Classification based on prototypes with spheres of influence

Martin Anthony

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

Joel Ratsaby

Electrical and Electronics Engineering Department, Ariel University, Ariel 40700, ISRAEL

Abstract

We present a family of binary classifiers and analyse their performance. Each classifier is determined by a set of 'prototypes', whose labels are given; and the classification of any other point depends on the labels of the prototypes to which it is sufficiently close, and on how close it is to these prototypes. More precisely, the classification of a given point is determined through the sign of a discriminant function. For each prototype, its sphere of influence is the largest sphere centred on it that contains no prototypes of opposite label, and, given a point to be classified, there is a contribution to the discriminant function at that point from precisely those prototypes whose spheres of influence contain the point, this contribution being positive from positive prototypes and negative from negative prototypes. Furthermore, these contributions are larger in absolute value the closer the point is (relative to the sphere's radius) to the prototype. We quantify the generalization error of such classifiers in a standard probabilistic learning model, and we do so in

Email addresses: m.anthony@lse.ac.uk (Martin Anthony), ratsaby@ariel.ac.il (Joel Ratsaby)

To appear in: Information and Computation

a way that involves the values of the discriminant function on the points of the random training sample.

Keywords: Classification, learning, generalisation error

---

1. Introduction

Learning Vector Quantization (LVQ) and its various extensions introduced by Kohonen [16] are used sucessfully 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. LVQ fails if the Euclidean representation is not well-suited for the data. To that end, several extensions of the LVQ algorithm exist which use a weighted Euclidean metric [13] that 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 [13, 20] and, as is usually the case for large-margin learning [1], the 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.

In the world of big data, which deals with a rich variety of learning domains, there is a huge potential in doing prototype-based learning over non-Euclidean spaces. In this paper we present a family of binary classifiers for learning on any metric input space. We analyse their performance and present generalization learning error bounds that are sample-dependent and hence take advantage of samples that can be classified with a large margin. Each classifier is determined by a set of 'prototypes', whose classifications are given; and the classification of any other point depends on the classifications of the prototypes to which it is sufficiently close, and on how close it is to these prototypes. Thus, in contrast to the above-mentioned works, here a classifier's decision is not based only on the nearest prototype. In many domains of application, data can no longer simply be considered to be in Euclidean space. As has been pointed out in [14], data can take diverse forms in areas such as linguistics and bioinformatics. For this reason, an approach

that analyses data in a general metric space (such as that taken here) might be more useful.

More precisely, the classification of a given point is determined through the sign of a discriminant function. For each prototype, its sphere of influence is defined to be the largest sphere centred on it that contains no prototypes of opposite label. Given a point to be classified, there is a contribution to the discriminant function at that point from precisely those prototypes whose spheres of influence contain the point, this contribution being positive from positive prototypes and negative from negative prototypes. These contributions are larger in absolute value the closer the point is (relative to the sphere's radius) to the prototype. We quantify the generalization error of such classifiers in a standard probabilistic learning model, and we do so in a way that involves the values of the discriminant function on the points of the random training sample.

We note in passing that the idea of a sphere of influence is not new. In fact, RCE networks [17] have a hidden layer of activation units associated with a spherical decision region in the input space. There are some differences between our classifier and the RCE. RCE is essentially a classifier whose decision regions are union of spheres, which may not cover all of the input space and hence the classifier can in some cases reject making a decision. The radii of the spheres are parameters to be learnt. Learning RCE involves adapting the size of the radii in an incremental manner in response to whether sample instances are included or not in spheres that are associated with a mismatching class label. New spherical units, that is, prototypes, can also be added when sample points are not covered and not classified. In contrast to RCE, our classifier is non-parametric and the region of influence of each prototype, in resemblance to Voronoi cells in the nearest-neighbor classifier [8], is determined directly from the sample without any parameter such as a radius. The classifier's definition is intentionally left very general in that the set of prototypes can be *any* set of k points, in particular a subset of the sample, and can be determined via *any* algorithm. The error bounds that we state in the paper apply regardless of the algorithm that is used to learn these prototypes.

2. Classifiers based on spheres of influence

The classifiers we consider are binary classifiers defined on a metric space $\mathcal{X}$; so, they are functions $h : \mathcal{X} \to \{-1, 1\}$. We shall assume that $\mathcal{X}$ is of finite diameter with respect to the metric d and, for the sake of simplicity, that its diameter is 1. (The analysis can easily be modified for any other value of the diameter.) Each classifier we consider is defined by a set of *labeled prototypes*. More precisely, a typical classifier is defined by a finite set Π^+ of *positive prototypes* and a disjoint set Π^- of *negative prototypes*, with Π^+ and Π^- both being subsets of $\mathcal{X}$. The idea is that the correct classifications of the points in Π^+ (Π^-, respectively) are $+1$ (-1). We define the *sphere of influence* of each prototype as follows. Suppose $p \in \Pi^+$ and let

$$r(p) = \min\{d(p, p^-) : p^- \in \Pi^-\},$$

the distance to the closest oppositely-labeled prototype; and define $r(p)$ analogously in the case where $p \in \Pi^-$. Then the open ball $B_{r(p)}(p) = B_{r(p)}(p; d)$, of radius $r(p)$ and centred on p, is the sphere of influence of p. Suppose that $\Pi = \Pi^+ \cup \Pi^- = \{p_1, p_2, \ldots, p_k\}$, where $\Pi^+ = \{p_1, \ldots, p_t\}$ and $\Pi^- = \{p_{t+1}, \ldots, p_k\}$, and let r_i denote $r(p_i)$ where $0 < r(p_i) \leq 1$. For $x \in \mathcal{X}$, let

$$\phi_i(x) = 1 - \frac{d(x, p_i)}{r_i}$$

and let

$$s_i(x) = [\phi_i(x)]_+ ,$$

where, for $z \in \mathbb{R}$, $[z]_+ = z$ if $z \geq 0$ and $[z]_+ = 0$ otherwise. Define the 'discriminant' function $f_\Pi : \mathcal{X} \to \mathbb{R}$ as follows:

$$f_\Pi(x) = \sum_{i=1}^{t} s_i(x) - \sum_{i=t+1}^{k} s_i(x). \tag{1}$$

The corresponding binary classifier defined by Π (and its labels) is $h_\Pi(x) = \mathrm{sgn}(f_\Pi(x))$ where $\mathrm{sgn}(z) = 1$ if $z \geq 0$ and $\mathrm{sgn}(z) = -1$ if $z < 0$. (Note that $|f_\Pi(x)| \leq k$ for all x.) We denote the class of all such f_Π by $\mathcal{F}$ and we denote by $\mathcal{H}$ the corresponding set of classifiers h_Π. In the context of learning, (1) defines the *margin* of h_Π at x.

To explain the idea behind this classifier, consider the contribution that a prototype p makes to the value $f_\Pi(x)$ of the discriminant function at x and suppose, without loss of generality, that p is a positive prototype. This prototype makes no contribution at all if x lies outside the sphere of influence of p. The rationale for this is simply that, in this case, there must be at least one negative prototype p^- whose distance from p is no more than the distance from x to p; and so there seems to be little justification for assuming x is close enough to p to derive some influence from the classification of p. If x does lie inside the sphere of influence of p, then there is a positive contribution to $f_\Pi(x)$ that is between 0 and 1 and is larger in absolute value the closer x is to p. The rationale here is that if x is deeply embedded in the sphere of influence of p (rather than being more on its periphery), and if we were considering how we should classify the point by taking into account only the prototype p, then, given its relative proximity to p, it would be reasonable to propose a positive classification. The overall classification is determined by the net effect of these contributions. So, if x lies closer to a prototype p than does any oppositely-labeled prototype, we can think of p as contributing an influence (signed the same as the label of p) to the discriminant (and hence an influence on the final value of the classification); and this influence depends on how relatively close x is to p within its sphere of influence. Although slightly reminiscent of nearest neighbor methods, this approach is quite different. It is likely to have less sensitivity to changes in the prototypes than nearest neighbor methods would. This is because the discriminant has a term for each prototype, not just a fixed number of nearest ones. Furthermore, for the introduction of a new prototype to change the classification of a point, that point would have to be sufficiently close to the new prototype, sufficiently within its sphere of influence. (Note that any point whose classification changes on the introduction of the new prototype must be in its sphere of influence, and it could be that few points change classification because the contribution to the the discriminant arising from the new prototype would have to be sufficiently large to change the sign of the discriminant: this is not a winner-takes-all classification in contrast to the standard nearest neighbor method.)

We should note that the particular form we take for ϕ_i could be modified: indeed, we could take $\phi_i(x) = \psi(d(x, p_i)/r_i)$ where $\psi(z)$ decreases with z and $\psi(z) \leq 0$ if $z \geq 1$. For many such ψ functions, the analysis that follows could be modified appropriately.

3. Generalization performance of the classifiers

3.1. Probabilistic modelling of learning

To quantify the performance of a classifier after training, we use a form of the popular 'PAC' model of computational learning theory (see [3], [21], [7]). This assumes that we have some training examples $z_i = (x_i, b_i) \in Z = \mathcal{X} \times \{-1, 1\}$, each of which has been generated independently at random according to some fixed probability measure P on Z. Then, we can regard a training sample of length m, which is an element of Z^m, as being randomly generated according to the product probability measure P^m. Suppose that $\mathcal{F}$ is the set of discriminant functions we are using to classify. (So, recall that $\mathcal{F}$ is a set of real-valued functions and that the corresponding binary classification functions are the functions $h = \text{sgn}(f)$ for $f \in \mathcal{F}$.)

The natural way to measure the predictive accuracy of $h = \text{sgn}(f)$ for $f \in \mathcal{F}$ in this context is by the probability that h agrees with the classification of future randomly drawn elements of Z. We therefore use the following error measure of the classifier $h = \text{sgn}(f)$:

$$\text{er}_P(h) := \text{er}_P(f) = P\left(\{(x, b) \in Z : \text{sgn}(f(x)) \neq b\}\right).$$

Of course, we do not know this error: we only know how well the classifier performs on the training sample. We could quantify how well h performs on the training sample by using the *sample error* of $h = \text{sgn}(f)$:

$$\text{er}_{\mathbf{z}}(h) = \frac{1}{m} |\{i : \text{sgn}(f(x_i)) \neq b_i\}|$$

(the proportion of points in the sample incorrectly classified by h or, equivalently, for which f gives the incorrect sign). We will also denote this sample error by $\text{er}_{\mathbf{z}}(f)$. We will find it more useful, however, to use a variant of this, involving a 'width' or 'margin' parameter γ. Much emphasis has been placed in practical machine learning techniques, such as Support Vector Machines [10], on 'learning with a large margin'. (See, for instance [19], [1], [2] and[18].) Related work involving 'width' (applicable to binary-valued classifiers directly rather than those obtained by taking the sign of real-valued

functions) has also been carried out [4] and, similarly, shows that 'definitive' classification is desirable. If $h = \mathrm{sgn}(f)$, we define

$$\mathrm{er}_{\mathbf{z}}^{\gamma}(h) = \mathrm{er}_{\mathbf{z}}^{\gamma}(f) = \frac{1}{m}|\{i : f(x_i)b_i < \gamma\}|.$$

This is the proportion of points $z_i = (x_i, b_i)$ in the sample for which *either* $\mathrm{sgn}(f(x_i)) \neq b_i$, *or* $\mathrm{sgn}(f(x_i)) = b_i$ *but* $|f(x_i)| < \gamma$. So it is the fraction of the sample that is either misclassified by the classifier, or is correctly classified but *not definitively so*, in the sense that the value of $f(x_i)$ is *only just* of the right sign (but not of absolute value at least γ).

A number of results give high-probability bounds on $\mathrm{er}_P(h)$ in terms of $\mathrm{er}_{\mathbf{z}}^{\gamma}(f)$. A typical such result would be of the following form: for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$\mathrm{er}_P(\mathrm{sgn}(f)) < \mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon(m, \gamma, \delta),$$

where ϵ decreases with m and δ. We obtain a bound of a similar, but slightly different, form in this paper for the set of classifiers we are considering.

3.2. Covering numbers and a generalization result

To deploy techniques from the theory of large-margin learning, we will need to 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 (or pseudo-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 $\|f\|_{l_\infty(S)} = \max_{x \in S} |f(x)|$ and we denote by $d_\infty(S)$ the corresponding metric, $d_\infty(f, g) = \|f - g\|$. For $\alpha > 0$, an α-cover of F with respect to $d_\infty(S)$ is then 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 $x \in S$, $|f(x) - \hat{f}(x)| \leq \alpha$. The covering number $\mathcal{N}(F, \alpha, d_\infty(S))$ is the smallest cardinality of a covering for F with respect to $d_\infty(S)$. We define the *uniform covering number* $\mathcal{N}_\infty(F, \alpha, m)$ to be the maximum of $\mathcal{N}(F, \alpha, d_\infty(S))$, over all S with $S \subseteq X$ and $|S| = m$.

We will make use of the following result from [5].

Theorem 3.1. *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 $\mathbf{z}$ of length m will be such that:* for all $f \in F$, *and* for all $\gamma \in (0, B]$,

$$\mathrm{er}_P(\mathrm{sgn}(f)) \leq 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \frac{4}{m}\left(\ln \mathcal{N}_\infty(F, \gamma/4, 2m) + \ln\left(\frac{4B}{\gamma\delta}\right)\right).$$

Note that, in Theorem 3.1, 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.

4. Covering numbers for the class of discriminants

Our approach to bounding the covering number of $\mathcal{F}$ with respect to the $d_\infty(S)$ metrics is to construct and bound the size of a covering with respect to the sup-norm on $\mathcal{X}$. (This is the norm given by $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$.) This clearly also serves as a covering with respect to $d_\infty(S)$, for any S, since if $\|f - \hat{f}\|_\infty \leq \gamma$ then, by definition of the sup-norm, $\sup_{x \in \mathcal{X}} |f(x) - \hat{f}(x)| \leq \gamma$ and, hence, for all $x \in \mathcal{X}$ (and, therefore, for all $x \in S$ where S is any subset of $\mathcal{X}$), $|f(x) - \hat{f}(x)| \leq \gamma$. The construction we use is based on one from [5].

4.1. A Lipschitz bound for the function class

We first show that the discriminant functions are 'smooth', meaning Lipschitz-continuous. Suppose $f = f_\Pi \in F$ is defined by prototypes $\Pi = \{p_1, p_2, \ldots, p_k\}$

and define $R(\Pi)$ to be

$$R(\Pi) = \sum_{i=1}^{k} \frac{1}{r_i}$$

where, as before, r_i denotes $r(p_i)$, the radius of the sphere of influence of p_i. Let $R > 0$ and suppose that $\mathcal{F}_R^k \subseteq \mathcal{F}$ is $\mathcal{F}_R^k = \{f_\Pi \in \mathcal{F} : |\Pi| = k, R(\Pi) \leq R\}$, the set of all $f_\Pi \in \mathcal{F}$ such that $|\Pi| = k$ and $\sum_{i=1}^{k} 1/r_i \leq R$.

We prove that the class $\mathcal{F}_R^k$ satisfies a Lipschitz condition, as follows (recall from (1) the definition of $f = f_\Pi$):

Theorem 4.1. *For every $f \in \mathcal{F}_R^k$,*

$$|f(x) - f(x')| \leq R\, d(x, x')$$

uniformly for any x, $x' \in \mathcal{X}$.

Proof: Suppose $f \in \mathcal{F}_R^k$ and consider two points x, $x' \in \mathcal{X}$. We have

$$\begin{aligned} |f(x) - f(x')| &= \left| \sum_{i=1}^{t} s_i(x) - \sum_{i=t+1}^{k} s_i(x) - \sum_{i=1}^{t} s_i(x') + \sum_{i=t+1}^{k} s_i(x') \right| \\ &= \left| \sum_{i=1}^{t} (s_i(x) - s_i(x')) - \sum_{i=t+1}^{k} (s_i(x) - s_i(x')) \right| \\ &\leq \sum_{i=1}^{k} |s_i(x) - s_i(x')|. \end{aligned}$$

Now, for any real numbers a, b, we have $|[a]_+ - [b]_+| \leq |a - b|$ (as can be easily checked), and so

$$\begin{aligned} |s_i(x) - s_i(x')| &= |[\phi_i(x)]_+ - [\phi_i(x')]_+| \\ &\leq |\phi_i(x) - \phi_i(x')| \\ &= \left| \left(1 - \frac{d(x, p_i)}{r_i}\right) - \left(1 - \frac{d(x', p_i)}{r_i}\right) \right| \\ &= \frac{1}{r_i} |d(x, p_i) - d(x', p_i)| \\ &\leq \frac{1}{r_i} d(x, x'). \end{aligned}$$

It follows, then, that

$$|f(x) - f(x')| \leq \sum_{i=1}^{k} |s_i(x) - s_i(x')| \leq \sum_{i=1}^{k} \frac{1}{r_i} d(x, x') \leq R\, d(x, x'),$$

as required. □

Note that this proof relies on d being a metric, because the triangle inequality is used when we assert that $|d(x, p_i) - d(x', p_i)| \leq d(x, x')$. (The argument would work also for a pseudo-metric space, but it is intuitively more satisfying to deal with a metric space so that the spheres of influence have positive radii.)

Next we use this 'smoothness' to obtain a cover.

4.2. Covering the function class

For now, let us fix R and k. Let the subset $C_\gamma \subseteq \mathcal{X}$ be a *minimal* size γ-cover for $\mathcal{X}$ with respect to the metric d of the metric space. So, for every $x \in \mathcal{X}$ 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

$$\Lambda_\gamma = \left\{ i\gamma : \; i = -\left\lceil \frac{k}{\gamma} \right\rceil, \ldots, -1, 0, 1, 2, \ldots, \left\lceil \frac{k}{\gamma} \right\rceil \right\}$$

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* $\hat{f}_{ext} : \mathcal{X} \to \Lambda_\gamma$ of $\hat{f}$ to the whole domain $\mathcal{X}$ as follows. For each $x \in \mathcal{X}$, let $\hat{x} \in C_\gamma$ be such that $d(x, \hat{x}) \leq \gamma$. There may be more than one possible choice of $\hat{x}$ with this property, but we assume a fixed choice of one such $\hat{x}$ is made for each x, so that we have a fixed mapping $x \mapsto \hat{x}$. Then, define $\hat{f}_{ext}(x) = \hat{f}(\hat{x})$. There

is a one-to-one correspondence between the functions $\hat{f}$ 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 $f \in \mathcal{F}_R^k$ there is $\hat{f}_{ext} \in \hat{F}_{ext}$ such that

$$\|f - \hat{f}_{ext}\|_\infty \leq (R+1)\gamma;$$

that is, such that

$$\sup_{x \in \mathcal{X}} |f(x) - \hat{f}_{ext}(x)| \leq (R+1)\gamma.$$

First for every point $\hat{x} \in C_\gamma$ consider the value $f(\hat{x})$ and find a corresponding value in Λ_γ, call it $\hat{f}(\hat{x})$, such that $|f(\hat{x}) - \hat{f}(\hat{x})| \leq \gamma$. (That there exists such a value follows by the design of Λ_γ.) By the above definition of extension, $\hat{f}_{ext}(x) = \hat{f}(\hat{x})$ where the mapping $x \mapsto \hat{x}$ is as above. Since $d(x, \hat{x}) \leq \gamma$, Theorem 4.1 shows that $|f(x) - f(\hat{x})| \leq R\,\gamma$. We therefore have

$$|f(x) - \hat{f}_{ext}(x)| = |f(x) - \hat{f}(\hat{x})| \leq |f(x) - f(\hat{x})| + |f(\hat{x}) - \hat{f}(\hat{x})| \leq R\,\gamma + \gamma,$$

from which the claim follows.

Hence the set $\hat{F}_{ext}$ forms an $(R+1)\gamma$-covering of the class $\mathcal{F}_R^k$ in the sup-norm over $\mathcal{X}$. Thus we have the following covering number bound (holding uniformly for all m) which we will use in Theorem 3.1 .

Theorem 4.2. *With the above notation, for all m, and $\gamma \in (0, k]$,*

$$\mathcal{N}_\infty \left(\mathcal{F}_R^k, \frac{\gamma}{4}, 2m \right) \leq \left(\frac{11k(R+1)}{\gamma} \right)^N,$$

where $N = N_{\gamma/(4(R+1))} = \mathcal{N}\left(\mathcal{X}, \frac{\gamma}{4(R+1)}, d \right)$.

Proof: The analysis above shows that

$$\begin{aligned} \mathcal{N}_\infty \left(\mathcal{F}_R^k, (R+1)\gamma, 2m \right) &\leq |\hat{F}_{ext}| \\ &\leq |\Lambda_\gamma|^{N_\gamma} \\ &= \left(2 \left\lceil \frac{k}{\gamma} \right\rceil + 1 \right)^{N_\gamma}. \end{aligned}$$

It follows (by scaling γ) that

$$\mathcal{N}_\infty\left(\mathcal{F}_R^k, \frac{\gamma}{4}, 2m\right) \leq \left(2\left\lceil\frac{4k(R+1)}{\gamma}\right\rceil + 1\right)^{N_{\gamma/(4(R+1))}}.$$

The result follows on noting that

$$2\left\lceil\frac{4k(R+1)}{\gamma}\right\rceil + 1 \leq 2\left(\frac{4k(R+1)}{\gamma} + 1\right) + 1 \leq \frac{11k(R+1)}{\gamma},$$

noting that $k(R+1)/\gamma \geq 1$. □

5. A generalization error bound for prototype-based classifiers

We now come to our main result. To keep it fairly general, we will work with two decreasing sequences of positive numbers: $(\delta_i)_{i=1}^\infty$ and $(\alpha_i)_{i=1}^\infty$, which are such that $\sum_{i=1}^\infty \delta_i = \delta$ and $\sum_{i=1}^\infty \alpha_i = 1$. For example, we could take $\delta_i = 6\delta/(\pi^2 i^2)$ and $\alpha_i = 6/(\pi^2 i^2)$.

Theorem 5.1. *Let P be any probability distribution on $Z = \mathcal{X} \times \{-1, 1\}$. For all m, the following holds with P^m probability at least $1 - \delta$ for a sample $\mathbf{z} \in Z^m$ randomly drawn according to P^m:*

–for any positive integer k,

– for any $R \geq 1$,

– for any $\gamma \in (0, k]$,

if Π is any set of k prototypes such that $\sum_{i=1}^k \frac{1}{r_i} \leq R$ and if $f = f_\Pi$ is the corresponding discriminator (and h_Π the corresponding classifier) then

$$\mathrm{er}_P(f) = \mathrm{er}_P(h) \leq 3\,\mathrm{er}^{\gamma}_{\mathbf{z}}(f) + \frac{4}{m}\left(N \ln\left(\frac{11k(2R+1)}{\gamma}\right) + \ln\left(\frac{4k}{\gamma\alpha_k\delta_{\lfloor \log_2(2R)\rfloor}}\right)\right)$$

where $N = \mathcal{N}\left(\mathcal{X}, \frac{\gamma}{4(2R+1)}, d\right)$.

As mentioned in section 1, the classifier h is non-parametric and the variables R and γ above are not parameters that need to be learnt. Rather, they are variables that describe its properties, and are measured after learning it.

Proof: For the moment, fix R and k and let E^k_R (a subset of Z^m) be the event that for $\mathbf{z} \in Z^m$, the following holds: for some $\gamma \in (0, k]$, there exists $f = f_\Pi$ where $|\Pi| = k$ and $R(\Pi) \leq R$, such that

$$\mathrm{er}_P(f) > 3\,\mathrm{er}^{\gamma}_{\mathbf{z}}(f) + \epsilon(k, 2R, m, \gamma, \delta_{\lfloor \log(2R)\rfloor}),$$

where

$$\epsilon(k, R, m, \gamma, \delta) = \frac{4}{m}\left(N \ln\left(\frac{11k(R+1)}{\gamma}\right) + \ln\left(\frac{4k}{\gamma\delta\alpha_k}\right)\right),$$

where $N = \mathcal{N}\left(\mathcal{X}, \frac{\gamma}{4(R+1)}, d\right)$. Then what we want to establish is that

$$P^m\left(\bigcup_{R\geq 1, k\geq 1} E^k_R\right) \leq \delta.$$

Next, we note that, for fixed R and k, Theorems 3.1 and 4.2 show that with probability no more than $\delta\alpha_k$, there will be some $\gamma \in (0, k]$ and some $f \in \mathcal{F}^k_R$ with

$$\mathrm{er}_P(f) > 3\,\mathrm{er}^{\gamma}_{\mathbf{z}}(f) + \epsilon(k, R, m, \gamma, \delta).$$

In particular, let $D^k_i \subseteq Z^m$ be the event that for $\mathbf{z} \in Z^m$, there exists $\gamma \in (0, k]$ and $f \in \mathcal{F}^k_{2^i}$ with

$$\mathrm{er}_P(f) > 3\,\mathrm{er}^{\gamma}_{\mathbf{z}}(f) + \epsilon(k, 2^i, m, \gamma, \delta_i).$$

Then it follows that $P^m(D^k_i) \leq \delta_i\alpha_k$.

Next, we claim that, for all $R \in [2^{i-1}, 2^i)$, $E_R^k \subseteq D_i^k$. First, we note that, clearly, if $R(\Pi) \in [2^{i-1}, 2^i)$, then $f = f_\Pi \in \mathcal{F}_{2^i}^k$. Next, if $R \in [2^{i-1}, 2^i)$, then $i = \lfloor \log(2R) \rfloor$, so that $\delta_{\lfloor \log(2R) \rfloor} = \delta_i$. Furthermore, given this, and since $R \in [2^{i-1}, 2^i)$ implies $2R \geq 2^i$, we have

$$\epsilon(k, 2R, m, \gamma, \delta_{\lfloor \log(2R) \rfloor}) \geq \epsilon(k, 2^i, m, \gamma, \delta_i).$$

Put together, these observations imply $E_R^k \subseteq D_i^k$.

We therefore have

$$\begin{aligned}
P^m\left(\bigcup_{k=1}^{\infty}\bigcup_{R\geq 1} E_R^k\right) &\leq \sum_{k=1}^{\infty} P^m\left(\bigcup_{R\geq 1} E_R^k\right) \\
&= \sum_{k=1}^{\infty} P^m\left(\bigcup_{i=1}^{\infty}\bigcup_{R\in[2^{i-1},2^i)} E_R^k\right) \\
&\leq \sum_{k=1}^{\infty} P^m\left(\bigcup_{i=1}^{\infty} D_i^k\right) \\
&\leq \sum_{k=1}^{\infty}\sum_{i=1}^{\infty} P^m(D_i^k) \\
&\leq \sum_{k=1}^{\infty}\sum_{i=1}^{\infty} \delta_i \alpha_k \\
&= \sum_{k=1}^{\infty} \alpha_k \left(\sum_{i=1}^{\infty} \delta_i\right) \\
&= \sum_{k=1}^{\infty} \alpha_k \delta = \delta.
\end{aligned}$$

□

Note that, in Theorem 5.1, Π could itself be taken to depend on the sample, for the result holds uniformly over the class of all f_Π. For instance, Π could be a subset of the sample, or could be a set of prototypes derived from the

sample by a clustering algorithm, for instance k-means, or by editing methods which start with the sample and produce a reduced set of more important points in the sample to be used as prototypes [15, 11].

As mentioned earlier, the Theorem is stated for general α_i, δ_i. By way of illustration, we state the special case corresponding to $\delta_i = 6\delta/(\pi^2 i^2)$ and $\alpha_i = 6/(\pi^2 i^2)$.

Corollary 5.2. *Let P be any probability distribution on $Z = \mathcal{X} \times \{-1, 1\}$. For all m, the following holds with P^m probability at least $1 - \delta$ for a sample $\mathbf{z} \in Z^m$: for any positive integer k, for any $R \geq 1$, for any $\gamma \in (0, k]$, if Π is any set of k prototypes such that $\sum_{i=1}^{k} \frac{1}{r_i} \leq R$ and if h_Π is the corresponding classifier, then*

$$\mathrm{er}_P(h_\Pi) \leq 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f_\Pi) + \frac{4}{m}\left(N \ln\left(\frac{11k(2R+1)}{\gamma}\right) + \ln\left(\frac{\pi^4 k^3 (\log(2R))^2}{9\gamma\delta}\right)\right)$$

where $N = \mathcal{N}\left(\mathcal{X}, \frac{\gamma}{4(2R+1)}, d\right)$.

Suppressing constants, this error bound is of the form

$$\mathrm{er}_P(f) \leq 3\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + O\left(\frac{1}{m}\left(N \ln\left(\frac{kR}{\gamma\delta}\right)\right)\right).$$

As another corollary to Theorem 5.1, since that result holds uniformly for all γ, we may take γ to be such that the error $\mathrm{er}_{\mathbf{z}}^{\gamma}(f_\Pi)$ is 0. We have:

Corollary 5.3. *Let P be any probability distribution on $Z = \mathcal{X} \times \{-1, 1\}$. For all m, the following holds with P^m probability at least $1 - \delta$ for a sample $\mathbf{z} \in Z^m$: for any positive integer k, for any $R \geq 1$, if Π is any set of k prototypes such that $\sum_{i=1}^{k} \frac{1}{r_i} \leq R$ and if $f = f_\Pi$ is the corresponding discriminator (and h_Π the corresponding classifier) then, for all $\gamma \in (0, k]$ such that*

$\mathrm{er}_{\mathbf{z}}^{\gamma}(h_{\Pi}) = 0$, *we have*

$$\mathrm{er}_P(h_{\Pi}) \leq \frac{4}{m}\left(N \ln\left(\frac{11k(2R+1)}{\gamma}\right) + \ln\left(\frac{4k}{\gamma \alpha_k \delta_{\lfloor \log_2(2R) \rfloor}}\right)\right) \tag{2}$$

where $N = \mathcal{N}\left(\mathcal{X}, \frac{\gamma}{4(2R+1)}, d\right)$.

The sample margin is the maximum value of γ such that $\mathrm{er}_{\mathbf{z}}^{\gamma}(f_{\Pi})$ is 0. We can compare the above bound with existing results on large-margin learning with prototypes. The paper [9] studies the problem of learning vector quantization (LVQ) in $\mathbb{R}^n$ where a point x is classified by the label of the nearest prototype to it (winner-take-all). In the current paper, the classifier h_{Π} decides based on contribution from all the prototypes whose sphere of influence contains x. With respect to the number k of prototypes [9] states a generalization learning-error bound which is $O(k\sqrt{\log k})$, compared to the $O(\ln(k))$ error bound of Corollary 5.2 which is exponentially smaller and yet holds for a much more general learning setting, applicable to *any* metric-space. In [9] the error bound depends on the dimension n and the margin parameter γ as $O\left(\min\left(n, 1/\gamma^2\right)\right)$ while in current paper, the dependence is $O\left(\mathcal{N}(\mathcal{X}, \gamma, d) \ln(1/\gamma)\right)$. In [13], an error bound is obtained for LVQ in a Euclidean space with squared Euclidean distance weighted by relevance parameters. With respect to the number k of prototypes, their generalization bound is $O(k^2)$, and is inversely proportional to the margin parameter γ.

6. Conclusions

We have studied the generalization error of classifiers defined in a particular way by a set of labeled prototypes in a metric space. The classifiers, through the use of a discriminant function, take into account the proximity to the prototypes of a point to be classified. Those prototypes involved in the classification of a point are those whose sphere of influence contains it, where the sphere of influence is the largest sphere centred on the prototype which contains no oppositely-labeled prototypes. Each of these prototypes then influences the classification of the point in a way that depends on how close the point is to the prototype, relative to the radius of its sphere of influence.

We have obtained bounds on the generalization error that involve the margin-based error on a training sample. One implication of the bounds is that it appears to be advantageous to use a classifier which involves a small number of prototypes and 'definitively' classifies the points of the training sample (in the sense that the discriminant takes large absolute value on the each sample point, which would be the case if, for instance, a sample point was deeply embedded within a large number of spheres corresponding to prototypes of a particular classification). We have worked in the context of a general metric space, but in future work would wish to investigate weakening this assumption (dealing with a general 'dissimilarity' measure d, which need not be a metric).

Acknowledgements

This work was supported in part by a research grant from the Suntory and Toyota International Centre for Economics and Related Disciplines at the London School of Economics. We are very grateful to the referees for their valuable comments.

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, 2000: 213–225.

[3] M. Anthony and N. L. Biggs (1992). *Computational Learning Theory: An Introduction*. Cambridge Tracts in Theoretical Computer Science, 30. Cambridge University Press, Cambridge, UK.

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

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

[6] 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), 1998: 525–536.

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

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

[9] 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]*, p. 462–469, 2002.

[10] N. Cristianini and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK.

[11] G.W. Gates. The reduced nearest neighbor rule. *IEEE Transactions on Information Theory* 18 (3), 1972: 431–433.

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

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

[14] B. Hammer, D. Hofmann, F-M. Schleif, and X. Zhu. Learning vector quantization for (dis-)similarities. *Neurocomputing* 131: 43–51, 2014.

[15] P. Hart. The condensed nearest neighbour rule (Corresp.). *IEEE Transactions on Information Theory* 14 (3), 1968: 515–516.

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

[17] D. L. Reilly and L. N. Cooper. *An Overview of Neural Networks: Early Models to Real World Systems*. Academic Press, San Diego, 1990.

[18] 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.

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

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

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