Neural Network Interest Group

Título/*Title*:
Introduction to Statistical Learning Theory
PART I – Basic Notions. Data Classification

Autor(es)/*Author(s)*:
J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 1 /2003

FEUP/INEB, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Título/*Title*:
Introduction to Statistical Learning Theory
PART I – Basic Notions. Data Classification

Autor(es)/*Author(s)*:
J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 1 /2003
Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Introduction to Statistical Learning Theory

Part I –Basic Notions. Data Classification

J.P. Marques de Sá (*)

Part II – Data Regression

F. Sereno (**), J.P. Marques de Sá (*)

(*) FEUP – Faculdade de Engenharia, Universidade do Porto
INEB – Instituto de Engenharia Biomédica (PSI)

(**) Escola Superior de Educação do Porto - IPP
INEB – Instituto de Engenharia Biomédica (PSI)

May, 2003

Foreword

The main purpose of the present tutorial text is to familiarize the reader with the main topics of Statistical Learning Theory. We then skip proofs of Theorems which can be found in the References and put the emphasis on illustrative examples.

Part I –Basic Notions. Data Classification

J.P. Marques de Sá

jmsa@fe.up.pt

Contents

1 Learning Problem

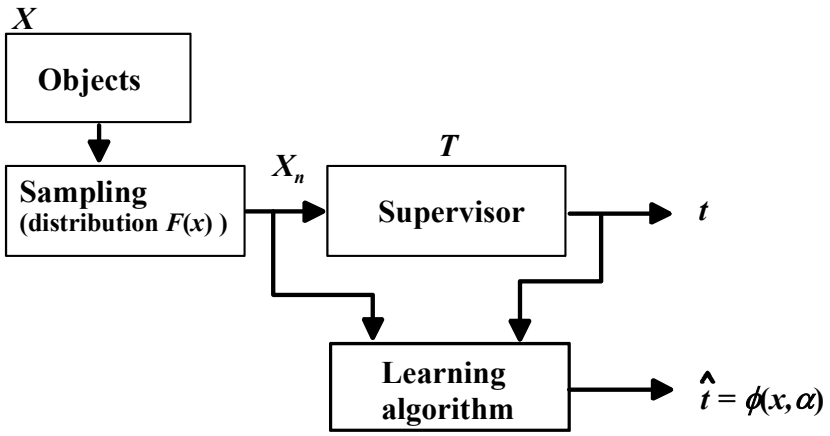


Figure 1.1

Figure 1.1 shows the general configuration of an algorithm/machine attempting to learn the target values t assigned by a supervisor to a set of objects, represented by a feature or predictor vector x. These are sampled according to a probability distribution $F(x)$, often unknown. The algorithm/machine issues estimates $\hat{t} = \phi(x, \alpha)$, where $\alpha \in \mathrm{A}$ is a parameter vector from a certain parameter vector set A. (When the machine is a neural network α is a weight vector.)

Let:

X- Object space (objects, cases, patterns, instances)
X_n- Sample with n objects
T- Target value domain (e.g. {0, 1} or [0, 1])

Consider the risk $$R(\alpha) = \int Q(z, \alpha) dF(z), \quad \alpha \in \mathrm{A}$$ [1] 1.1

with
$z = (t, x)$, Target-object data pair
$Q(z, \alpha) = L(z, \phi(x, \alpha))$ Loss function
$\phi(x, \alpha)$ Approximating function
$dF(z)$ Joint probability distribution of (t, x)

Learning problem:

Choose in the function set $\{Q(z, \alpha), \ \alpha \in \mathrm{A}\}$ a function $Q(z, \alpha_0)$ (therefore, an optimal parameter α_0) which minimizes $R(\alpha)$ when $F(z)$ is unknown but a sample with n random i.i.d. (independent and identically distributed) observations $Z_n = \{z_1, \ldots, z_n\}$ is given.

The loss function $Q(z, \alpha)$ can be suitably chosen in order to encompass the classification, regression and pdf learning problems.

[1] Note that it is an expectation of the loss function $Q(z, \alpha)$.

In this Part I we will restrict ourselves to the data classification ("pattern recognition") problem, with:

$z_i = (\omega_i, x_i)$ (where x_i is a d-dimensional vector and ω_i is a class label; thus, z_i has $d + 1$ coordinates)

Furthermore, for classification problems, one usually uses:

$$Q(z_i, \alpha) = L(\omega_i, \phi(x_i, \alpha)) = \begin{cases} 0 & \omega_i = \phi(x_i, \alpha) \\ 1 & \omega_i \neq \phi(x_i, \alpha) \end{cases},$$

i.e., the loss function is an *indicator function* ($T = \{0, 1\}$), assigning zero loss to correct classifications and equal unitary losses to misclassifications.

2 Empirical Risk Minimization (ERM) Principle

Determination of the α_0 minimizing 1.1 is generally impossible. In practice, one has available a *training set* Z_n with n objects, and approximates $Q(z, \alpha_0)$ by the function $Q(z, \alpha_n)$ which minimizes the *empirical risk* in the training set Z_n:

$$R_{\text{emp}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i, \alpha) \qquad 2.1$$

For the classification problem[2]:

$$R_{\text{emp}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i, \alpha) = \frac{1}{n}\sum_{i=1}^{n} L(\omega_i, \phi(x_i, \alpha)) = \hat{P}_e \qquad 2.2$$

Therefore: "Minimize $R_{\text{emp}}(\alpha)$" $\equiv$ "determine the function $\phi(x,\alpha_n)$ that achieves the smallest error rate (estimate of the probability of error, $\hat{P}_e$), in the training set Z_n". The empirical risk minimization is widely used in practice. Instead of an optimal (usually unknown) α_0 we are only able to determine (empirically) a "best" α_n.

3 Consistency of the Learning Process

The following issues can be raised when using the empirical risk minimization principle:

1. **Do the empirical risks converge to the optimal risk when the training-set size increases to arbitrarily large values ($n \to \infty$)?**
2. **Do the true risks of the machine designed with a training set also converge to the optimal risk when $n \to \infty$ (i.e. does the machine generalize its performance to any new independent set of cases)?**

[2] Note that $R_{\text{emp}}(\alpha)$ is the relative frequency (or frequency for short) of the correct classification event, whereas $R(\alpha)$ is the (true) probability of correct classification.

Classical definition of consistency:

The ERM principle (method) is *consistent* for the set of functions $\{Q(z, \alpha), \ \alpha \in A\}$, and for the probability distribution $F(z)$, if the following two sequences converge in probability to the same limit:

1: $$R_{\text{emp}}(\alpha_n) \xrightarrow[n\to\infty]{P} \inf_{\alpha\in A} R(\alpha).$$

2: $$R(\alpha_n) \xrightarrow[n\to\infty]{P} \inf_{\alpha\in A} R(\alpha).$$

Figure 3.1 depicts the general behavior of learning curves. Often – but not necessarily for all n -, $R_{\text{emp}}(\alpha_n)$ is "optimistic" and $R(\alpha_n)$ is "pessimistic". For consistent learning the stochastic processes $R_{\text{emp}}(\alpha_n)$ and $R(\alpha_n)$ should both converge in probability to inf $R(\alpha)$.

Note that for a given sample, one expects that $R_{emp}(\alpha_n) < R(\alpha_n)$, because the function $\phi(x,\alpha_n)$ obtained by ERM (equivalently, the particular value of the parameter α_n obtained by ERM) constitutes a biased estimate of the functions minimizing true risk.

The above first condition states that the empirical risks, computed with 2.2, should converge to the lowest attainable (optimal) true risk, for all the set of functions. After all, without this condition, the ERM principle would be quite useless.

The second condition states that the expected risks, computed with 1.1, taking into account the *sample distribution* and the determined α_n, should also converge to the optimal risk. This is a generalization condition, because the expected risk depends not only on the particular α_n (thus, on the particular sample) determined by the ERM principle, but depends also on the probability of every possible sample.

The above consistency conditions include trivial cases of consistency[3]. A stricter, nontrivial consistency condition is needed (see VN Vapnik, 1999, for details; see also the Appendix for notions on convergence of stochastic processes).

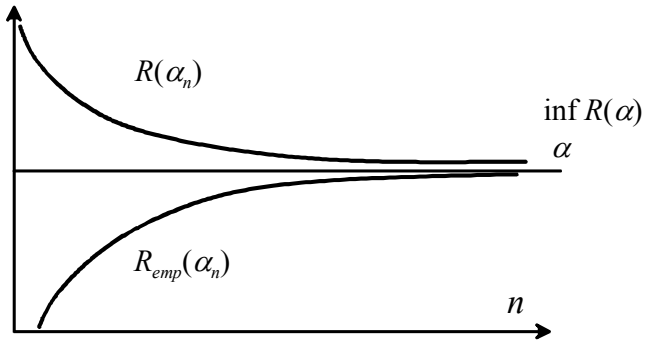


Figure 3.1

[3] For instance, by adding to the $Q(z,\alpha)$ family a function $\phi(z) < \inf_{\alpha\in A} Q(z,\alpha)$, we obtain a new trivially consistent set, because true and empirical risks will converge to $\phi(z)$. Such sets with a minorizing function are said to be trivially consistent.

Key Theorem for consistent learning (Vapnik and Chervonenkis, 1989):

Let $\{Q(z, \alpha), \quad \alpha \in \mathrm{A}\}$ be a set of functions that have a bounded loss for the probability measure $F(z)$:

$$A \le \int Q(z,\alpha) dF(z) \le B \qquad \forall \alpha \in \mathrm{A}$$

Then, for the ERM principle to be nontrivially consistent, it is necessary and sufficient that the empirical risk converges uniformly[4] to the actual risk:

$$\lim_{n\to\infty} \quad P\left\{ \sup_{\alpha\in\mathrm{A}} (R(\alpha) - R_{\mathrm{emp}}(\alpha)) > \varepsilon \right\} = 0 \qquad \forall \varepsilon > 0 \tag{3.1}$$

Equivalently:

$$\Delta(\alpha_{\mathrm{worst}}) = \sup_{\alpha\in\mathrm{A}} (R(\alpha) - R_{\mathrm{emp}}(\alpha)) \text{ converges in probability to zero.}$$

Figure 3.2 shows learning curves for two different sets of parameters, α_1 and α_2. For consistent learning the worst case difference between $R_{\mathrm{emp}}(\alpha_n)$ and $R(\alpha_n)$ must converge in probability to zero.

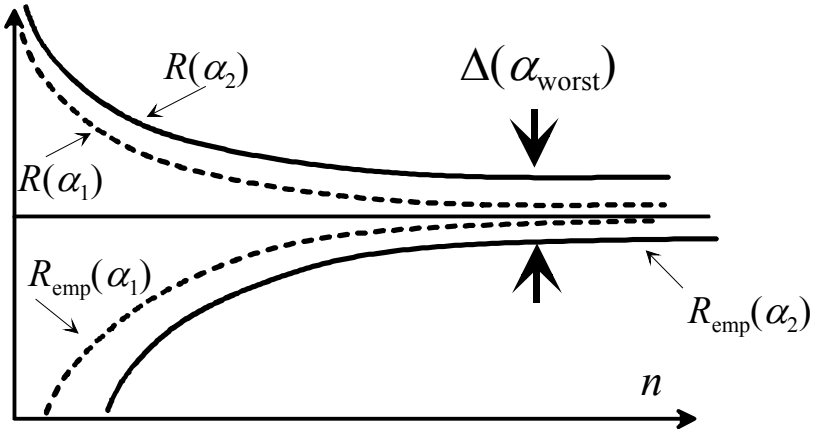


Figure 3.2

Example 3.1

Consider the classification of two classes of points in [0, 1], i.e., $X \times T = [0,1] \times \{0, 1\}$. The data distribution is as follows (see Figure 3.3a):

$P(\omega_1) = P(\omega_2) = ½; \quad p(x \mid \omega_1) = \mathrm{unif}(0, 0.47); \qquad p(x \mid \omega_2) = \mathrm{unif}(0.53, 1)$

[4] This is called one-sided uniform convergence. The two-sided uniform convergence corresponds to $\lim_{n\to\infty} \quad P\left\{ \sup_{\alpha\in\mathrm{A}} | R(\alpha) - R_{\mathrm{emp}}(\alpha) | > \varepsilon \right\} = 0$. Using two-sided convergence in the key theorem is also allowed (see section 5).

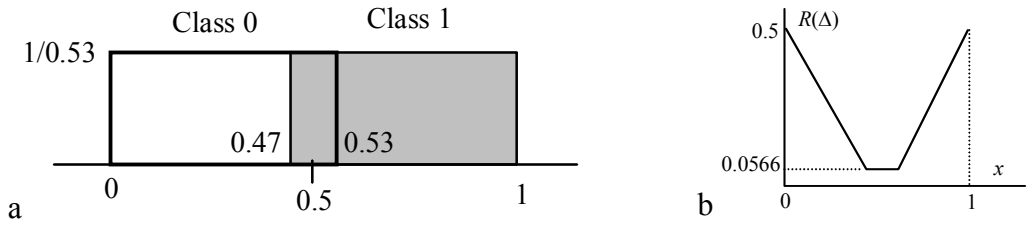


Figure 3.3

A 10-case sample, Z_{10}, sorted according to the values of x and obtained with this distribution could be:

#	1	2	3	4	5	6	7	8	9	10
ω_i	0	0	0	0	1	0	1	1	1	1
x	0.0138	0.0401	0.1337	0.2092	0.5074	0.5130	0.8035	0.8762	0.9259	0.9742

Assume that, in order to classify a data sample we use the following set of approximating (classifying) functions:

$$\phi(x,\alpha) = \{\theta(x-\alpha);\ \alpha\in\Re\},\ \text{where } \theta \text{ is the Heaviside function.}$$

Thus, the learning procedure consists of choosing a threshold that achieves the dichotomy with minimal empirical error. For that purpose, we may just scan the ordered set from left to right and choose a best threshold. In the sample above, if we choose $\alpha\in$] 0.2092, 0.5074] we obtain one misclassified case (#6). If $\alpha\in$] 0.5074, 0.5130] we obtain two misclassified cases (#5, #6). If $\alpha\in$] 0.5130, 0.8035] again we obtain one misclassified case. A minimum error situation corresponds to the first interval and we set α_{10} as the middle point of the interval $\alpha_{10} = \Delta = 0.3583$.
For any Δ (i.e. the ERM derived α_n for a given Z_n) the empirical risk (equivalently, the training set error rate) is:

$$R_{\text{emp}}(\Delta) = \frac{1}{n}\sum_{i=1}^{n} L(\omega_i, \phi(x_i,\Delta)) = \frac{1}{10} = 0.1\ .$$

The true risk (probability of error) is computed as follows:

$$R(\Delta) = \int Q(z,\Delta)dF(z) = \sum_{i=0}^{1} P(\omega_i)\int_0^1 L(\omega_i,\phi(x,\Delta)p(x\mid\omega_i)dx$$

We now distinguish three situations (see Figure 3.3b):

$\Delta < 0.47$: $$R(\Delta) = \frac{1}{2}\int_{\Delta}^{0.53}\frac{1}{0.53}dx = \frac{1}{2}\left(1-\frac{\Delta}{0.53}\right)$$

$0.47 \le \Delta < 0.53$: $$R(\Delta) = \frac{1}{2}\int_{\Delta}^{0.53}\frac{1}{0.53}dx + \frac{1}{2}\int_{0.47}^{\Delta}\frac{1}{0.53}dx = \frac{1}{2}\left(1-\frac{0.47}{0.53}\right) = 0.0566$$

$\Delta \ge 0.47$: $$R(\Delta) = \frac{1}{2}\int_{0.47}^{\Delta}\frac{1}{0.53}dx = \frac{1}{2}\left(\frac{\Delta-0.47}{0.53}\right)$$

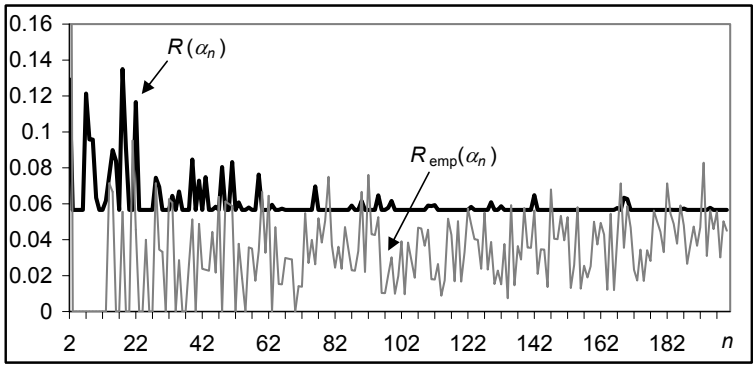


Figure 3.4

Figure 3.4 shows an experiment of randomly drawing a sample with n cases (points) and computing the empirical and true risk as described above (simulation in MATLAB). The true risk quickly converges to the optimal risk, $\inf(R(\Delta)) = 0.0566$ (Δ_{opt}=0.5). The empirical risk has a slower convergence. Figure 3.5 is similar to the previous one showing coarser details up to $n = 600$.

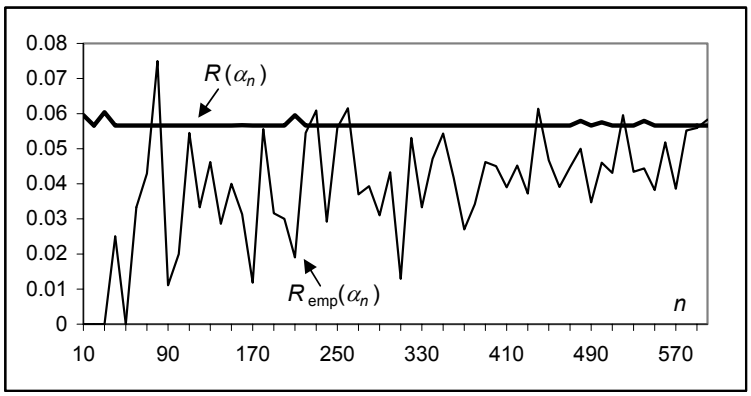


Figure 3.5

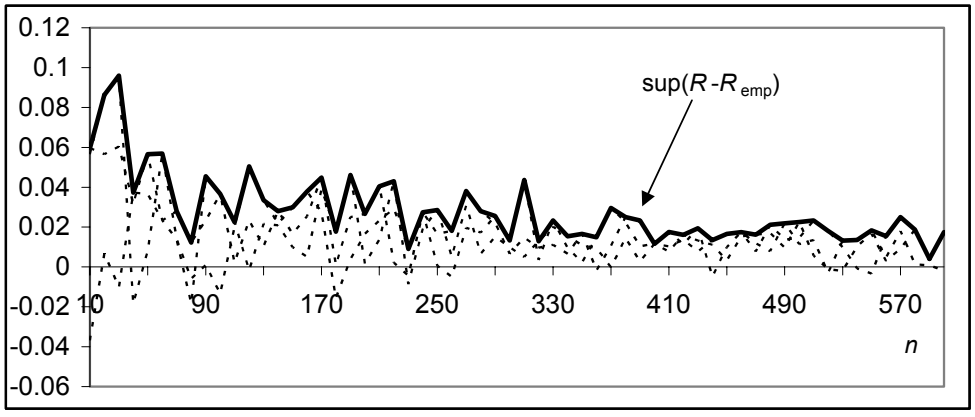


Figure 3.6

Finally, Figure 3.6 shows with dotted lines the $R(\Delta) - R_{emp}(\Delta)$ differences for 3 experiments, and with solid line the $\sup(R(\Delta) - R_{emp}(\Delta))$. This last stochastic series should converge in probability to zero in order to guarantee that the above described learning process is nontrivially consistent. As a matter of fact, it does, as will be shown in section 5. □

Example 3.2

We now analyze the convergence of $\sup(R(\Delta) - R_{emp}(\Delta))$ as in the previous example, considering that the data distributions are exponential (see Figure 3.7):

$$p(x \mid \omega_1) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}; \qquad p(x \mid \omega_2) = \begin{cases} \lambda e^{-\lambda(1-x)} & x \leq 1 \\ 0 & x > 1 \end{cases}$$

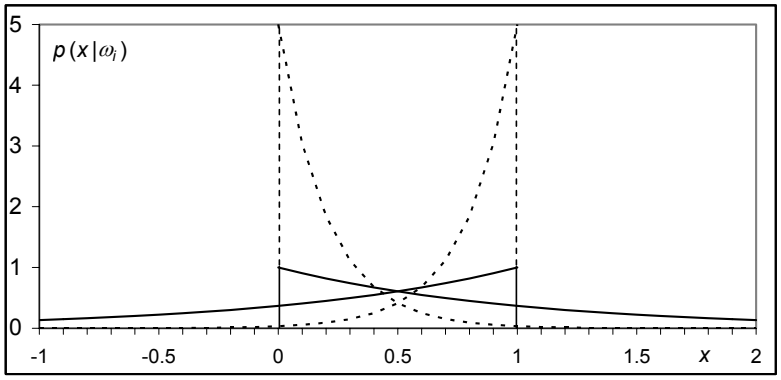


Figure 3.7. Data distributions for λ =1 (solid line) and λ =5 (dotted line).

We use the same "family of classifying functions" as in the previous example. For any $\Delta \in]-\infty, 0.5]$ the (true) error probability of the classifier is:

$$R(\Delta) = P(\omega_1) \int_{-\infty}^{\Delta} \lambda e^{-\lambda(1-x)} dx + P(\omega_2) \int_{\Delta}^{+\infty} \lambda e^{-\lambda x} dx = \left(e^{-\lambda(1-\Delta)} + e^{-\lambda\Delta} \right) / 2 ,$$

and symmetrically for $\Delta \in [0.5, +\infty[$ (see Figure 3.8). The optimal error corresponds to $\Delta = 0.5$ with $\inf(R(\Delta)) = e^{-\lambda/2}$.

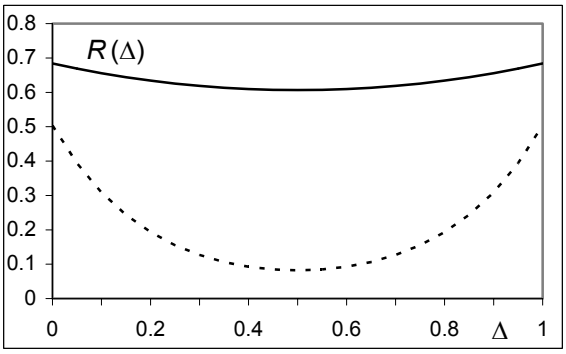


Figure 3.8. True error curves: Dotted curve: λ=5; Solid curve: λ=1.

Figure 3.9 shows the true and empirical error curves for one experiment with λ=1 (left) and λ=5 (right). Figure 3.10 shows $\sup(R(\Delta) - R_{emp}(\Delta))$ for three experiments with λ=1 (left; $R_{opt} = e^{-0.5}$=0.6065) and λ=5 (right; $R_{opt} = e^{-2.5}$=0.0821).

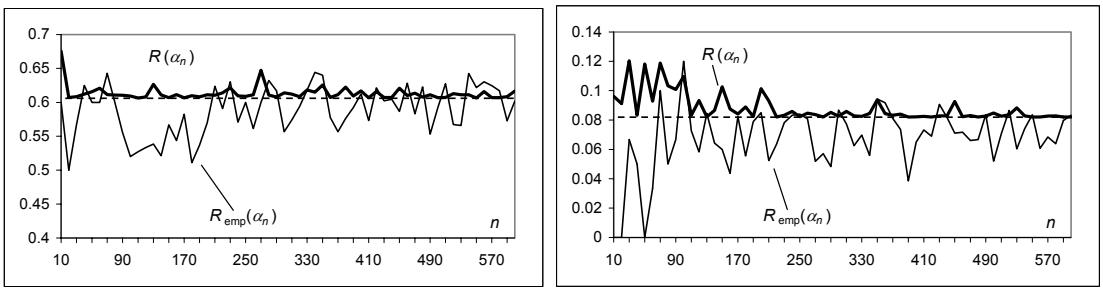


Figure 3.9

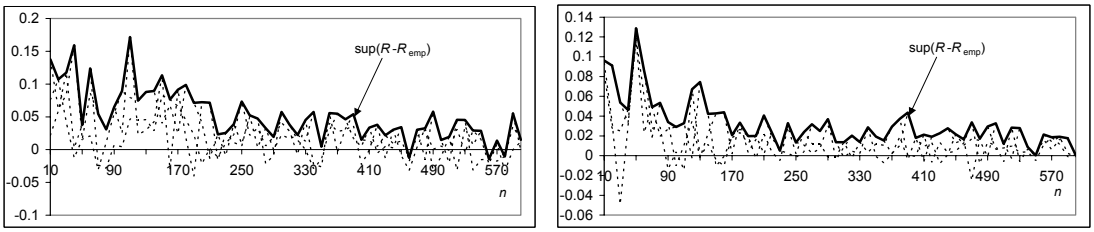


Figure 3.10

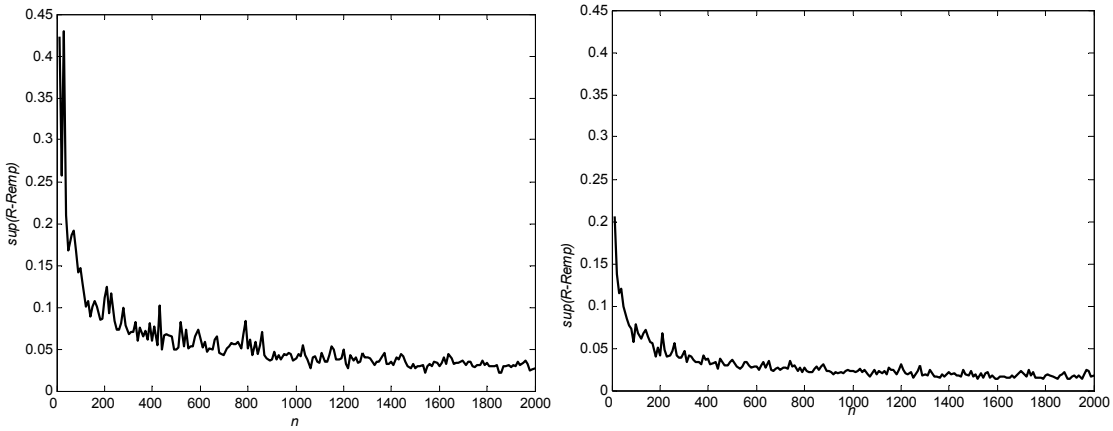


Figure 3.11

Figure 3.11 shows $\sup(R(\Delta) - R_{\text{emp}}(\Delta))$ for 50 experiments with λ=1 (left) and λ=5 (right). We notice that for λ=5 a faster convergence is obtained as could be expected by looking at the true error curves (Figure 3.8), which reflect the fact that for λ=5 the classes are "better" separated. □

4 Diversity of a set of indicator functions

The Key Theorem 3.1 does not afford a constructive procedure for assessing the consistency of the learning process based on the approximating functions. In order to derive practical useful conditions one needs to characterize the "expressiveness" of the loss functions. This can be done using a measure of the diversity of separations in relation to the sample complexity and the classifier $\phi(x, \alpha)$ complexity. For this

purpose, let $Q(z,\alpha)$, $\alpha \in$ A, be a set of indicator functions and consider a sample of size n:

$$Z_n = \{z_1, \ldots, z_n\}$$

The *diversity* of the set of functions $Q(z,\alpha)$ on the given sample is characterized by:

$$N^{\mathrm{A}}(Z_n) \equiv \{\text{Nr. of different separations of } Z_n \text{ achieved with } Q(z,\alpha),\ \alpha \in \mathrm{A}\}$$ [5]

Consider the set of n-dimensional binary vectors for $\alpha \in$ A:

$$q(\alpha) = (Q(z_1,\alpha), \ldots, Q(z_n,\alpha))$$

$N(Z_n)$ is the number of different vertices of the n-dimensional cube that can be obtained on the basis of the sample Z_n and the set of functions $Q(z,\alpha)$.

Example 4.1

Consider a one-dimensional (d = 1) set of objects and the following family of classification functions into two classes ($T \equiv \Omega = \{\omega_i\} \equiv \{0, 1\}$):

$$\underset{\alpha=(a,b)\in\Re^2}{\phi(x,\alpha)} = \begin{cases} 1 & ax+b \geq 0 \\ 0 & \text{otherwise} \end{cases};$$

Equivalently:

$$\phi(x,\alpha) = \theta(ax+b) \text{ where } \theta \text{ is the Heaviside function.}$$ [6]

The loss function for classification is:

$$Q(z,\alpha) = L(\omega, \phi(x,\alpha)) = \begin{cases} 0 & \omega = \phi \\ 1 & \omega \neq \phi \end{cases}$$

For n = 3 points with the classifications shown below (solid circle for ω = 1 and open circle for ω = 0) in Figure 4.1 and according to the values of the parameters (a, b), we have:

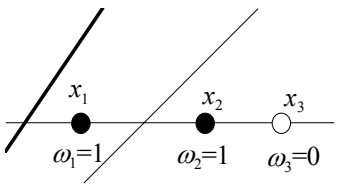


Figure 4.1

$a > 0$, decreasing b	$a < 0$, increasing b
110 (bold line)	001
010	101
000	111
001	110

[5] In general $N^{\mathrm{A}}(Z_n)$ will depend on the set of parameters A (see following Example 4.2). However, for the sake of simplicity we will from now on denote $N(Z_n)$.

[6] Note that this classifying family is equivalent to the one used in Examples 3.1 and 3.2, with $\alpha = -b/a$.

The set of 3-dimensional binary vectors $q(\alpha) = (Q(z_1,\alpha), \ldots, Q(z_n,\alpha))$ is as shown in Figure 4.2. We see that $N(Z_3) = 6$.

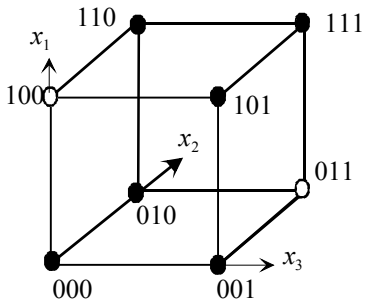


Figure 4.2

For the other distinct configurations of 3 points (excluding the complementary configurations) we obtain the cubes shown in Figure 4.3.

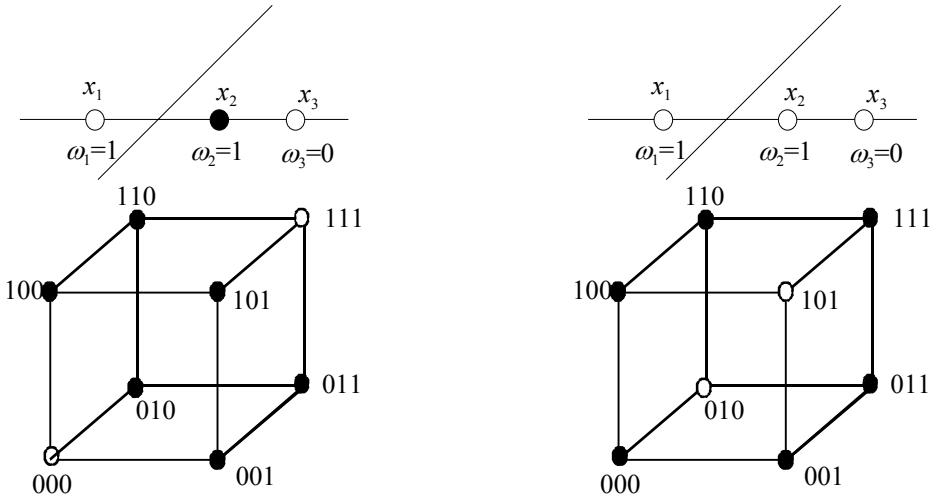


Figure 4.3

In general, for this set of linear functions we get: $N(Z_n) = 2n$

Proof:
For the $Q(z,\alpha)$ set of functions, $N(Z_n)$ corresponds to the number of achievable dichotomies. Let us consider the number of *runs* (one-symbol sequences) in an n-dimensional sequence of 0s and 1s. Then, the only sequences corresponding to achievable dichotomies are those that have 1 or 2 runs. For 1 run there are 2 sequences. For 2 runs there are $2(n-1)$ sequences.
E.g. for $n = 4$:

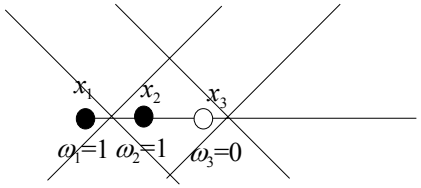
1 runs	2 runs		
0000 1111	0001 1000	0011 1100	0111 1110

In this example $N(Z_n)$ varies with n, but not with the particular sample.

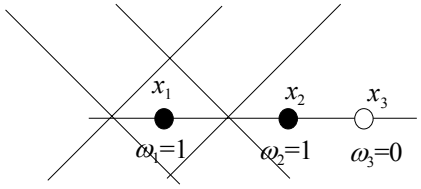
□

Example 4.2

Same as before but now the family of functions ϕ is constrained with $b \in [-B, B]$.
Then we get situations such as shown in Figure 4.4: on the left, the straight lines can "walk beyond" x_3 and all $2n$ dichotomies are achievable; on the right the straight lines can only go until somewhere between x_1 and x_2.



$N(Z_3) = 6$



$N(Z_3) = 4$

Figure 4.4

Thus, in this example $N(Z_n)$ depends on the particular sample. □

Example 4.3

Again a one-dimensional case but with a more "expressive" family of classifying functions:

$$\phi(x, \alpha) = \theta(ax^2 + bx + c)$$

$$\alpha = \{ (a, b, c) \in \Re^3 \}$$

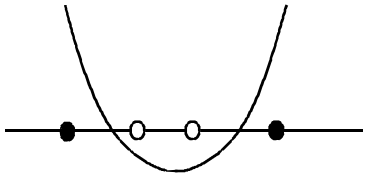


Figure 4.5

$N(Z_n) = 2n$ + "Nr of distinct sequences with 3 runs"

n	1,2 runs	3 runs	N	ln(N)
1	2	0	2	0.693
2	4	0	4	1.386
3	6	2	8	2.079
4	8	6	14	2.639
5	10	12	22	3.091
6	12	20	32	3.466
7	14	30	44	3.784

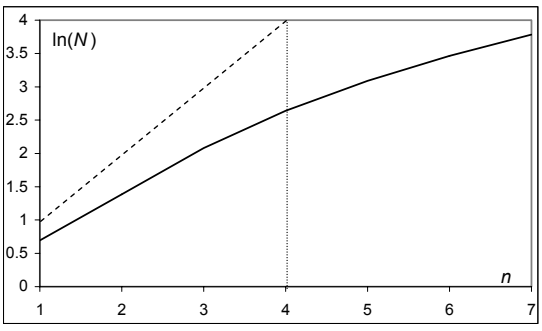


Figure 4.6

Counting the number of distinct sequences with 3 runs, one can build the table above. As shown in Figure 4.6. $N(Z_n)$ grows less than linearly with n.

□

Example 4.4

Now, the family of classifying functions is the following "very expressive" family shown below.

$$\phi(x,\alpha) = \theta(\sin(\alpha x)) \quad \text{with } x \in [0, 2\pi],\ \alpha \in \Re^+$$

Furthermore we consider the set of n points located in $[0, 2\pi]$ such that $x_i = 2\pi 10^{-i}$. Then, for any class-label assignment of the points, $\{\omega_1,\ldots\omega_i,\ldots,\omega_n\} \in \{0,1\}^n$, one can find the following value of α, achieving the classification:

$$\alpha^* = \frac{1}{2}\left(\sum_{i=1}^{n}(1-\omega_i)10^i + 1\right)$$

Thus, $N(Z_n) = 2^n$. □

5 Entropies and Growth Function

The quantity $N(Z_n)$, reflecting the expressiveness of the set of classifying functions, influences the way on how $R_{emp}(\alpha_n)$ and $R(\alpha_n)$ may converge to the optimal risk. To see how, we first need the following:

Definitions:

Random entropy for $\{Q(z, \alpha),\ \alpha \in A\}$ and Z_n sample:

$$H(Z_n) = \ln N(Z_n); \qquad H(Z_n) \text{ is a r.v. (trivial or nontrivial)}$$

Entropy for $\{Q(z, \alpha),\ \alpha \in A\} \equiv$ Expectation of the random entropy:

$$H(n) = E\,[\ln N(Z_n)] \qquad \text{(depends on the distribution law)}$$

Theorem (Vapnik):

For uniform two-sided convergence of the frequencies $R_{emp}(\alpha)$ to their probabilities $R(\alpha)$:

$$\lim_{n\to\infty} P\left\{\sup_{\alpha\in A}\left|R(\alpha) - R_{emp}(\alpha)\right| > \varepsilon\right\} = 0, \qquad \forall \varepsilon > 0$$

it is necessary and sufficient that:

$$\lim_{n\to\infty} \frac{H(n)}{n} = 0, \qquad \forall \varepsilon > 0$$

Remark 1: For a finite number N of functions the condition reduces to:

$$\lim_{n\to\infty} \frac{N}{n} = 0, \qquad \forall \varepsilon > 0$$

Thus, for a finite number of functions uniform two-sided convergence holds.

Remark 2: The condition is also necessary and sufficient for almost sure convergence (see definition of almost sure convergence in the Appendix).

Example 5.1

Let us consider the classification problem of Example 4.1. We have:

$N(Z_n) = 2n$; $H(Z_n) = H(n) = \ln(2n)$, since it is independent of $F(z)$.

Thus: $\lim_{n\to\infty} \frac{H(n)}{n} = \lim_{n\to\infty} \frac{\ln(2n)}{n} = 0$. ERM is consistent for that set of functions, as well as for the set in Examples 3.1 and 3.2.

□

Example 5.2

Let us now consider the classification problem of Example 4.4. We have:

$N(Z_n) = 2^n$; $H(Z_n) = H(n) = n\ln(2)$, since it is independent of $F(z)$.

We have:

$\lim_{n\to\infty} \frac{H(n)}{n} = \lim_{n\to\infty} \frac{n\ln(2)}{n} = \ln(2)$. ERM is inconsistent for that set of functions.

□

6 Three Milestones in Learning Theory

Definitions:

Entropy for sets of indicator functions:	$H(n) = \mathrm{E}\,[H(Z_n)]$
Annealed VC-entropy:	$H_{\mathrm{ann}}(n) = \ln\{\mathrm{E}\,[N(Z_n)]\}$
Growth function:	$G(n) = \ln[\sup_{Z_n} N(Z_n)]$

Note that:

a) Both $H(n)$ and $H_{\mathrm{ann}}(n)$ depend on a probability measure $P(z)$. $G(n)$ is independent of $P(z)$.

b) These quantities satisfy: $H(n) \le H_{\mathrm{ann}}(n) \le G(n)$ 6.1

The three milestones are:

1. **Necessary and sufficient condition for the consistency of the ERM principle:**

$$\lim_{n\to\infty} \frac{H(n)}{n} = 0$$

2. **Sufficient condition for fast convergence:**

$$\lim_{n\to\infty} \frac{H_{\text{ann}}(n)}{n} = 0 \quad \Rightarrow \quad \forall n > n_0, \quad P\{R(\alpha_n) - R(a_0) > \varepsilon\} < e^{-c\varepsilon^2 n} \quad (c > 0 \text{ constant})$$

3. **Necessary and sufficient condition for the consistency of the ERM principle, independently of the probability measure (independently of the problem to be solved):**

$$\lim_{n\to\infty} \frac{G(n)}{n} = 0 \qquad 6.2$$

It is also a sufficient condition for fast convergence.

Computation of $H(n)$ in the case of a sample with only one object (!):

$$H(1) = \text{E}\,[H(Z_1)] \text{ with } Z_1 = \{(\omega, x)_1\}$$

But: $dF(z) = dF(\omega, x) = P(\omega \mid x)dF(x) = P(\omega \mid x)p(x)dx = p(x \mid \omega)P(\omega)dx$

Therefore: $\text{E}\,[H(Z_1)] = \int H((\omega, x))dF((\omega, x)) = \sum_{\omega\in\Omega} P(\omega) \int_X H((\omega, x))p(x \mid \omega)dx$

Computation of $H(n)$ in the case of a sample with two i.i.d. objects:

$$\text{E}[H(Z_2)] = \sum_{\omega_1\in\Omega} \sum_{\omega_2\in\Omega} P(\omega_1)P(\omega_2) \iint_{X_1 x X_2} H((\omega_1, x_1), (\omega_2, x_2))p(x_1 \mid \omega_1)p(x_2 \mid \omega_2)dx_1 dx_2 . \qquad 6.3$$

Example 6.1

Assume that we have two points in $d = 1$, distributed in [0, 1], and two classes, with:

$$P(\omega_1) = P(\omega_2) = ½ \qquad \text{(equal prevalences)}$$

$$p(x_2 \mid \omega_i) = 1 \qquad \text{(uniform for all classes)}$$

$$p(x_1 \mid \omega_1) = \begin{cases} 2 & 0 \le x_1 \le ½ \\ 0 & \text{otherwise} \end{cases}; \qquad p(x_1 \mid \omega_2) = \begin{cases} 0 & 0 \le x_1 \le ½ \\ 2 & \text{otherwise} \end{cases}$$

Figure 6.1 depicts how the pairs of points (x_1, x_2) are distributed in $X_1 x X_2 = [0,1]^2$ with the respective conditional distributions for the two classes.

Assume further that $\{Q(z,\alpha),\ \alpha \in A\}$ is:

$$N(Z_n) = \begin{cases} 2^n & \max(|x_i - x_j|) > ½ \\ 2^{\sqrt{n}} & \text{otherwise} \end{cases}$$

Thus, the number of different separations depends on the data.

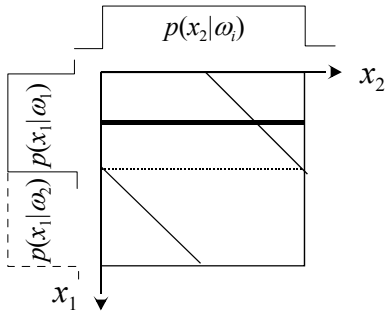


Figure 6.1

Since in 6.3 we have two equal terms for ω_2, independent of ω_2, we get:

$$\text{E}\,[H(Z_2)] = \sum_{\omega_1 \in \Omega} P(\omega_1) \iint_{X_1 x X_2} H(x_1, x_2)\, p(x_1 \mid \omega_1)\, dx_1 dx_2$$

Therefore:

$$H(2) = 0.5 \sum_{\omega_1 \in \Omega} \iint_{X_1 x X_2} H(x_1, x_2)\, p(x_1 \mid \omega_1)\, dx_1 dx_2$$

Given the symmetry for x_1 around ½ (see Figure 6.1), we have:

$$H(2) = 2\underbrace{\left(0.5 \sum \iint H(x_1, x_2)\, p(x_1 \mid \omega_1)\, dx_1 dx_2\right)}_{1 \text{ slice}}$$

Now, we distinguish the two cases of $N(Z_n)$ (triangular region):

$$H(2) = \frac{2}{\log_2(e)} \int_0^{1/2} \left\{ \int_0^{z_1+1/2} 2\, dz_2 + \int_{z_1+1/2}^{1} \sqrt{2}\, dz_2 \right\} dz_1 =$$

$$\int_0^{1/2} \left[(2z_1 + 1) + \sqrt{2}\left(\frac{1}{2} - z_1\right) \right] dz_1 = 1.285$$

Working along the same lines, we obtain:

$$H_{\text{ann}}(2) = \ln\left\{ 2 \int_0^{1/2} \left\{ \int_0^{z_1+1/2} 4\, dz_2 + \int_{z_1+1/2}^{1} 2^{\sqrt{2}}\, dz_2 \right\} dz_1 \right\} =$$

$$\ln\left\{ 2 \int_0^{1/2} \left[(4z_1 + 2) + 2^{\sqrt{2}}\left(\frac{1}{2} - z_1\right) \right] dz_1 \right\} = 1.299$$

Finally:

$$G(2) = \ln \sup_{Z_2}(N(Z_2)) = \ln(4) = 1.386$$

Thus, we then confirm 6.1: $H(2) < H_{\text{ann}}(2) < G(2)$

□

7 Bounds on the Rate of Convergence

The entropy and annealed VC-entropy can be used to establish distribution-dependent rates of convergence of $R_{\text{emp}}(\alpha_n)$ and $R(\alpha_n)$. Usually $F(z)$ is unknown; therefore, we are more interested in establishing distribution-independent rates of convergence using the growth function.

7.1 VC-dimension

Theorem about the structure of the growth function (Vapnik):

Any growth function either satisfies

$$G(n) = n \ln 2 \quad \text{if } n \le h$$

or is bounded by:

$$G(n) \le \ln\left(\sum_{i=0}^{h} \binom{n}{i}\right) \quad \text{if } n > h. \tag{7.1}$$

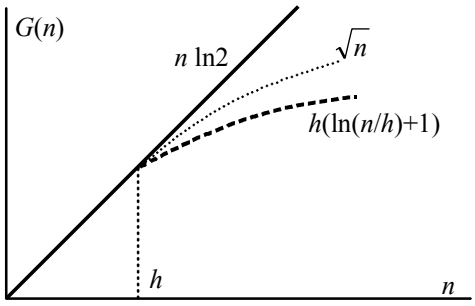


Figure 7.1

The structure of the growth function is shown in Figure 7.1. Note that for $n > h$:

$$G(n) \le \ln\left(\sum_{i=0}^{h} \binom{n}{i}\right) \le \ln\left(\frac{en}{h}\right)^{h} = h\left(1 + \ln\frac{n}{h}\right). \tag{7.2}$$

Thus, for $n > h$, $G(n)$ is bounded by a logarithmic function with coefficient h. It cannot be, for example, $G(n) = \sqrt{n}$.
The quantity h, separating the two different behaviors of the growth function, is called Vapnik-Chervonenkis dimension, and denoted d_{VC}:

$$h = d_{VC}, \text{ integer such that } \begin{cases} G(h) = h \ln 2 \\ G(h+1) \ne (h+1) \ln 2 \end{cases}$$

Alternative definition of the VC-dimension for a set of indicator functions:

The VC-dimension of a set $\{Q(z, \alpha), \ \alpha \in A\}$ is the maximum number h of vectors $z_1,\ldots,z_h$, which can be separated in all 2^h possible ways using functions of the set (*shattered* by the set).

Remark

Note that the VC-dimension is defined in terms of a family of loss functions $Q(z,\alpha)$. In the case of data classification, one has:

$$Q(z, \alpha) = L(\omega, \phi(x,\alpha)) = \begin{cases} 0 & \omega = \phi(x,\alpha) \\ 1 & \omega \neq \phi(x,\alpha) \end{cases}.$$

For two-class classification $\phi(x,\alpha)$ is an indicator function. Therefore:

$$Q(z,\alpha) = \begin{cases} \phi(x,\alpha) & \text{if} \quad \omega = 0 \\ 1-\phi(x,\alpha) & \text{if} \quad \omega = 1 \end{cases}$$

In this case the VC-dimension of the loss function equals the VC-dimension of the set of approximating functions $\phi(x,\alpha)$. Thus, for data classification, it makes no difference to talk about one or the other.

Example 7.1

Let $Z = \{z_1, z_2, \ldots\}$ be an arbitrary set, e.g., $Z = \{a, b, c, d, e\}$ (the elements could be any points on a d-dimensional space). Let S represent the set of subsets of Z, which have at most h elements. For the example of Z, we have for $h = 3$: $S = \{\varnothing, \{a\}, \{b\}, \ldots, \{a, b\}, \ldots, \{a, b, c\}, \ldots\}$. Finally, assume we had a family of functions defined on S, $Q(z, A)$, $A \in S$, such that:

$$Q(z, A) = \begin{cases} 1 & z \in A \\ 0 & z \in Z - A \end{cases}$$

Then, $\max N(z_1,\ldots,z_n) = 2^h$ if $n \leq h$. For instance, for the previous example of Z and h, one can obtain any dichotomy of a subset with 1, 2 or 3 elements.

On the other hand, $\max N(z_1,\ldots,z_n) = \sum_{i=0}^{h} \binom{n}{i}$ if $n > h$. for the previous example of Z and h, one can only obtain the dichotomies that correspond to subsets with 1, 2 or 3 elements, which correspond to the combinations in the formula.
Thus, formula 7.1 is a tight bound.

□

Example 7.2

Consider a set of linear indicator functions in d-dimensional space:

$$Q(z,\alpha) = \theta\left\{\sum_{i=1}^{d} \alpha_i z_i + \alpha_0\right\}$$

The VC-dimension is equal to the number of parameters: $d_{VC} = d + 1$ (we skip the proof).

□

Example 7.3

Consider a set of points in $d = 2$ dichotomized by a straight line. According to the result in Example 7.2, $d_{VC} = 3$.

Now consider the following Theorem (Cover,1965):

The number of linearly separable dichotomies (i.e. by a linear discriminant) of n points regularly distributed [7] in $\Re^d$, is:

$$D(n,d) = \begin{cases} 2\sum_{i=0}^{d}\binom{n-1}{i} & , \quad n > d+1; \\ 2^n & , \quad n \le d+1. \end{cases}$$

According to this Theorem, 3 points regularly distributed in $\Re^2$ can be shattered (see Figure 7.2).

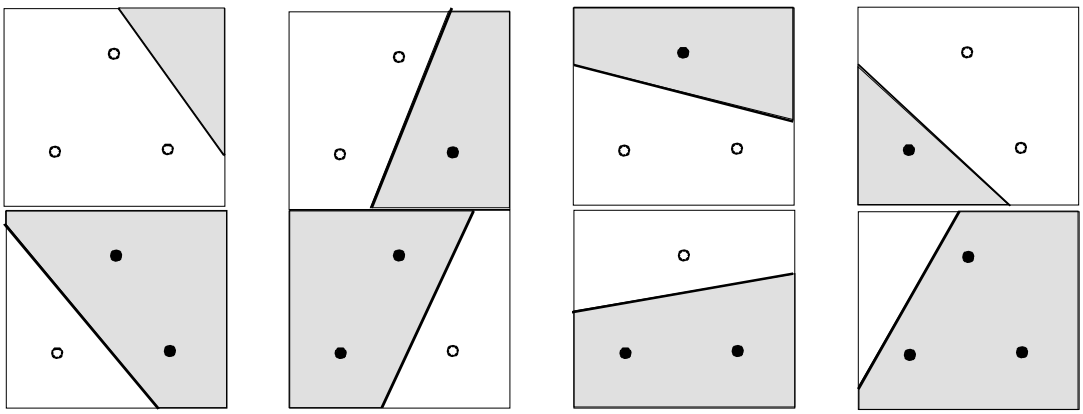


Figure 7.2

Figure 7.3 shows the upper bounds for the growth function when $d_{VC} = 3$. These upper bounds are independent of the family of classifying function used. They depend only on the particular value of d_{VC}. Figure 7.3 also shows the ln$D(n,2)$ curve, which reflects the type of evolution expected for the growth function in the conditions of the example. (Note that the definition of the VC-dimension does not require the set of points being regularly distributed.)

□

[7] A set of n points is regularly distributed in $\Re^d$ if no d+1 points lie on a linear variety of $\Re^d$.

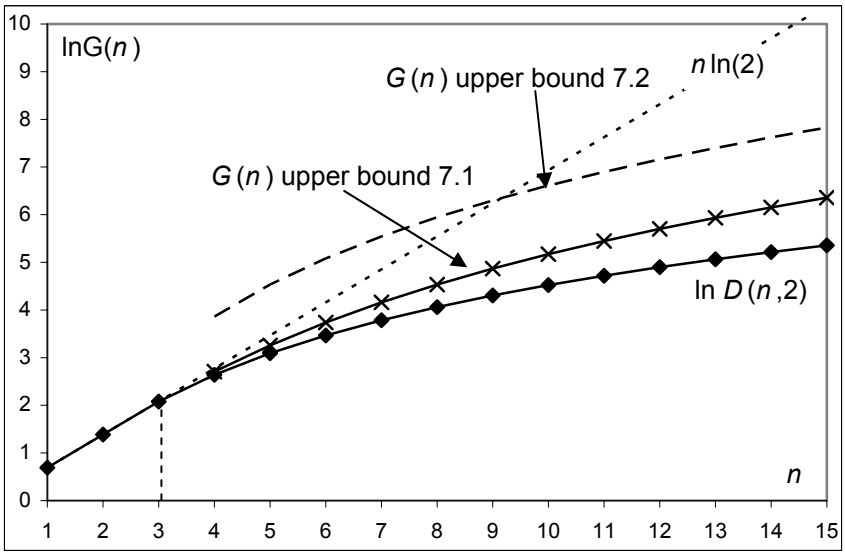


Figure 7.3

7.2 Bounds on the VC-Dimension for Neural Networks

We consider Multi Layer Perceptrons (MLPs) with (see Figure 7.4):

- Two layers
- A hidden layer with m neurons
- One output
- Neuronal activation function: step (threshold) function.

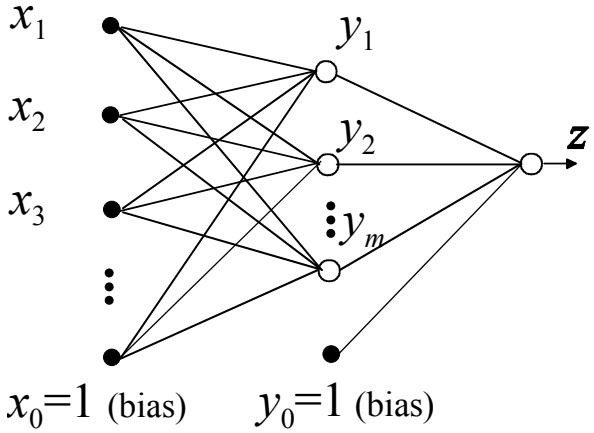


Figure 7.4

Model complexity:

Number of neurons (processing units): $u = m + 1$

Number of weights (model parameters): $w = (d+1)m + m + 1$

Model representation capability:

Each neuron of the first layer implements a linear discriminant, dividing the space into half-spaces:

$$y_j = \theta(\mathbf{w}'\mathbf{x} + w_0)$$

The output layer can be thought of as implementing logical combinations of the half-spaces.

Example 7.4

Figure 7.5a illustrates the classic XOR example with the linear discriminants obtained with an MLP2:2:1 (2 inputs; 2 hidden neurons; one output; see Figure 7.5b), and using θ as a threshold function in [-1, 1]. The classification table is shown below,. Notice that the two discriminants can originate 4 linearly separable regions.

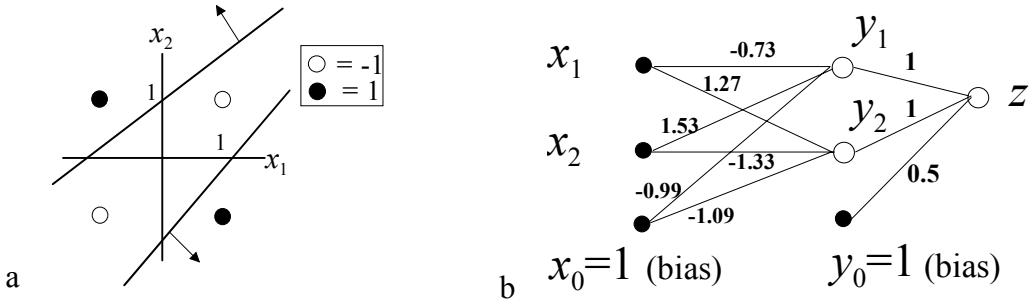


Figure 7.5

x_1	x_2	y_1	y_2	z = y_1 OR y_2
1	1	-1	-1	-1
1	-1	-1	1	1
-1	1	1	-1	1
-1	-1	-1	-1	-1

□

Theorem (Mirchandani and Cao, 1989):

The maximum number of regions linearly separable in $\Re^d$, by a MLP (satisfying the mentioned conditions) with m hidden neurons, is:

$$R(m,d) = \sum_{i=0}^{\min(m,d)} \binom{m}{i}. \qquad 7.3$$

Note that: $R(m, d) = 2^m$ for $m \le d$.
Number of linearly separable regions for $d = 2$ inputs:

m	1	2	3	4	5	6	7	8
$R(m, 2)$	2	4	7	11	16	22	29	37

Corolary 1:

Lower bound for the number of training set objects: $n \geq R(m, d)$. Since we need at least one example to learn one of the possible regions.

Corolary 2:

The lower bound on the VC-dimension for the MLP is: $d_{VC}(\text{MLP}) \geq R(m, d)$ (see Figure 7.6). Since, according to the definition, in order to find a lower bound for d_{VC} it is sufficient to find a configuration of points that can be shattered by the MLP. By placing each of the $R(m, d)$ points in a distinct linearly separable region we can achieve every possible dichotomy (provided the final layer implements the correct Boolean combination of regions).

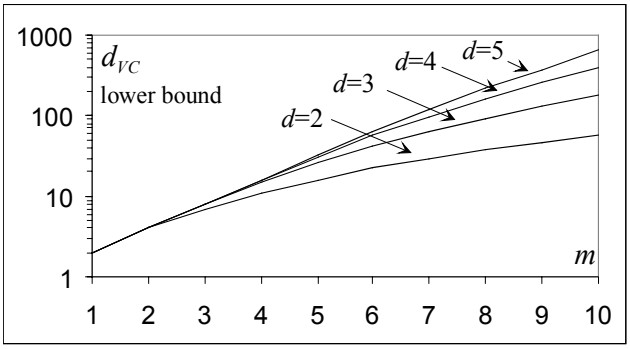


Figure 7.6

An upper bound for d_{VC} is difficult to find:

d_{VC} (MLP) $\leq k$: Prove that no set of $k + 1$ points can be shattered by the MLP.

The following upper bound for an MLP with u neurons and w weights is due to (Baum and Haussler, 1989):

$$d_{VC} \leq 2w \log_2(eu). \qquad 7.4$$

For d=2 inputs we have:

m	1	2	…	10
lower bound	2	4	…	56
upper bound	6*	54	…	402

* Using only one unit (neuron)

Notice the wide range between the two bounds.
In some simple cases, it may be possible by enumeration to derive the d_{VC} value, as in the following:

Example 7.5

Assume that we want to derive the d_{VC} value for an MLP2:2:1 as in Example 7.4, constrained to the sample points being regularly distributed. Notice that if all the n points are the vertices of a convex hull we need trunc(n/2)-1 lines to shatter the set (since in the most unfavorable case we have alternating 0, 1 class-label sequences,

with one possible repeating label; see also Huang S-C and Huang Y-F, 1991). Thus with 2 lines we shatter a pentagon (5-gon) as shown in Figure 7.7a. On the other hand, it is easy to find a 6-point configuration which can also be shattered: the pentagon with one point inside as shown in Figure 7.7b. As a matter of fact, whatever the class label of the inside point is, one can always move the lines discriminating the vertices in order to include it appropriately.

Figure 7.7

Thus, we know that the d_{VC} for a MLP2:2:1 is at least 6. Now, to confirm that $d_{VC} = 6$, we need to show that no regularly distributed configuration of 7 points can be shattered. For this purpose, we enumerate these configurations of 7 points by enclosures of n-gons as shown in Figure 7.8, and try to find at least one class labeling that cannot be achieved by two lines. Figure 7.8 also shows examples of such configurations. Thus, indeed $d_{VC} = 6$.

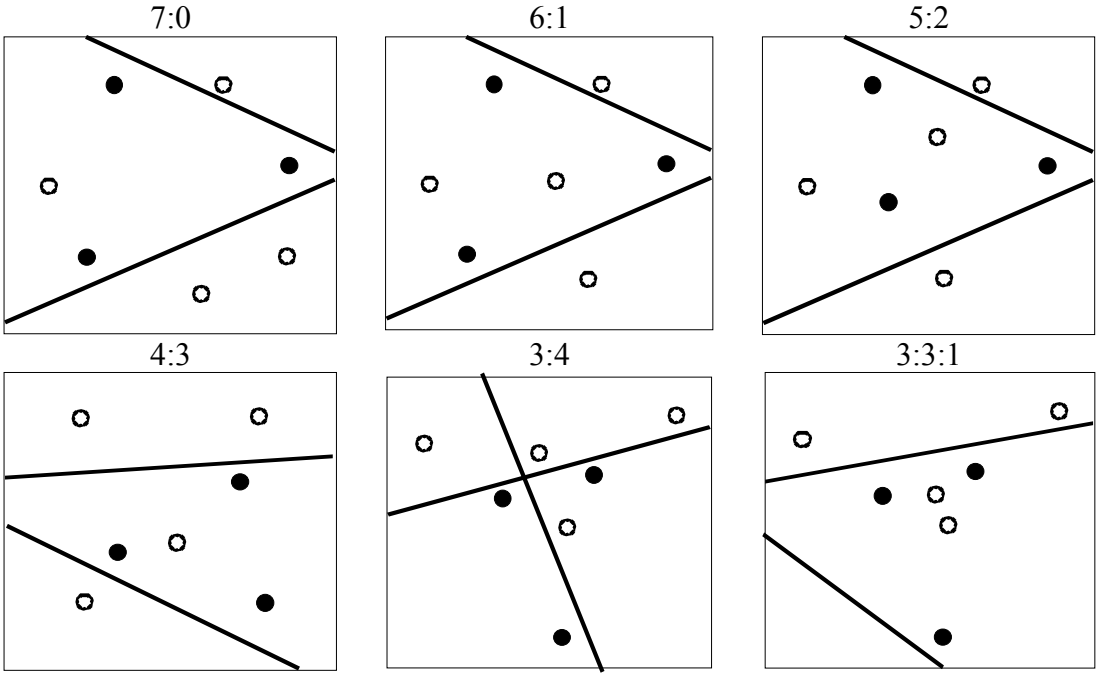


Figure 7.8

7.3 VC-Dimension for a Δ-Margin Separating Hyperplane

Consider a hyperplane:

$$(\mathbf{w}'\mathbf{x}) - b = 0, \quad |\mathbf{w}| = 1$$

The Δ-margin separating hyperplane classifies **x** vectors as follows:

$$y = \begin{cases} 1 & \mathbf{w}'\mathbf{x} - b \geq \Delta \\ -1 & \mathbf{w}'\mathbf{x} - b \leq -\Delta \end{cases}$$

(Classifications of vectors that fall into the (-Δ, Δ)-margin are undefined.)

Theorem (Vapnik):

Let the vectors **x** belong to a sphere of radius R (see Figure 7.7). Then, the set of Δ-margin separating hyperplanes has the VC-dimension bounded by:

$$h \leq \min\left(\frac{R^2}{\Delta^2}, d\right) + 1. \qquad 7.5$$

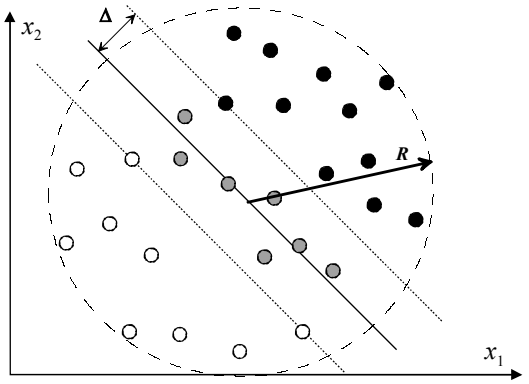


Figure 7.9

The table below compares the d_{VC} lower bound for a "normal" d-dimensional hyperplane (Example 7.2), with the lower bound for a Δ-margin separating hyperplane (formula 7.5) for two values of R/Δ. Notice how increasing Δ lowers d_{VC} (relative to d + 1).

d	1	2	3	4	5	…	10	…	30
for hyperplane	2	3	4	5	6	…	11	…	31
for Δ-margin hyperplane with$(R/\Delta) = \sqrt{3}$	2	3	4	4	4	…	4	…	4
for Δ-margin hyperplane with$(R/\Delta) = 5$	2	3	4	5	6	…	11	…	26

7.4 Distribution Independent Bounds for Convergence Rates

Theorem (Vapnik)
For a set of indicator functions with finite VC-dimension h, the following inequality holds true:

$$P\left\{\sup_{\alpha\in\Lambda}\frac{R(\alpha)-R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}}>\varepsilon\right\}<4\exp\left\{\left(\frac{h(1+\ln(2n/h)}{n}-\frac{\varepsilon^2}{4}\right)n\right\}$$

This is a *relative uniform convergence* of $R_{\text{emp}}(\alpha)$ to $R(\alpha)$ [8].

Let:

$$\delta=4\exp\left\{\left(\frac{h(1+\ln(2n/h)}{n}-\frac{\varepsilon^2}{4}\right)n\right\}$$

Then:

$$\varepsilon^2=4\frac{h(1+\ln(2n/h))-\ln(\delta/4)}{n}$$

Then, as a consequence of the Theorem, with probability at least 1-δ the following inequality holds true simultaneously for all indicator functions:

$$\frac{R(\alpha)-R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}}\le\varepsilon$$

Therefore (denoting $R(\alpha)$, $R_{\text{emp}}(\alpha)$ by R, R_{e}, resp., for simplification):

$$R-R_{\text{e}}\le\varepsilon\sqrt{R_{\text{e}}}\Rightarrow\quad R^2-(2R_{\text{e}}+\varepsilon^2)R+R_{\text{e}}^2\le 0\Rightarrow\quad R\le R_{\text{e}}+\frac{\varepsilon^2}{2}\left(1+\sqrt{1+\frac{4R_{\text{e}}}{\varepsilon^2}}\right)$$

Let $Q(z, \alpha_n)$ be the function which minimizes the empirical risk. For *this* function the following bound holds true with probability1-δ:

$$R(\alpha_n)\le R_{\text{emp}}(\alpha_n)+\frac{\varepsilon(n)}{2}\left(1+\sqrt{1+\frac{4R_{\text{emp}}(\alpha_n)}{\varepsilon(n)}}\right)$$

$$\text{with}\qquad \varepsilon(n)\equiv\varepsilon^2=4\frac{h\left(\ln\frac{2n}{h}+1\right)-\ln(\delta/4)}{n}. \qquad 7.6$$

[8] The proof of this Theorem is based on the inequality $H_{ann}(n)\le G(n)<h\left(1+\ln\frac{n}{h}\right)$. See formulas 6.1 and 7.2.

Now, consider that α_0 is the optimum parameter set; therefore, $R(\alpha_0)$ denotes the minimum expected risk (optimal risk).

Then, the following can be proved using the well-known result of the additive Chernoff bound:

With probability at least 1-δ the following inequality holds true:

$$R(\alpha_0) > R_{\text{emp}}(\alpha_0) - \sqrt{\frac{-\ln \delta}{2n}} \qquad (R_{\text{emp}}(\alpha_0) \geq R_{\text{emp}}(\alpha_n))$$

Using this result together with formula 7.6, one can conclude that: with probability at least 1-2δ the following inequality holds true:

$$\Delta(\alpha_n) = R(\alpha_n) - R(\alpha_0) \leq \sqrt{\frac{-\ln \delta}{2n}} + \frac{\varepsilon(n)}{2}\left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_n)}{\varepsilon(n)}}\right). \qquad 7.7$$

Remarks

- n/h large: $R(\alpha_n) \approx R_{\text{emp}}(\alpha_n) \approx R_{\text{emp}}(\alpha_0)$ (the network generalises)
- n/h small: a small $R_{\text{emp}}(\alpha_n)$ does not guarantee a small $R(\alpha_n)$

Example 7.6

$Q(z, \alpha)$ corresponds to a MLP2:2:1 (Example 7.5) with $h = 6$. Which n will guarantee an $R(\alpha_n)$–$R_{\text{emp}}(\alpha_n)$ deviation below 0.1, with 95% confidence ($\delta = 0.05$), for an $R_{\text{emp}}(\alpha_n) = 0.05$?

Figure 7.10 dotted line shows the evolution of $R(\alpha_n)$–$R_{\text{emp}}(\alpha_n)$ with n (formula 7.6), for the given specifications. It turns out that only for $n \geq 3125$ we obtain a deviation below 0.1.

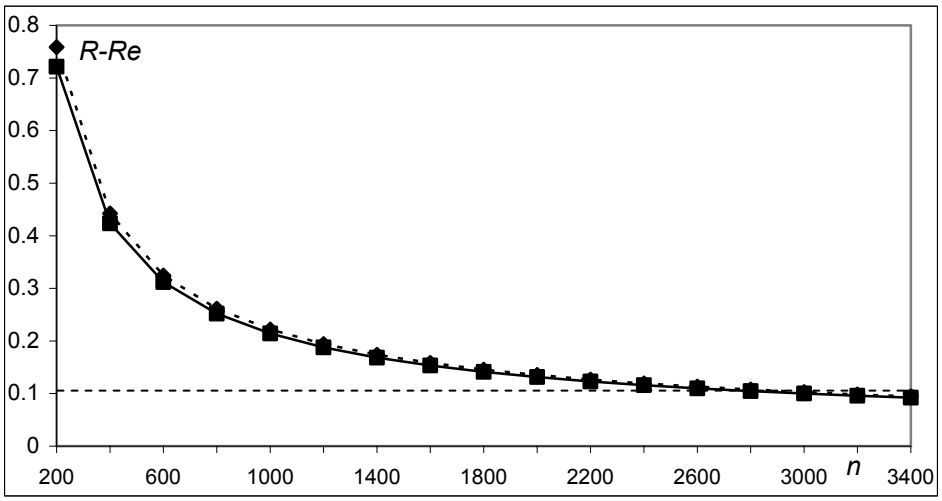


Figure 7.10

Note, however, that formula 7.6 is based on the growth function upper bound given by formula 7.2. A tighter upper bound is given by formula 7.1. Using this tighter upper bound the value of $\varepsilon(n)$ of formula 7.6 becomes:

$$\varepsilon(n) \equiv \varepsilon^2 = 4\frac{\ln\left(\sum_{i=0}^{h}\binom{2n}{i}\right) - \ln(\delta/4)}{n}. \qquad 7.8$$

One then obtains the solid curve of Figure 7.10 and an equivalent $n \geq 3000$. We see that the difference between the two curves is small.

□

Figure 7.11 shows the evolution of formula 7.6 with d_{VC} using the lower bounds given by formula 7.3, the other conditions of Example 7.6 remaining true. We see that for more complex classifying functions (higher h) we need increasing training set sizes in order to guarantee a given deviation $R(\alpha_n)-R_{\text{emp}}(\alpha_n)$ bound with the same confidence.

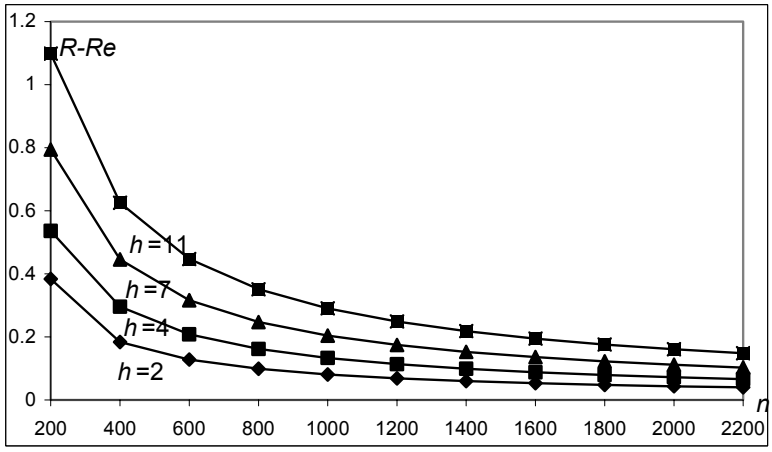


Figure 7.11

Figure 7.12 shows the training set size n that will guarantee with 95% confidence the upper bound of $R(\alpha_n)-R_{\text{emp}}(\alpha_n)$ shown at the right (from 0.3 through 0.1), for various values of d_{VC} and with $R_{\text{emp}}(\alpha_n)=0.05$. We observe that, the value of n increases dramatically for small deviations of $R(\alpha_n)-R_{\text{emp}}(\alpha_n)$ and large d_{VC}.

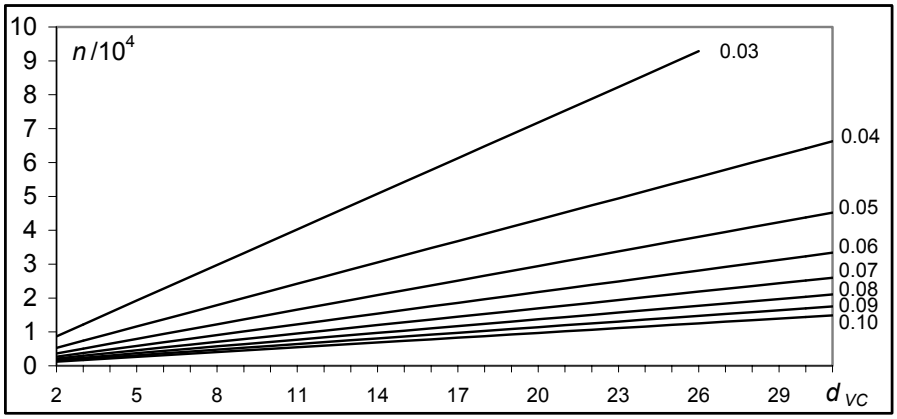


Figure 7.12

Example 7.7

Compute for the previous Example 7.6 the upper bound deviation of the empirical classifier risk from the optimal risk with a probability at least a 90%:

$$\Delta(\alpha_{1900}) \le \sqrt{\frac{-\ln \delta}{2n}} + 0.1 = 0.122$$

□

Final Note

The previous formulas 7.6 and 7.7 look unrealistically pessimistic. For instance, reported experiments with MLP2:2:1 do not indicate the need of such high values of n in order to obtain a generalization. However, one must take into account that these formulas are distribution-free bounds, based on general statistical laws, such as the Chernoff bound. Something similar occurs when using the well-known Chebyshev inequality.

It might be instructive to recall this issue.

Example 7.8

What is the probability that a r.v. deviates from the mean less than 1.5σ?

Distribution-free (Chebyshev):

$$P(|x - \mu| > k\sigma) \le \frac{1}{k^2} \qquad \Rightarrow \qquad P(|x - \mu| \le k\sigma) \le 1 - \frac{1}{k^2} = 0.5556$$

Uniform: $P = 0.8660$
Normal: $P = 0.8664$
t_3-Student: $P = 0.9195$

□

Example 7.9

How many observations of a r.v. x must one have in order to obtain a sample mean estimate that deviates less than 0.2σ from the true mean with 95% confidence?

Distribution-free (Chebyshev): $P(|x - \mu| > k\sigma) \le \frac{1}{k^2}$.

Hence: $P(|\bar{x} - \mu| > k\sigma_{\bar{x}}) \le \frac{1}{k^2} \qquad \Rightarrow \qquad P(|\bar{x} - \mu| \le k\frac{\sigma}{\sqrt{n}}) \le 1 - \frac{1}{k^2}$

Therefore: $1/k^2 = 0.05 \quad \Rightarrow \quad k^2 = 20$

$$k\frac{\sigma}{\sqrt{n}} = 0.2\sigma \quad \Rightarrow \quad \sqrt{n} = \frac{k}{0.2} \quad \Rightarrow \quad n = k^2/0.04 = 500$$

Normal distribution: 95% confidence $\Rightarrow \quad 1.96\,\sigma_{\bar{x}} = 0.2\sigma \Rightarrow \quad n = 96$

Observe the disparity between the two values of n.

□

References

Anthony M, Bartlett P (1999) Neural Network Learning: Theoretical Foundations.Cambridge University Press.

Brady MJ (1990) Guaranteed Learning Algorithm for Network with Units Having Periodic Threshold Output Function. Neural Computation, 2:405-408.

Carter MA, Oxley ME (1999) Evaluating the Vapnik-Chervonenkis Dimension of Artificial Neural Networks Using the Poincaré Polynomial. Neural Networks, 12: 403-408.

Carter MA, Oxley ME (2001) Letter to the Editor. Neural Networks, 14: 1467-1470.

Cherkassky V, Mulier F (1998) Learning From Data. John Wiley & Sons, Inc.

Cover TM (1965) Geometrical and Statistical Properties of Systems with Linear Inequalities with Applications in Pattern Recognition. IEEE Tr. Electronic Computers, 14:326-334.

Gaynier RJ, Downs T (1995) Sinusoidal and Monotonic Transfer Functions: Implications for VC Dimension. Neural Networks, 8:901-904.

Huang S-C and Huang Y-F (1991) Bounds on the Number of Hidden Neurons in Multilayer Perceptrons. IEEE Tr. NN 2:47-55.

Vapnik VN (1998) Statistical Learning Theory. John Wiley & Sons, Inc.

Vapnik VN (1999) An Overview of Statistical Learning Theory. IEEE Tr. Neural Networks, 10:988-999.

Appendix - Stochastic Convergence

Definitions

1. Limit of a sequence of numbers x_n:

$$x_n \underset{n\to\infty}{\to} x \text{ sse } \quad \forall \varepsilon > 0, \quad \exists\, n_0, \quad \forall n > n_0, \quad |x_n - x| < \varepsilon$$

2. Stochastic process = sequence of random variables:

$$x(n) \equiv \{x(1), x(2), \ldots, x(n)\}$$

The *ensemble* of $x(n)$ is the collection of the *sample paths* $\{x(1), x(2), \ldots, x(n)\}$.

Example A.1
Bernoulli process of coin tossing (see Figure A.1)

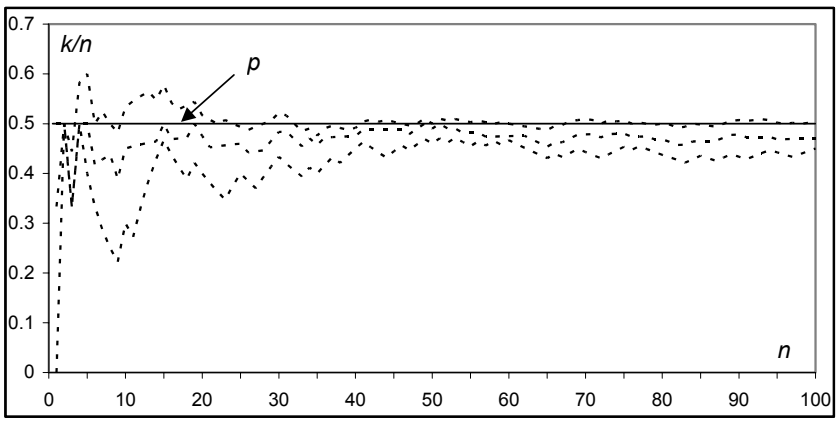


Figure A.1

Convergence Types

1. Convergence in distribution

A stochastic process $x(n)$, $n = 1,2, \ldots$ with distribution functions

$$F_{x(n)}(x) \equiv F_n(x) = P(x(n) \le x)$$

is said to converge in distribution if there is a distribution function $F(x)$ such that $F_n(x) \underset{n\to\infty}{\to} F(x)$ for all x where $F(x)$ is continuous.

Example A.2
The following stochastic process represented by a family of distribution functions, converges in distribution.

$$F_{x(n)}(x) = \begin{cases} 1 - \dfrac{1}{n+1} e^{-n(x-1)} & x \ge 1 \\ \dfrac{n}{n+1} x & 0 \le x < 1 \\ 0 & x < 0 \end{cases} \quad \to \quad \text{uniform distribution } F_x(x) = \begin{cases} 1 & x \ge 1 \\ x & 0 \le x < 1 \\ 0 & x < 0 \end{cases}$$

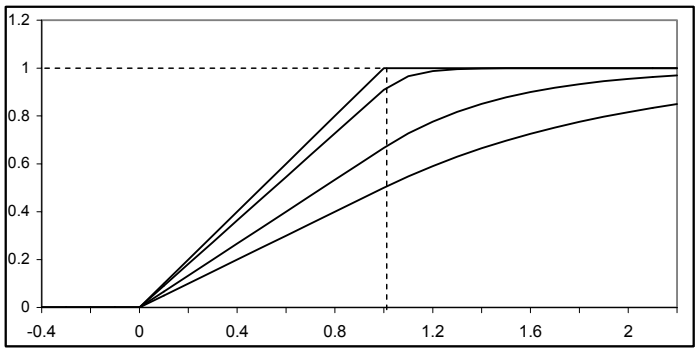


Figure A.2

This is a *weak* type of convergence, that has very little to do with convergence of the sample paths of the process.

Example A.3
Bernoulli process with $P(x(n) = 0) = P(x(n) = 1) = ½$.
Convergence in distribution is obviously verified:

$$F_{x(n)}(x) = F(x) = \begin{cases} 1 & x \ge 1 \\ ½ & 0 \le x < 1 \\ 0 & x < 0 \end{cases}$$

However, the probability that a sample path converges is zero.

2. Convergence in probability

A stochastic process $x(n)$, $n = 1,2, \ldots$ is said to *converge in probability* if there exists a r.v. x (possibly a constant) such that:

$$\forall \varepsilon > 0, \quad P\{|x(n) - x| \geq \varepsilon\} \underset{n\to\infty}{\rightarrow} 0$$

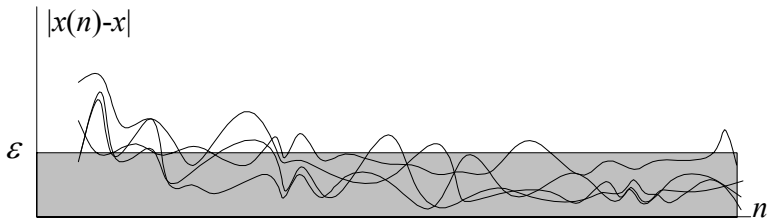


Figure A.3

Given a certain ε, it is possible to find n_0 such that for $n > n_0$ the probability of finding excursions above ε is zero. It may happen that *all* sample paths go on having excursions above ε, but they become rarer and rarer.

Example A.4
Weak Law of Large Numbers (Bernoulli)

$$\forall \varepsilon > 0, \quad P\{|k / n - p| \leq \varepsilon\} \underset{n\to\infty}{\rightarrow} 1$$

k/n tends to p in probability.

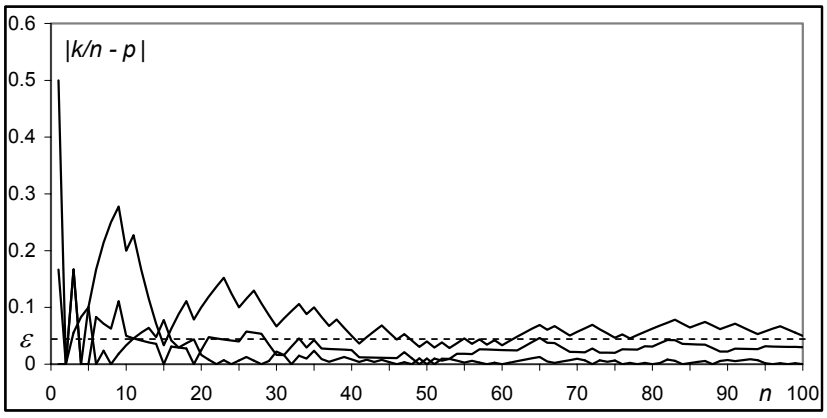


Figure A.4

Assume, for Example A.1, we want an estimate of p with $\varepsilon = 0.1$, for $n \geq 1000$. Applying Chebyshev Inequality:

$$P\{|k / n - p| < 0.1\} \geq \frac{39}{40}$$

Thus in 39 out of 40 of the trials we get $\varepsilon < 0.1$.
Imagine we repeat the experiment 40 times with $n > 1000$ (say 2000). Then, for a specific n we expect that at the most only one bad run will exceed $\varepsilon = 0.1$. However,

we cannot conclude that all good runs will have $\varepsilon < 0.1$ for all $n > 1000$ (say between 1000 and 2000).

Consistency of the learning process:

$$R(\alpha_n) \underset{n\to\infty}{\overset{P}{\to}} \inf_{\alpha\in A} R(\alpha)$$

means:

$$\forall \varepsilon > 0, \quad \forall \delta > 0, \quad \exists\, n_0 = n_0(\varepsilon,\delta), \quad \forall n > n_0 \quad P\{R(a_n) - R(\alpha_0) < \varepsilon\} \geq 1-\delta$$

Uniform Convergence (see Vapnik, 1998)

$$\forall \varepsilon > 0 \quad P\left\{\sup_{\alpha\in A}\left|R(\alpha) - R_{\text{emp}}(\alpha)\right| \geq \varepsilon\right\} \underset{n\to\infty}{\to} 0$$

Glivenko-Cantelli Theorem
The convergence

$$\forall \varepsilon > 0 \quad P\left\{\sup_{x}\left|F(x) - F_n(x)\right| \geq \varepsilon\right\} \underset{n\to\infty}{\to} 0$$

takes place.

3. Convergence in mean square

A stochastic process $x(n)$, $n = 1,2, \ldots$ is said to *converge in mean square* if there exists a r.v. χ (possibly a constant) such that:

$$\mathrm{E}[(x(n) - \chi)^2] \underset{n\to\infty}{\to} 0$$

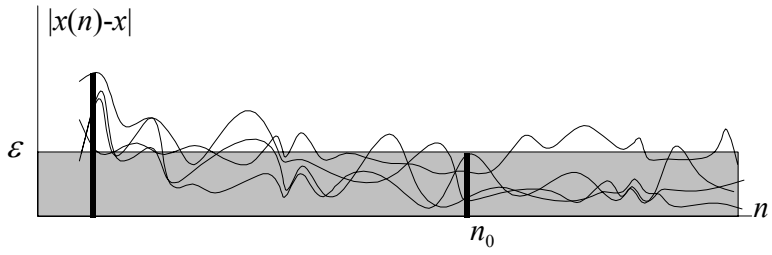


Figure A.5

The expected value of the squared deviations tends to zero with increasing n. Above n_0 one may find with constant probability excursions above ε, but still keeping the probability associated to large deviations sufficiently low in such a way that the expected value of the squared deviations is below ε.
If $x(n)$ converges in probability it will also converge in mean square if:

$$\mathrm{E}[|x(n)|^2] \underset{n\to\infty}{\to} \mathrm{E}[|\chi|^2]$$

4. Convergence almost surely

This is a strong type of convergence defined as:

A stochastic process $x(n)$, $n = 1,2, \ldots$ is said to *converge almost surely* if there exists a r.v. χ (possibly a constant) such that:

$$P(x(n) \underset{n\to\infty}{\to} \chi) = 1$$

Equivalently (Vapnik):

$$\forall \varepsilon > 0 \quad P\left\{ \sup_{n>n_0} |x(n) - x| > \varepsilon \right\} \underset{n_0\to\infty}{=} 0$$

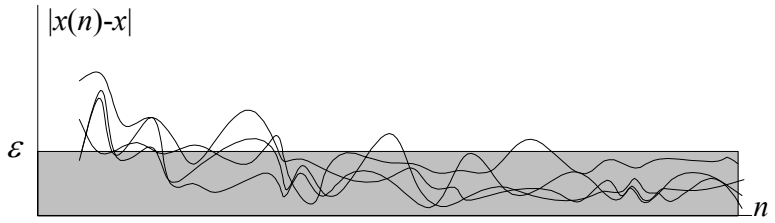


Figure A.6

Given a certain ε, it is possible to find n_0 such that for $n > n_0$ most of the sample paths are below ε. In the limit there are infinitely many paths all below e.

Example A.5
Let $y(n)$ be an arbitrary stochastic sequence such that $y(n)$ takes on only values 0 or 1, and let x_0 be a r.v. Consider a stochastic sequence defined by:

$$x(n) = x_0\left(1 - \frac{1}{n}\right)^{y(n)}$$

For each value $x_0 = x_0$ all the sample paths:

$$x(n) = x_0\left(1 - \frac{1}{n}\right)^{y(n)},$$

converge to x_0 regardless of the sequence $y(n)$.

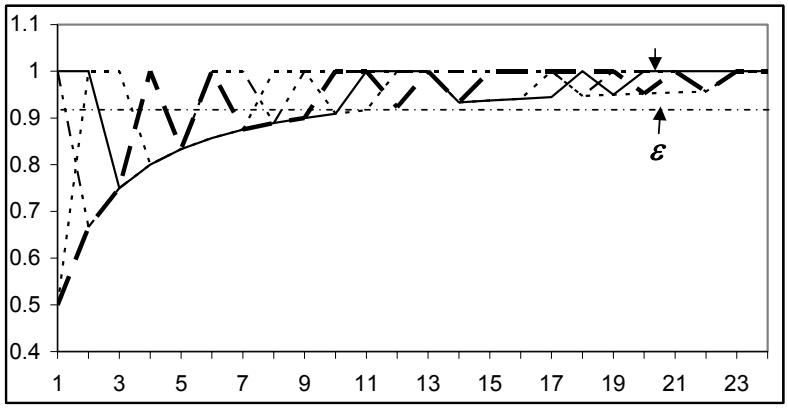


Figure A.7

Example A.6
Strong Law of Large Numbers (Borel)

$$P(k / n \underset{n\to\infty}{\to} p) = 1$$

k/n tends almost surely to p.

In the previous Bernoulli-process example the Strong Law of Large Numbers says that all good runs will have excursions below 0.1, for n above 1000.

Result due to a Borel-Cantelli Lemma:
In order for $x(n)$ to converge almost surely to x, it is sufficient (and necessary if the $x(n)$ r.v are independent) that for any $\varepsilon > 0$ the following holds:

$$\sum_{n=1}^{\infty} P\left\{\left|x(n) - x\right| > \varepsilon\right\} < \infty$$

Relation among the Convergence Types

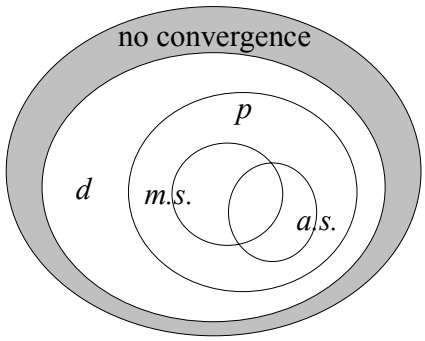


Figure A.8

References

HJ Larson, BO Shubert (1979) Random Variables and Stochastic Processes (vol I). John Wiley & Sons.
A Papoulis (1965) Probability, Random Variables and Stochastic Processes. McGraw Hill Pub. Co.
VN Vapnik (1998) Statistical Learning Theory. John Wiley & Sons, Inc.