) &= P\left(h_{R,\sigma}(X) \ne Y\right) && (2.6) \\ &= P\left(Y f_{R,\sigma}(X) < 0\right) + P\left(Y = 1, f_{R,\sigma}(X) = 0\right) \\ &\le P\left(Y f_{R,\sigma}(X) \le 0\right). && (2.7) \end{aligned}$$

Our aim is to show that the upper bound (2.7) on the generalization misclassification error is not much greater than $\hat{P}_m\left(Y f_{R,\sigma}(X) \le \gamma\right)$. Explicitly, we aim for bounds of the form: for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $\gamma \in (0, \mathrm{diam}(\mathcal{X})]$, we have

$$\mathrm{er}_P(h_{R,\sigma}) \le \hat{P}_m(Y f_{R,\sigma}(X) \le \gamma) + \epsilon(m, \gamma, \delta).$$

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.

The advantage of working with the notion of width is that it is possible to have such a uniform bound over a very large family of classifiers. For instance, in [7] we obtained such bounds for learning the family of *all* possible binary classifiers on any finite metric space and in [8] we did the same for multi-category classifiers over infinite metric spaces. As mentioned above, in the current work, we consider particular kinds of classifiers $h_{R,\sigma}$ that are defined on the nearest-prototype rule based on a fixed number n of prototypes. Thus we expect that the bound that we obtain is tighter than the one in [7], which holds for the family of all binary classifiers.

To obtain a uniform bound, we are interested in showing that the probability of the 'bad event' — namely, that there exists some value of γ and some classifier $h_{R,\sigma}$ such that the generalization error is *not* bounded from above by some small deviation ϵ from the empirical γ-margin error — is small. That is, we aim to bound the following failure probability:

$$P^m_{X,Y}\left(\left\{\xi : \exists\gamma,\, \exists\sigma,\, \exists R,\, P\left(Yf_{R,\sigma}(X) \le 0\right) > \frac{1}{m}\sum_{j=1}^{m}\mathbb{I}\left\{Y_j f_{R,\sigma}(X_j) \le \gamma\right\} + \epsilon\right\}\right). \tag{2.8}$$

This can be expressed as follows:

$$P^m_{X,Y}\left(\left\{\xi : \exists\gamma,\, \exists\sigma,\, \exists R,\, P\left(Yf_{R,\sigma}(X) > 0\right) < \frac{1}{m}\sum_{j=1}^{m}\mathbb{I}\left\{Y_j f_{R,\sigma}(X_j) > \gamma\right\} - \epsilon\right\}\right). \tag{2.9}$$

Let us fix γ for now, and deal with bounding the probability

$$P^m_{X,Y}\left(\left\{\xi : \exists\sigma,\, \exists R,\, P\left(Yf_{R,\sigma}(X) > 0\right) < \frac{1}{m}\sum_{j=1}^{m}\mathbb{I}\left\{y_j f_{R,\sigma}(x_j) > \gamma\right\} - \epsilon\right\}\right). \tag{2.10}$$

3 Towards bounding the probability

3.1 Representing the bad event by related sets

Define the set $M_{R,\sigma,\gamma} \subset \mathcal{X} \times \mathcal{Y}$ as follows,

$$M_{R,\sigma,\gamma} := \{(x,y) : yf_{R,\sigma}(x) > \gamma\}$$

and let

$$M^+_{R,\sigma,\gamma} := \{(x,1) : f_{R,\sigma}(x) > \gamma\} \tag{3.1}$$

$$M^-_{R,\sigma,\gamma} := \{(x,-1) : f_{R,\sigma}(x) < -\gamma\}. \tag{3.2}$$

Note that

$$\begin{aligned} M_{R,\sigma,\gamma} &= \left\{M_{R,\sigma,\gamma} \bigcap \{(x,y) : y = 1\}\right\} \bigcup \left\{M_{R,\sigma,\gamma} \bigcap \{(x,y) : y = -1\}\right\} \\ &= M^{+}_{R,\sigma,\gamma} \bigcup M^{-}_{R,\sigma,\gamma}. \end{aligned} \tag{3.3}$$

We can write

$$P\left(Y f_{R,\sigma}(X) > \gamma\right) = P\left(M_{R,\sigma,\gamma}\right)$$

and

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\left\{Y_j f_{R,\sigma}(X_j) > \gamma\right\} = P_m\left(M_{R,\sigma,\gamma}\right),$$

where P_m denotes the empirical measure based on a sample of length m. Thus (2.9) is expressed as

$$P^m_{X,Y}\left(\{\xi : \exists\gamma,\, \exists\sigma,\, \exists R,\, P\left(M_{R,\sigma,0}\right) < P_m\left(M_{R,\sigma,\gamma}\right) - \epsilon\}\right).$$

Let $\epsilon(\gamma)$ be any function depending on γ, in a way to be specified later and define the set $E(\gamma) \subseteq (\mathcal{X} \times \mathcal{Y})^m$ as

$$E(\gamma) := \{\xi : \exists\sigma\exists R,\, P\left(M_{R,\sigma,0}\right) < P_m\left(M_{R,\sigma,\gamma}\right) - \epsilon(\gamma)\}. \tag{3.4}$$

Then substituting $\epsilon(\gamma)$ for ϵ in (2.10) implies that (2.10) equals the probability $P^m_{X,Y}\left(E(\gamma)\right)$. It follows that (2.8) equals

$$P^m_{X,Y}\left(\bigcup_{\gamma\in(0,\mathrm{diam}(\mathcal{X})]} E(\gamma)\right). \tag{3.5}$$

Let L be an integer (to be specified in a section further below). For integer $0 \le l \le L+1$, let γ_l be a decreasing sequence such that the following conditions hold:

1. $0 \le \gamma_l \le 1$
2. $\gamma_0 = 1$, $\gamma_{L+1} = 0$.

Denote by

$$C := \sum_{l=1}^{L} \gamma_l.$$

While all the above quantities L, γ_l and C may depend on $\mathcal{X}$, we keep this dependence implicit in the notation.

Define $\Gamma_l := (\gamma_l, \gamma_{l-1}]$ for $1 \le l \le L+1$. Then (3.5) equals

$$P^m_{X,Y}\left(\bigcup_{l=1}^{L+1}\bigcup_{\gamma\in\Gamma_l} E(\gamma)\right) \le \sum_{l=1}^{L+1} P^m_{X,Y}\left(\bigcup_{\gamma\in\Gamma_l} E(\gamma)\right). \tag{3.6}$$

Define the set $E_l \subseteq (\mathcal{X} \times \mathcal{Y})^m$ as

$$E_l := \{\xi : \exists\sigma\exists R,\, P\left(M_{R,\sigma,\gamma_l}\right) < P_m\left(M_{R,\sigma,\gamma_l}\right) - \epsilon(\gamma_{l-1})\}.$$

Henceforth, assume that $\epsilon(\gamma)$ is a non-increasing function over each interval Γ_l.

Proposition 1. *For any $\gamma \in \Gamma_l$, $E(\gamma) \subseteq E_l$.*

Proof. We have $M_{R,\sigma,0} \supseteq M_{R,\sigma,\gamma_l}$; thus $P\left(M_{R,\sigma,0}\right) \geq P\left(M_{R,\sigma,\gamma_l}\right)$. And $M_{R,\sigma,\gamma_l} \supseteq M_{R,\sigma,\gamma}$ since $\gamma_l \leq \gamma$. Therefore $P_m\left(M_{R,\sigma,\gamma}\right) \leq P_m\left(M_{R,\sigma,\gamma_l}\right)$. For $\gamma \leq \gamma_{l-1}$, by the above assumption on ϵ, $\epsilon(\gamma) \geq \epsilon(\gamma_{l-1})$. It follows that $E(\gamma) \subseteq E_l$. □

The event that there exists σ and R such that $P\left(M_{R,\sigma,\gamma_l}\right) < P_m\left(M_{R,\sigma,\gamma_l}\right) - \epsilon(\gamma_{l-1})$ holds, together with (3.3), implies that either of the following events occurs: there exists σ and R such that

$$P\left(M^+_{R,\sigma,\gamma_l}\right) < P_m\left(M^+_{R,\sigma,\gamma_l}\right) - \epsilon(\gamma_{l-1})/2$$

or there exists a σ and R such that

$$P\left(M^-_{R,\sigma,\gamma_l}\right) < P_m\left(M^-_{R,\sigma,\gamma_l}\right) - \epsilon(\gamma_{l-1})/2.$$

Let

$$E_l^+ := \left\{\xi : \exists\sigma\exists R,\, P\left(M^+_{R,\sigma,\gamma_l}\right) < P_m\left(M^+_{R,\sigma,\gamma_l}\right) - \epsilon(\gamma_{l-1})/2\right\} \tag{3.7}$$

and

$$E_l^- := \left\{\xi : \exists\sigma\exists R,\, P\left(M^-_{R,\sigma,\gamma_l}\right) < P_m\left(M^-_{R,\sigma,\gamma_l}\right) - \epsilon(\gamma_{l-1})/2\right\}. \tag{3.8}$$

Then

$$P^m_{X,Y}\left(E_l\right) \leq P^m_{X,Y}\left(E_l^+\right) + P^m_{X,Y}\left(E_l^-\right).$$

3.2 Bounding the probability in terms of the growth function

We now aim to bound from above the first probability $P^m_{X,Y}\left(E_l^+\right)$. We briefly first recall the definitions of growth function and VC-dimension [28]. Suppose that $\mathcal{C}$ is a collection of subsets of a set Z. Let S be any (finite) subset of Z. Then a *dichotomy* of S by $\mathcal{C}$ is a set of the form $S \cap C$ where $C \in \mathcal{C}$. We denote the number of dichotomies of S by $\mathcal{C}$ as $\#(\mathcal{C}; S)$. Thus,

$$\#(\mathcal{C}; S) = \left|\{S \cap C : C \in \mathcal{C}\}\right|.$$

Then the *growth function* of $\mathcal{C}$ is the function $\Pi_{\mathcal{C}} : \mathbb{N} \to \mathbb{N}$ defined as follows: for $m \in \mathbb{N}$,

$$\Pi_{\mathcal{C}}(m) = \max\{\#(\mathcal{C}; S) : S \subseteq Z, |S| = m\}.$$

The *VC-dimension* of $\mathcal{C}$ is (infinity, or) the largest value of m such that $\Pi_{\mathcal{C}}(m) = 2^m$. (A set S of size m such that $\#(\mathcal{C}; S) = 2^m$ is said to be *shattered* by $\mathcal{C}$.)

Define the classes

$$\mathcal{M}^+_{\gamma_l} := \left\{M^+_{R,\sigma,\gamma_l} : \sigma \in \mathcal{Y}^n,\, R \subset \mathcal{X}\right\},\ \mathcal{M}^-_{\gamma_l} := \left\{M^-_{R,\sigma,\gamma_l} : \sigma \in \mathcal{Y}^n,\, R \subset \mathcal{X}\right\}.$$

Denote by $\Pi_{\mathcal{M}^+_{\gamma_l}}(m)$ the growth function of the class $\mathcal{M}^+_{\gamma_l}$. By [28] (see also Theorem 4.3 of [1]), it follows that

$$P^m_{X,Y}\left(E_l^+\right) \;\leq\; 4\Pi_{\mathcal{M}^+_{\gamma_l}}(2m)\exp\left(-m\epsilon^2(\gamma_{l-1})/32\right)$$

and

$$P^m_{X,Y}\left(E^-_l\right) \leq 4\Pi_{\mathcal{M}^-_{\gamma_l}}(2m)\exp\left(-m\epsilon^2(\gamma_{l-1})/32\right).$$

Define $G(m,\gamma)$ to be an upper bound on $\ln\left(\Pi_{\mathcal{M}^+_\gamma}(2m)\right)$ and $\ln\left(\Pi_{\mathcal{M}^-_\gamma}(2m)\right)$, to be specified later, such that choosing

$$\epsilon(\gamma) := \sqrt{\frac{32}{m}\left(G(m,\gamma)+\ln\left(\frac{8(C+1)}{\gamma\delta}\right)\right)} \tag{3.9}$$

makes the inequality $\epsilon(\gamma) \geq \epsilon(\gamma_{l-1})$ hold for all $\gamma \in \Gamma_l$, as required for Proposition 1 and for the definition of $\epsilon(\gamma)$ in (3.4). Then substituting γ_{l-1} for γ in (3.9) and letting (3.9) be the choice for $\epsilon(\gamma_{l-1})$ in (3.7), it follows from Theorems 3.7, 4.3 of [1] that both $P\left(E^+_l\right)$ and $P(E^-_l)$ are bounded from above by $\gamma_{l-1}\delta/2(C+1)$. From Proposition 1, it follows that (3.6) is bounded as follows:

$$\begin{aligned}
\sum_{l=1}^{L+1} P^m_{X,Y}\left(\bigcup_{\gamma\in\Gamma_l} E(\gamma)\right) &\leq \sum_{l=1}^{L+1} P^m_{X,Y}(E_l) && (3.10)\\
&\leq \sum_{l=1}^{L+1} P^m_{X,Y}\left(E^+_l\right) + \sum_{l=1}^{L+1} P^m_{X,Y}\left(E^-_l\right) && (3.11)\\
&\leq 2\left(\frac{\delta}{2(C+1)}\right)\sum_{l=1}^{L+1}\gamma_{l-1}\\
&= \frac{\delta}{C+1}\left(\sum_{l=1}^{L+1}\gamma_{l-1}\right)\\
&= \frac{\delta}{C+1}\left(\sum_{l=0}^{L}\gamma_l\right)\\
&= \frac{\delta}{C+1}\left(\sum_{l=1}^{L}\gamma_l + 1\right)\\
&= \delta. && (3.12)
\end{aligned}$$

In the next section, we derive a value of $G(m,\gamma)$ which bounds from above the logarithm of the growth functions of $\mathcal{M}^+_\gamma$ and $\mathcal{M}^-_\gamma$.

4 Bounding the growth function

In this section we bound the growth functions of the classes $\mathcal{M}^+_\gamma$ and $\mathcal{M}^-_\gamma$.

4.1 Half-spaces of $\mathcal{X}$

Define $ind(p_i) \in [N]$ to be the index of a point x such that $p_i = x_{ind(p_i)}$, that is, $ind(p_i)$ is the index of a prototype with respect to the pre-determined ordering of the distance space $\mathcal{X}$. Clearly, each p_i has a unique $ind(p_i)$ value.

For any i, $j \in [N]$, define an *affined half-space* set as follows:

$$W_\gamma^{(i,j)} := \{x : d(x, x_j) - d(x, x_i) > \gamma\}. \tag{4.1}$$

Let the class of such sets be

$$\mathcal{W}_\gamma := \left\{W_\gamma^{(i,j)} : 1 \leq i, j \leq N, j \neq i\right\}.$$

4.2 Matrix representation

Recall that $\mathcal{X} = \{x_1, \ldots, x_N\}$. Since a prototype may be any point in $\mathcal{X}$ then, in general, for any pair of prototypes p, $q \in \mathcal{X}$ there is some $1 \leq i \neq j \leq N$, such that $p = x_i$, $q = x_j$. Write $W_0^{(p,q)} := W_0^{(i,j)}$. Then $W_0^{(i,j)}$ corresponds to the positive elements of the following vector:

$$F_j^{(i)} := \begin{bmatrix} d(x_1, x_j) - d(x_1, x_i) \\ \vdots \\ d(x_N, x_j) - d(x_N, x_i) \end{bmatrix}. \tag{4.2}$$

Note that taking the sign of a vector $F_j^{(i)}$ yields a partition of $\mathcal{X}$ into two parts, referred to as half-spaces. For a real vector $v \in \mathbb{R}^N$ define $\operatorname{sgn}(v) = [\operatorname{sgn}(v_1), \ldots, \operatorname{sgn}(v_N)]$. Hence $\operatorname{sgn}(F_j^{(i)})$ corresponds to a half-space on $\mathcal{X}$.

Fix any point $x_i \in \mathcal{X}$ and let the $N \times (N-1)$ matrix F_i be defined by

$$F^{(i)} = \left[F_1^{(i)}, \ldots, F_{i-1}^{(i)}, F_{i+1}^{(i)}, \ldots, F_N^{(i)}\right] \tag{4.3}$$

where the j^{th} column is $F_j^{(i)}$, $j \neq i$, $1 \leq j \leq N$.

Define the $N \times N(N-1)$ matrix

$$F := \left[F^{(1)}, \ldots, F^{(N)}\right]. \tag{4.4}$$

The binary matrix

$$\operatorname{sgn}(F) := \left[\operatorname{sgn}(F^{(1)}), \ldots, \operatorname{sgn}(F^{(N)})\right]$$

where

$$\operatorname{sgn}(F^{(i)}) := \left[\operatorname{sgn}(F_1^{(i)}), \ldots, \operatorname{sgn}(F_N^{(i)})\right],$$

represents the class of all half-spaces on $\mathcal{X}$.

4.3 Thresholding by γ

The set $W_0^{(i,j)}$ corresponds to some column of the matrix F. We now define a more general matrix whose columns corresponds to the sets $W_\gamma^{(i,j)}$ defined in (4.1), for any fixed $\gamma > 0$. Bounding the VC dimension of this matrix means that we obtain a bound on the VC

dimension of the class $\mathcal{W}_\gamma$. (By the VC-dimension of a binary matrix, we mean that of the set system in which the indicator functions of the sets correspond to the columns of a matrix, with a 1-entry denoting inclusion in the set.)

For any $1 \leq i \neq j \leq N$, a set $W_\gamma^{(i,j)}$ corresponds to the positive elements of the vector $F_j^{(i)} - \gamma\mathbf{1}$ where $\mathbf{1}$ is an $N \times 1$ vector of all ones. Denote by J an $N \times N(N-1)$ matrix of all ones. For any $\gamma > 0$, let us consider the $N \times N(N-1)$ matrix

$$F_\gamma := F - \gamma J = \left[F_2^{(1)} - \gamma\mathbf{1}, \ldots, F_{N-1}^{(N)} - \gamma\mathbf{1}\right]. \tag{4.5}$$

The matrix F_γ corresponds to the class $\mathcal{W}_\gamma$ of sets, where for column $F_j^{(i)} - \gamma\mathbf{1}$, the positive elements of the vector correspond to the elements of the set $W_\gamma^{(i,j)}$.

The binary matrix $\operatorname{sgn}(F_\gamma)$ corresponds to a class of 'affined' half-spaces on $\mathcal{X}$ (the columns of $\operatorname{sgn}(F_\gamma)$). We now choose the constant L of section 3 to be the number of distinct positive entries of F, denoted as $0 < a_L < a_{L-1} < \cdots < a_1 < 1$ where a_1 and a_L are the maximum and minimum positive entries of F, respectively. For $1 \leq i \leq L$, define the multiplicity of a_i, denoted by m_i, $1 \leq i \leq L$, as the number of times that a_i appears in F. We refer to $S_F = \{a_1, a_2, \ldots, a_L\}$ as the *positive set* of F and we set $a_{L+1} = 0$, $a_0 = 1$.

We henceforth choose for γ_l (defined in section 3) the value $\gamma_l := a_l$ thus we have

$$\Gamma_l := (a_l, a_{l-1}],\ 1 \leq l \leq L$$

and

$$\Gamma_{L+1} := [0, a_L].$$

From [10] we have the following bound on the VC-dimension of $\operatorname{sgn}(F_\gamma)$,

$$VC\left(\operatorname{sgn}\left(F_\gamma\right)\right) \leq \mathsf{w}(\gamma) \tag{4.6}$$

where

$$\mathsf{w}(\gamma) := \mathsf{w}_{l-1}, \text{ for } \gamma \in \Gamma_l,\ 1 \leq l \leq L+1 \tag{4.7}$$

is a non-increasing step function taking the constant value

$$\mathsf{w}_l := \log_2\left(\lambda\left(\nu_l\right) + 1\right)$$

over the interval Γ_l and the $\lambda(\nu_l)$ (defined further below in section 5) are based on the multiplicity values m_l of the positive entries a_l. The value of ν_0 is 0 and $\lambda(0) = 0$ so $\mathsf{w}_0 = 0$.

Let $T \subset \mathcal{X}$ and denote by $(F_\gamma)_{|T}$ the sub-matrix of F_γ restricted to the rows that correspond to the elements of T. Sauer's Lemma (see for instance, Theorem 3.6 in [1]) implies that the number of distinct columns of $\operatorname{sgn}((F_\gamma)_{|T})$, denoted by $\left|\operatorname{sgn}((F_\gamma)_{|T})\right|$, is bounded as follows:

$$\begin{aligned}\left|\operatorname{sgn}((F_\gamma)_{|T})\right| &\leq \sum_{i=0}^{\mathsf{w}(\gamma)} \binom{|T|}{i} \\ &\leq \left(\frac{e\,|T|}{\mathsf{w}(\gamma)}\right)^{\mathsf{w}(\gamma)}.\end{aligned}$$

Since F_γ corresponds to the class $\mathcal{W}_\gamma$ then the number of dichotomies of the class of sets $\mathcal{W}_\gamma$ on T is bounded as follows

$$\#(\mathcal{W}_\gamma; T) \leq \left(\frac{e\,|T|}{\mathsf{w}(\gamma)}\right)^{\mathsf{w}(\gamma)}. \tag{4.8}$$

Because $\mathsf{w}(\gamma)$ is a step function over the intervals Γ_l, then the right side of (4.8) is also a step function over these intervals and it suffices to derive its values at the interval boundaries a_l. Note that $a_l \in \Gamma_{l+1}$ so for $\gamma = a_l$, $1 \leq l \leq L+1$,

$$\#(\mathcal{W}_{a_l}; T) \leq \left(\frac{e\,|T|}{\mathsf{w}_l}\right)^{\mathsf{w}_l}. \tag{4.9}$$

For $l = 0$, since $\mathsf{w}_0 = 0$, we have $\#(\mathcal{W}_{a_0}; T) = 1$.

4.4 Bounding the growth function of $\mathcal{M}^+_{a_l}$ and $\mathcal{M}^-_{a_l}$

We bound the growth function of the class $\mathcal{M}^+_{a_l}$. We first fix the prototype-label vector σ and let R run over all possible n-prototype sets. We denote by $\mathcal{M}^+_{\sigma,a_l}$ the corresponding class of sets M^+_{R,σ,a_l}, $R \subset \mathcal{X}$.

Proposition 2. *For any fixed $\sigma \in \mathcal{Y}^n$, for any subset $S \subset \mathcal{X} \times \{1\}$, the number of dichotomies $\#\left(\mathcal{M}^+_{\sigma,a_l}; S\right)$ obtained by $\mathcal{M}^+_{\sigma,a_l}$ on S is bounded as follows:*

$$\#\left(\mathcal{M}^+_{\sigma,a_l}; S\right) \leq \left(\#(\mathcal{W}_{a_l}; S^+)\right)^{N_+(\sigma)N_-(\sigma)},$$

where $S^+ := \{x \in \mathcal{X} : (x,1) \in S\}$.

Proof. The number of dichotomies that the class $\mathcal{M}^+_{\sigma,a_l}$ of sets M^+_{R,σ,a_l} gets on S is the same as the number of dichotomies that the class $\mathcal{V}^+_{\sigma,a_l}$ of sets $V^+_{R,\sigma,a_l} := \{x : (x,1) \in M^+_{R,\sigma,a_l}\}$ obtains on S^+. Hence it suffices to find an upper bound on $\#\left(\mathcal{V}^+_{\sigma,a_l}; S^+\right)$. Fix any set of prototypes $R \subset \mathcal{X}$ of cardinality n. We have

$$\begin{aligned}
V^+_{R,\sigma,a_l} &= \left\{x : \min_{j \in N_-(\sigma)} d(x,p_j) - \min_{i \in N_+(\sigma)} d(x,p_i) > a_l\right\} \\
&= \left\{x : \exists i \in N_+(\sigma), \forall j \in N_-(\sigma), d(x,p_j) - d(x,p_i) > a_l\right\} \\
&= \bigcup_{i \in N_+(\sigma)} \bigcap_{j \in N_-(\sigma)} \{x : d(x,p_j) - d(x,p_i) > a_l\} \\
&= \bigcup_{i \in N_+(\sigma)} \bigcap_{j \in N_-(\sigma)} W_{a_l}^{(ind(p_i),ind(p_j))}.
\end{aligned}$$

Define the $N_-(\sigma)$-fold intersection set

$$A_i^{(R)} := \bigcap_{j \in N_-(\sigma)} W_{a_l}^{(ind(p_i),ind(p_j))}$$

and the class of such sets by

$$\mathcal{A}_R := \left\{A_i^{(R)} : i \in N_+(\sigma)\right\}.$$

From Theorem 13.5(iii) of [17] it follows that

$$\#\left(\mathcal{A}_R; S^+\right) \leq \left(\#\left(\mathcal{W}_{a_l}^+; S^+\right)\right)^{N_-(\sigma)}.$$

Now, let

$$B_R := \bigcup_{i \in N_+(\sigma)} A_i^{(R)}$$

and denote the class of all such $N_+(\sigma)$-fold unions by

$$\mathcal{B} := \{B_R : R \subset \mathcal{X}\}.$$

From Theorem 13.5(iv) of [17] it follows that

$$\#\left(\mathcal{B}; S^+\right) \leq \left(\#\left(\mathcal{A}_R^+; S^+\right)\right)^{N_+(\sigma)}.$$

Hence we have

$$\#\left(\mathcal{B}; S^+\right) \leq \left(\#\left(\mathcal{W}_{a_l}^+; S^+\right)\right)^{N_-(\sigma)N_+(\sigma)}.$$

Finally, we have

$$V_{R,\sigma,a_l}^+ = B_R$$

and

$$\mathcal{V}_{\sigma,a_l}^+ = \mathcal{B}.$$

Combining all of the above we obtain

$$\begin{aligned}\#\left(\mathcal{M}_{\sigma,a_l}^+; S\right) &= \#\left(\mathcal{V}_{\sigma,a_l}^+; S^+\right) \\ &= \#\left(\mathcal{B}; S^+\right) \\ &\leq \left(\#\left(\mathcal{W}_{a_l}^+; S^+\right)\right)^{N_-(\sigma)N_+(\sigma)}.\end{aligned}$$

□

Denote the set of dichotomies of S by all sets M_{R,σ,a_l}^+ as

$$\mathcal{U}_{R,\sigma,a_l}(S) := \left\{u\left(M_{R,\sigma,a_l}^+\right) : M_{R,\sigma,a_l}^+ \in \mathcal{M}_{\sigma,a_l}^+\right\}.$$

Letting σ be unfixed, the number of dichotomies obtained by $\mathcal{M}_{a_l}^+$ satisfies

$$\begin{aligned}\left|\{\mathcal{U}_{R,\sigma,a_l}(S) : \sigma \in \mathcal{Y}^n,\, R \subseteq \mathcal{X},\, |R| = n\}\right| &= \left|\bigcup_{\sigma \in \mathcal{Y}^n} \{\mathcal{U}_{R,\sigma,a_l}(S) : R \subseteq \mathcal{X},\, |R| = n\}\right| \\ &\leq \sum_{\sigma \in \mathcal{Y}^n} \left|\{\mathcal{U}_{R,\sigma,a_l}(S) : R \subseteq \mathcal{X},\, |R| = n\}\right| \qquad (4.10) \\ &\leq \sum_{\sigma \in \mathcal{Y}^n} \left(\#(\mathcal{W}_{a_l}; S^+)\right)^{N_+(\sigma)N_-(\sigma)} \qquad (4.11) \\ &= \sum_{k=1}^{n-1} \binom{n}{k} \left(\#(\mathcal{W}_{a_l}; S^+)\right)^{k(n-k)}, \qquad (4.12)\end{aligned}$$

where (4.11) follows from Proposition 2.

Remark 3. The expression $2^n \left(\#(\mathcal{W}_{a_l}; S^+)\right)^{n(n-1)}$ is a simple yet trivial bound on (4.11).

We now obtain a much tighter bound than the one in Remark 3.

Proposition 4. *Define w by $w := \#(\mathcal{W}_{a_l}; S^+)$. Then the following bound holds:*

$$\sum_{k=1}^{n-1} \binom{n}{k} w^{k(n-k)} \leq w^{\frac{n^2}{4}} n \binom{n}{\lfloor n/2 \rfloor}. \tag{4.13}$$

Proof: We have that $k(n-k)$ has a maximum at $k = n/2$ and therefore

$$k(n-k) \leq \frac{n^2}{4}.$$

Furthermore, for all k,

$$\binom{n}{k} \leq \binom{n}{\lfloor n/2 \rfloor}.$$

Hence

$$\sum_{k=1}^{n-1} \binom{n}{k} w^{k(n-k)} \leq n \binom{n}{\lfloor n/2 \rfloor} w^{n^2/4}.$$

□

From (4.12) and Proposition 4, it follows that for any $S \subseteq \mathcal{X} \times \{1\}$ of cardinality m,

$$\#\left(\mathcal{M}_{a_l}^+; S\right) \leq w^{\frac{n^2}{4}} n \binom{n}{\lfloor n/2 \rfloor}. \tag{4.14}$$

We now consider the class $\mathcal{M}_{a_l}^-$.

Proposition 5. *For any fixed $\sigma \in \mathcal{Y}^n$, for any subset $S \subset \mathcal{X} \times \{-1\}$, the number of dichotomies $\#\left(\mathcal{M}_{\sigma,a_l}^-; S\right)$ obtained by $\mathcal{M}_{\sigma,a_l}^-$ on S is bounded from above by the right side of (4.14).*

Proof. We follow the proof of Proposition 2, and instead of the class $\mathcal{M}_{\sigma,a_l}^+$ and a set M_{R,σ,a_l}^+ we consider the class $\mathcal{M}_{\sigma,a_l}^-$ and a set M_{R,σ,a_l}^-. By definition this set equals

$$M_{R,\sigma,a_l}^- = \left\{(x,-1) : \min_{j \in N_-(\sigma)} d\left(x, p_j\right) - \min_{i \in N_+(\sigma)} d\left(x, p_i\right) < -a_l \right\}$$

and can be written as

$$M_{R,\sigma,a_l}^- = \left\{(x,-1) : \min_{i \in N_+(\sigma)} d\left(x, p_i\right) - \min_{j \in N_-(\sigma)} d\left(x, p_j\right) > a_l \right\}. \tag{4.15}$$

Thus as in the proof of Proposition 2, the set $M^-_{R,\sigma,\gamma}$ corresponds to a set V^-_{R,σ,a_l} which is defined as

$$\begin{aligned} V^-_{R,\sigma,a_l} &= \left\{ x : \min_{i\in N_+(\sigma)} d(x,p_i) - \min_{j\in N_-(\sigma)} d(x,p_j) > a_l \right\} \\ &= \left\{ x : \exists j \in N_-(\sigma), \forall i \in N_+(\sigma), d(x,p_i) - d(x,p_j) > a_l \right\} \\ &= \bigcup_{j\in N_-(\sigma)} \bigcap_{i\in N_+(\sigma)} \{x : d(x,p_i) - d(x,p_j) > a_l\} \\ &= \bigcup_{j\in N_-(\sigma)} \bigcap_{i\in N_+(\sigma)} W_{a_l}^{(ind(p_j),ind(p_i))}. \end{aligned}$$

From here the proof proceeds as the proof of Proposition 2, just swapping: the indices i with j, the sets N_- with N_+ and the sets S^+ with S^-. □

4.5 Finalizing

In the previous section we derived an upper bound on the number of dichotomies on any set of cardinality m obtained by $\mathcal{M}^+_{\sigma,a_l}$ or $\mathcal{M}^-_{\sigma,a_l}$ in terms of the number of dichotomies by the classes $\mathcal{W}_{a_l}$. For $T \subset \mathcal{X}$, let $w(T) := \#(\mathcal{W}_{a_l}; T)$ and define the growth function of $\mathcal{W}_{a_l}$ as

$$w_m := \max_{T:|T|=m} w(T).$$

From (4.14) and from the fact that $w_m \geq 1$, it follows that the growth function $\Pi_{\mathcal{M}^+_{\sigma,a_l}}(m)$ of $\mathcal{M}^+_{\sigma,a_l}$ is bounded as follows,

$$\Pi_{\mathcal{M}^+_{\sigma,a_l}}(m) \leq w_m^{\frac{n^2}{4}} n \binom{n}{\lfloor n/2 \rfloor}. \tag{4.16}$$

Using a standard bound for the central binomial coefficient, we have

$$\binom{n}{\lfloor n/2 \rfloor} < \sqrt{\frac{2}{\pi n}} 2^n$$

and so

$$\Pi_{\mathcal{M}^+_{\sigma,a_l}}(m) \leq {w_m}^{\frac{n^2}{4}} \sqrt{\frac{2n}{\pi}} 2^n.$$

From (4.9) we have

$$w_m \leq \left(\frac{em}{\mathsf{w}(a_l)} \right)^{\mathsf{w}(a_l)}$$

and therefore

$$\begin{aligned} \ln\left(\Pi_{\mathcal{M}^+_{\sigma,a_l}}(m)\right) &\leq \ln\left({w_m}^{\frac{n^2}{4}} \sqrt{\frac{2n}{\pi}} 2^n \right) \\ &\leq \frac{n^2}{4} \ln\left(\left(\frac{em}{\mathsf{w}(a_l)} \right)^{\mathsf{w}(a_l)} \right) + n \ln 2 + \frac{1}{2} \ln\left(\frac{2n}{\pi} \right) \\ &\leq \frac{n^2 \mathsf{w}(a_l)}{4} \ln\left(\frac{em}{\mathsf{w}(a_l)} \right) + n \ln 2 + \frac{1}{2} \ln\left(\frac{2n}{\pi} \right). \end{aligned} \tag{4.17}$$

From (4.7), we have $\mathsf{w}(a_l) = \mathsf{w}_l$. Therefore (4.17) equals

$$\frac{n^2\mathsf{w}_l}{4}\ln\left(\frac{em}{\mathsf{w}_l}\right) + n\ln 2 + \frac{1}{2}\ln\left(\frac{2n}{\pi}\right)$$

We now define $G(m,\gamma)$ in section 3.2 as follows (— recall that $G(m,\gamma)$ bounds the growth function evaluated at $2m$): for any $\gamma \in \Gamma_{l+1}$,

$$G(m,\gamma) := \frac{n^2\mathsf{w}_l}{4}\ln\left(\frac{2em}{\mathsf{w}_l}\right) + n\ln 2 + \frac{1}{2}\ln\left(\frac{2n}{\pi}\right).$$

Note that for all $\gamma \in \Gamma_l$, $G(m,\gamma) = G(m, a_{l-1})$ and the second term inside the square root in (3.9) satisfies $\ln(8(C+1)/\gamma\delta) \geq \ln(8(C+1)/a_{l-1}\delta)$, so it follows that $\epsilon(\gamma) \geq \epsilon(a_{l-1})$ and therefore the requirement on $\epsilon(\gamma)$ of section 3.2, namely, that it is non-decreasing as γ decreases, is satisfied.

5 Main result

To obtain the main result, we draw on some results and notations from [10]. We define a few quantities, leaving the dependence on N implicit.

Let us denote by the i^{th} *shell* of the binary N-dimensional cube $\{-1,1\}^N$ (N-cube) the set of all vertices that have i components that are 1.

Denote by

$$c_j := \binom{N}{j}$$

the number of vertices in the j^{th} shell.

Denote by

$$b_n := \sum_{j=1}^{n} c_j \tag{5.1}$$

the number of vertices of the cube contained in the first n shells, and let $b_0 := 0$

Let us define

$$\ell_n := \sum_{j=1}^{n} j c_j, \tag{5.2}$$

the total 'weight' (number of 1-entries) in all of the vertices of the cube that are in the first n shells.

For a positive integer m define

$$Q(m) := \min\{q : \ell_q \geq m\}. \tag{5.3}$$

For instance, if $m = 17$, $N = 4$, then $1\binom{4}{1} + 2\binom{4}{2} + 1 = 17$ so $Q(17) = 3$.

Let

$$\Delta := m - \ell_{Q(m)-1},$$

and then define $\lceil m \rceil$ as the following 'rounded' value:

$$\lceil m \rceil := \begin{cases} m & \text{if } Q(m) = 1 \text{ or } \Delta \bmod Q(m) = 0 \\ m + (Q(m) - \Delta \bmod Q(m)) & \text{otherwise.} \end{cases}$$

So, $\lceil m \rceil$ is the smallest integer greater than or equal to m which is the total weight (number of 1-entries) in a set of vertices of the cube, where that set is formed by first populating the first shell, then the second, and so on. So, m 1-entries might not be enough to form all the vertices of shells 1 to $Q(m)-1$ and then some vertices in shell $Q(m)$: an additional (at most $Q(m)-1$) 1-entries might be necessary (to 'complete' a vertex in shell $Q(m)$).

For instance (continuing the above example), since $Q(17) = 3 \geq 2$ then $\lceil 17 \rceil = 17 + (3 - 1 \bmod 3) = 19$, and indeed 19 1-entries are required in the vertices in shells 1 and 2, and in the first vertex of the 3rd shell.

For positive integer m let us define

$$\lambda(m) := b_{Q(m)-1} + \frac{\lceil m \rceil - \ell_{Q(m)-1}}{Q(m)} \tag{5.4}$$

and define $\lambda(0) := 0$. Define the numbers ν_i as follows (where m_i is the multiplicity of a_i, as earlier):

$$\begin{aligned} \nu_0 &= 0 \\ \nu_i &= \lceil \nu_{i-1} + m_i \rceil, \quad 1 \leq i \leq L. \end{aligned} \tag{5.5}$$

Note that ν_i, $1 \leq i \leq L$, depend only on m_i and hence can be evaluated directly from the matrix F.

The following is the main result of the paper.

Theorem 6. *Let $N \geq 1$ and $n \geq 2$ be fixed integers and $\mathcal{X} = \{x_i\}_{i=1}^N$ be a finite distance space with a distance function $d(x_i, x_j)$, normalized such that diam$(\mathcal{X}) = \max_{1 \leq i,j \leq N} d(x_i, x_j) = 1$. Let*

$$f_j^{(i)} := \begin{bmatrix} d(x_1, x_j) - d(x_1, x_i) \\ \vdots \\ d(x_N, x_j) - d(x_N, x_i) \end{bmatrix},$$

and

$$F^{(i)} = \left[f_1^{(i)}, \ldots, f_{i-1}^{(i)}, f_{i+1}^{(i)}, \ldots, f_N^{(i)} \right]$$

and define the $N \times N(N-1)$ matrix

$$F := \left[F^{(1)}, \ldots, F^{(N)} \right].$$

Let $0 = a_{L+1} < a_L < \cdots < a_1 < a_0 = 1$ be the values of the positive entries of F and let $m_l \geq 1$ be the number of times that a_l appears in F, $1 \leq l \leq L$. Define $\Gamma_l := (a_l, a_{l-1}]$, $1 \leq l \leq L$, $\Gamma_{L+1} := [0, a_L]$ and $C := \sum_{l=1}^L a_l$. For $1 \leq l \leq L$, let $\mathsf{w}_l = \log_2(\lambda(\nu_l) + 1)$.

Let $\mathcal{Y} = \{-1, 1\}$ *and let* $(R, \sigma) \subseteq (\mathcal{X} \times \mathcal{Y})^n$, $R := \{p_i\}_{i=1}^n$ *denote any set of* n *prototypes* p_i *with* $N_+(\sigma)$ *of them that are labeled by* 1 *and* $N_-(\sigma)$ *that are labeled by* -1. *Let* $h_{R,\sigma}$ *be a* nearest-prototype binary classifier*, given by*

$$h_{R,\sigma}(x) = \begin{cases} 1 & \text{if } argmin_{1 \le i \le n} d(x, p_i) \in N_+(\sigma) \\ -1 & \text{otherwise} \end{cases} \tag{5.6}$$

and define its signed width function $f_{R,\sigma}$ *as*

$$f_{R,\sigma}(x) = \min_{j \in N_-(\sigma)} d(x, p_j) - \min_{i \in N_+(\sigma)} d(x, p_i).$$

Let $P^m := P_{XY}^m$ *be a probability measure over* $(\mathcal{X} \times \mathcal{Y})^m$. *For any* $0 < \delta < 1$, *with* P^m*-probability at least* $1-\delta$ *the following holds for an i.i.d. sample* $\xi := \{(X_i, Y_i)\}_{i=1}^m \subseteq (\mathcal{X} \times \mathcal{Y})^m$ *drawn according to* P^m *:*

- *for any set of* n *labeled prototypes* $(R, \sigma) \subseteq (\mathcal{X} \times \mathcal{Y})^n$, *where* (R, σ) *may depend on the sample* ξ *(and in particular, may be a subset of* ξ*)*
- *for all* $\gamma > 0$,

$$P\left(Y f_{R,\sigma}(X) \le 0\right) \le \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}\left\{Y_j f_{R,\sigma}(X_j) \le \gamma\right\} + \epsilon(m, \gamma, \delta) \tag{5.7}$$

where for $\gamma \in \Gamma_{l+1}$, $0 \le l \le L$,

$$\epsilon(m, \gamma, \delta) := \sqrt{\frac{32}{m}\left(\frac{n^2 \mathsf{w}_l}{4} \ln\left(\frac{2em}{\mathsf{w}_l}\right) + n \ln 2 + \frac{1}{2} \ln\left(\frac{2n}{\pi}\right) + \ln\left(\frac{8(C+1)}{\gamma\delta}\right)\right)} \tag{5.8}$$

where

$$\mathsf{w}_l := \log_2\left(\lambda\left(\nu_l\right) + 1\right).$$

The size N of the distance space does not enter the bound of Theorem 6 but because the value of w_l and C depend on the distance space through the positive entries of the matrix F then these quantities may grow with N (depending on the distance space $\mathcal{X}$). The value of w_l decreases as $l := l(\gamma)$ decreases (because the intervals Γ_l are situated more to the right as l decreases). Thus w_l decreases as a step function over the intervals Γ_l as γ increases such that the larger the value of γ, the lower the interval index l and the lower the value of w_l. Thus the upper bound (5.8) decreases as γ increases (assuming that the change in w_l dominates the change in the $\ln\left(\frac{1}{\gamma}\right)$ term). To get a feel for the rate of decrease of w_l with respect to γ, see example on p. 23 of [10].

Since γ is a parameter that may be chosen after the random sample is drawn, it can depend on the sample, which makes the bound data-dependent.

5.1 Comparison with other results

Theorem 6 states an upper bound on the generalization error for learning on any finite distance space decreases with respect to m like $O\left(\sqrt{\frac{n^2 \ln m}{m}}\right)$. Let us compare this rate to other works.

If the distance space is $\mathcal{X} = \mathbb{R}^d$ and if the prototype set R is a subset of the sample then Theorem 19.6 of [17] gives the following error bound for learning nearest-neighbor classifiers that are based on R:

$$\epsilon = O\left(\sqrt{\frac{1}{m}\left(\frac{(d+1)n(n-1)}{2}\ln m + n + \ln\left(\frac{1}{\delta}\right)\right)}\right)$$

which is also $O\left(\sqrt{\frac{n^2 \ln m}{m}}\right)$, but in contrast to Theorem 6 is not data-dependent, which in general makes it looser.

In [15], Theorem 1 presents an error bound for learning LVQ on $\mathbb{R}^d$ which has a dependence on the sample margin of the following form $O\left(\sqrt{\frac{an^2}{m}}\right)$, where $a = \min\left(d+1, \left(\frac{\rho}{\gamma}\right)^2\right)$ and ρ bounds the magnitude of each sample point, $0 < \gamma < \frac{1}{2}$ is the sample margin (which is defined similar to the width (2.4)). They claim that this bound is independent of the dimension d, presumably because if $\left(\frac{\rho}{\gamma}\right)^2$ is smaller than $d+1$ then the $d+1$ factor disappears from the bound. Their proof is not included; however, it appears to be a direct application of fat-shattering error bounds, see for instance, Theorem 4.18 of [16] which bounds the generalization error of learning linear classifiers and has the same dependence on the margin parameter, namely, $O\left(\frac{\rho}{\gamma}\right)^2$. These bounds are based on a bound on the log of the γ-covering number by $O(d_\gamma)$ and a bound on the fat-shattering number for linear classifiers $d_\gamma \leq \left(\frac{\rho}{\gamma}\right)^2$. In comparison, the bound of Theorem 6 also depends on γ but in a non-direct way through w_l (as discussed above, $l = l(\gamma)$). Depending on the distance space's positive set S_F, w_l may decrease even faster than $\frac{1}{\gamma^2}$.

Although we assume that the distance space is finite of cardinality N, the bound (5.8) may or may not grow with N and this depends on the matrix F. In comparison to the above works, in (5.8) there is an implicit complexity quantity which enters through w_l and is defined as $\lambda(\nu_l)$. This is an upper bound on the pseudo-rank of the perturbed matrix F, denoted by F_{a_l}, (see [10]), which is the number of distinct columns of a matrix $\mathrm{sgn}(F_{a_l})$ and may depend on N (depending on the definition of the distance space).

6 Conclusions

We use the concept of width to learn a family of classifiers based on the nearest-prototype decision rule over an arbitrary finite distance spaces (a significantly more general and perhaps

more applicable setting than that of metric spaces). In this setting a classifier is represented by a set of n prototypes, that can be any labeled points in the distance space, and even a subset of the sample. We obtain an upper bound on the generalization error of any such nearest-prototype classifier. Using γ as the width parameter, the error bound depends on the γ-empirical error and holds uniformly over the family of such classifiers. Ignoring the dependence on γ, the bound is $O\left(\sqrt{\frac{n^2 \ln m}{m}}\right)$. The dependence on γ is more subtle because it enters through a complexity quantity $\lambda(\nu_l)$ which bounds from above the pseudo-rank of a γ-perturbed version of a matrix that represents all half spaces in the distance space.

Acknowledgements

This work was supported in part by a research grant from the Suntory and Toyota International Centres for Economics and Related Disciplines at the London School of Economics. The authors thank the referees for their comments which resulted in shorter proofs of Propositions 2 and 4.

References

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

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

[3] M. Anthony and J. Ratsaby. Classification based on prototypes with spheres of influence. *Information and Computation*, 256:372–380, 2017.

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

[5] M. Anthony and J. Ratsaby. A hybrid classifier based on boxes and nearest neighbors. *Discrete Applied Mathematics*, 172:1–11, 2014.

[6] M. Anthony and J. Ratsaby. Analysis of a multi-category classifier. *Discrete Applied Mathematics*, 160(16-17):2329–2338, 2012.

[7] M. Anthony and J. Ratsaby. Learning bounds via sample width for classifiers on finite metric spaces. *Theoretical Computer Science*, 529:2–10, 2014.

[8] M. Anthony and J. Ratsaby. Multi-category classifiers and sample width. *Journal of Computer Systems and Sciences*, 82(8):1223–1231, 2016.

[9] M. Anthony and J. Ratsaby. A probabilistic approach to case-based inference. *Theoretical Computer Science*, 589:61–75, 2015.

[10] M. Anthony and J. Ratsaby. Large-width bounds for learning half-spaces on distance spaces. *Submitted*, 2017.

[11] A. Belousov and J. Ratsaby. Massively parallel computations of the LZ-complexity of strings,. In *Proc. of the 28th IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI'14)*, pages pp. 1–5, Eilat, Dec. 3-5 2014.

[12] A. Belousov and J. Ratsaby. A parallel distributed processing algorithm for image feature extraction. In *Advances in Intelligent Data Analysis XIV - 14th International Symposium, IDA 2015, Saint-Etienne, France, October 22 - 24, 2015. Proceedings*, volume 9385 of *Lecture Notes in Computer Science*. Springer, 2015.

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

[14] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, January 1967.

[15] K. Crammer, R. Gilad-Bachrach, A. Navot, and N. Tishby. Margin analysis of the LVQ algorithm. In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 462–469, 2002.

[16] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based learning methods*. Cambridge University Press, 2000.

[17] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, 1996.

[18] E. Deza M. Deza. *Encyclopedia of Distances*, volume 15 of *Series in Computer Science*. Springer-Verlag, 2009.

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

[20] B. Hammer, M. Strickert, and T. Villmann. On the generalization ability of GRLVQ networks. *Neural Processing Letters*, 21(2):109–120, 2005.

[21] T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

[22] E. Pekalska and R. P. W. Duin *The Dissimilarity Representation for Pattern Recognition : Foundations and Applications*. World Scientific Publishing Co Pte Ltd, Singapore, SINGAPORE, 2005.

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

[24] P. Schneider, M. Biehl, and B. Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532 – 3561, 2009.

[25] 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.

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

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

[28] 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.