Measuring the VC-dimension of a Learning Machine

Vladimir Vapnik, Esther Levin, Yann Le Cun
AT&T Bell Laboratories
101 Crawfords Corner Road, Holmdel, NJ 07733

Abstract

A method for measuring the capacity of learning machines is described. The method is based on fitting a theoretically derived function to empirical measurements of the maximal difference between the error rates on two separate data sets of varying sizes. Experimental measurements of the capacity of various types of linear classifiers are presented.

1 Introduction.

Many theoretical and experimental studies have shown the influence of the *capacity* of a learning machine on its generalization ability (Vapnik, 1982; Baum and Haussler, 1989; Le Cun et al., 1990; Weigend, Rumelhart and Huberman, 1991; Guyon et al., 1992; Abu-Mostafa, 1993). Learning machines with a small capacity may not require large training sets to approach the best possible solution (lowest error rate on test sets). High-capacity learning machines, on the other hand, may provide better asymptotical solutions (i.e. lower test error rate for very large training sets), but may require large amounts of training data to reach acceptable test performance. For a given training set size, the *difference* between the training error and the test error will be larger for high-capacity machines. The theory of learning based on the VC-dimension predicts that the behavior of the difference between training error and test error as a function of the training set size is characterized by a single quantity –the VC-dimension– which characterizes the machine's capacity (Vapnik, 1982).

In this paper, we introduce an empirical method for measuring the capacity of a learning machine. The method is based on a formula for the maximum deviation between the frequency of errors produced by the machine on two separate data sets, as a function of the capacity of the machine and the size of the data sets. The main idea is that the capacity of a learning machine can be measured by finding the capacity that produces the best fit between the formula and a set of experimental measurements of the frequency of errors on data sets of varying sizes.

In the paradigm of learning from examples, the learning machine must learn to approximate as well as possible an unknown target rule, or input-output relation, given a training set of labeled examples. The l input-output pairs composing the training set $(x, \omega), x \in X \subset R^n, \omega \in \{0, 1\}$,

$$(x_1, \omega_1), ..., (x_l, \omega_l)$$

are assumed to be drawn independently from an unknown distribution function $P(x, \omega) = P(\omega|x)P(x)$. Here $P(x)$ describes the region of interest in the input space, and the distribution $P(\omega|x)$ describes the target input-output relation. The learning machine is characterized by the set of binary classification functions $f(x, \alpha), \alpha \in \Lambda$ (α is a parameter which specifies the function, and Λ is the set of all admissible parameters)

To appear in Neural Computation. 1994

that it can realize (indicator functions). The goal of learning is to choose a function $f(x,\alpha^*)$ within that set which minimizes the probability of error, i.e., the probability of disagreement between the value of ω and the output of the learning machine $f(x,\alpha)$

$$p(\alpha) = E|\omega - f(x,\alpha)|,$$

where the expectation is taken with respect to the probability distribution $P(x,\omega)$.

The problem is that this distribution is unknown, and the only way to assess $p(\alpha)$ is through the frequency of errors computed on the training set

$$\nu_l(\alpha) = \frac{1}{l}\sum_{i=1}^{l} |\omega_i - f(x_i,\alpha)|.$$

Many learning algorithm are based on the so-called *"principle of empirical risk minimization"*, which consists in picking the function $f(x,\alpha_l)$ that minimizes the number of error on the training set.

In (Vapnik, 1982) it is shown that for algorithms that minimize the empirical risk, the knowledge of three quantities allows us to establish an upper bound on the probability of error $p(\alpha)$. The first quantity is the *capacity* of the learning machine, as measured by the *VC-dimension* h of the set of functions it can implement. The second one is the frequency of errors on the training set (empirical risk) $\nu(\alpha)$. The third one is the size of the training set l. With probability $1-\eta$,, the bound

$$p(\alpha) \leq \nu(\alpha) + D(l,h,\nu(\alpha),\eta) \tag{1}$$

is simultaneously valid for all $\alpha \in \Lambda$ (including the α_l that minimizes the empirical risk $\nu(\alpha)$). The function D is of the form

$$D(l,h,\nu(\alpha),\eta) = c\frac{h(\ln(2l/h)+1)-\ln\eta}{2l}\left(\sqrt{1+\frac{4l\nu(\alpha)}{c(h(\ln(2l/h)+1)-\ln\eta)}}+1\right) \tag{2}$$

where c is a universal constant less than 1. It can be interpreted as a confidence interval on the difference between the training error and the "true" error.

There are two regimes in which the behavior of the function D simplifies. The first regime is when the training error happens to be small (large capacity, or small training set size, or easy task), then D can be approximated by

$$D(l,h,\nu(\alpha_l),\eta) \sim \frac{h(\ln(2l/h)+1)-\ln\eta}{l}.$$

When the training error is large (near 1/2), D can be approximated by

$$D(l,h,\nu(\alpha_l),\eta) \sim \sqrt{\frac{h(\ln(2l/h)+1)-\ln\eta}{l}}.$$

Unfortunately, theoretical estimates of the VC-dimension have been obtained for only handful of simple classes of functions, most notably the class of linear discriminant functions. The set of linear discriminant functions is defined by

$$f(x,\alpha) = \theta((x\cdot\alpha)+\alpha_0),$$

where $(x\cdot\alpha)$ is dot-product of the vectors x and α, and θ is the threshold function:

$$\theta(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{otherwise} \end{cases},$$

The VC-dimension of the set of linear discriminant functions with n inputs is equal to $n+1$ (Vapnik, 1982). Attempts to obtain the exact value of the VC-dimension for other classes of decision rules encountered substantial difficulties. Most of the estimated values appear to exceed the real one by a large amount. Such poor estimates result in rather crudely overestimated evaluations of the confidence interval D.

This article considers the possibility of empirically estimating the VC-dimension of a learning machine. The idea that this might be possible was inspired by following observations.

It was shown in (Vapnik and Chervonenkis, 1989) that a learning algorithm that minimizes the empirical risk (i.e. minimizes the error on the training set) will be consistent *if and only if* the following one-sided uniform convergence condition holds

$$\lim_{l\to\infty} P\{\sup_{\alpha\in\Lambda}(p(\alpha)-\nu(\alpha))>\varepsilon\}=0$$

In other words the one-sided uniform convergence (within a given set of function) of frequencies to probabilities is a *necessary and sufficient* condition for the consistency of the learning process.

In the 30s Kolmogorov and Smirnov found the law of distribution of the maximal deviation between a distribution function and an empirical distribution function for any random variable. This result can be formulated as follows. For the set of functions

$$f_*(x,\alpha)=\theta(x-\alpha),\ \ \alpha\in(-\infty,\infty)$$

the equality

$$P\{\sup_{\alpha\in\Lambda}(p_*(\alpha)-\nu_*(\alpha))>\varepsilon\}=\exp\{-2\varepsilon^2 l\}-2\sum_{n=2}^{\infty}(-1)^n\exp\{-2\varepsilon^2 n^2 l\}$$

holds for sufficiently large l, where $p_*(\alpha)=Ef_*(\alpha)$ and $\nu_*(\alpha)=\frac{1}{l}\sum_{i=1}^{l}f_*(x_i,\alpha)$. This equality is independent of the probability measure on x. Note that the second term is very small compared to the first one.

In (Vapnik and Chervonenkis, 1971) it was shown that for any set of indicator functions $f(x,\alpha),\alpha\in\Lambda$ with finite VC-dimension h, the rate of uniform convergence is bounded as follows

$$P\{\sup_{\alpha\in\Lambda}(p(\alpha)-\nu(\alpha))>\varepsilon\}<\min\left(1,\exp\left\{\left(c_1\frac{\ln 2l/h+1}{l/h}-c_2\varepsilon^2\right)l\right\}\right)$$

where c_1 and c_2 are universal constants such that $c_1\leq 1$ and $c_2>1/4$. As above, this inequality is independent of the probability measure. A direct consequence is that for any given δ there exists an $l_0=l_0(h,\delta,\varepsilon)$ such that for any $l>l_0$, the inequality

$$P\{\sup_{\alpha\in\Lambda}(p(\alpha)-\nu(\alpha))>\varepsilon\}<\exp\{-(c_2-\delta)\varepsilon^2 l\}$$

holds. In (Vapnik and Chervonenkis, 1971) it was shown that c_2 is at least $1/4$. The result was later improved to $c_2=2$ by (Devroye, 1982). Interestingly, for $c_2=2$, the above *inequality* is close to the Kolmogorov-Smirnov *equality* obtained for a simple set of functions. This means that, although the above is an upper bound, it is asymptotically close to the exact value. Tightening the upper bound on the value of c_1 is theoretically very difficult, because the term it multiplies is not the main term in the exponential for large values of l/h.

Now, suppose that there exists a value for the constant c_1, $0<c_1\leq 1$, for which the above upper bound (which is independent of the probability measure) is tight. Furthermore, suppose that the bound is tight not only for large numbers of observations, but also for smaller numbers (statisticians commonly use the asymptotic Kolmogorov-Smirnov test starting with $l=30$ and the learning processes usually involves several hundreds of observations). In this case, one can expect the function $\Phi^*(l/h)$ defined by

$$\Phi^*(\frac{l}{h})=E(\sup_{\alpha\in\Lambda}(p(\alpha)-\nu(\alpha))=\int_0^{\infty}P\{\sup_{\alpha\in\Lambda}(p(\alpha)-\nu(\alpha))>\varepsilon\}d\varepsilon$$

to be independent of the probability measure, even for rather small values of l/h.

Now, assuming we know a functional form for Φ^*, then we can experimentally estimate the expected maximal deviation between the empirical risk and expected risk for various values of l, and measure h by fitting Φ^* to the measurements. The remainder of the paper is concerned with finding an appropriate functional form for Φ^*, and applying it to measuring the VC-dimension of various learning machines.

A learning algorithm is said to be consistent if the probability of error on the training set converges to the true probability of error when the size of the training set goes to infinity.

In practice, rather than measuring the maximum difference between the training set error and the error on an infinite test set, it is more convenient to measure the maximum difference between the error rates measured on two separate sets

$$\xi_l = \{\sup_{\alpha \in \Lambda}(\nu_1(\alpha) - \nu_2(\alpha))\},$$

where $\nu_1(\alpha)$ and $\nu_2(\alpha)$ are frequencies of error calculated on two different samples of size l.

We will introduce an approximate functional form for the right hand side $\Phi(l/h)$ of the relation

$$E\xi_l \approx \Phi(l/h)$$

which we will use for a wide range of values of l/h. To construct this approximation we will determine in Section 2 two different bounds for $E\xi_l$: the bound valid for large l/h and the bound valid for small l/h.

In section 3, we introduce the notion of effective VC-dimension, which reflects some weak properties of the unknown probability distribution. The effective VC-dimension does not exceed the VC-dimension, and we show that the functional form of the bounds on $E\xi_l$ obtained for the VC-dimension are valid for the effective VC-dimension. As a consequence, the bounds obtained with the effective VC dimension are tighter. In section 4 we introduce a single functional form inspired by bound (2), and by the bounds on $E\xi_l$ for both small and large l/h. We hypothesize that this function describes well the expectations $E\xi_l$ in term of effective VC-dimension. Assuming this hypothesis is true, section 5 proposes a method for measuring the effective VC-dimension of a learning machine. In section 6 we demonstrate the proposed method of measuring VC-dimension for various classes of linear decision rules.

2 Estimation of the Maximum Deviation.

In this section we first establish the formal definition of the VC-dimension, and then give the bounds on the maximum deviation between frequencies of error on two separate sets.

Definition: The VC-dimension of the set of indicator functions $f(x, \alpha)$, $\alpha \in \Lambda$, is the maximal number h of vectors

$$x_1, \ldots, x_h$$

from the set X that can be shattered by $f(x, \alpha)$, $\alpha \in \Lambda$.

The vectors $x_1, \ldots, x_h$ are said to be shattered by $f(x, \alpha)$, $\alpha \in \Lambda$ if, for any possible partition of these vectors into two classes A and B, (there are 2^h such partitions) there exists a function $f(x, \alpha^*)$ that implements the partition, in the sense that $f(x, \alpha^*) = 0$ if $x \in A$ and $f(x, \alpha^*) = 1$ if $x \in B$.

To estimate a bound on the expectation of the random variable ξ_l we need to define

$$Z^{2l} = x_1, \omega_1; \ldots; x_{2l}, \omega_{2l}$$

as a random independent sample of vectors (the x's) and their class (the ω's, ($\omega = 0, 1$)). We denote by $Z(2l)$ the set of all samples of size $2l$.

Let us denote by $\nu_1^l(Z^{2l}, \alpha)$ the frequency of erroneous classifications of the vectors

$$x_1, \ldots, x_l$$

of the first half-sample of Z^{2l}, obtained by using the decision rule $f(x, \alpha)$:

$$\nu_1^l(Z^{2l}, \alpha) = \frac{1}{l} \sum_{i=1}^{l} |\omega_i - f(x_i, \alpha)|.$$

Let us denote by $\nu_2^l(Z^{2l}, \alpha)$ the frequency of erroneous classification of the vectors

$$x_{l+1}, \ldots, x_{2l}$$

obtained using the same decision rule $f(x, \alpha)$:

$$\nu_2^l(Z^{2l}, \alpha) = \frac{1}{l} \sum_{i=l+1}^{2l} |\omega_i - f(x_i, \alpha)|.$$

We study the properties of the random variable

$$\xi_{2l} = \xi(Z^{2l}) = \sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha)), \tag{3}$$

which is the maximal deviation of error frequencies on the two half-samples over a given set of functions. There exists an upper bound for the expectation of this random value $\xi(Z^{2l})$ for varying sample size l. Consider three cases.

Case 1 ($l/h \leq 0.5$). For this case we use the trivial bound:

$$E\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha)) \leq 1.$$

Case 2 ($0.5 < l/h \leq g$, g is small). Let the supremum of the difference

$$\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha)$$

be attained on the function $f(x,\alpha^\star)$, where $\alpha^\star = \alpha^\star(Z^{2l})$. Consider the event

$$B_\delta = \{Z^{2l}: \frac{(\nu_1^l(Z^{2l},\alpha^\star) - \nu_2^l(Z^{2l},\alpha^\star))}{(\nu(Z^{2l},\alpha^\star) + 1/2l)(1 + 1/2l - \nu(Z^{2l},\alpha^\star))} > \delta\}, \tag{4}$$

where

$$\nu(Z^{2l},\alpha^\star) = \frac{\nu_1^l(Z^{2l},\alpha^\star) + \nu_2^l(Z^{2l},\alpha^\star)}{2}.$$

We denote the probability of this event by $P(B_\delta)$. In Appendix 1 we prove the following theorem :

Theorem 1. *Let the set of indicator functions $f(x,\alpha)$, $\alpha \in \Lambda$ have VC-dimension h. Then the following bound of the conditional expectation of $\xi(Z^{2l})$*

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha))|B_\delta\} \leq \frac{4}{\delta}\left(\frac{\ln(2l/h)+1}{l/h} + \frac{1-\ln P(B_\delta)}{l}\right). \tag{5}$$

is valid.

From (5) and from the fact that the deviation of the frequencies on two half-samples does not exceed 1 we obtain

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z_{2l},\alpha))\} \leq \frac{4P(B_\delta)}{\delta}\left(\frac{\ln(2l/h)+1}{l/h} + \frac{1-\ln P(B_\delta)}{l}\right) + (1 - P(B_\delta)). \tag{6}$$

This bound should be used when δ is not too small (e.g. $\delta > 0.5$) and when the probability $P\{B_\delta\}$ is close to one. Such a situation seems to arise when the ratio l/h is not too large and when l is large. In this case, when $P(B_\delta)$ is close to one, l/h is not large and l is large, the bound

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha))\} \leq C_1\frac{\ln(2l/h)+1}{l/h}$$

is valid, where C_1 is a constant.

In the following, we shall make use of this bound for small l/h when constructing an empirical estimate of the expectation.

Case 3 (l/h is large). In appendix 1 we show that for the set of indicator functions with VC-dimension h the following bound holds:

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha))\} \leq C_2\sqrt{\frac{\ln(2l/h)+1}{l/h}}. \tag{7}$$

where C_2 is a constant. This bound is true for all l/h. We shall use it only for large l/h, where the bound of case 2 is not valid.

From the conducted experiments described in the section 6 we find that $g \approx 8$.

3 Effective VC-dimension.

According to the definition, the VC-dimension does not depend on the input probability measure $P(x)$. Our purpose here is to introduce a concept of *effective VC-dimension* which reflects a weak dependence on the properties of the probability measure. Let the probability measure P be given on X and let the subset $X^\star$ of the set X have a probability measure which is close to one, i.e.

$$P(X^\star) = 1 - \eta$$

for small η.

Let the set of indicator functions $f(x,\alpha)$, $\alpha \in \Lambda$ with VC-dimension h be defined on X, and let the same set of functions on $X^\star \subset X$ have VC-dimension $h^\star$. In Appendix 2 we prove the following theorem.

Theorem 2. *For all $l > h$ for which the inequality*

$$\eta < \frac{1}{2l}\left(\frac{2el}{h^\star}\right)^{h^\star}\left(\frac{h}{2el}\right)^{h} \tag{8}$$

is fulfilled, the following bounds

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha))|B_\delta\} < 4/\delta\left(\frac{\ln(2l/h^\star)+1}{l/h^\star} + \frac{1-\ln P(B_\delta)}{l}\right), \tag{9}$$

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha))\} < \sqrt{\frac{\ln(2l/h^\star)+1}{l/h^\star}} + \frac{3}{\sqrt{lh^\star(\ln(2l/h^\star)+1)}} \tag{10}$$

are valid.

Remark. If $\eta \to 0$ the inequalities are true for all $l > h$. Note that in this case the left hand side of inequalities (9) and (10) do not depend on X^*, but the right hand side does. Therefore the tightest inequality will be achieved for the X^* with the smallest h^*.

Definition. The effective VC-dimension of the set $f(x,\alpha), \alpha \in \Lambda$, (for the given measure P) is the minimal VC-dimension of this set of functions defined on all subsets $X^\star \subset X$ whose measure is almost one ($\mu(X^\star) > 1 - \eta^\star$, where $\eta^\star > 0$ is a small value).

As in the previous section, from (9) we obtain that for the effective VC-dimension $h^\star$ the following inequality is true

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha))\} < \frac{4P(B_\delta)}{\delta}\left(\frac{\ln(2l/h^\star)+1}{l/h^\star} + \frac{1-\ln P(B_\delta)}{l}\right) + (1 - P(B_\delta)). \tag{11}$$

For the case when $P(B_\delta)$ is close to one, the quantity $l/h^\star$ is small, and l is large we obtain

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha))\} < C\frac{\ln(2l/h^\star)+1}{l/h^\star}.$$

where C is a constant. Since the bounds (9) and (10) have the same form as the bounds (5) and (7), to simplify our notation we use h in the next sections to denote the *effective* VC-dimension.

4 The Law of the largest deviations.

In section 2 we gave several bounds, for different cases, on the expectation of the largest deviation over the set of decision rules. In this section we first give a *single* functional form that reflects the properties of the bounds. Then we conjecture that the same functional form approximates the expectation of the largest deviation.

The bounds given in the previous section are summarized as follows

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_2^l(Z^{2l},\alpha))\} \le \begin{cases} 1 & \text{if } l/h \le 0.5 \\ C_1\frac{\ln(2l/h)+1}{l/h} & \text{if } l/h \text{ is small} \\ C_2\sqrt{\frac{\ln(2l/h)+1}{l/h}} & \text{if } l/h \text{ is large.} \end{cases} \tag{12}$$

Consider the following continuous approximation of the right side of the bound (12)

$$\Phi(\tau) = \begin{cases} 1 & \text{if } \tau < 0.5 \\ a\frac{\ln(2\tau)+1}{\tau-k}\left(\sqrt{1+\frac{b(\tau-k)}{\ln(2\tau)+1}}+1\right) & \text{otherwise} \end{cases}, \tag{13}$$

where $\tau = l/h$. The function $\Phi(\tau)$ has two free parameters a and b. The third parameter $k < 0.5$ is chosen from the conditions of continuity at point $\tau = 0.5$: $\Phi(0.5) = 1$. Note that the function $\Phi(\tau)$ has the same structure as the confidence interval (2), except that the frequency $\nu(\alpha)$ is replaced by a constant.

For small τ, $(0.5 < \tau < g)$, $\Phi(\tau)$ behaves as

$$C_1^* \frac{\ln(2\tau)+1}{\tau - k},$$

and for large τ as

$$C_2^* \sqrt{\frac{\ln(2\tau)+1}{\tau - k}}.$$

The constants a and b determine the regions of "large" and "small" values of τ.

We make the hypothesis that there exist constants a, b that are only weakly dependent upon the properties of the input probability measure, such that the function $\Phi(\tau)$ approximates the expectation sufficiently well for all $l > h$:

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^l(Z^{2l},\alpha) - \nu_1^l(Z^{2l},\alpha))\} \approx \Phi(\tau). \tag{14}$$

The constants can be found empirically by fitting Φ to experimental data using a machine whose VC-dimension is known. If we assume that the constants obtained are universal, we can then use the function Φ obtained this way to measure the capacity of other learning machines.

5 Measuring the Effective VC-Dimension of a Classifier

If the function $\Phi(\tau)$ approximates the expectation with sufficient accuracy, then the effective VC-dimension of the set of indicator functions realizable by a given classifier can be estimated on the basis of the following experiment. First, we generate a random independent set of size $2l$,

$$Z_i^{2l} = x_1^i, \omega_1^i; \ldots; x_l^i, \omega_l^i; x_{l+1}^i, \omega_{l+1}^i; \ldots; x_{2l}^i, \omega_{2l}^i. \tag{15}$$

using a generator of random vectors $P(x)$ and a (possibly non-deterministic) generator of labels $P(\omega|x)$. Using this sample, we measure the quantity

$$\xi(Z_i^{2l}) = \sup_{\alpha\in\Lambda}(\nu_1^l(Z_i^{2l},\alpha) - \nu_2^l(Z_i^{2l},\alpha)). \tag{16}$$

by approximating the expectation by an average over N independently generated sets of size $2l$.

$$\xi(l) = \frac{1}{N}\sum_{i=1}^{N}\xi(Z_i^{2l})$$

Repeating the above procedure for various values of l, we obtain the set of estimates $\xi(l_1), \ldots, \xi(l_k)$.

The effective VC-dimension of the set of functions $f(x,\alpha), \alpha \in \Lambda$, can then be approximated by finding the integer parameter $h^\star$ which provides the best fit between $\Phi(l/h)$ and the $\xi(l_i)$'s:

$$h^\star = \arg\min_h \sum_{i=1}^{k}(\xi(l_i) - \Phi(l_i/h))^2.$$

The accuracy of the obtained VC-dimension estimate, and in fact, the validity of the approach presented in this paper, depend crucially on how well the function $\Phi(l/h)$ describes the expectation (14).

To estimate the expectation of the largest deviation empirically one must be able to define for every fixed sample (15) the value of the largest deviation (16). This can be done by considering a modified training set were the labels of the first half of the set have been reversed

$$x_1, \varpi_1; \ldots; x_l, \varpi_l; x_{l+1}, \omega_{l+1}; \ldots; x_{2l}, \omega_{2l}$$

where ϖ denotes the opposite class of ω. As shown in Appendix 3, evaluating (16) can be done by minimizing the functional

$$R(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(\varpi_i - f(x_i, \alpha))^2 + \frac{1}{l}\sum_{i=l+1}^{2l}(\omega_i - f(x_i, \alpha))^2 \tag{17}$$

6 Empirical measurement of the effective VC dimension

In this section, we illustrate the method of measuring the VC-dimension by applying it various types of linear classifiers. Our method relies on several assumptions which must be checked experimentally:

- The expected deviation $E\{\xi_l\}$ is largely independent of the generator distribution $P(\omega|x)$.
- The expected deviation $E\{\xi_l\}$ depends on the choice of a learning machine only through the parameter h.
- The expected deviation $E\{\xi_l\}$ can be well described by the function $\Phi(l/h)$ in equation (13) with fixed parameters a and b.
- The values of the free parameters a and b have constant values over wide classes of learning machines.

Experiments with learning machine that implement various subsets of the set of linear classification functions were conducted to check these hypotheses. The VC-dimension of the set of linear classifiers (without bias) is known from theory to be equal to the dimension of the input space n. The experiments described here were conducted with $n = 50$, $x = (x^1, ..., x^{50})$. Additional experiments with various values of n ranging from 25 to 200 were performed with similar results.

Since no efficient algorithm is known for minimizing the classification error of linear threshold units, we trained the machine by minimizing the mean squared error between the labels and the output of a unit where the hard threshold was replaced by a sigmoid function This procedure does not ensure that the classification error is minimized, but it is known empirically to give good approximate solutions in the separable and non-separable cases. After training, the output labels were computed by thresholding the output of the sigmoid.

6.1 Independence of Average Deviations from the Task Difficulty

A set of experiments was performed to assess the influence of the difficulty of the task (an important property of the conditional probability distribution $P(\omega|x)$) on the deviation. Three ensembles of training sets were generated such that the expected frequency of errors using a linear classifier would be respectively 0, 0.25 and 0.5. First, we randomly selected a set of 50-dimensional input vectors with coordinates independently drawn with a uniform distribution over the $[-1, 1]$ interval. Then, the labels for the three experiments were generated by picking a 50-dimensional vector of coefficients α^* at random, and by generating the labels according to the following three conditional probability distributions

$$P_1(\omega|x) = \begin{cases} 1 & \text{for } \omega = \theta(\alpha^* \cdot x), \\ 0 & \text{for } \omega = 1 - \theta(\alpha^* \cdot x), \end{cases}$$

$$P_2(\omega|x) = \begin{cases} 0.75 & \text{for } \omega = \theta(\alpha^* \cdot x), \\ 0.25 & \text{for } \omega = 1 - \theta(\alpha^* \cdot x), \end{cases}$$

$$P_3(\omega|x) = \begin{cases} 0.5 & \text{for } \omega = \theta(\alpha^* \cdot x), \\ 0.5 & \text{for } \omega = 1 - \theta(\alpha^* \cdot x), \end{cases}$$

P_1 corresponds to a linearly separable task (no noise), P_3 to a random classification task (random labels, no generalization is expected), P_2 to a task of intermediate difficulty.

Figure 1 shows the average values of the deviations as a function of l, the size of the half-set, for the three conditional probabilities. Each average value was obtained from 20 realizations of the deviations. As can be seen on the figure, the results for the three cases practically coincide, demonstrating the independence of the $E\xi_i$ from the difficulty of the task. In further experiments only the conditional probability $P_3(\omega|x)$ was used.

6.2 Estimation of the free parameters a and b in $\Phi(\frac{l}{h})$

The value of the free parameters a and b in $\Phi(\frac{l}{h})$ (eq 13) can be determined if we fit equation (13) to a set of experimentally obtained values of the average maximal deviation produced by a learning machine with a *known* capacity h. Assuming the optimal values a^* and b^* thereby obtained are universal, we can simply consider them constant, and use them for the measurement of the capacity of other learning machines.

Figure 2 shows the approximation of the empirical data by $\Phi(\frac{l}{h})$, given in (13) with parameters set to $a^* = 0.16$, $b^* = 1.2$. Note how well the function Φ describes the average maximal deviation throughout the full range of training set sizes used in our experiments ($0.5 < \frac{l}{h} < 32$). Experiments with various input sizes yielded consistent values for the parameters.

A simpler functional form for the deviation $\Phi_1(\frac{l}{h})$,

$$\Phi_1(\frac{l}{h}) = d\frac{\ln(2\frac{l}{h}) + 1}{\frac{l}{h} + d - 0.5},$$

inspired by the bound for small l/h, describes the data well for small l/h (up to $l/h < 8$). Figure 3 shows the approximation of the empirical data by Φ_1 using $d = 0.39$.

The values $a = 0.16$ and $b = 1.2$ obtained with the experiment of figure 2 were used for all the experiments described below.

6.3 Control Experiments

For further validation of the proposed method, a series of control experiments were conducted. In these experiments we used the above-described method to measure the effective VC-dimension of several learning machines, and compared it to their theoretically known values.

The learning machine we used was composed of a fixed preprocessor which transformed the n-dimensional input vector x into another n-dimensional vector y using a linear projection onto a subspace of dimension k. The resulting vector y was then fed to a linear classifier with n inputs.

It is easy to see that the theoretical value of the effective VC-dimension is k, as this can be reduced to a linear classifier with k inputs. Table 1, which shows the estimated VC-dimension for $k = 10, 20, 30, 40$, indicates that there is a strong agreement between the estimated effective VC-dimension, and the theoretically predicted dimension.

Effective VC dim.	40	30	20	10
Estimated VC dim.	40	31	20	11

Table 1: The true and the measured values of VC-dimension in 4 control experiments

Figures 4a-d demonstrate how well the function $\Phi(\frac{l}{h^*})$, where h^* is the estimate of the effective VC-dimension, describes the experimental data.

6.4 Smoothing experiments

In this section, we measure the effect of "smoothing" the input on the capacity. Contrary to the previous section, this effect was not predicted theoretically. As in the previous section, the learning machine incorporated a linear preprocessor which transformed x into y by a smoothing operation:

$$y^i = \sum_{j=1}^{n} x^j exp\{-\beta|j-i|_n\},$$

where

$$|j-i|_n = \begin{cases} j-i & \text{for } 1 \leq i < j \\ j+n-i & \text{for } i \geq j \end{cases}$$

As described previously, the classifier was trained with gradient descent, but a very small "weight decay" term was added (the reason for the weight decay will be explained below).

The parameter β determines the amount of smoothing: for very large β, the components of the vector y are virtually independent, and the VC-dimension of the learning machine is close to n. As the value of β decreases, strong correlations between input variables start to appear. This causes the distribution of preprocessed vectors to have a very small variance in some directions. Intuitively, an appropriate weight decay term will make these dimensions effectively "invisible" to the linear classifier, thereby reducing the measured effective VC-dimension. The measured VC-dimension decreases down to 1 for $\beta = 0$. When β takes non-zero values, there is no theoretically known evaluation of the VC-dimension, and the success of the method can be measured only by how well the experimental data is described by the functional Φ.

Figures 5(a-c) show the average maximal deviation as a function of l, for different smoothing parameters β, and demonstrates how well the function $\Phi(\frac{l}{h^*})$, where h^* is the estimate of the effective VC-dimension, approximates the experimental data. Figure 6 shows the estimated value of the VC-dimension as a function of the smoothing parameter β.

7 Conclusion

We have shown that there exist three different regimes for the behavior of the maximum possible deviation between the frequencies of error on two different sets of examples. For very small values of l/h ($< 1/2$), the maximal deviation is ~ 1, for small values of l/h (from 1/2 up to about 8), it behaves like $\log(2l/h)/(l/h)$, for large values it behaves like $\sqrt{\log(2l/h)/(l/h)}$.

We have introduced the concept of effective VC-dimension, which takes into account some weak properties of probability distribution over the input space. We prove that the effective VC-dimension can be used in place of the VC-dimension in all the formulae obtained previously. This provides us with tighter bounds on the maximal deviation.

Based on the functional forms of the bounds for the three regimes, we propose a single formula that contains two free parameters. We have shown how the value of these parameters can be evaluated experimentally. We show how the formula can be used to estimate the effective VC-dimension.

We illustrate the method by applying it to various learning machines based on linear classifiers. These experiments show good agreement between the model and the experimental data in all cases.

Interestingly, it appears that, at least within the set of linear classifiers, the values of the parameters are universal. This universality was confirmed by recent results obtained since the first version of this paper with linear classifiers trained on real-life tasks (image classification). Excellent fit with the same values of the constants were obtained even when the classifiers were trained by minimizing the empirical risk subject to various types of constraints. This included simple constraints, such as a limit on the norm of the weight vector (equivalent to weight decay), and more complex ones which improve the invariance of the classifier to distortions of the input image by limiting the norm of some Lie derivatives (Cortes, 1993; Guyon et al., 1992).

The extension of the present work to multilayer networks faces the following difficulties. First, the theory is derived for methods that minimize the empirical risk. However, existing learning algorithms for multilayer nets cannot be viewed as minimizing the empirical risk over entire set of functions implementable by the network. Because the energy surface has multiple local minima, it is likely that, once the initial parameter value is picked, the search will be confined to a subset of all possible functions realizable by the network. The capacity of this subset can be much less than the capacity of the whole set (which may explain why neural nets with large numbers of parameters work better than theoretically expected). Second, because the number of local minima may change with the number of examples, the capacity of the search subset may change with the number of observations. This may require a theory which considers the notion of non-constant capacity associated with an "active" subset of functions.

8 Appendix 1. Proof of Theorem 1.

Let the random independent sample

$$Z^{2l} = x_1, \omega_1, \ldots, x_l, \omega_l, x_{l+1}, \omega_{l+1}, \ldots, x_{2l}, \omega_{2l} \tag{18}$$

be given.

Call the classes of equivalence of the set $f(x,\alpha), \alpha \in \Lambda$ on the sample (18) the subset of indicator functions $f(x,\alpha), \alpha \in \Lambda^\star$ that take the same value on the vectors x from (18). Denote the number of different classes of equivalence given on the sample (18) by

$$\Delta^\Lambda(Z^{2l}) = \Delta^\Lambda(x_1, \omega_1, \ldots, x_{2l}, \omega_{2l}).$$

Consider the event

$$A_\epsilon = \{Z^{2l} : \sup_{\alpha \in \Lambda}(\nu_1^l(Z^{2l}, \alpha) - \nu_2^l(Z^{2l}, \alpha)) > \epsilon\} = \{Z^{2l} : \ \nu_1^l(Z^{2l}, \alpha^\star) - \nu_2^l(Z^{2l}, \alpha^\star) > \epsilon\}$$

and the event B_δ defined as in equation (4). The following lemma is true

Lemma.

$$P(A_\epsilon B_\delta) \leq E\Delta^\Lambda(x_1, \omega_1, \ldots, x_{2l}\omega_{2l}) \exp\{-\frac{\delta \epsilon l}{4}\}$$

The proof of this lemma will follow the same scheme as the proof of Theorem A2 in (Vapnik, 1982) pp. 170-172. Write the probability in the evident form

$$P(A_\epsilon B_\delta) = \tag{19}$$

$$\int_{Z(2l)} \theta\{(\nu_1^l(Z^{2l}, \alpha^\star) - \nu_2^l(Z^{2l}, \alpha^\star)) - \epsilon\} \theta\{\frac{(\nu_1^l(Z^{2l}, \alpha^\star) - \nu_2^l(Z^{2l}, \alpha^\star))}{(\nu(Z^{2l}, \alpha^\star) + \frac{1}{2l})(1 + \frac{1}{2l} - \nu(Z^{2l}, \alpha^\star))} - \delta\} dP(Z^{2l})$$

where$Z(2l)$ is the space of samples Z^{2l},

$$\nu(Z^{2l}, \alpha^\star) = \frac{1}{2}[\nu_1^l(Z^{2l}, \alpha^\star) + \nu_2^l(Z^{2l}, \alpha^\star)],$$

and $f(x,\alpha^\star)$ is the function which attains the maximum deviation of frequencies in two half-samples. Denote by T_i $(i = 1, 2, \ldots, (2l)!)$ the set of all possible permutations of the elements of the sample (18). Denote also

$$\chi(T_i Z^{2l}, \alpha^\star) = \nu_1^l(T_i Z^{2l}, \alpha^\star) - \nu_2^l(T_i Z^{2l}, \alpha^\star),$$

$$\sigma^2(T_i Z^{2l}, \alpha^\star) = (\nu(T_i Z^{2l}, \alpha^\star) + 1/2l)(1 + 1/2l - \nu(T_i Z^{2l}, \alpha^\star)).$$

It is obvious that the equality

$$P(A_\epsilon B_\delta) = \int_{Z(2l)} \theta\{\chi(T_i Z^{2l}, \alpha^\star) - \epsilon\} \theta\{\frac{\chi(T_i Z^{2l}, \alpha^\star)}{\sigma^2(T_i Z^{2l}, \alpha^\star)} - \delta\} dP(Z^{2l}) = \tag{20}$$

$$\int_{Z(2l)} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\{(\chi(T_i Z^{2l}, \alpha^\star) - \epsilon\} \theta\{\frac{\chi(T_i Z^{2l}, \alpha^\star)}{\sigma^2(T_i Z^{2l}, \alpha^\star)} - \delta\} dP(Z^{2l}).$$

is true.

Note that for any fixed Z^{2l} and T_i the following inequality is true:

$$\theta\{\chi(T_i Z^{2l}, \alpha^\star) - \epsilon\} \theta\{\frac{\chi(T_i Z^{2l}, \alpha^\star)}{\sigma^2(T_i Z^{2l}, \alpha^\star)} - \delta\} \leq$$

$$\sum_{\alpha_j \in \Delta^{\Lambda}(Z^{2l})} \theta\{\chi(T_i Z^{2l}, \alpha_j) - \epsilon\}\theta\{\frac{\chi(T_i Z^{2l}, \alpha_j)}{\sigma^2(T_i Z^{2l}, \alpha_j)} - \delta\}$$

Using this bound, one can estimate the integrand of the right-hand side of the equality (20). It does not exceed the value

$$\sum_{\alpha_j \in \Delta^{\Lambda}(Z^{2l})} \left[\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\{(\chi(T_i Z^{2l}, \alpha_j - \epsilon\}\theta\{\frac{\chi(T_i Z^{2l}, \alpha_j)}{\sigma^2(T_i Z^{2l}, \alpha_j)} - \delta\} \right]. \quad (21)$$

The expression in the brackets is the fraction of the number of permutations T_i for which the two events

$$(\nu_1^l(T_i Z^{2l}, \alpha_j) - \nu_2^l(T_i Z^{2l}, \alpha_j) > \epsilon,$$

and

$$\frac{\nu_1^l(T_i Z^{2l}, \alpha_j) - \nu_2^l(T_i Z^{2l}, \alpha_j)}{(\nu(T_i Z^{2l}, \alpha_j) + 1/2l)(1 + 1/2l - \nu(T_i Z^{2l}, \alpha_j))} > \delta$$

hold simultaneously, among all possible $(2l)!$ permutations. It is equal to the following value

$$\Gamma = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

where

$$m = l(\nu_1^l(Z^{2l}, \alpha) + \nu_2^l(Z^{2l}, \alpha))$$

and the summation is performed on all k satisfying the inequalities

$$\frac{k}{l} - \frac{m-k}{l} > \epsilon, \quad (22)$$

$$\frac{k}{l} - \frac{m-k}{l} > \delta\left(\frac{m+1}{2l}(1 - \frac{m-1}{2l})\right). \quad (23)$$

We now estimate the value Γ. The derivation of the estimate of Γ repeats the derivation of similar estimates in (Vapnik, 1982) pp. 173-176.

$$\ln \Gamma/2 < \begin{cases} -\frac{4(l+1)}{(m+1)(2l-m+1)}(k^* - \frac{m}{2} + 1)^2 & \text{if } m \text{ is an even number} \\ -\frac{4(l+1)}{(m+1)(2l-m+1)}(k^* - \frac{m-1}{2} + 1)(k^* - \frac{m-1}{2}) & \text{if } m \text{ is an add number} \end{cases} \quad (24)$$

where k^* is the least integer satisfying the condition

$$\frac{k^*}{l} - \frac{m}{2l} > \frac{\epsilon}{2}. \quad (25)$$

From (23) and (24) we obtain

$$\ln \Gamma/2 < -\frac{\delta}{2}(k^* - m/2).$$

And from this inequality and (25) we obtain

$$\ln \Gamma/2 < -\delta\frac{\epsilon l}{4}.$$

Substitute now the estimate of the value Γ into right-hand side of the inequality (20). We obtain

$$P\{A_\epsilon B_\delta\} \leq \int \sum_{\alpha_j \in \Delta^{\Lambda}(Z^{2l})} \exp\{-\frac{\delta\epsilon l}{4}\} dP(Z^{2l}) = E\Delta^{\Lambda}(Z^{2l}) \exp\{-\frac{\delta\epsilon l}{4}.\} \quad (26)$$

The lemma is proved.

Note now that the inequality

$$E\Delta^{\Lambda}(Z^{2l}) < \max_{Z^{2l}} \Delta^{\Lambda}(Z^{2l}) = m^{\Lambda}(2l)$$

is true. The function $m^{\Lambda}(2l)$ known as the *growth function* is estimated by means of the VC-dimension (Vapnik, 1982):

$$m^{\Lambda}(2l) < (\frac{2le}{h})^{h}.$$

Therefore

$$P\{A_{\epsilon}B_{\delta}\} < (\frac{2le}{h})^{h} e^{-\frac{\delta\epsilon l}{4}}.$$

Then

$$P\{A_{\epsilon}|B_{\delta}\} < (\frac{2le}{h})^{h} \frac{e^{-\frac{\delta\epsilon l}{4}}}{P(B_{\delta})}.$$

Estimate now the conditional expectation

$$E(A_{\epsilon}|B_{\delta}) = \int_{0}^{\infty} P\{\sup_{\alpha\in\Lambda}(\nu_1^l(\alpha, Z^{2l}) - \nu_2^l(\alpha, Z^{2l})) > \epsilon|B_{\delta}\}d\epsilon.$$

For this estimation divide the interval area into two parts: (0,u), where we use the trivial estimate of conditional probability $P(A_{\epsilon}|B_{\delta}) = 1$, and the area (u, ∞), were we use the estimate (26). We obtain

$$E(A_{\epsilon}|B_{\delta}) < \int_{0}^{u} d\epsilon + (\frac{2le}{h})^{h} \frac{1}{P(B_{\delta})} \int_{u}^{\infty} \exp\{-\frac{\delta\epsilon l}{4}\}d\epsilon = u + (\frac{2le}{h})^{h} \frac{4}{l\delta P(B_{\delta})} \exp\{-\frac{\delta u l}{4}\}$$

The minimum of the right-hand side of the inequality is attained when

$$u = 4\frac{h(\ln(2l/h) + 1) - \ln P(B_{\delta})}{\delta l}.$$

In such a way we obtain the estimate

$$E(A_{\epsilon}|B_{\delta}) < 4\frac{h(\ln(2l/h) + 1) - \ln P(B_{\delta}) + 1}{\delta l}.$$

The theorem is proved.

Now we want to obtain the estimate of the expectation of the random value (4) for an arbitrary l/h. For this we use the inequality (Vapnik, 1982) p.172

$$P\{\sup_{\alpha\in\Lambda} |\nu_1^l(\alpha, Z^{2l}) - \nu_2^l(\alpha, Z^{2l})| > \epsilon\} \; < \; 3(\frac{2le}{h})^{h} \exp\{-\epsilon^2 l\}$$

As in the case above we get

$$E\{\sup_{\alpha\in\Lambda} |\nu_1^l(\alpha, Z^{2l}) - \nu_2^l(\alpha, Z^{2l})|\} < \int_{0}^{u} d\epsilon + 3(\frac{2le}{h})^{h} \int_{u}^{\infty} \exp\{-\epsilon^2 l\}d\epsilon <$$

$$u + 3(\frac{2le}{h})^{h} \int_{u}^{\infty} \exp\{-\epsilon u l\}d\epsilon = u + 3(\frac{2le}{h})^{h} \frac{1}{lu} \exp\{-ul\}$$

For

$$u = \sqrt{\frac{\ln(2l/h) + 1}{l/h}}$$

we obtain the estimate

$$E\{\sup_{\alpha\in\Lambda} |\nu_1^l(\alpha, Z^{2l}) - \nu_2^l(\alpha, Z^{l})|\} < \sqrt{\frac{\ln(2l/h) + 1}{l/h}} + \frac{3}{\sqrt{lh(\ln(2l/h) + 1)}} < C\sqrt{\frac{\ln(2l/h) + 1}{l/h}}.$$

9 Appendix 2. Proof of Theorem 2.

We first prove the first part of the theorem. By the lemma proved in Appendix 1, the inequality

$$P(A_\epsilon B_\delta) \leq E\Delta^\Lambda(Z^{2l}) \exp\{\frac{-\epsilon\delta l}{4}\} \tag{27}$$

is true. Then it follows that

$$P(A_\epsilon|B_\delta) \leq \frac{E\Delta^\Lambda(Z^{2l})}{P(B_\delta)} \exp\{-\frac{\epsilon\delta l}{4}\} = \left(\int_{Z(2l)} \Delta(Z^{2l}) dP(Z^{2l})\right) \frac{\exp\{-\frac{\epsilon\delta l}{4}\}}{P(B_\delta)}. \tag{28}$$

Divide the integration domain $Z(2l)$ into two parts: Γ_{2l} and $Z(2l)/\Gamma_{2l}$, where Γ_{2l} is a sample space all of whose elements belong to $X^\star$. Then

$$\exp\{-\frac{\epsilon\delta l}{4}\} \int_{Z(2l)} \Delta^\Lambda(Z^{2l}) dP(Z^{2l}) = \tag{29}$$

$$\exp\{-\frac{\epsilon\delta l}{4}\} \left(\int_{\Gamma_{2l}} \Delta^\Lambda(Z^{2l}) dP(Z^{2l}) + \int_{Z(2l)/\Gamma_{2l}} \Delta^\Lambda(Z^{2l}) dP(Z^{2l})\right).$$

According to the conditions of the theorem, in area Γ_{2l} the bound

$$\Delta^\Lambda(Z^{2l}) < \left(\frac{2le}{h^*}\right)^{h^*} \tag{30}$$

holds, and in the area $Z(2l)/\Gamma_{2l}$ the bound

$$\Delta^\Lambda(Z^{2l}) < \left(\frac{2le}{h}\right)^{h} \tag{31}$$

holds. Thus from (29), (30) and (31) it follows that

$$P(A_\epsilon|B_\delta) < \frac{\exp\{-\frac{\epsilon\delta l}{4}\}}{P(B_\delta)} \left((\frac{2le}{h^*})^{h^*} + \mu(\frac{2le}{h})^h\right) \tag{32}$$

where μ is the measure of $Z(2l)/\Gamma_{2l}$.

Note that

$$\mu < 1 - (1-\eta)^{2l} < 2l\eta$$

where $1-\eta$ is the measure of the set X^*. According to the condition of the theorem the inequality

$$\eta < \frac{1}{2l}(\frac{2le}{h^*})^{h^*}(\frac{h}{2le})^h$$

holds. We then derive

$$\mu < (\frac{2le}{h^*})^{h^*}(\frac{h}{2le})^h \tag{33}$$

Substituting (33) into (32) we obtain

$$P(A_\epsilon|B_\delta) < (\frac{2le}{h^*})^{h^*} \frac{\exp\{-\frac{\epsilon\delta l}{4}\}}{P(B_\delta)}$$

Then, as in Appendix 1, we obtain the estimate of the expectation (10).

We can derive the estimate (11) similarly.

10 Appendix: Estimation of the Maximum Error Rate Difference on Two Half-Samples

We would like to estimate the value of the maximal error rate difference of a learning machine on two half-sets (maximal deviation). Denote by $\bar{Z}^{2l}$ a sequence obtained from the set Z^{2l} of equation (15) by changing the values of ω_i for the first half set.

$$\bar{Z}^{2l} = x_1, \bar{\omega}_1; x_2, \bar{\omega}_2, ..., x_l, \bar{\omega}_l \ ; \ x_{l+1}, \omega_{l+1}; ...; x_{2l}, \omega_{2l}, \tag{34}$$

$$\bar{\omega}_i = 1 - \omega_i.$$

We will use $\bar{Z}^{2l}$ as a training set for the learning machine. In this case, training results in the minimization of

$$R(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(\bar{\omega}_i - f(x_i,\alpha))^2 + \frac{1}{l}\sum_{i=1}^{l}(\omega_i - f(x_i,\alpha))^2 \tag{35}$$

It is easy to see that the minimum of (36) is obtained for the same function $f(x, \alpha^*)$ that maximizes the deviation. Since ω and $f(x, \alpha)$ have binary values,

$$(\bar{\omega}_i - f(x_i,\alpha))^2 = 1 - (\omega_i - f(x_i,\alpha))^2. \tag{36}$$

From (35) and (36) it follows that

$$R(\alpha) = 1 - \frac{1}{l}\sum_{i=1}^{l}(\omega_i - f(x_i,\alpha))^2 + \frac{1}{l}\sum_{i=1}^{l}(\omega_i - f(x_i,\alpha))^2. \tag{37}$$

Therefore, minimizing $R(\alpha)$ is equivalent to maximizing

$$\bar{R}(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(\omega_i - f(x,\alpha))^2 - \frac{1}{l}\sum_{i=1}^{l}(\omega_i - f(x_i,\alpha))^2 = (\nu_1(\alpha, Z^{2l}) - \nu_2(\alpha, Z^{2l})). \tag{38}$$

To summarize, the maximization of deviation is obtained by training the learning machine with the training set $\bar{Z}^{2l}$. The value of the maximal deviation is related to the minimum of the training functional through

$$\bar{R}(\alpha) = 1 - R(\alpha).$$

References

Abu-Mostafa, Yaser, S. (1993). Hints and the VC Dimension. *Neural Computation*, 5(2).

Baum, E. B. and Haussler, D. (1989). What Size Net Gives Valid Generaliztion? *Neural Computation*, 1:151–160.

Cortes, C. (1993). personnal communication.

Devroye, L. (1982). Bounds for the uniform deviations of empirical measures. *Journal of Multivariate Analysis*, 12:72–79.

Guyon, I., Vapnik, V., Boser, B., Bottou, L., and Solla, S. (1992). Structural Risk Minimization for Character Recognition. In Moody, J., Hanson, S., and Lippmann, R., editors, *Advances in Neural Information Processing Systems 4*, Denver, CO. Morgan Kaufman.

Le Cun, Y., Denker, J. S., Solla, S., Howard, R. E., and Jackel, L. D. (1990). Optimal Brain Damage. In Touretzky, D., editor, *Advances in Neural Information Processing Systems 2*, Denver, CO. Morgan Kaufman.

Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York.

Vapnik, V. and Chervonenkis, D. (1971). 1971] On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Applications*, 16:264–280.

Vapnik, V. and Chervonenkis, D. (1989). The necessary and sufficient conditions for consistency of empirical risk minimization method. *Pattern recognition and Image Analysis*, 1(3):283–305.

Weigend, A., Rumelhart, D., and Huberman, B. (1991). Generalization by weight elimination with application to forecasting. In Lippmann, R., Moody, J., and Touretzky, D., editors, *Advances in Neural Information Processing Systems 3*, Denver, CO. Morgan Kaufman.

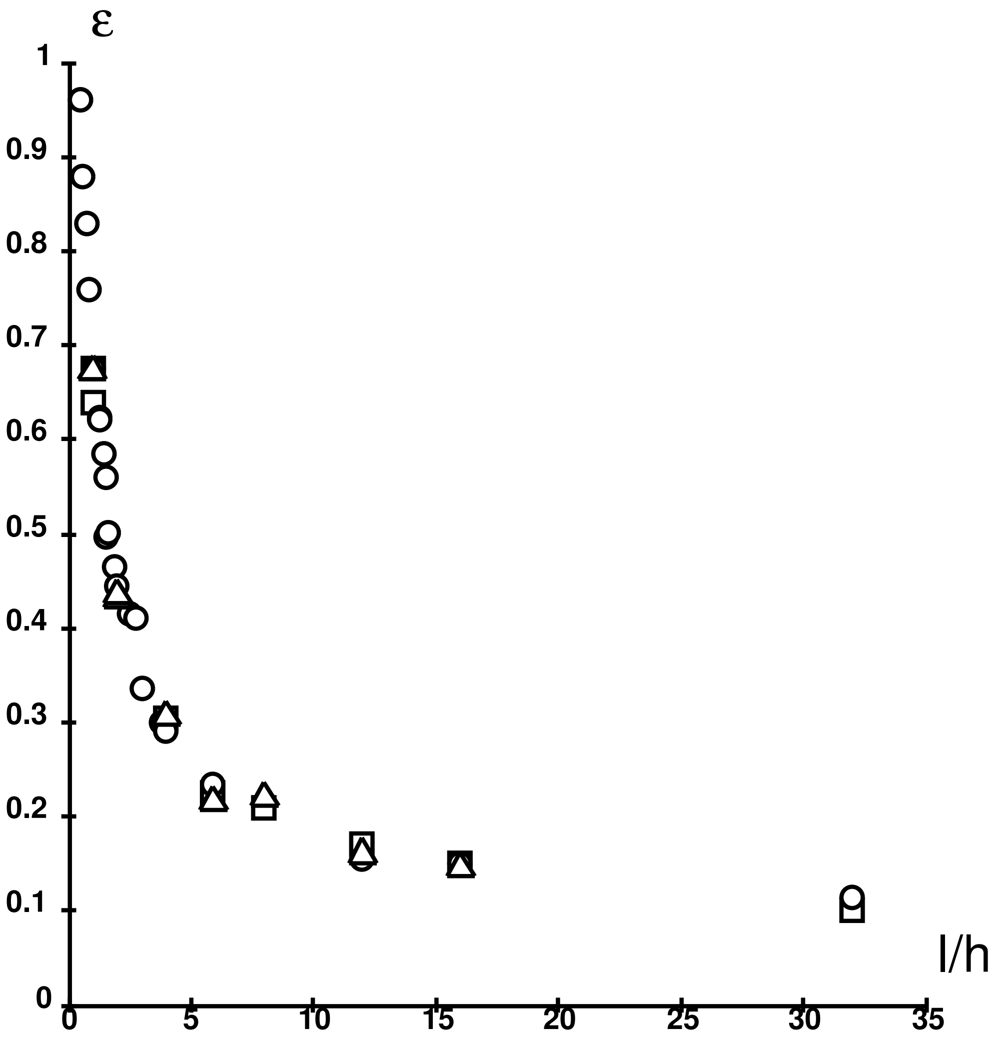

Figure 1: The measured values for average deviation $\bar{\xi}_l$ are shown as a function of l/h, the size of the half-sample normalized by the VC-dimension. These values were obtained in experiments with three different conditional probabilities $P(\omega|x)$. The symbol $\triangle$ denotes values measured for P_1. The symbol $\square$ denotes values measured for P_2. The symbol $\bigcirc$ denotes values measured for P_3.

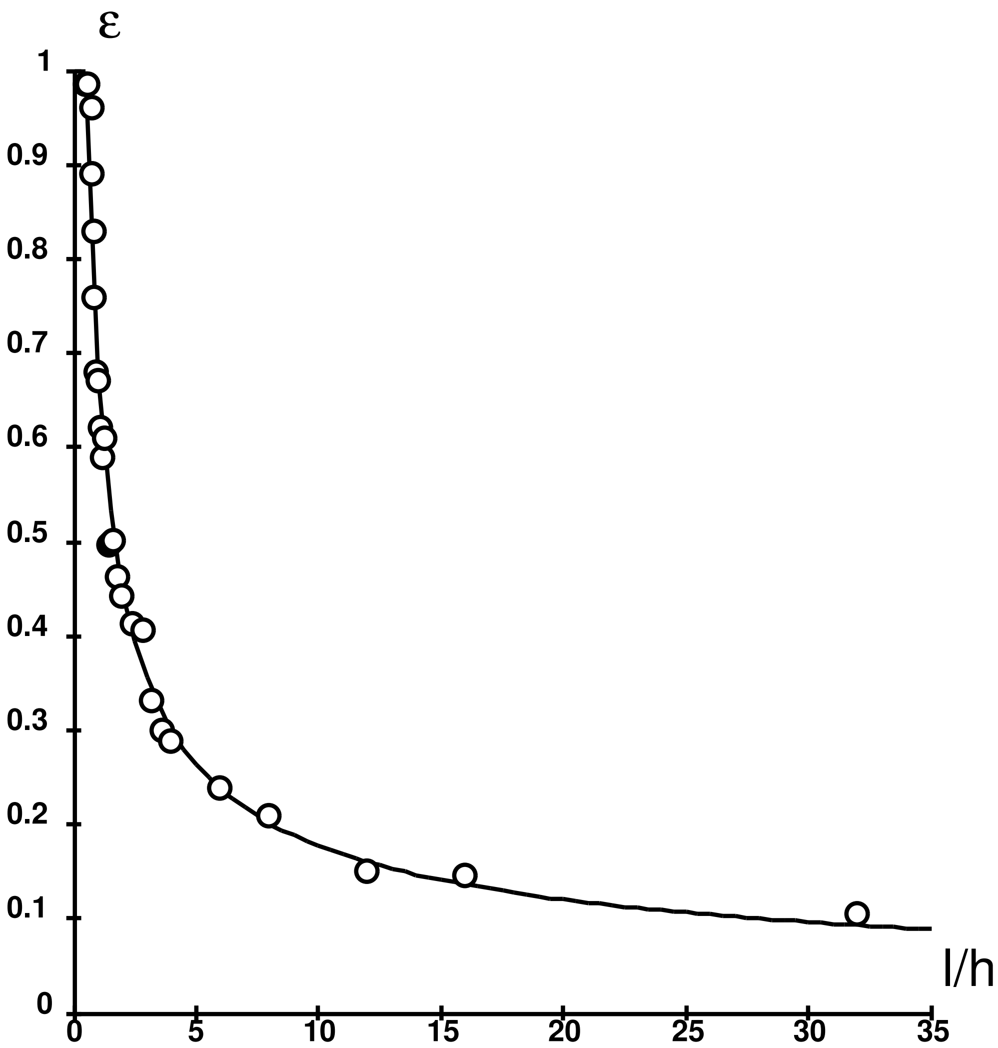

Figure 2: The values for average deviation $\bar{\xi}_l$ are fitted by $\Phi(l/h)$ with parameter values $a = 0.16, b = 1.2$.

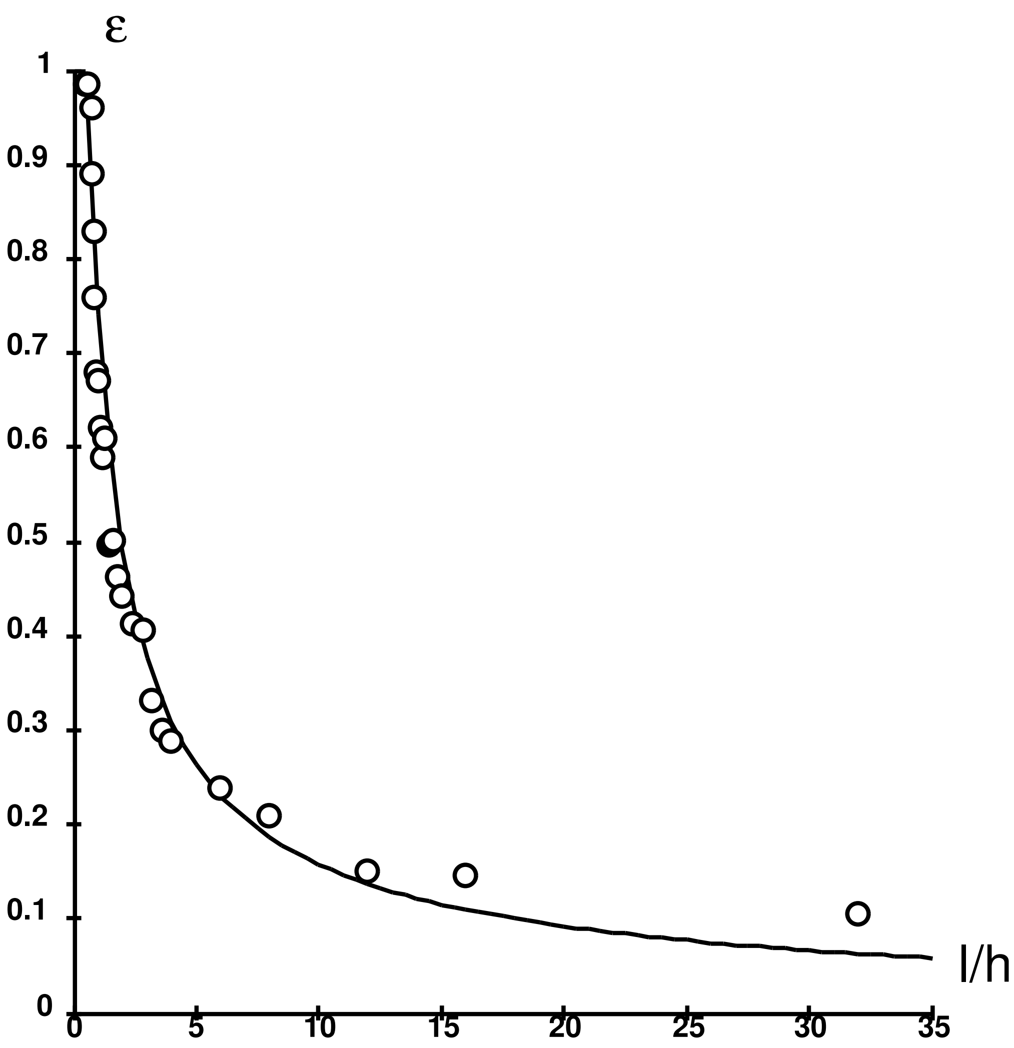

Figure 3: The values for average deviation $\bar{\xi}_l$ are fitted by $\Phi_1(l/h)$ with $d = 0.39$.

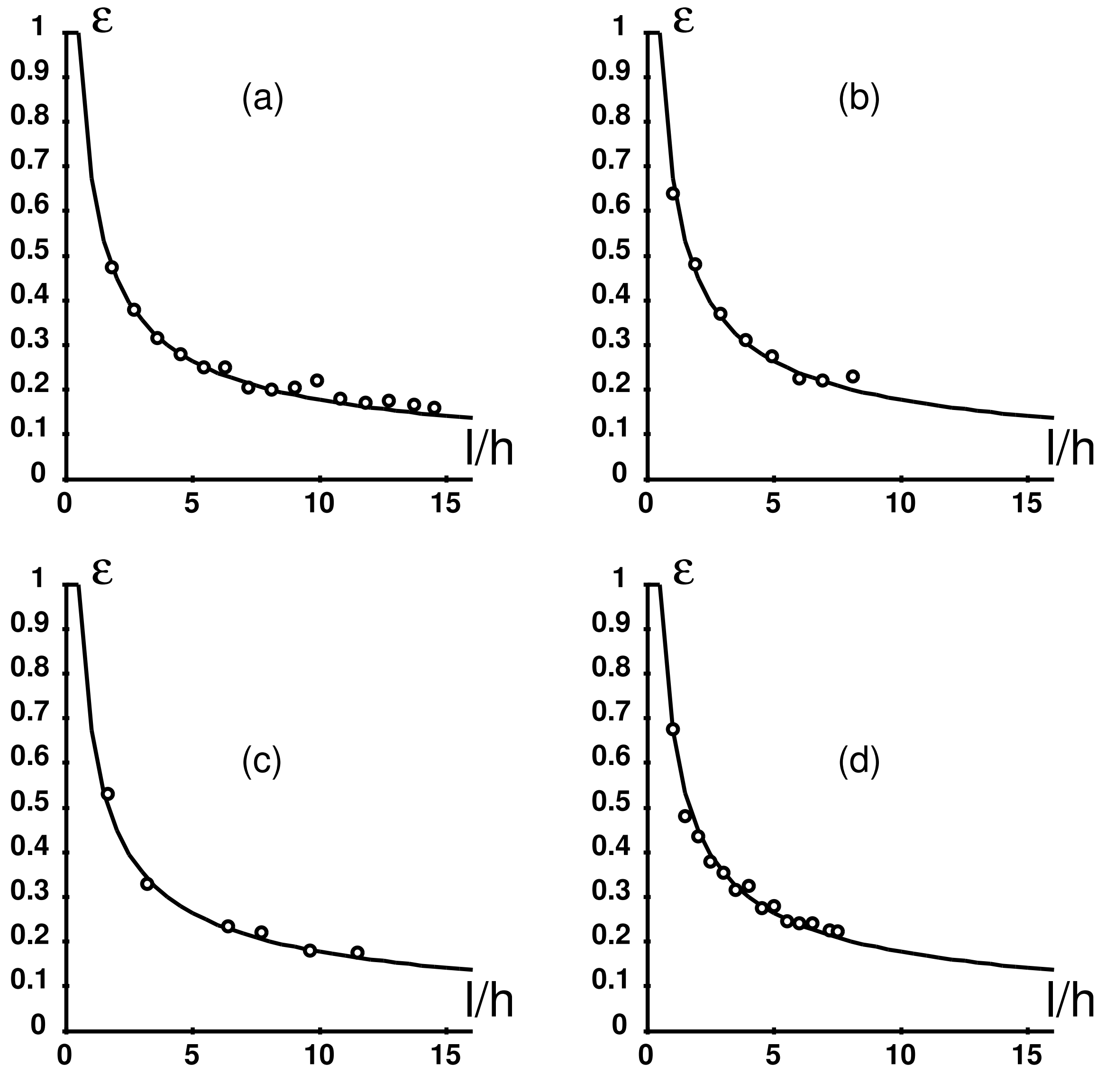

Figure 4: The values of average deviation $\bar{\xi}_l$, measured in control experiments, are fitted by $\Phi(l/h^*)$ (h^* is the measured value of VC-dimension), and are shown as a function of l/h^*. (a) h=40, (b) h=30, (c) h=20 (d) h=10.

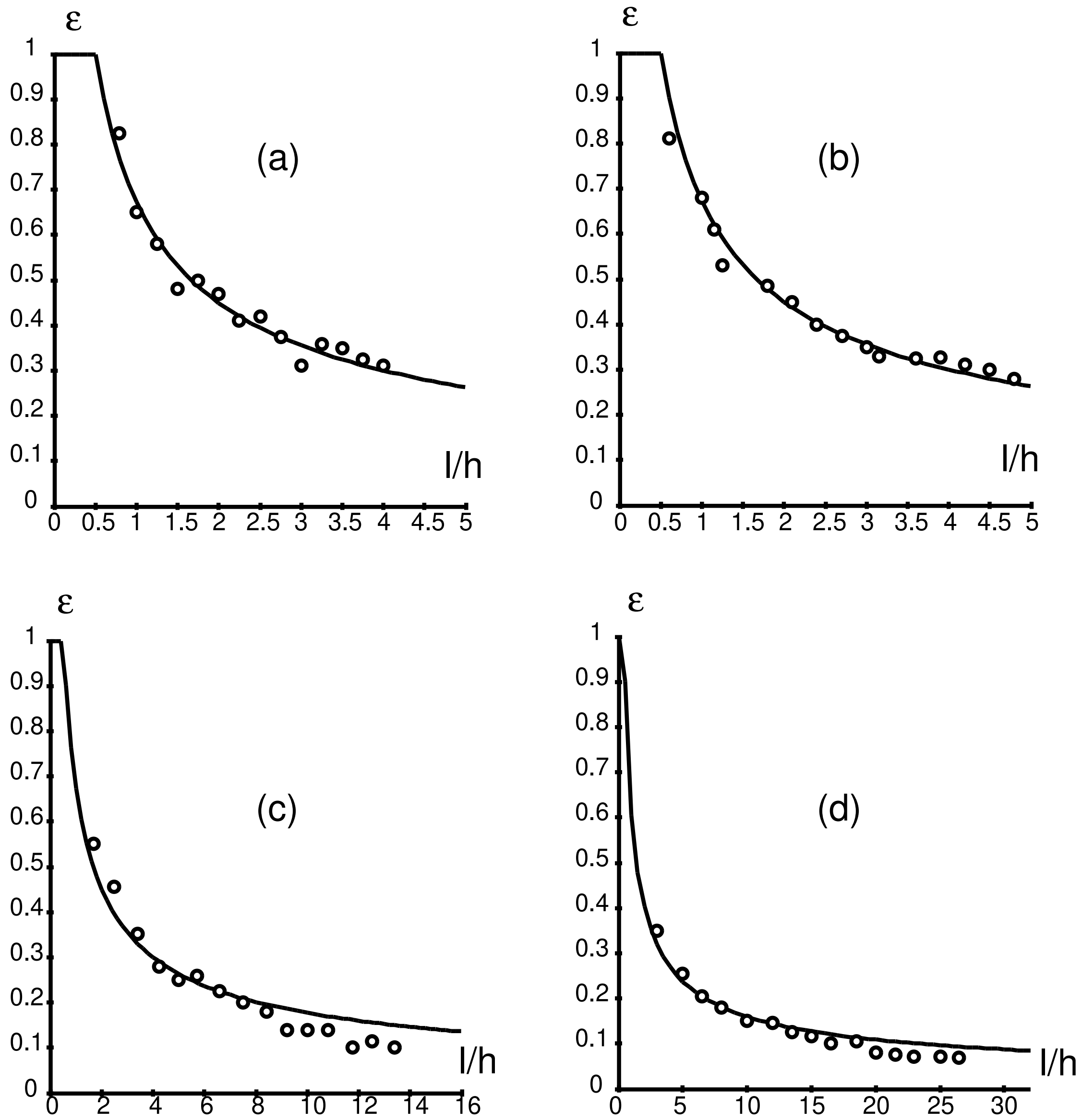

Figure 5: The values of average deviation $\bar{\xi}_l$, measured in experiments with different values of the smoothing parameter β, are fitted by $\Phi(l/h^*)$ (h^* is the measured value of VC-dimension), and are shown as a function of l/h^*. (a) For $\beta = 0.1$. The estimated VC-dimension is 40. (b) For $\beta = 0.05$. The estimated VC-dimension is 33. (c) For $\beta = 0.02$. The estimated VC-dimension is 12. For $\beta = 0.01$. The estimated VC-dimension is 6.

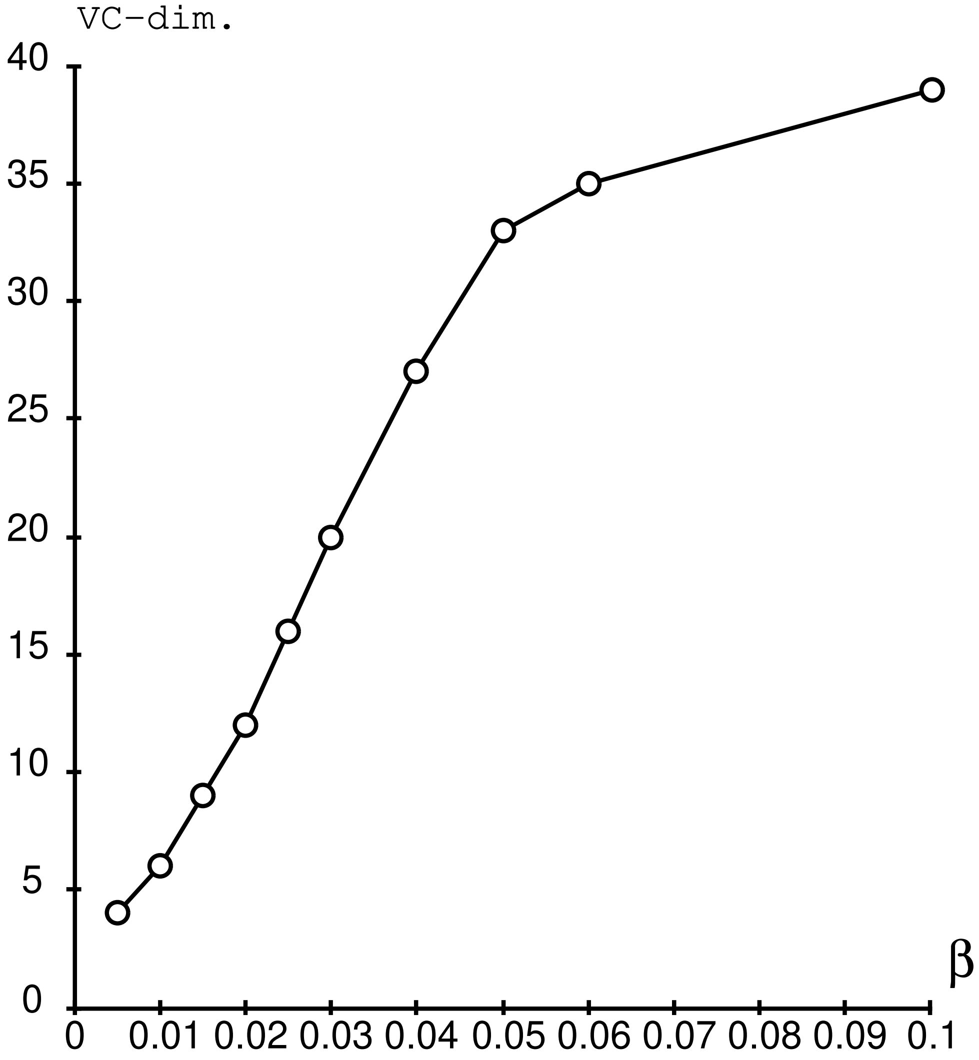

Figure 6: The estimated effective VC-dimension as a function of the smoothing parameter β.