A data-dependent skeleton estimate and a scale-sensitive dimension for classification

Márta Horváth
Department of Mathematics
and Computer Science,
Technical University of Budapest,
1521 Stoczek u. 2, Budapest, Hungary.
email: s4094pin@sun10.vsz.bme.hu.

Gábor Lugosi
Department of Economics,
Pompeu Fabra University,
Ramon Trias Fargas, 25-27,
08005 Barcelona, Spain.
email: lugosi@upf.es

December, 1996

Abstract

The classical binary classification problem is investigated when it is known in advance that the posterior probability function (or regression function) belongs to some class of functions. We introduce and analyze a method which effectively exploits this knowledge. The method is based on minimizing the empirical risk over a carefully selected "skeleton" of the class of regression functions. The skeleton is a covering of the class based on a data-dependent metric, especially fitted for classification. A new scale-sensitive dimension is introduced which is more useful for the studied classification problem than other, previously defined, dimension measures. This fact is demonstrated by performance bounds for the skeleton estimate in terms of the new dimension. [1]

[1] Parts of the paper were presented at COLT'96 [15].

1 Introduction

The following pattern classification problem is investigated: let (X, Y) be a pair of random variables, taking their values from some set $\mathcal{X}$ and $\{0,1\}$, respectively. The value of the *label* Y is to be predicted upon observing the *feature vector* X. The *prediction rule* or *classifier* g is a function $\mathcal{X} \to \{0,1\}$, whose performance is measured by the *probability of error*

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

The joint distribution of (X, Y) is determined by the *regression function*

$$\eta^*(x) = \mathbf{P}\{Y = 1 | X = x\}$$

(also known as the *a posteriori probability function*) and the measure μ of X on $\mathcal{X}$, that is,

$$\mu(A) = \mathbf{P}\{X \in A\} \quad \text{for each measurable set } A \subseteq \mathcal{X}.$$

The Bayes classifier

$$g^*(x) = \begin{cases} 0 & \text{if } \eta^*(x) < 1/2 \\ 1 & \text{otherwise} \end{cases}$$

is well-known to have minimal probability of error among all possible classifiers. Its error probability $L(g^*)$ is called the Bayes risk, and is denoted by L^*.

Recall that if $\eta : \mathcal{X} \to [0,1]$ is an arbitrary measurable function, and we define the corresponding classifier by

$$g(x) = \begin{cases} 0 & \text{if } \eta(x) < 1/2 \\ 1 & \text{otherwise,} \end{cases}$$

then the following elementary property holds:

$$\begin{aligned} L(g) - L^* &= 2\mathbf{E}\left\{ I_{\{g(X) \neq g^*(X)\}} \left| \eta^*(X) - \frac{1}{2} \right| \right\} \\ &\leq 2\mathbf{E}\left\{ I_{\{g(X) \neq g^*(X)\}} \left| \eta^*(X) - \eta(X) \right| \right\}, \end{aligned}$$

see, for example, [10, p.16]. (I_A denotes the indicator of an event A.)

Assume that n independent copies of (X, Y) form the available data sequence:

$$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n)).$$

These data may be used to obtain the classification rule $g_n(x)$, whose probability of error is the random variable

$$L(g_n) = \mathbf{P}\{g_n(X) \neq Y | D_n\}.$$

Very often, apart from the training sequence, some prior information is available about the joint distribution of (X, Y). For example, in some applications with $\mathcal{X} = \mathcal{R}^d$, it is known that η^* is a monotone function in all components of x. In other situations it may be known that η^* is a smooth function. In the basic PAC-learning setup [8], η^* is known to be the indicator function of one of the sets in a given class of sets. We assume throughout that η^* is a member of a known class of functions $\mathcal{F}$. In the next section we present a classification rule that first forms a finite skeleton of $\mathcal{F}$ based on a part of the training data, and then uses the other half of the data to select the empirically best candidate from the skeleton.

An upper bound for the performance of the skeleton estimate is given in Theorem 1. The bound is formulated in terms of some covering numbers of $\mathcal{F}$, specifically suited for the classification problem.

In Section 3 we introduce a new "scale-sensitive" dimension for classes of functions, and relate it to the covering numbers appearing in Theorem 1. The new dimension is closely related to a dimension introduced by Kearns and Shapire [13] whose usefulness have been demonstrated for learning "probabilistic concepts" and for more general regression function estimation problems, see also Alon, Ben-David, Cesa-Bianchi, and Haussler [1], Bartlett, Long, and Williamson [6], Bartlett and Long [5], Bartlett [4], Anthony and Bartlett [2], Shawe-Taylor, Bartlett, Williamson, and Anthony [19]. Other dimensions of similar type have also been introduced, see [1] for a survey. However, these dimensions are not quite adequate for the binary classification problem as they do not capture the particular properties of the classification problem when the probability of error is used as the measure of loss. We point out in Section 3 that the new dimension is more useful in the particular situation we are investigating.

2 A data-based skeleton estimate

In this section we describe the proposed classification rule. First, the data sequence D_n is split into two parts:

$$D_m = ((X_1, Y_1), \ldots, (X_m, Y_m))$$

and

$$T_{n-m} = ((X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)).$$

The first part, D_m, is used to create a skeleton of $\mathcal{F}$, that is, a finite subset $\mathcal{F}_\epsilon$ of $\mathcal{F}$ such that each member of $\mathcal{F}$ is closely approximated by a function from $\mathcal{F}_\epsilon$ in a carefully chosen metric. More precisely, given $\epsilon > 0$, let $\mathcal{F}_\epsilon \subset \mathcal{F}$ be a class of functions of minimal cardinality satisfying the property that for each $\eta \in \mathcal{F}$ there exists an $\bar{\eta} \in \mathcal{F}_\epsilon$ such that

$$\frac{1}{m} \sum_{i=1}^{m} 2 I_{\{g(X_i) \neq \bar{g}(X_i)\}} |\eta(X_i) - \bar{\eta}(X_i)| < \epsilon, \tag{1}$$

where

$$g(x) = \begin{cases} 0 & \text{if } \eta(x) < 1/2 \\ 1 & \text{otherwise} \end{cases}$$

and

$$\bar{g}(x) = \begin{cases} 0 & \text{if } \bar{\eta}(x) < 1/2 \\ 1 & \text{otherwise} \end{cases}$$

are the corresponding classifiers. The second part of the data, T_{n-m}, is used to test all classifiers defined by functions in $\mathcal{F}_\epsilon$, and to select one with minimal empirical error. In other words, we select the classifier $\bar{g} = I_{\{\bar{\eta} \geq 1/2\}}$ with $\bar{\eta} \in \mathcal{F}_\epsilon$, if

$$L_{n-m}(\bar{g}) = \frac{1}{n-m} \sum_{i=m+1}^{n} I_{\{\bar{g}(X_i) \neq Y_i\}}$$

is minimal among all rules in $\mathcal{F}_\epsilon$. The choice of the metric used in the empirical covering is motivated by (1). We take the liberty of using $L(g)$ and $L(\eta)$ (and similarly $L_{n-m}(g)$ and $L_{n-m}(\eta)$) interchangeably. Denote the obtained classification rule by g_n (and the corresponding regression function by η_n). Note that g_n ignores the values of $Y_1, \ldots, Y_m$, and therefore it may make efficient use of additional unlabeled samples, if available. The first half of the sample is only used to obtain information about μ. We have the following key property.

Theorem 1 *Assume that $\eta^* \in \mathcal{F}$. Then for all n, m, ϵ, and $\delta \geq 6\epsilon$,*

$$\mathbf{P}\{L(g_n) - L^* > 3\delta\} \leq (\mathbf{E}\{|\mathcal{F}_\epsilon|\} + 1)\, e^{-\frac{3}{8}(n-m)\delta^2/(L^*+3\delta)} + 8\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\} e^{-m\delta/256}.$$

Corollary 1 *Assume that $\eta^* \in \mathcal{F}$. If $m = n/2$, then for all n*

$$\mathbf{E}\{L(g_n)\} - L^* \leq \max\left(\sqrt{\frac{6L^* \log\left(8n\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\}\right)}{n}}, \frac{512 \log\left(8n\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\}\right)}{n} \right) + \frac{1}{n}$$

provided that ϵ satisfies the inequality

$$\epsilon \leq \frac{1}{6} \max\left(\sqrt{\frac{6L^* \log\left(8n\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\}\right)}{n}}, \frac{512 \log\left(8n\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\}\right)}{n} \right).$$

Proofs are given in Section 4.

Thus, the rate of convergence of the error of the selected classifier is determined by the logarithm of the expected value of the covering number $|\mathcal{F}_\epsilon|$. The condition on ϵ is always satisfied if we take $\epsilon = 85/n$, but this may not be a good choice in some situations: if $|\mathcal{F}_\epsilon|$ is very large, much larger values of ϵ may be advantageous. Better values of ϵ may be obtained by bounding the covering numbers. However, is not the purpose of this paper to explore this direction in depth. The main message is that the expected size of the error is of the order of $\sqrt{L^* \log\left(\mathbf{E}\left\{|\mathcal{F}_\epsilon|\right\}\right)/n}$, unless the Bayes error L^* is very small (of the order of $\log\left(\mathbf{E}\left\{|\mathcal{F}_\epsilon|\right\}\right)/n$), in which case an even smaller bound, of the order of $\log\left(\mathbf{E}\left\{|\mathcal{F}_\epsilon|\right\}\right)/n$ is achievable.

It is worth comparing $\mathbf{E}\left\{|\mathcal{F}_\epsilon|\right\}$ to other quantities relevant for analyzing the performance of certain classification rules. Perhaps the most natural way of selecting a classifier from a class is empirical risk minimization: one selects a classifier g_n by minimizing, over all $g = I_{\{\eta>1/2\}}$, $\eta \in \mathcal{F}$, the empirical error $(1/n)\sum_{i=1}^n I_{\{g(X_i)\neq Y_i\}}$. Then an inequality of Vapnik and Chervonenkis [20] (see also [3]) implies a bound analogous to that of Theorem 1, but with $\mathbf{E}\left\{|\mathcal{F}_\epsilon|\right\}$ replaced by $\mathbf{E}\left\{S_{\mathcal{F}}(n)\right\}$, where $S_{\mathcal{F}}(n)$ is the random shatter coefficient of the class $\mathcal{F}$, that is, the number of different ways the members of $\mathcal{F}$ can classify the n i.i.d. random variables $X_1, \ldots, X_n$. But it is easy to see that for all $\epsilon \geq 2/m$,

$$\mathbf{E}\left\{|\mathcal{F}_\epsilon|\right\} \leq \mathbf{E}\left\{S_{\mathcal{F}}(m)\right\}.$$

(Note that the recommended choice for ϵ is always larger, and in some cases much larger than $2/m$.) Thus, for the data-dependent skeleton classifier, Theorem 1 guarantees better performance than the above inequalities for empirical risk minimization. The reason why it is possible to improve on these bounds is that one can make use of the additional information provided by the knowledge of the form of the possible regression functions. That the improvement may be essential, can be seen from the simple example in which $\mathcal{F}$ is the class of all Lipschitz functions $\eta : [0,1]^d \to [0,1]$. Then it is easy to see that for all distributions $\log\left(\mathbf{E}\left\{|\mathcal{F}_\epsilon|\right\}\right) = O\left(\epsilon^{-d}\right)$, while for each absolutely continuous distribution of X, $\mathbf{E}\left\{S_{\mathcal{F}}(n)\right\} = 2^n$ for each n. For example, in Theorem 1, ϵ may be chosen of the order of $n^{-1/(2+d)}$, which results in an upper bound of the order of $n^{-1/(2+d)}$ for the error of the data-dependent skeleton estimate. At the same time, no nontrivial bound can be obtained for empirical risk minimization.

We must remark here, however, that in order to obtain Theorem 1, we needed to assume that $\eta^* \in \mathcal{F}$, that is, the "true" regression function is the member of the class $\mathcal{F}$. The bounds obtained for empirical risk minimization do not require this assumption. Empirical risk minimization is, therefore, much more robust than the skeleton estimate introduced here.

Another approach to the classification problem is to directly estimate the regression function η^* by a function η_n. For the probability of error of the corresponding classifier $g_n(x) = I_{\{\eta_n(x) \geq 1\}}$, we have from (1) that

$$L(g_n) - L^* \leq 2\mathbf{E}\left\{ |\eta^*(X) - \eta_n(X)| \,|\, D_n \right\}.$$

The bounds for the probability of error of classifiers based on the empirical squared or L_1 error typically involve L_1 or related covering numbers, see, for example, [2, 4, 1, 16], and [10, Chapters 28,29]. In all cases, $\mathbf{E}\{|\mathcal{F}_\epsilon|\}$ compares favourably to these covering numbers. The reason is the presence of the factor $I_{\{g(X_i) \neq \bar{g}(X_i)\}}$ in the definition of the distance (1) according to which the covering number $|\mathcal{F}_\epsilon|$ is defined. Instead of detailing such (trivial) inequalities, we refer to Section 3, where $|\mathcal{F}_\epsilon|$ is estimated in terms of a new dimension. Bounds for the probability of error obtained through bounds for regression function estimation are generally loose, and results of function learning and regression function estimation have little to say in this situation. Consider, for example, the following trivial case: let $\mathcal{F}$ contain *all* functions $\eta : \mathcal{R} \to [0,1]$ such that $\eta(x) < 1/2$ if $x < 0$ and $\eta(x) \geq 1/2$ if $x \geq 0$. Then there is only one classifier induced by these functions, and accordingly, $|\mathcal{F}_\epsilon| = 1$. On the other hand, this class is clearly too large for obtaining meaningful bounds for regression function estimation. In the next section we quantify this observation in terms of a dimension of $\mathcal{F}$, which is always smaller than those that have been proved useful in the regression function estimation scenario.

3 A new scale-sensitive dimension

Next we define a dimension for a class $\mathcal{F}$ of functions $\mathcal{X} \to [0,1]$.

Definition 1 *Let $0 < \gamma \leq 1/2$. We say that $\mathcal{F}$* γ-shatters *a finite set $A \subset \mathcal{X}$ if there exists some function $s : A \to [(1-\gamma)/2, (1+\gamma)/2]$ such that for every subset $E \subset A$ there is a function $\eta_E \in \mathcal{F}$ such that $\eta_E(x) \geq s(x) + \gamma$ if $x \in E$ and $\eta_E(x) \leq s(x) - \gamma$ if $x \in A - E$. The largest positive integer n for which there exists a set A of cardinality n which is γ-shattered by $\mathcal{F}$ is denoted by d_γ. If for every n there is a set A which is γ-shattered by $\mathcal{F}$ then we say that $d_\gamma = \infty$. The* γ-dimension *of $\mathcal{F}$ is defined as $D_\gamma = \inf_{\epsilon \leq \gamma} d_\epsilon$.*

Our main result concerning D_γ is an upper bound for the covering numbers appearing in Theorem 1 in terms of the γ-dimension:

Theorem 2 *For any value of* $X_1, \ldots, X_m$*, if* $36m \geq \lceil 4/\epsilon \rceil$*,* $\epsilon \leq 1$*, then*

$$|\mathcal{F}_\epsilon| \leq 2(72m)^{D_{\epsilon/5} \log_2(2m)}.$$

The proof of this theorem is based on the proof of Lemma 3.4 of Alon, Ben-David, Cesa-Bianchi, and Haussler [1], and it is given in Section 5.

We may combine the above result with Theorem 1 to obtain the following sample-size bound for the data-dependent skeleton estimate defined in Section 2:

Corollary 2 *Assume that* $\eta^* \in \mathcal{F}$*. Let* $m = \lfloor n/2 \rfloor$*. For any* $\epsilon, \delta > 0$*,*

$$\mathbf{P}\{L(g_n) - L^* > \epsilon\} \leq \delta$$

if

$$n = O\left(\left(D_{\epsilon/80} \log^2\left(\frac{D_{\epsilon/80}}{\epsilon}\right) + \log\frac{1}{\delta}\right) \max\left(\frac{L^*}{\epsilon^2}, \frac{1}{\epsilon}\right)\right).$$

d_γ is closely related to the so-called P_γ-dimension introduced by Kearns and Shapire [13]. The only difference is that in the definition of P_γ the range of the shattering function s is not restricted, it can take any value in $[0,1]$. Therefore, clearly, for every γ,

$$D_\gamma \leq d_\gamma \leq P_\gamma.$$

In fact, D_γ may be finite even if $P_\gamma = \infty$ for every γ. (Just consider the class of all functions $\eta : \mathcal{R} \to [0,1]$ such that $\eta(x) < 1/2$ if $x \leq 0$ and $\eta(x) \geq 1/2$ if $x > 0$.) The restriction of the range of s is motivated by the fact that from the point of view of classification, only the behavior of the functions in $\mathcal{F}$ around 1/2 matters. For some discussion on this we refer to Section 6.7 of [10]. It is clear that P_γ is a monotonically decreasing function of γ. Because of the restriction of the range of s in the definition of γ-shattering, as the next example shows, this monotonicity is no longer true for d_γ, which justifies the definition of the γ-dimension D_γ. D_γ is obviously monotonically decreasing in γ.

Example. Let $0 < a < 1/3$, and let $\mathcal{F}$ be the class of all functions $\eta : \mathcal{R} \to [0,1]$ such that $\eta(x) > 1/2 + 3a/2$ if $x > 0$ and $\eta(x) < 1/2 - a/2$ if $x \leq 0$. Then it is clear that if $\gamma \leq a$, then $d_\gamma < 2$, while for $\gamma > a$, $d_\gamma = \infty$. □

We define the VC dimension V of $\mathcal{F}$ as the VC dimension of the class of classifiers induced by $\mathcal{F}$, that is, as the VC dimension of the class of sets of the form $\{x : \eta(x) > 1/2\}$, $\eta \in \mathcal{F}$. Then clearly,

$$D_\gamma \leq V$$

for each γ. Again, D_γ may be finite even if $V = \infty$. As a simple example, consider the class of Lipschitz functions on $[0,1]$. Then $D_\gamma \leq P_\gamma = O(1/\gamma)$, but obviously $V = \infty$. (Note that V is different from the "V-dimension" discussed in [1].) In a distribution-free setting, V basically describes the minimax behavior of the probability of error, see, for example, [20, 10]. For example, it is shown by Vapnik and Chervonenkis [20] that there exists a classification rule g_n and a constant c such that for any distribution of the pair (X, Y),

$$\mathbf{E}L(g_n) - \inf_{\eta \in \mathcal{F}} L(\eta) \leq c \max\left(\sqrt{\frac{V \inf_{\eta \in \mathcal{F}} L(\eta) \log n}{n}}, \frac{V \log n}{n}\right).$$

On the other hand, for any classification rule g_n, there exists a distribution of (X, Y) such that

$$\mathbf{E}L(g_n) - \inf_{\eta \in \mathcal{F}} L(\eta) \geq c' \max\left(\sqrt{\frac{V \inf_{\eta \in \mathcal{F}} L(\eta)}{n}}, \frac{V}{n}\right)$$

for some other constant c', see [11, 10]. The reason why we can improve on the above upper bound is that we no longer work in a completely distribution-free setting but we assume $\eta^* \in \mathcal{F}$, and we are able to exploit this additional information.

It is easy to see that no universal relationship exists between V and P_γ. In fact, if $\mathcal{F}$ is the class of all functions on $\mathcal{R}$ whose value is in $[0,1/2]$ if $x < 0$ and in $(1/2, 1]$ if $x > 0$ then $V = 1$ but $P_\gamma = \infty$ for every γ. On the other hand, if $\mathcal{F}$ contains every function defined on the positive integers such that $|\eta(x) - 1/2| \leq e^{-x}/2$, then $P_\gamma = \lfloor -\log(2\gamma) \rfloor$, but $V = \infty$. (The latter example is taken from [1].) Since $D_\gamma \leq \max(V, P_\gamma)$, we may interpret the new dimension as one that unifies the advantages of V and the scale-sensitive dimension P_γ.

Finally, we indicate some points where the bound of Theorem 2 is lose. First of all, the theorem provides an upper bound for the maximal possible value of $|\mathcal{F}_\epsilon|$, whereas the interesting quantity in Theorem 1 is its expected value $\mathbf{E}\{|\mathcal{F}_\epsilon|\}$. In certain cases, the difference may be significant: [10, Theorem 13.13] provides such an example.

If $\mathcal{F}$ is a class of indicator functions, then $|\mathcal{F}_\epsilon|$ is just the random shatter coefficient $S_{\mathcal{F}}(m)$ for $\epsilon = 2/m$. In this case Sauer's lemma [18] implies that $\log|\mathcal{F}_\epsilon| \leq V \log m$, whereas Theorem 2 only gives $\log|\mathcal{F}_\epsilon| = O(V \log^2 m)$.

If $\mathcal{F}$ is the class of all Lipschitz functions (with Lipschitz constant 1) on $[0,1]$, then it is easy to see that $\log|\mathcal{F}_\epsilon| = O(1/\epsilon)$. However, it is also easy to see that $D_\epsilon = O(1/\epsilon)$, and therefore Theorem 2 only implies $\log|\mathcal{F}_\epsilon| = O\left(\frac{\log^2 m}{\epsilon}\right)$. However, the practical importance of the log factors is minor, so Theorem 2 may be a useful tool to bound $|\mathcal{F}_\epsilon|$.

4 Proof of Theorem 1

In the proof of Theorem 1 we apply an inequality of Pollard [17], sharpened by Haussler [12]. In particular, the following corollary is used, which was obtained by Buescher and Kumar [9] in a slightly different form. The form given here is found in Lugosi and Nobel [14].

Lemma 1 *Let $\mathcal{H}$ be class of functions on $\mathcal{X}$ such that $h(x) \in [0, A]$ for every $h \in \mathcal{H}$ and every $x \in \mathcal{X}$. Let $X_1, \ldots, X_m \in \mathcal{X}$ be i.i.d. random vectors. Then for each $\delta > 0$ and, $\epsilon > 0$,*

$$\mathbf{P}\left\{ \sup_{h \in \mathcal{H}: \frac{1}{m}\sum_{i=1}^{m} h(X_i) < \epsilon} \mathbf{E}\{h(X)\} > \delta + 3\epsilon \right\} \leq 4\mathbf{E}\left\{ |H_{\epsilon/16}| \right\} e^{-m(\delta+\epsilon)/(64A)},$$

where H_ϵ is any set of functions satisfying the property that for each $h \in \mathcal{H}$ there is a $h' \in \mathcal{H}_\epsilon$ with

$$\frac{1}{m}\sum_{i=1}^{m} |h(X_i) - h'(X_i)| < \epsilon.$$

The basic idea of the next lemma may be found in Vapnik and Chervonenkis [20].

Lemma 2 *Let $\widehat{\eta}$ be an element of $\mathcal{F}_\epsilon$ such that $L(\widehat{\eta}) = \min_{\eta \in \mathcal{F}_\epsilon} L(\eta)$. Then*

$$\begin{aligned} &\mathbf{P}\{L(g_n) - L(\widehat{\eta}) > 2\delta | D_m\} \\ &\leq \mathbf{P}\{L_{n-m}(\widehat{\eta}) - L(\widehat{\eta}) > \delta | D_m\} + \mathbf{P}\left\{ \max_{\eta \in \mathcal{F}_\epsilon} \frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \,\middle|\, D_m \right\}. \end{aligned}$$

Proof. If $L(g_n) - L(\widehat{\eta}) > 2\delta$, then there exists an $\eta \in \mathcal{F}_\epsilon$ such that $L(\eta) > L(\widehat{\eta}) + 2\delta$ and $L_{n-m}(\eta) \leq L_{n-m}(\widehat{\eta})$. Thus,

$$\begin{aligned} &\mathbf{P}\{L(g_n) - L(\widehat{\eta}) > 2\delta | D_m\} \\ &\leq \mathbf{P}\left\{ \min_{\eta: L(\eta) > L(\widehat{\eta}) + 2\delta} L_{n-m}(\eta) < L_{n-m}(\widehat{\eta}) \,\middle|\, D_m \right\} \\ &\leq \mathbf{P}\left\{ \min_{\eta: L(\eta) > L(\widehat{\eta}) + 2\delta} L_{n-m}(\eta) < L(\widehat{\eta}) + \delta \,\middle|\, D_m \right\} + \mathbf{P}\{L_{n-m}(\widehat{\eta}) > L(\widehat{\eta}) + \delta | D_m\}, \end{aligned}$$

so we need to show that

$$\mathbf{P}\left\{ \min_{\eta: L(\eta) > L(\widehat{\eta}) + 2\delta} L_{n-m}(\eta) < L(\widehat{\eta}) + \delta \,\middle|\, D_m \right\} \leq \mathbf{P}\left\{ \max_{\eta \in \mathcal{F}_\epsilon} \frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \,\middle|\, D_m \right\}. \tag{2}$$

But if

$$\max_{\eta \in \mathcal{F}_\epsilon} \frac{L(\eta) - L_{n-m}(\eta)}{\sqrt{L(\eta)}} \leq \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}},$$

then for each $\eta \in \mathcal{F}_\epsilon$

$$L_{n-m}(\eta) \geq L(\eta) - \delta\sqrt{\frac{L(\eta)}{L(\widehat{\eta}) + 2\delta}}.$$

If, in addition, η is such that $L(\eta) > L(\widehat{\eta}) + 2\delta$, then by the monotonicity of the function $x - c\sqrt{x}$ (for $c > 0$ and $x > c^2/4$),

$$L_{n-m}(\eta) \geq L(\widehat{\eta}) + 2\delta - \delta\sqrt{\frac{L(\widehat{\eta}) + 2\delta}{L(\widehat{\eta}) + 2\delta}} = L(\widehat{\eta}) + \delta,$$

and (2) follows. □

Proof of Theorem 1. Write

$$L(g_n) - L^* = \left(L(g_n) - \min_{\eta\in\mathcal{F}_\epsilon} L(\eta)\right) + \left(\min_{\eta\in\mathcal{F}_\epsilon} L(\eta) - L^*\right),$$

so that

$$\mathbf{P}\{L(g_n) - L^* > 3\delta\} \leq \mathbf{P}\left\{L(g_n) - \min_{\eta\in\mathcal{F}_\epsilon} L(\eta) > 2\delta\right\} + \mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon} L(\eta) - L^* > \delta\right\}. \quad (3)$$

First we bound the second probability. Introduce the notation

$$J_m(\eta) = \frac{2}{m}\sum_{i=1}^{m} I_{\{g(X_i)\neq g^*(X_i)\}} \left|\eta^*(X) - \eta(X)\right|,$$

and

$$J(\eta) = \mathbf{E}\{J_m(\eta)\} = 2\mathbf{E}\left\{I_{\{g(X)\neq g^*(X)\}} \left|\eta^*(X) - \eta(X)\right|\right\}$$

for all $\eta \in \mathcal{F}$. Recall that by (1), $L(\eta) - L^* \leq J(\eta)$, and that $\min_{\eta\in\mathcal{F}_\epsilon} J_m(\eta) < \epsilon$ by the definition of $\mathcal{F}_\epsilon$ and by the assumption $\eta^* \in \mathcal{F}$. Therefore,

$$\min_{\eta\in\mathcal{F}_\epsilon} L(\eta) - L^* \leq \min_{\eta\in\mathcal{F}_\epsilon} J(\eta) \leq \sup_{\eta\in\mathcal{F}: J_m(\eta)<\epsilon} J(\eta).$$

Therefore, for $\delta \geq 6\epsilon$,

$$\begin{aligned}
\mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon} L(\eta) - L^* > \delta\right\} &\leq \mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon} L(\eta) - L^* > \frac{\delta}{2} + 3\epsilon\right\} \\
&\leq \mathbf{P}\left\{\sup_{\eta\in\mathcal{F}: J_m(\eta)<\epsilon} J(\eta) > \frac{\delta}{2} + 3\epsilon\right\} \\
&\leq 4\mathbf{E}\left\{\left|\mathcal{F}_{\epsilon/16}\right|\right\} e^{-m\delta/256} \qquad (4)
\end{aligned}$$

by Lemma 1.

To bound the first probability on the right-hand side of (3), we apply Lemma 2. First, since given D_m, the conditional distribution of $(n-m)L_{n-m}(\widehat{\eta})$ is binomial with parameters $n-m$ and $L(\widehat{\eta})$, we have by Berstein's inequality [7] that

$$\mathbf{P}\{L_{n-m}(\widehat{\eta}) - L(\widehat{\eta}) > \delta | D_m\} \le e^{-(n-m)\delta^2/(2L(\widehat{\eta})+\frac{2}{3}\delta)}.$$

(Recall that $\widehat{\eta}$ minimizes the probability of error in $\mathcal{F}_\epsilon$.) On the other hand, clearly

$$\begin{aligned}
&\mathbf{P}\left\{ \max_{\eta\in\mathcal{F}_\epsilon} \frac{L(\eta)-L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta})+2\delta}} \right| \left. D_m \vphantom{\frac{\delta}{\sqrt{L}}} \right\} \\
&\le \; |\mathcal{F}_\epsilon| \max_{\eta\in\mathcal{F}_\epsilon} \mathbf{P}\left\{ \frac{L(\eta)-L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta})+2\delta}} \right| \left. D_m \vphantom{\frac{\delta}{\sqrt{L}}} \right\}.
\end{aligned}$$

For any fixed η, the probability on the right-hand side is zero if $\gamma \overset{\text{def}}{=} \delta/\sqrt{L(\widehat{\eta})+2\delta} > \sqrt{L(\eta)}$. Otherwise, if $\gamma \le \sqrt{L(\eta)}$, then again by Bernstein's inequality,

$$\begin{aligned}
&\mathbf{P}\left\{ \frac{L(\eta)-L_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta})+2\delta}} \right| \left. D_m \vphantom{\frac{\delta}{\sqrt{L}}} \right\} \\
&= \; \mathbf{P}\left\{ L(\eta)-L_{n-m}(\eta) > \gamma\sqrt{L(\eta)} \right| \left. D_m \right\} \le e^{-\frac{(n-m)\gamma^2 L(\eta)}{2L(\eta)+2\sqrt{L(\eta)}\gamma/3}} \\
&\le \; e^{-\frac{3}{8}(n-m)\gamma^2} \\
&= \; e^{-\frac{3}{8}(n-m)\delta^2/(L(\widehat{\eta})+2\delta)}.
\end{aligned}$$

Therefore, by Lemma 2,

$$\mathbf{P}\left\{ L(g_n) - L(\widehat{\eta}) > 2\delta |\, D_m \right\} \le \left(|\mathcal{F}_\epsilon|+1\right) e^{-\frac{3}{8}(n-m)\delta^2/(L(\widehat{\eta})+2\delta)}. \qquad (5)$$

Finally,

$$\begin{aligned}
&\mathbf{P}\left\{L(g_n) - L(\widehat{\eta}) > 2\delta\right\} \\
&\le \; \mathbf{P}\left\{ L(g_n) - L(\widehat{\eta}) > 2\delta |\, L(\widehat{\eta}) - L^* \le \delta\right\} + \mathbf{P}\left\{L(\widehat{\eta}) - L^* > \delta\right\} \\
&\le \; \left(\mathbf{E}\left\{|\mathcal{F}_\epsilon|\right\}+1\right) e^{-\frac{3}{8}(n-m)\delta^2/(L^*+3\delta)} + 4\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\} e^{-m\delta/256},
\end{aligned}$$

where we used (4) and (5). Collecting bounds, the proof is finished. □

PROOF OF COROLLARY 1. Assume $\delta \ge 6\epsilon$. Observe that if $\delta \ge L^*/93$, then

$$\mathbf{P}\{L(g_n) - L^* > 3\delta\} \le 8\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\} e^{-n\delta/512},$$

and otherwise

$$\mathbf{P}\{L(g_n) - L^* > 3\delta\} \le 8\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\} e^{-n\delta^2/(6L^*)}.$$

Thus, for any $u > 6\epsilon$,

$$\begin{aligned} \mathbf{E}L(g_n) - L^* &\le 3u + \mathbf{P}\{L(g_n) - L^* > 3u\} \\ &\le 3u + \max\left(8\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\} e^{-nu/512}, 8\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\} e^{-nu^2/(6L^*)}\right). \end{aligned}$$

Chosing

$$u = \max\left(\frac{512\log\left(8n\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\}\right)}{n}, \sqrt{\frac{6L^*\log\left(8n\mathbf{E}\left\{|\mathcal{F}_{\epsilon/16}|\right\}\right)}{n}}\right),$$

yields the corollary. □

5 Proof of Theorem 2

The line of the proof of Theorem 2 is analogous to that of Lemma 3.4 in [1]. Just like there, we also begin with "discretizing." First we introduce a discrete analogue of the γ-dimension, related to the "strong dimension" of [1].

Let b be a positive even integer, and let $\mathcal{G}$ be a class of functions $\mathcal{X} \to \{1, 2, \ldots, b\}$. We say that $\mathcal{G}$ *b-shatters* a finite set $A \subset \mathcal{X}$ according to a function $s : A \to \{b/2-1, b/2, b/2+1\}$ if to every subset E of A, there exists a function $f_E \in \mathcal{G}$ such that

$$f_E(x) \begin{cases} \le s(x) - 3 & \text{if } x \in E \\ \ge s(x) + 3 & \text{if } x \in A - E. \end{cases}$$

We say that $\mathcal{G}$ *b-shatters A* if $\mathcal{G}$ *b-shatters A* according to some s. The *b-dimension* Δ_b of $\mathcal{G}$ is the largest integer n such that there exists a set A, shattered by $\mathcal{G}$, with $|A| = n$. If there is no such larger integer, then we say that $\Delta_b = \infty$.

Let $\rho > 0$. The ρ-discretization of a function $\eta : \mathcal{X} \to [0, 1]$ is defined by $\eta^\rho(x) = \lfloor \eta(x)/\rho \rfloor$.

Lemma 3 *Let $\mathcal{F}^\gamma = \{\eta^\gamma : \eta \in \mathcal{F}\}$ denote the class of γ-discretizations of functions in $\mathcal{F}$. Assume that $\gamma = 1/b$ for some positive even integer b. If Δ_b denotes the b-dimension of $\mathcal{F}^\gamma$, then*

$$\Delta_b \le d_{2\gamma},$$

where $d_{2\gamma}$ is defined in Definition 1 for $\mathcal{F}$.

PROOF. We show that if $\mathcal{F}^\gamma$ b-shatters a set A, then $\mathcal{F}$ 2γ-shatters A. Let $s : A \to \{b/2-1, b/2, b/2+1\}$ be the function which is used by $\mathcal{F}^\gamma$ to b-shatter A. Then for every $E \subset A$ there is a function $\eta_E \in \mathcal{F}$ such that

$$\eta_E^\gamma(x) \left\{ \begin{array}{ll} \leq s(x) - 3 & \text{if } x \in E \\ \geq s(x) + 3 & \text{if } x \in A - E. \end{array} \right.$$

Then clearly,

$$\eta_E(x) \left\{ \begin{array}{ll} \leq \gamma s(x) - 2\gamma & \text{if } x \in E \\ \geq \gamma s(x) + 3\gamma & \text{if } x \in A - E. \end{array} \right.$$

Then with $s'(x) = \gamma s(x)$, $\mathcal{F}$ clearly 2γ-shatters A. □

Next we relate $|\mathcal{F}_\epsilon|$ to certain packing numbers of the class of discretizations of functions in $\mathcal{F}$. Let $\{x_1, \ldots, x_m\} \subset \mathcal{X}$. We say that a subset $\mathcal{F}'$ of $\mathcal{F}$ is *ϵ-separated* if for any $\eta_1, \eta_2 \in \mathcal{F}'$,

$$\frac{1}{m} \sum_{i=1}^{m} 2 I_{\{g_1(x_i) \neq g_2(x_i)\}} |\eta_1(x_i) - \eta_2(x_i)| \geq \epsilon.$$

The maximal size $M(\epsilon, \mathcal{F})$ of such an ϵ-separated set is called the ϵ-packing number of $\mathcal{F}$. Now consider a class $\mathcal{G}$ of functions $\mathcal{X} \to \{1, \ldots, b\}$, where b is an even positive integer. We say that $\mathcal{G}' \subset \mathcal{G}$ is *4-separated* if for any $f_1, f_2 \in \mathcal{G}'$,

$$\max_{i=1,\ldots,m} |f_1(x_i) - f_2(x_i)| I_{\{u(f_1(x_i)) \neq u(f_2(x_i))\}},$$

where the function u is defined by

$$u(a) = \left\{ \begin{array}{ll} 1 & \text{if } a > b/2 \\ \frac{1}{2} & \text{if } a = b/2 \\ 0 & \text{if } a < b/2. \end{array} \right.$$

The maximal size of a 4-separated subset of $\mathcal{G}$ is denoted by $M_b(4, \mathcal{G})$. The proof of the next lemma is trivial.

Lemma 4 *Let $\gamma \leq \epsilon/8$ such that $b = 1/\gamma$ is a positive even integer. Then*

$$|\mathcal{F}_\epsilon| \leq M(\epsilon, \mathcal{F}) \leq M_b(4, \mathcal{F}^\gamma).$$

The key of the proof of Theorem 2 is the following combinatorial lemma:

Lemma 5 *Let $\mathcal{X}$ be a set of cardinality m, and let $\mathcal{G}$ be a class of functions $\mathcal{X} \to \{1, \ldots, b\}$, where b is an even positive integer and $m \geq b/72$. Then*

$$M_b(4, \mathcal{G}) \leq 2(72m)^{\log_2 y}, \qquad \text{where} \quad y = \sum_{i=1}^{\Delta_b} \binom{m}{i} 3^i$$

and Δ_b is the b-dimension of $\mathcal{G}$.

PROOF. We may assume that $b \geq 5$ since otherwise there are no two 4-separated functions in $\mathcal{G}$ and the statement is trivial. Let $A \subset \mathcal{X}$ and $s : A \to \{b/2-1, b/2, b/2+1\}$. We say that $\mathcal{G}$ b-shatters the pair (A,s) if it b-shatters A according to s. To any $k \geq 2$ and $m \geq 1$, define $t(k,m)$ as the largest integer t such that if $\mathcal{H}$ is any 4-separated class of functions with $|\mathcal{H}| = k$, then $\mathcal{H}$ b-shatters at least t distinct pairs (A,s). If no such $\mathcal{H}$ exists, then we say that $t(k,m) = \infty$. (Recall that $m = |\mathcal{X}|$.)

Clearly, the number of possible pairs (A,s) such that $|A| \leq d$ is at most $y = \sum_{i=1}^{d} \binom{m}{i} 3^i$. Thus, if $t(k,m) > y$ for some k, then $M_b(4,\mathcal{G}) < k$ whenever $\Delta_b \leq d$. Therefore, we need to show that $t\left(2(72m)^{\lfloor \log_2 y \rfloor}, m\right) > y$ for all $d \geq 1$, $m \geq 1$.

We see immediately that $t(2,m) = 1$ for all $m \geq 1$. Next we show that

$$t(144km, m) \geq 2t(2k, m-1). \tag{6}$$

If there is no 4-separated class with size $144km$, then the left-hand side of (6) is ∞, and the inequality is trivially true. Thus, assume that there is a 4-separated class $\mathcal{H}$ with size $144km$. Split $\mathcal{H}$ into $72km$ pairs of functions. For each such pair (h_1, h_2),

$$\max_{x \in \mathcal{X}} |h_1(x) - h_2(x)| I_{\{u(h_1(x)) \neq u(h_2(x))\}} \geq 4$$

that is, for each such pair there exists an $x \in \mathcal{X}$ such that $|h_1(x) - h_2(x)| \geq 4$ and $u(h_1(x)) \neq u(h_2(x))$. Since $|\mathcal{X}| = m$, there exists an $x \in \mathcal{X}$ such that this property holds for at least $72k$ pairs. For $j \in \{1, \ldots, b\}$, define

$$\tau(j) = \begin{cases} 1 & \text{if } j \leq b/2 - 4 \\ i & \text{if } j = b/2 - 5 + i,\ i = 2, 3, \ldots, 8 \\ 9 & \text{if } j \geq b/2 + 4. \end{cases}$$

By the pigeonhole principle, there are at least $72k/\binom{9}{2} = 2k$ pairs (h_1, h_2) for which the set $\{\tau(h_1(x), h_2(x))\}$ is the same. Then it follows that there are two subclasses $\mathcal{H}_1, \mathcal{H}_2 \subset \mathcal{H}$ and indeces $i, j \in \{1, \ldots, 9\}$ with $|\mathcal{H}_1| = |\mathcal{H}_2| = 2k$ such that for each $h_1 \in \mathcal{H}_1$, $\tau(h_1(x)) = i$, for each $h_2 \in \mathcal{H}_2$, $\tau(h_2(x)) = j$, and $i \geq j + 4$. Clearly, the members of $\mathcal{H}_1$ are 4-separated on $\mathcal{X} - \{x\}$, and the same is true for $\mathcal{H}_2$. Thus, according to the definition of $t(k,m)$, $\mathcal{H}_1$ and $\mathcal{H}_2$ both b-shatter $t(2k, m-1)$ pairs (A,s) with $A \subset \mathcal{X} - \{x\}$.

Clearly, $\mathcal{H}$ b-shatters every pair (A,s) which is b-shattered by either $\mathcal{H}_1$ or $\mathcal{H}_2$. Also, if a pair (A,s) is b-shattered by both $\mathcal{H}_1$ and $\mathcal{H}_2$, then it is easy to see that $\mathcal{H}$ b-shatters the pair $(A \cup \{x\}, s')$, where $s'(z) = s(z)$ if $z \in A$, and

$$s'(x) = \begin{cases} b/2 + 1 & \text{if } j = 5 \\ b/2 - 1 & \text{if } i = 5 \\ b/2 & \text{otherwise.} \end{cases}$$

Therefore, $\mathcal{H}$ b-shatters at least as many (A, s) pairs as the sum of the numbers of pairs shattered by $\mathcal{H}_1$ and $\mathcal{H}_2$, so (6) is proved.

Let now $n = 2(72m)(72(m-1)) \cdots (72(m-r+1))$, where $r \leq m$. Then by repeated application of (6), we obtain

$$t(n, m) \geq 2^r t(2, n-r) = 2^r.$$

Since t is monotone in its first argument, for all $r \leq m$,

$$t(2(72m)^r, m) \geq 2^r.$$

Take $r = \lceil \log_2 y \rceil$. If $r \leq m$, then

$$t(2(72m)^{\lceil \log_2 y \rceil}, m) \geq 2^{\lceil \log_2 y \rceil} > y,$$

as desired. If $r > m$, then $2(72m)^r > 2(72m)^m > b^m$ by the condition $72m \geq b$. But b^m is the number of all functions from $\mathcal{X}$ to $\{1, \ldots, b\}$, so there is no 4-separated class larger than this, hence $t(2(72m)^{\lceil \log_2 y \rceil}, m) = \infty > y$, establishing the lemma. □

Proof of Theorem 2. Let $\gamma = 1/(2\lceil 4/\epsilon \rceil)$ and $b = 1/\gamma$. Then

$$\begin{aligned} |\mathcal{F}_\epsilon| &\leq M_b(4, \mathcal{F}^\gamma) \quad \text{(by Lemma 4)} \\ &\leq 2(72m)^{\log_2(\sum_{i=1}^{\Delta_b} \binom{m}{i} 3^i)} \quad \text{(by Lemma 5)} \\ &\leq 2(72m)^{\log_2\left(\sum_{i=1}^{d_{2\gamma}} \binom{m}{i} 3^i\right)} \quad \text{(by Lemma 3)} \\ &\leq 2(72m)^{d_{\epsilon/5} \log_2(2m)} \end{aligned}$$

by bounding with Stirling's formula. But clearly,

$$|\mathcal{F}_\epsilon| = \inf_{\delta \leq \epsilon} |\mathcal{F}_\delta| \leq \inf_{\delta \leq \epsilon} 2(72m)^{d_{\delta/5} \log_2(2m)} = 2(72m)^{D_{\epsilon/5} \log_2(2m)},$$

as desired. □

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Technical Report 143-95, DSI, University of Milan, Italy*, 1993. An extended abstract appeared in the Proceedings of the 1993 IEEE Symposium on the Foundations of Computer Science IEEE Press.

[2] M. Anthony and P. L. Bartlett. Function learning from interpolation. Technical report, Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia, 1994.

[3] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.

[4] 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. Technical report, Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia, 1996.

[5] P.L. Bartlett and P.M. Long. More theorems about scale-sensitive dimensions and learning. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, pages 392–401. Association for Computing Machinery, New York, 1995.

[6] P.L. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the 7th Annual ACM Workshop on Computational Learning Theory*, pages 299–310. Association for Computing Machinery, New York, 1994.

[7] S.N. Bernstein. *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow, 1946.

[8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[9] K.L. Buescher and P.R. Kumar. Learning by canonical smooth estimation, Part II: Learning and choice of model complexity. *IEEE Transactions on Automatic Control*, 41:557–569, 1996.

[10] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[11] L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28:1011–1018, 1995.

[12] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[13] M. Kearns and R.E. Shapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer Systems Sciences*, 48:464–497, 1994.

[14] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *submitted*, 1996.

[15] G. Lugosi and M. Pintér. A data-dependent skeleton estimate for learning. In *Proceedings of the Nineth Annual ACM Conference on Computational Learning Theory*, pages 51–56. Association for Computing Machinery, New York, 1996.

[16] G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41:677–678, 1995.

[17] D. Pollard. Rates of uniform almost sure convergence for empirical processes indexed by unbounded classes of functions, 1986. Manuscript.

[18] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147, 1972.

[19] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. A framework for structural risk minimization. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 68–76. Association of Computing Machinery, New York, 1996.

[20] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition.* Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.