Neural Network Interest Group

Título/*Title*:
Learning Bounds

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

Relatório Técnico/*Technical Report* No. 1 /2004

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

Título/*Title*:
Learning Bounds
Autor(es)/*Author(s)*:
J.P. Marques de Sá
Relatório Técnico/*Technical Report* No. 1 /2004
Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/
Abril 2004

Learning Bounds

J.P. Marques de Sá, FEUP/DEEC, 2004

Table of Contents

1 General Learning Model ... 3
1.1 Basic Definitions ... 3
1.2 Bayes error ... 4
1.3 Learning Algorithm ... 4
2 Error Estimation ... 6
2.1 Deviations of Empirical Errors from True Errors ... 6
2.2 Deviations of ERM Errors from Optimal Errors ... 10
3 ERM Learning with Finite Classes ... 13
4 Vapnik-Chervonenkis Theory ... 15
4.1 Growth Function ... 17
4.1.1 Definitions ... 17
4.1.2 Growth Function Properties ... 19
4.1.3 VC-Dimension of Some Classes ... 22
4.1.4 Growth Function of Perceptrons with Linear Thresholds ... 23
4.1.5 Growth Function of Perceptrons with Sigmoids ... 29
4.2 Learning Bounds for Infinite Classes of Classifiers ... 30
4.2.1 Upper Bounds ... 30
4.2.2 Lower Bounds ... 33
5 Restricted Learning Model ... 34
5.1 Basic Definitions ... 34
5.2 Consistent Learning ... 35
5.3 Learning Bounds ... 36
6 Appendix ... 37
6.1 The Glivenko-Cantelli Theorem ... 37
6.2 Useful Formulas ... 37
6.2.1 Markov's inequality ... 37
6.2.2 Logarithms ... 38
6.2.3 Binomial Formulas ... 38
6.2.4 Exponentials ... 39
6.2.5 Stirling Formula ... 39
References ... 39

1 General Learning Model

1.1 Basic Definitions

Object (instance or input) set: $X \subseteq \Re^d$ $X = \{\mathbf{x}\}$; input (feature) vector $\mathbf{x}$.

Target set: $T = \{0,1\}$ (or $\{-1,1\}$; a two-class problem)

Problem space: $Z = X \times T$. There is a fixed but unknown probability measure P defined on Z, for the r.v. pair $(\mathbf{x},t)$. Note that we consider that for any given $\mathbf{x} \in X$, both $(\mathbf{x},0)$ and $(\mathbf{x},1)$ may have a non-null probability (so, neither 0 nor 1 is the "correct" classification).

Training (design) sample: $D_n = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\} \in Z^n$ randomly drawn, with each $z_i = (\mathbf{x}_i,t_i)$ pair – *labelled example* - i.i.d. ($D_n \sim P^n$).

Decision function (or classifier): $\phi : X \to T$

Class of classifiers (or machine):

$C = \{\phi : X \to T\}$ (NN: $C_W = \{\phi_w : X \to T; w \in W\}$; W is the weight space)

Note that even if there is a correct "classification" function, f, i.e., with $P(\{(\mathbf{x}, f(\mathbf{x})); \mathbf{x} \in X\}) = 1$ (thus with zero probability of error), it may happen that $f \notin C$.

Classifier designed on D_n: ϕ_n

Risk (or error) of a classifier, ϕ:

$$R(\phi) = P(\phi(\mathbf{x}) \neq t) \equiv P((\mathbf{x},t) \in Z; \phi(\mathbf{x}) \neq t)$$

Error of ϕ_n: $R(\phi_n) = P((\mathbf{x},t) \in Z; \phi_n(\mathbf{x}) \neq t)$

Note that $R(\phi_n)$ is a [0, 1]-valued r.v., dependent on D_n.

Empirical error[1] of ϕ (in D_n):

$$\hat{R}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(\mathbf{x}_i)\neq t_i;(\mathbf{x}_i,t_i)\in D_n\}}(\mathbf{x}_i) \equiv \frac{1}{n}\left|\{i : \phi(\mathbf{x}_i) \neq t_i, (\mathbf{x}_i,t_i) \in D_n\}\right|$$

Note that $\hat{R}_n(\phi)$ is a [0, 1]-valued r.v., dependent on D_n. I_A is the indicator of set A.

Optimal error[2] of the class C: $R_{\text{opt}} = \inf_{\phi \in C} R(\phi)$ (this is a constant)

[1] Also called *sample error*, *observed error* or *error-count estimate*.

[2] Also called *approximation error*. We use *inf* and not *min* because $\{R(\phi)\}$ may be infinite.

We may also express the risk as an expectation of a *loss function*:

$$R(\phi) = \int_Z L(t, \phi(\mathbf{x}))dF(z),$$

where the loss function is:

$$L(t_i, \phi(\mathbf{x}_i)) = \begin{cases} 0 & t_i = \phi(\mathbf{x}_i) \\ 1 & t_i \neq \phi(\mathbf{x}_i) \end{cases},$$

Whenever the data distribution is continuous distribution we can also write:

$$R(\phi) = \sum_{i=0}^{1} P_i \int_X L(t_i, \phi(\mathbf{x})) f(\mathbf{x}, t_i) d\mathbf{x}$$

where $f(\mathbf{x},t_i)$ is the pdf for class t_i and P_i are the prior probabilities.

1.2 Bayes error

Let us assume that the probability measure P for the the r.v. pair (χ,t) taking value in Z corresponds to:

- The distribution of χ: $\mu(A) = P(\chi \in A;\ A \subseteq X^d)$;
- The "a posteriori" probability of class 1: $\eta(\mathbf{x}) = P(t = 1 \mid \chi = \mathbf{x}) = \mathrm{E}[t \mid \chi = \mathbf{x}]$

We then define the Bayes classifier:

$$\phi^*(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad \eta(\mathbf{x}) > 1/2 \\ 0 & \text{otherwise} \end{cases},$$

which can be proved to be optimal, i.e.:

$$R^* = R(\phi^*) = \mathrm{E}[I_{\{\eta(\mathbf{x}) \leq 1/2\}}(\mathbf{x})\eta(\mathbf{x}) + I_{\{\eta(\mathbf{x}) > 1/2\}}(\mathbf{x})(1 - \eta(\mathbf{x}))] \leq R(\phi) \text{ for any } \phi : X \to T$$

Hence: $\forall C, \quad R^* \leq \inf_{\phi \in C} R(\phi)$.

1.3 Learning Algorithm

Definition 1.1
Given $C = \{\phi : X \to Y\}$ a *learning algorithm L* for C is a function

$$L : \bigcup_{n=1}^{\infty} Z^n \to C$$

from the set of all training sets to C, such that given $\varepsilon, \delta \in]0, 1[$ ([3]), there is an integer $n_0(\varepsilon, \delta)$ – *sufficient sample size* - such that for $n \geq n_0(\varepsilon, \delta)$ and every training sample D_n (as above), then $\phi_n = L(D_n)$ satisfies

$$P^n\left(R(\phi_n) < \varepsilon + \inf_{\phi \in C} R(\phi) \right) \geq 1 - \delta$$

for any probability distribution P on Z (therefore, P^n on Z^n).
C is learnable if there is a learning algorithm for C.

■

[3] $1-\varepsilon$ and $1-\delta$ are known as *accuracy* and *confidence*, respectively. In practice $\varepsilon, \delta \in]0,0.5]$

Equivalent formulations:

1 $\exists\, n_0(\varepsilon, \delta)$ such that for $n \geq n_0(\varepsilon, \delta)$, $P^n\left(R(\phi_n) - \inf_{\phi \in C} R(\phi) \geq \varepsilon \right) < \delta$ (see Figure 1.1)

2 $\forall\ \varepsilon \in]0,1[,\ P^n\left(R(\phi_n) - \inf_{\phi \in C} R(\phi) \geq \varepsilon \right) \underset{n \to \infty}{\to} 0$ (convergence in probability)

3 $\exists \varepsilon_0(n, \delta)$ such that $\forall\, n, \delta, P, D_n$

$P^n\left(R(\phi_n) - \inf_{\phi \in C} R(\phi) < \varepsilon_0(n, \delta) \right) \geq 1 - \delta$ with $\varepsilon_0(n, \delta) \underset{n \to \infty}{\to} 0$

$\varepsilon_0(n, \delta)$ is the *estimation error bound*.

4 $\mathrm{E}\left[R(\phi_n) - \inf_{\phi \in C} R(\phi) \right] < \varepsilon\delta$ (or $\mathrm{E}[R(\phi_n)] < \inf_{\phi \in C} R(\phi) + \varepsilon\delta$)

with $\mathrm{E} \equiv \underset{Z^n}{\mathrm{E}}$.

As a matter of fact, by Markov's inequality,

$$\mathrm{E}\left[R(\phi_n) - \inf_{\phi \in C} R(\phi) \right] < \varepsilon\delta \quad \Rightarrow \quad P^n\left(R(\phi_n) - \inf_{\phi \in C} R(\phi) \geq \varepsilon \right) < \frac{\varepsilon\delta}{\varepsilon} = \delta$$

Conversely, assuming that

$$P^n\left(R(\phi_n) - \inf_{\phi \in C} R(\phi) \geq \alpha/2 \right) < \alpha/2$$

since $R(\phi_n) - \inf_{\phi \in C} R(\phi) \leq 1$, we have

$$\mathrm{E}\left[R(\phi_n) - \inf_{\phi \in C} R(\phi) \right] < \frac{\alpha}{2} P^n\left(R(\phi_n) - \inf_{\phi \in C} R(\phi) < \frac{\alpha}{2} \right) + P^n\left(R(\phi_n) - \inf_{\phi \in C} R(\phi) \geq \frac{\alpha}{2} \right) < \alpha \quad ([4])$$

Thus, in terms of convergence of the mean deviations we have a sample complexity $n'_0(\alpha) = n_0(\alpha/2, \alpha/2)$.

From now on we will simplify the notation using P instead of P^n.

Definition 1.2

Sample complexity of L: $n_L(\varepsilon, \delta) = \min_{\{n_0\}} n_0(\varepsilon, \delta)$

Estimation error of L: $\varepsilon_L(n, \delta) = \min_{\{n_0\}} \varepsilon_0(n, \delta)$

The sample complexity sets a lower bound on the sample size needed by L for learning C. ■

[4] Let $d_n = R(\phi_n) - \inf_{\phi \in C} R(\phi) \leq 1$. Then, note that $\mathrm{E}[d_n] = \int_0^{\alpha/2} d_n dP(d_n) + \int_{\alpha/2}^1 d_n dP(d_n) < \alpha/2 + \int_{\alpha/2}^1 dP(d_n)$

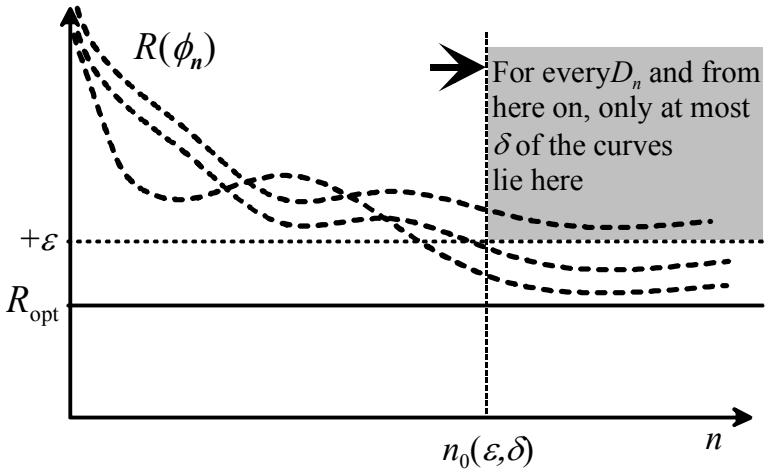

Figure 1.1. Error curves for a class C.

Definition 1.3

Inherent complexity needed by any learning algorithm:

$$n_C(\varepsilon,\delta) = \min_{\{L\}} n_L(\varepsilon,\delta)$$

■

2 Error Estimation

2.1 Deviations of Empirical Errors from True Errors

We now consider the error-count estimate or *empirical error*, instead of the "true" error $R(\phi)$:

$$\hat{R}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(\mathbf{x}_i)\neq t_i;(\mathbf{x}_i,t_i)\in D_n\}}(\mathbf{x}_i) \qquad \textbf{2.1}$$

Notes:

1. The distribution of the r.v. $k = n\hat{R}_n(\phi)$ (obtaining k errors in D_n) is binomial with parameters n and $R(\phi)$ [5].
2. Formula 2.1 can be written as $\hat{R}_n(\phi) = \sum_{i=1}^{n} x_i$, where the x_i are n i.i.d. Bernoulli r.v. (see Figure 2.1).

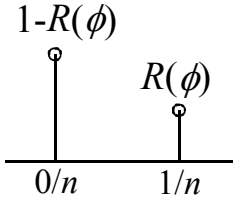

Figure 2.1

3. $\hat{R}_n(\phi)$ is an unbiased estimate of the true error: $\mathrm{E}[\hat{R}_n(\phi)] = R(\phi)$ (convergence of averages to the true mean).

[5] $k \cong N(nR(\phi), R(\phi)(1-R(\phi)))$; $\quad \hat{R}_n(\phi) \cong N(R(\phi), R(\phi)(1-R(\phi))/n)$

How does $\hat{R}_n(\phi)$ converges with n?

Theorem 2.1 (Hoefding's Inequality)

Let $x_1, \ldots, x_n$ be independent and bounded r.v. such that each x_i falls in the $[a_i, b_i]$ interval with probability one. Consider the sum $S_n = \sum_{i=1}^{n} x_i$. Then,for any $\varepsilon > 0$:

$$P\left(S_n - \mathrm{E}[S_n] \geq \varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2} \quad \text{and} \quad P\left(S_n - \mathrm{E}[S_n] \leq -\varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}$$

$$\left(\text{thus, } P\left(\left|S_n - \mathrm{E}[S_n]\right| \geq \varepsilon\right) \leq 2e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}\right) \quad \blacksquare$$

Corolary 2.1

When the x_i take value in $[-c, c]$ and have *zero mean*, Hoefding's inequality can be written as:

$$P\left(\left|S_n\right| / n \geq \varepsilon\right) \leq 2e^{-n\varepsilon^2/(2c^2)}$$

Proof:
Since $\mathrm{E}[S_n] = 0$, we get for the first Hoefding's inequality:

$$P\left(S_n \geq \varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}$$

We now consider: $S_n / n = \sum_{i=1}^{n} x_i / n$. But $x_i / n \in [-c/n, c/n]$; therefore:

$$P\left(S_n / n \geq \varepsilon\right) \leq \exp\left(-2\varepsilon^2 / \left[n\left(\frac{2c}{n}\right)^2\right]\right) = \exp\left(-n\varepsilon^2 / \left(2c^2\right)\right)$$

Using the second inequality we obtain the result above.

$\blacksquare$

We now proceed to bound $P\left(\left|\hat{R}_n(\phi) - R(\phi)\right| > \varepsilon\right)$.

Theorem 2.2

For any $\varepsilon > 0$, n and P,

$$P\left(\left|\hat{R}_n(\phi) - R(\phi)\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2} \qquad \textbf{2.2}$$

Proof:
As previously seen, $\hat{R}_n(\phi)$ is the sum of n independent $\{0/n, 1/n\}$-valued random.variables. Thus, we may apply Hoefding's inequality obtaining the above result. (Notice that $\sum_{i=1}^{n}(b_i - a_i)^2 = n(1/n)^2 = 1/n$). $\blacksquare$

Comments:

- Formula 2.2 can also be obtained from *additive Chernoff bounds*, applicable to Bernoulli variables. As a matter of fact, Hoefding's formula is a kind of generalization of Chernoff bounds.
- From formula 2.2 we have:

$$\delta \geq 2e^{-2n\varepsilon^2} \quad \Rightarrow \quad n \geq n_0(\varepsilon, \delta) = \frac{1}{2\varepsilon^2} \ln\left(\frac{2}{\delta}\right), \qquad \textbf{2.3}$$

which shows an $O(1/\varepsilon^2, 1/\delta)$ behavior of $n_0(\varepsilon,\delta)$. These bounds are independent of the data distribution and the class C.

- As illustrated in Figure 2.2, for the same δ we get higher bounds for n when the accuracy increases (which makes sense).
- The variance of $\hat{R}_n(\phi) - R(\phi)$ can be computed taking into account that $n\hat{R}_n(\phi)$ has a binomial distribution. Thus:

$$\mathrm{E}\left(\left|\hat{R}_n(\phi) - R(\phi)\right|^2\right) = \frac{R(\phi)(1-R(\phi))}{n} \le \frac{1}{4n}$$

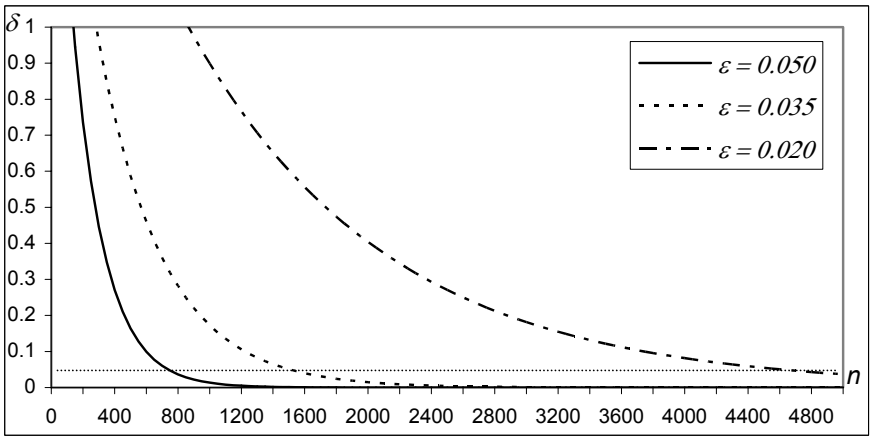

Figure 2.2. Hoefding-based bound. For the same δ larger values of n are required for smaller ε. The usual $\delta = 0.05$ is marked.

We may write the $\hat{R}_n(\phi) - R(\phi)$ deviations as:

$$\hat{R}_n(\phi) - R(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(\mathbf{x}_i)\neq t_i\}}(\mathbf{x}_i) - R(\phi) = \frac{1}{n}\sum_{i=1}^{n}\left(I_{\{\phi(x_i)\neq t_i\}}(\mathbf{x}_i) - R(\phi)\right) = \frac{1}{n}\sum_{i=1}^{n}\chi_i$$

The χ_i random variables are now Bernoulli $\{-R(\phi), 1-R(\phi)\}$-valued r.v. with zero mean and variance $R(\phi)(1-R(\phi))$ (see Figure 2.3).

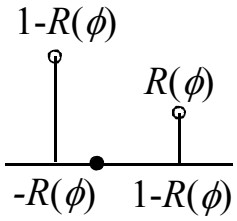

Figure 2.3

Note that:

- $\mathrm{E}[\chi_i] = 0$
- $\mathrm{E}[\chi_i^2] = Var[\chi_i] = R(\phi)(1-R(\phi))$
- χ_i, $|\chi_i| \le 1 - R(\phi)$ (assuming $R(\phi) \le ½$)

Using Corolary 2.1 (denoting $S_n = \sum_{i=1}^{n}\chi_i$) we have:

$$P\left(\left|\hat{R}_n(\phi) - R(\phi)\right| \ge \varepsilon\right) = P\left(|S_n|/n \ge \varepsilon\right) \le 2\exp\left(-n\varepsilon^2/(2(1-R(\phi))^2)\right). \qquad \textbf{2.4}$$

which has the advantage of expressing the bound in terms of $R(\phi)$. (However, it does not outperform the Hoefding's bound.)
There are several formulas for bounding the sum of independent r.v., namely in the form of *exponential inequalities*.Here is one of them:

Theorem 2.3 (Bernstein's inequality)

Let $x_1, \ldots, x_n$ be independent r.v.. with $|x_i| \le c$, zero mean and such that $\sigma^2 = \mathrm{E}[x_i^2]$.
Then, for any $\varepsilon > 0$

$$P\left(\frac{1}{n}\left|\sum_{i=1}^{n} x_i\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + 2c\varepsilon/3}\right) \qquad \textbf{2.5}$$

■

When applying this formula as in 2.4, we take into account that:

$$c = 1 - R(\phi); \qquad \sigma^2 = R(\phi)(1 - R(\phi))$$

Thus:

$$P\left(\left|\hat{R}_n(\phi) - R(\phi)\right| \ge \varepsilon\right) = P\left(\left|S_n\right|/n \ge \varepsilon\right) \le 2\exp\left(-\frac{n\varepsilon^2}{2R(\phi)(1-R(\phi)) + 2(1-R(\phi))\varepsilon/3}\right)$$

Example 2.1

Figure 2.4 shows how formulas 2.3 and 2.5 behave for several values of ε and $R(\phi)$. We see that the increase of ε, keeping $R(\phi)$ constant, leads to a drastic decrease of n for the same P; the increase of $R(\phi)$, keeping ε constant, leads to the decrease of n for the same P, but the difference between the bounds becomes smaller.
Let us consider a classifier ϕ with $R(\phi) = 0.05$. We want to determine the number of cases for which the probability of an estimate $\hat{R}_n(\phi)$ deviating from $R(\phi)$ more than $\varepsilon = 0.02$ is less than 5%. We see that this occurs for $n > 1000$ cases (Bernstein).

□

Corolary 2.2 (*)

The number of cases for which the probability is not greater than δ of an empirical estimate $\hat{R}_n(\phi)$ deviating from $R(\phi)$ more than $\pm\varepsilon$ is bounded as:

$$n \ge n_0(\varepsilon, \delta) = \frac{2R(\phi)(1-R(\phi)) + 2(1-R(\phi))\varepsilon/3}{\varepsilon^2} \ln\left(\frac{2}{\delta}\right). \qquad \textbf{2.6}$$

■

Note:
The bounds $P\left(\left|\hat{R}_n(\phi) - R(\phi)\right| > \varepsilon\right)$ apply to any $\phi \in C$. Thus, they also apply to any classifier of C designed with D_n, i.e. to $P\left(\left|\hat{R}_n(\phi_n) - R(\phi_n)\right| > \varepsilon\right)$.

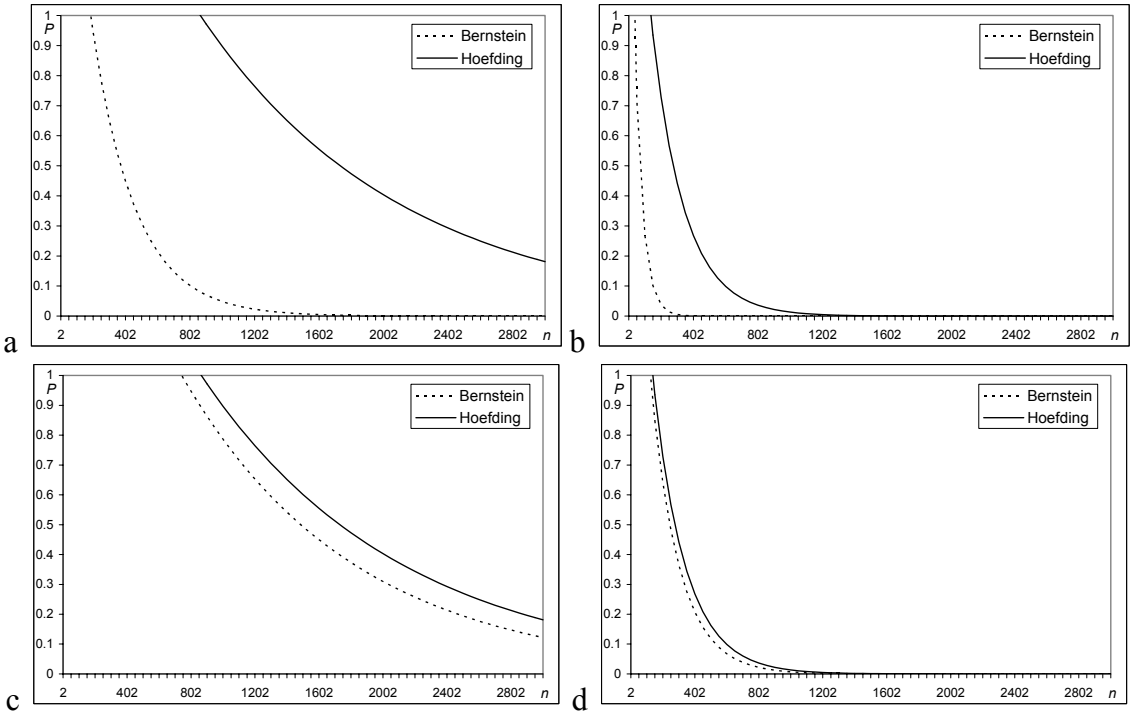

Figure 2.4. Formulas 2.3 and 2.5 for: a) $R(\phi) = 0.05$; $\varepsilon = 0.02$; b) $R(\phi) = 0.05$; $\varepsilon = 0.05$; c) $R(\phi) = 0.3$; $\varepsilon = 0.02$; d) $R(\phi) = 0.3$; $\varepsilon = 0.05$. Note that Hoefding-based formula is independent of $R(\phi)$.

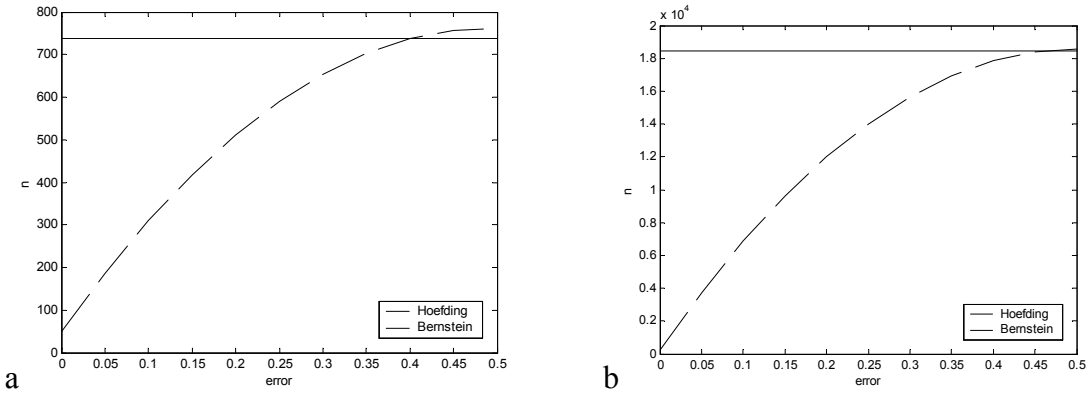

Figure 2.5. Values of bounding n for $\delta = 0.05$ and several values of $R(\phi)$ with: a) $\varepsilon = 0.05$ b) $\varepsilon = 0.01$.

2.2 Deviations of ERM Errors from Optimal Errors

We now assume an algorithm L that picks up the *classifier minimizing the empirical error* (ERM principle):

$$\phi_n^* = L(D_n) \quad \text{such that} \quad \hat{R}_n(\phi_n^*) = \min_{\phi \in C} \hat{R}_n(\phi)$$

Using the preceding results we will see in later sections that it is possible to bound:

$$P\left(\sup_{\phi \in C} \left| \hat{R}_n(\phi) - R(\phi) \right| > \varepsilon \right)$$

The so-called probability of an uniform two-sided distance[6].

[6] Given two probability distributions F and G, their *uniform distance* is $\rho(F,G) = \sup_x \left| F(x) - G(x) \right|$.

We then proceed to determine if this uniform convergence guarantees that the ERM algorithm is a learning algorithm. Before we do that we present an important Lemma.

Lemma 2.1 (Vapnik and Chervonenkis, 1974)

$$R(\phi_n^*) - \inf_{\phi \in C} R(\phi) \le 2 \sup_{\phi \in C} \left| \hat{R}_n(\phi) - R(\phi) \right|$$

$$\left| \hat{R}_n(\phi_n^*) - R(\phi_n^*) \right| \le \sup_{\phi \in C} \left| \hat{R}_n(\phi) - R(\phi) \right|$$

Proof:
The second inequality is trivial (see Figure 2.6).
For the first inequality we have:

$$\begin{aligned} R(\phi_n^*) - \inf_{\phi \in C} R(\phi) &= R(\phi_n^*) - \hat{R}_n(\phi_n^*) + \hat{R}_n(\phi_n^*) - \inf_{\phi \in C} R(\phi) \\ &\le R(\phi_n^*) - \hat{R}_n(\phi_n^*) + \sup_{\phi \in C} \left| \hat{R}_n(\phi) - R(\phi) \right| \\ &\le 2 \sup_{\phi \in C} \left| \hat{R}_n(\phi) - R(\phi) \right| \end{aligned}$$

■

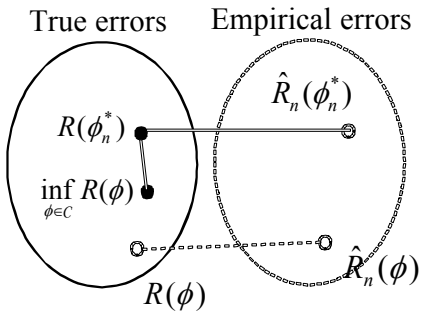

Figure 2.6

This Lemma shows that an upper bound of $\sup_{\phi \in C} \left| \hat{R}_n(\phi) - R(\phi) \right|$ also bounds:

- The suboptimality of the selected classifier ϕ_n^*, i.e, $R(\phi_n^*) - \inf_{\phi \in C} R(\phi)$.
- The error deviation, $\left| \hat{R}_n(\phi_n^*) - R(\phi_n^*) \right|$, due to using the error count estimate.

In what concerns the Bayes error, even if we use a Bayes-consistent rule[7] there isn't any estimation method assuring the convergence of $\hat{R}_n(\phi) - R(\phi^*)$ towards zero rapidly for all distributions, as shows the following:

Theorem 2.4

For any n, any estimate $\hat{R}_n$ of the Bayes probability of error R^*, and for every $\varepsilon > 0$, there is a distribution of (x,t), such that

[7] The Bayes consistency of a classifier implies either $\mathrm{E}[R_n(\phi)] \underset{n \to \infty}{\to} R(\phi^*)$ (weak consistency) or $P(R_n(\phi) - R(\phi^*) \ge \varepsilon) \underset{n \to \infty}{\to} 0$ (strong consistency).

$$\mathrm{E}\left[\left|\hat{R}_n - R^*\right|\right] \geq \frac{1}{4} - \varepsilon .$$

■

We now determine how to bound $P\left(R(\phi_n^*) - \inf_{\phi \in C} R(\phi) < \varepsilon_0\right)$ using a bound on $P\left(\sup_{\phi \in C}\left|\hat{R}_n(\phi_n^*) - R(\phi_n^*)\right| > \varepsilon\right)$.

Theorem 2.5

Suppose C is a finite class. Let $L : \bigcup_{n=1}^{\infty} Z^n \to C$ be such that, for any n and D_n, L picks up the classifier $\phi_n^* = L(D_n)$ with $\hat{R}_n(\phi_n^*) = \min_{\phi \in C} \hat{R}_n(\phi)$. Then L is a learning algorithm for C,

We first note that $P\left(\sup_{\phi \in C}\left|\hat{R}_n(\phi_n^*) - R(\phi_n^*)\right| > \varepsilon\right) \leq \delta$ means that with probability at least $1 - \delta$, the following holds:

$$R(\phi_n^*) - \varepsilon \leq \hat{R}_n(\phi_n^*) \leq R(\phi_n^*) + \varepsilon$$

Equivalently:

$$\hat{R}_n(\phi_n^*) - \varepsilon \leq R(\phi_n^*) \leq \hat{R}_n(\phi_n^*) + \varepsilon$$

Therefore, with probability at least $1 - \delta$:

$$R(\phi_n^*) \leq \hat{R}_n(\phi_n^*) + \varepsilon$$

But, by the ERM definition of L:

$$R(\phi_n^*) \leq \hat{R}_n(\phi_n^*) + \varepsilon = \inf_{\phi \in C} \hat{R}_n(\phi) + \varepsilon$$

Now, $\inf_{\phi \in C} \hat{R}_n(\phi)$ is surely less than or equal than the empirical estimate of any other classifier of C, namely the optimum ϕ^* (note that ϕ_n^* is an optimum for sets of size n). Thus, with probability at least $1 - \delta$:

$$R(\phi_n^*) \leq \inf_{\phi \in C} \hat{R}_n(\phi) + \varepsilon \leq \hat{R}_n(\phi^*) + \varepsilon \leq (R(\phi^*) + \varepsilon) + \varepsilon = R_{\text{opt}} + 2\varepsilon$$

Hence, we can bound $P\left(R(\phi_n^*) - \inf_{\phi \in C} R(\phi) < \varepsilon_0\right)$ with probability at least $1 - \delta$ by choosing $\varepsilon_0 = 2\varepsilon$.

■

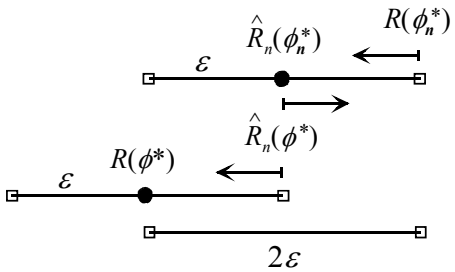

Figure 2.7. The 2ε-configuration of $P\left(R(\phi_n^*) - \inf_{\phi \in C} R(\phi) < \varepsilon_0\right)$.

3 ERM Learning with Finite Classes

We start by presenting an uniform two-sided convergence theorem.

Theorem 3.1

Suppose C is a finite class. Then, for any P, ε and n:

$$P\left(\max_{\phi\in C}\left|\hat{R}_n(\phi)-R(\phi)\right|>\varepsilon\right)\le 2|C|e^{-2n\varepsilon^2} \quad \text{(using Hoefding inequality)} \qquad \textbf{3.1}$$

$$P\left(\max_{\phi\in C}\left|\hat{R}_n(\phi)-R(\phi)\right|>\varepsilon\right)\le 2|C|\exp\left(-\frac{n\varepsilon^2}{2\sigma^2+2c\varepsilon/3}\right) \quad \text{(using Bernstein inequality)} \qquad \textbf{3.2}$$

Proof:
The proof is based on the *union bound* result.
Assume C has N functions ϕ_i and consider the sets:

$$A_i=\left\{z\in Z^d:\left|\hat{R}_n(\phi_i)-R(\phi_i)\right|\ge\varepsilon\right\} \quad \text{with} \quad P(A_i)\le p$$

We have:

$$P\left(\max_{\phi\in C}\left|\hat{R}_n(\phi)-R(\phi)\right|>\varepsilon\right)=P\left(\bigcup_{i=1}^{N}A_i\right)\le\sum_{i=1}^{N}P(A_i)\le Np$$

The probability p is given either by Hoefding or Bernstein inequality.

■

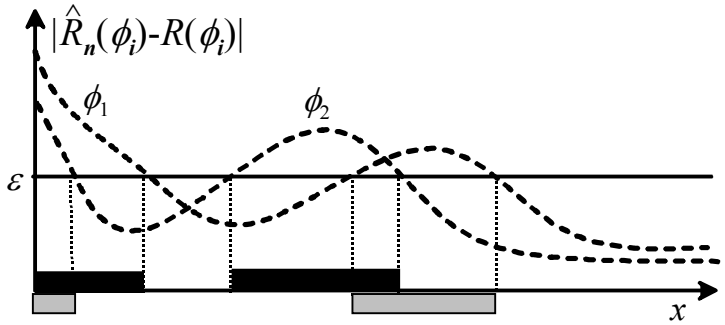

Figure 3.1. A_1 and A_2 are the black and grey intervals, respectively. The maximum deviation corresponds to the union of these intervals.

Corolary 3.1

The ERM algorithm is a learning algorithm with estimation error bound

$$\varepsilon_L(n,\delta)\ge\left(\frac{2}{n}\ln\left(\frac{2|C|}{\delta}\right)\right)^{1/2}, \qquad \textbf{3.3}$$

and sample complexity bound

$$n_L(\varepsilon,\delta)\ge\frac{2}{\varepsilon^2}\ln\left(\frac{2|C|}{\delta}\right). \qquad \textbf{3.4}$$

The proof is the direct application of Theorem 2.5 and Theorem 3.1.

■

Corolary 3.2 (*)

Using Bernstein inequality we obtain the following sample complexity bound from 3.2 by using $\varepsilon/2$ instead of ε:

$$n_L(\varepsilon,\delta) \geq 4\frac{2R(\phi_n^*)(1-R(\phi_n^*))+(1-R(\phi_n^*))\varepsilon/3}{\varepsilon^2}\ln\left(\frac{2|C|}{\delta}\right). \quad \textbf{3.5}$$

■

Application to Perceptrons

We consider perceptrons with weight vectors (including bias) represented with b bits. Thus, each weight can have 2^b values and $|C| \leq 2^{b(d+1)}$.

Theorem 3.2

Let C be the class of perceptrons with $d + 1$ weights represented with b bits. Let $L : \bigcup_{n=1}^{\infty} Z^n \to C$ be such that for any n and D_n L picks up the perceptron $\phi_n^* = L(D_n)$ such that $\hat{R}_n(\phi_n^*) = \min_{\phi \in C} \hat{R}_n(\phi)$.

Then L is a learning algorithm for C, with sample complexity

$$n_L(\varepsilon,\delta) \geq \frac{2}{\varepsilon^2}\ln\left(b(d+1)\ln 2 + \ln\left(\frac{2}{\delta}\right)\right). \quad \textbf{3.6}$$

Proof:
Direct application of Corolary 3.2.

■

One can obtain a more general result for the class of perceptrons with $d + 1$ integer weights represented in $[-k, k]$:

$$n_L(\varepsilon,\delta) \geq \frac{2}{\varepsilon^2}\ln\left(\frac{2(2k+1)^{d+1}}{\delta}\right). \quad \textbf{3.7}$$

Result 3.6 is then the special case for $b = \log_2(2k+1)$.

Theorem 3.3 (*)

Let C be the class of perceptrons with $d + 1$ integer weights represented in $[-k, k]$. Then, for any P, ε and δ, the following bound for the sample complexity holds:

$$n_L(\varepsilon,\delta) \geq 4\frac{2R(\phi_n^*)(1-R(\phi_n^*))+(1-R(\phi_n^*))\varepsilon/3}{\varepsilon^2}\ln\left(\frac{2N(k,d)}{\delta}\right). \quad \textbf{3.8}$$

Proof:
Based on the previous Theorems and C. Felgueiras result. ■

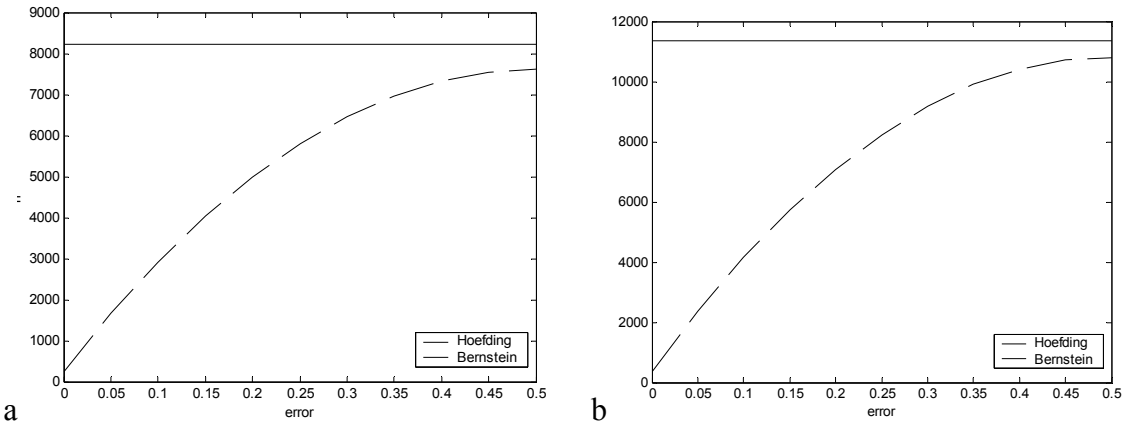

Figure 3.2. Sample complexity for $d = 2$ and $\delta = 0.05$: a) $\varepsilon = 0.05$, $k = 4$; b) $\varepsilon = 0.05$, $k = 16$.

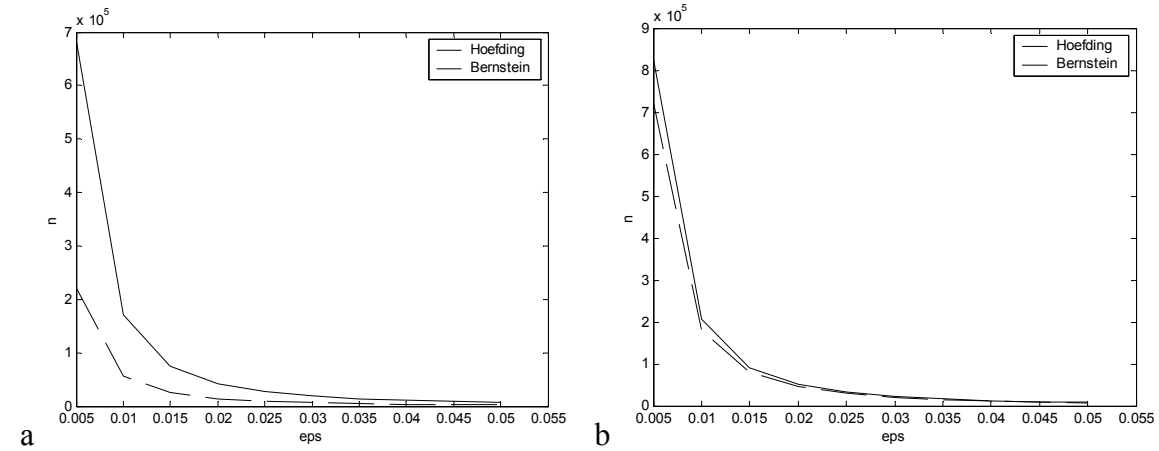

Figure 3.3. Sample complexity for several values of $R(\phi)$ with $d = 2$ and $\delta = 0.05$: a) $R(\phi) = 0.1$, $k = 2$; b) $R(\phi) = 0.4$, $k = 4$.

4 Vapnik-Chervonenkis Theory

Vapnik Chervonenkis (VC) Theory concerns the consistency of the ERM principle (not the Bayes error consistency) for a given class of functions (finite or *infinite*).
The empirical error is:

$$\hat{R}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(\mathbf{x}_i)\neq t_i\}}(\mathbf{x}_i)$$ [8]

The rule empirically selected is $\phi_n^* = \arg\inf_{\phi\in C}\left(\hat{R}_n(\phi)\right)$.

One expects that the true error, $R(\phi_n^*)$, is near the optimal error, $\inf_{\phi\in C} R(\phi)$, for the given class of functions.

[8] VC Theory texts often make explicit the parametrization of the classifier class, using a parameter vector $\alpha \in$ A. The empirical and true errors (risks) are then denoted $R_{emp}(\alpha_n)$ and $R(\alpha_n)$. We will keep the previous notation, with the understanding that: $\hat{R}_n(\phi_n^*) \equiv R_{emp}(\alpha_n)$, $R(\phi_n^*) \equiv R(\alpha_n)$, $\inf_{\phi\in C} R(\phi) \equiv \inf_{\alpha\in \mathrm{A}} R(\alpha)$.

VC Theory allows us to bound $R(\phi_n^*) - \inf_{\phi \in C} R(\phi)$ independently of the data distribution and with a convergence rate that only depends on the structure of C (which can either be finite or infinite).
However, $\inf_{\phi \in C} R(\phi)$ may be far away of the Bayes error (see Figure 4.1)…

$$R(\phi_n^*) - R^* = (R(\phi_n^*) - \inf_{\phi \in C} R(\phi)) + (\inf_{\phi \in C} R(\phi) - R^*)$$

$R(\phi_n^*) - \inf_{\phi \in C} R(\phi)$: *Estimation error* (can be controlled with a learning algorithm; often small)

$\inf_{\phi \in C} R(\phi) - R^*$: *Approximation error* (cannot be controlled; often larger than the estimation error)

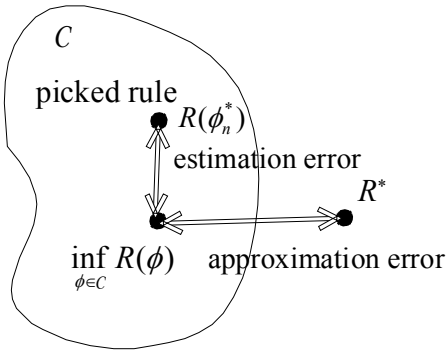

Figure 4.1

$|C|$ small: When the cardinality of C is small we can expect a small estimation error (compare with the discrete weight perceptron class), but the approximation error will probably be very large.

$|C|$ large: As we enlarge C we expect to reach a better approximation to the Bayes error, but the estimation error will probably be very large (due to the richness of class C).

Assuming C the class of all (!) decision functions, we may then expect to find a classifier in C with zero empirical error; however, this classifier may have arbitrary large errors outside D_n. An example is:

$$\phi_n^*(\mathbf{x}) = \begin{cases} t_i & \text{if} \quad \mathbf{x} = \mathbf{x}_i, \quad i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

We obtain the *overfitting* behaviour: the too large class C overfits the data. VC Theory stipulates precise conditions on C in order to avoid this.
We have two choices:

1. Select C such that $\inf_{\phi \in C} R(\phi)$ is near R^*. This corresponds to the Bayes error consistency issue. It can be shown for several classification rules that Bayes error consistency is assured as far as C *grows with n* in a certain way (e.g., *k-NN*).
2. Assume that C is fixed and minimize the estimation error $R(\phi_n^*) - \inf_{\phi \in C} R(\phi)$. This is the approach we take.

4.1 Growth Function

4.1.1 Definitions

Definition 4.1

Let $\mathcal{A}$ represent a collection of measurable sets (class) defined on X^n. For $(\mathbf{x}_1,\ldots,\mathbf{x}_n) \in X^n \subseteq (\Re^d)^n$ we define the *diversity of the collection* $\mathcal{A}$ on the sample $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, and we will represent it by $N_{\mathcal{A}}(\mathbf{x}_1,\ldots,\mathbf{x}_n)$, as the number of different sets in $\{\{\mathbf{x}_1,\ldots,\mathbf{x}_n\} \cap A; A \in \mathcal{A}\}$:

$$N_{\mathcal{A}}(\mathrm{x}_1,\ldots,\mathrm{x}_n) = |\{\{\mathrm{x}_1,\ldots,\mathrm{x}_n\} \cap A; A \in \mathcal{A}\}|. \quad \textbf{4.1}$$

Notes: ■

- Some authors call *concept class* to the collection $\mathcal{A}$.
- The original formulation in SLT (see e.g.,Vapnik,1998) is in terms of Z^n, where Z is the problem space (see SLT-I Tutorial, section 4). However, for classification problems both formulations are equivalent (see Note in page 21).

Example 4.1

Consider the class of semi-closed rectangles $\mathcal{A} = \{]a, b]\times]c, d]\}$ such that $a - b$ and $c - d$ are less than w.
For the point configurations of Figure 4.2 we have:
a) $N_{\mathcal{A}}(\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}) = |\{\varnothing, \{\mathbf{x}_1\}, \{\mathbf{x}_2\}, \{\mathbf{x}_3\}, \{\mathbf{x}_1, \mathbf{x}_3\}, \{\mathbf{x}_2, \mathbf{x}_3\}\}| = 6$
b) $N_{\mathcal{A}}(\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}) = |\{\varnothing, \{\mathbf{x}_1\}, \{\mathbf{x}_2\}, \{\mathbf{x}_3\}, \{\mathbf{x}_2, \mathbf{x}_3\}\}| = 5$

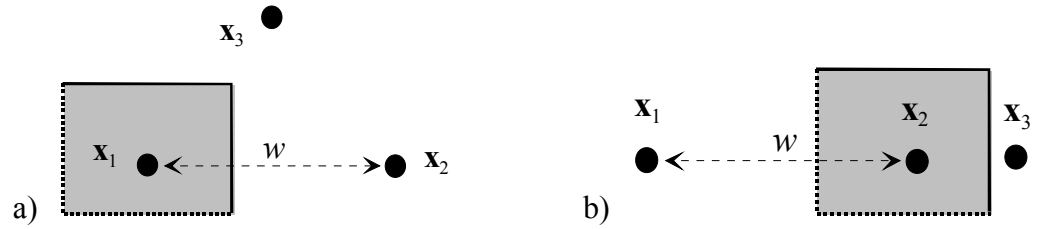

Figure 4.2

□

Definition 4.2

The *n-th shatter coefficient of* $\mathcal{A}$ is

$$S_{\mathcal{A}}(n) = \max_{(\mathbf{x}_1,\ldots,\mathbf{x}_n)\in X^n} N_{\mathcal{A}}(\mathbf{x}_1,\ldots,\mathbf{x}_n) \quad \textbf{4.2}$$

■

Example 4.2
Class $\mathcal{A}$ of Example 4.1 shatters any 3-point set in the equilateral triangle configuration of Figure 4.2a) when $\|\mathbf{x}_1-\mathbf{x}_2\| < w$
On the other hand, the class $Q = \{]-\infty, b]\times] -\infty, d]\}$ (see Figure 4.3) doesn't shatter any 3-point set. We have $\max N_Q(\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}) = 5$.

□

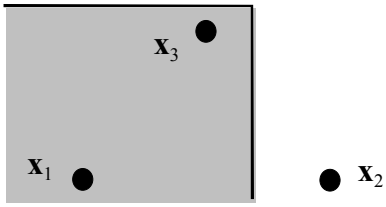

Figure 4.3

Notes:

- $S_{\mathcal{A}}(n)$ is the maximum number of different subsets of n points that can be picked out by the collection $\mathcal{A}$. It measures the richness of class $\mathcal{A}$.
- An alternative formulation considers a function class $\mathcal{F}$ consisting of binary functions $X \to \{0,1\}$. To $\mathcal{F}$ we associate a concept class $C_{\mathcal{F}} = \{C_f, f \in \mathcal{F}\}$, where $C_f = \{\mathbf{x} \in X; f(\mathbf{x}) = 1\}$. This functional definition is equivalent to the set definition, since one can always take the indicator function of the subsets; in other words, one can always think in terms of the subsets induced in X by the function class $\mathcal{F}$.
- Clearly $S_{\mathcal{A}}(n) \leq 2^n$. If $N_{\mathcal{A}}(\mathbf{x}_1,\ldots,\mathbf{x}_n) = 2^n$ for some $(\mathbf{x}_1,\ldots,\mathbf{x}_n) \in X^n$ we say that $\mathcal{A}$ *shatters* $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$.
- If $|\mathcal{A}|$ is finite, then clearly $S_{\mathcal{A}}(n) \leq |\mathcal{A}|$ for all n, and $S_{\mathcal{A}}(n) = |\mathcal{A}|$ for sufficiently large n. Therefore, $S_{\mathcal{A}}(n)$ can be considered a refinement of the cardinality notion applicable to classes of infinite sets.
- Some authors call growth function to $S_{\mathcal{A}}(n)$. Keeping with Vapnik original definition we will use:

Definition 4.3

$G_{\mathcal{A}}(n) = \ln S_{\mathcal{A}}(n)$ is called the *growth function* of the concept class $\mathcal{A}$.

■

Example 4.3

Let $\mathcal{L}$ be the collection of the left half-lines $]-\infty, x]$, $x \in \Re$.
Then $s_{\mathcal{L}}(2) = 3$. As a matter of fact, given any set $\{ z_1, z_2;\ z_1 < z_2, z_i \in \Re\}$, $\mathcal{L}$ only produces the three intersections in $\{\varnothing,\ \{z_1\}, \{z_1, z_2\}\ \}$, instead of the maximum 4 intersections. As a matter of fact, it is easy to see $\mathcal{L}$ only shatters 1-point sets and:

$$S_{\mathcal{L}}(n) = n + 1 = \binom{n}{0} + \binom{n}{1}.$$

□

Example 4.4

If $\mathcal{I}$ is the class of all intervals in $\Re$, then

$$S_{\mathcal{I}}(n) = 1 + \sum_{k=1}^{n} (n - k + 1) = \frac{n(n+1)}{2} + 1 = \binom{n}{0} + \binom{n}{1} + \binom{n}{2}$$

(Proof in Devroye et al., 1996) □

Example 4.5

Let $\mathcal{H} = \{]-\infty, x]; x \in \Re\ \} \cup \{[y, +\infty[; y \in \Re\ \}$ be the collection of half-lines.

Then $S_{\mathcal{H}}(2) = 4$ and $S_{\mathcal{H}}(3) = 6$ (see Example 4.1 of SLT-I). $\mathcal{H}$ shatters any 2-point set and it can be shown that:

$$S_{\mathcal{H}}(n) = 2n$$

□

Example 4.6

Let $\mathcal{P} = \mathcal{H} \cup \{[x_1,x_2];\ x_1 < x_2,\ x_1,x_2 \in \Re\}$ be the collection of half-lines and closed line intervals. (In Example 4.3 of SLT-I this collection was induced by a family of parabolic classifiers.)
Then $S_{\mathcal{P}}(3) = 8$. $\mathcal{P}$ shatters any 3-point set. (see page 22 of SLT-I) and it can be shown that:

$$S_{\mathcal{P}}(n) = n^2 - n + 2$$

Comment
$S_{\mathcal{P}}(n)$ is related to counting the number of *bitonic sequences*. These are binary sequences that have at most two transitions ($0 \to 1$ or $1 \to 0$).

□

Definition 4.4

Let $\mathcal{A}$ be a concept class with $|\mathcal{A}| \geq 2$. The largest k for which $S_{\mathcal{A}}(k) = 2^k$, is denoted $VCD_{\mathcal{A}}$ and called *Vapnik-Chervonenkis dimension* (or simply, *VC-dimension*) *of the class* $\mathcal{A}$. If, for every n, $S_{\mathcal{A}}(n) = 2^n$, then by definition $VCD_{\mathcal{A}} = \infty$. A class $\mathcal{A}$ such that $VCD_{\mathcal{A}} < \infty$ is called a *VC class*.

■

Example 4.7

Example 4.3: $VCD_{\mathcal{L}} = 1$. Example 4.5: $VCD_{\mathcal{H}} = 2$. Example 4.6: $VCD_{\mathcal{P}} = 3$.

□

Note that:

a) In order to prove that l is a VCD lower bound we only need to show that there is at least *one* l-sized dataset that can be shattered.
b) In order to prove that u is a VCD upper bound we need to show that *all* u-sized datasets cannot be shattered.

4.1.2 Growth Function Properties

How fast grows $S_{\mathcal{A}}(n)$? The interesting thing is that no matter which concept class $\mathcal{A}$ we consider it can be shown that $S_{\mathcal{A}}(n)$ grows only ploynomially with n, instead of exponentially.

For two nonnegative integers n, d with $n \geq d$ consider the function $\Phi(n,d)$ computing the number of possible subsets of an n-element set with at most d elements:

$$\Phi(n,d) = \sum_{i=0}^{d} \binom{n}{i} \qquad \textbf{4.3}$$

Lemma 4.1

The function Φ satisfies

$$\Phi(n,d) = \Phi(n-1, d) + \Phi(n-1, d-1). \qquad 4.4$$

Proof:
Based on $\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$

■

Lemma 4.2

The function Φ satisfies

$$\Phi(n,d) = \sum_{i=0}^{d} \binom{n}{i} \le 2\frac{n^d}{d!} \le \left(\frac{en}{d}\right)^d . \qquad \textbf{4.5}$$

The first inequality is established by double induction on n and d, as follows:

Case $d = 1$
Then $\Phi(n,d) = n + 1 \le 2n$ for all $n \ge 1$. Hence the inequality holds.

Case $n = d$
Then $\Phi(n,d) = 2^d$. But, by the binomial expansion:

$$2 \le \left(1 + \frac{1}{d-1}\right)^{d-1} = \left(\frac{d}{d-1}\right)^{d-1}$$

Now suppose that the inequality is verified for $d-1$ with $(n = d-1)$:

$$\Phi(n-1, d-1) = 2^{d-1} \le 2\frac{(d-1)^{d-1}}{(d-1)!}$$

Combining these two inequalities:

$$2^d \le \left(\frac{d}{d-1}\right)^{d-1} 2\frac{(d-1)^{d-1}}{(d-1)!} = 2\frac{d^{d-1}}{(d-1)!} = 2\frac{d^d}{d!}$$

This establishes the inductive step for $n = d > 1$.

Case $n > d > 1$.
From the previous Lemma we have:

$$\Phi(n,d) \le 2\frac{(n-1)^d}{d!} + 2\frac{(n-1)^{d-1}}{(d-1)!}$$

Thus, we have to show that:

$$2\frac{(n-1)^d}{d!} + 2\frac{(n-1)^{d-1}}{(d-1)!} \le 2\frac{n^d}{d!}$$

Multiplying both sides by $d!/2$:

$$(n-1)^d + d(n-1)^{d-1} \le n^d \Rightarrow (d+n-1)(n-1)^{d-1} \le n^d \Rightarrow 1 + \frac{d}{n-1} \le \left(1 + \frac{1}{n-1}\right)^d$$

The last inequality follows from the binomial expansion.
The second inequality is proved using Stirling's approximation.:

$$d! \geq \sqrt{2\pi d}\; d^d e^{-d}$$

■

Theorem 4.1 (Vapnik-Chervonenkis, 1971; Sauer-Shelah, 1972)

Suppose that $VCD_{\mathcal{A}} = h < \infty$. Then, for each $n \geq h$ and all sequences $x_1,\ldots,x_n$, we have

$$S_{\mathcal{A}}(n) \leq \Phi(n, h). \qquad \textbf{4.6}$$

The proof can be found e.g in Sontag (1999).

■

Thus $S_{\mathcal{A}}(n)$ grows polynomially with n.

Theorem 4.2 (Vapnik, 1974)

Any growth function either satisfies

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

or is bounded by:

$$G(n) \leq \ln\left(\sum_{i=0}^{h} \binom{n}{i}\right) \leq h\left(1 + \ln\frac{n}{h}\right) \quad \text{if } n > h, \qquad 4.7$$

where h is the VC-dimension.
The proof is based on the previous Theorem 4.1. The structure of the growth function is shown in Figure 4.2.

■

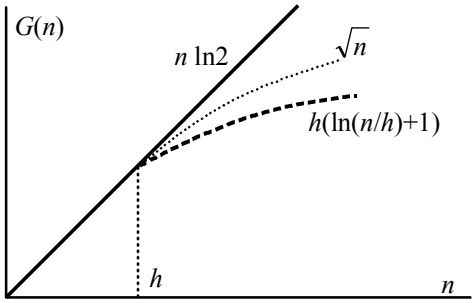

Figure 4.4 For $n > h$, $G(n)$ is bounded by a logarithmic function with coefficient h. It cannot be, for example, $G(n) = \sqrt{n}$ .The quantity h, separating the two different behaviors of the growth function, is the Vapnik-Chervonenkis dimension.

Example 4.8

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

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

Then, $\max N(x_1,\ldots,x_n) = 2^h$ if $n \le h$. For instance, for the previous example of X 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 X 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 4.6 is a tight bound.

□

4.1.3 VC-Dimension of Some Classes

1.

$$S_{\mathcal{A}}(n) \le \sum_{i=0}^{VCD_{\mathcal{A}}}\binom{n}{i} \le \sum_{i=0}^{VCD_{\mathcal{A}}}\binom{VCD_{\mathcal{A}}}{i} n^i = (1+n)^{VCD_{\mathcal{A}}} \quad \text{because} \quad \frac{n!}{i!(n-i)!} \le \frac{h!n^i}{i!(h-i)!}$$

Thus:

$VCD_{\mathcal{A}}$ finite: $S_{\mathcal{A}}(n) \le (n+1)^{VCD_{\mathcal{A}}}$

$VCD_{\mathcal{A}} = \infty$: $S_{\mathcal{A}}(n) = 2^n$ (Note that $\sum_{i=0}^{n}\binom{n}{i} = 2^n$).

2. For $n > 2VCD_{\mathcal{A}}$:

$$S_{\mathcal{A}}(n) \le \sum_{i=0}^{VCD_{\mathcal{A}}}\binom{n}{i} \le \left(\frac{en}{VCD_{\mathcal{A}}}\right)^{VCD_{\mathcal{A}}} \le n^{VCD_{\mathcal{A}}} + 1$$

In particular, for $VCD_{\mathcal{A}} > 2$, we have $S_{\mathcal{A}}(n) \le n^{VCD_{\mathcal{A}}}$

3. Boolean Combinations

i. $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \Rightarrow S_{\mathcal{A}}(n) \le S_{\mathcal{A}1}(n) + S_{\mathcal{A}2}(n)$

ii. $\mathcal{A} = \{A_1 \cup A_2;\ A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2\} \Rightarrow S_{\mathcal{A}}(n) \le S_{\mathcal{A}1}(n)\, S_{\mathcal{A}2}(n)$

iii. Given $\mathcal{A}$ let $\overline{\mathcal{A}} = \{\overline{A};\ A \in \mathcal{A}\}$. Then, $S_{\overline{\mathcal{A}}}(n) = S_{\mathcal{A}}(n)$

iv. $\mathcal{A} = \{A_1 \cap A_2;\ A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2\} \Rightarrow S_{\mathcal{A}}(n) \le S_{\mathcal{A}1}(n) S_{\mathcal{A}2}(n)$

4.
If $\mathcal{A} = \{\ ]-\infty,x_1] \times]-\infty,x_2] \times \ldots \times\ ]-\infty,x_d]\ \}$, then $VCD_{\mathcal{A}} = d$.
If $\mathcal{A}$ is the class of all rectangles in $\Re^d$, then $V_{\mathcal{A}} = 2d$.

5.
If $\mathcal{A}$ is the class of all convex polygons in $\Re^2$, then $VCD_{\mathcal{A}} = \infty$.

6.
$VCD = 0$ implies that the class has only one function.

Results on Boolean combinations can be found with

Lemma 4.3

Let $\mathcal{F}_1,\ldots,\mathcal{F}_k$ represent k classes of classifiers and $b:\{0,1\}^k \rightarrow \{0,1\}$ a fixed Boolean function. Consider the following new class of classifiers:

$$\mathcal{B} = b(\mathcal{F}_1,...,\mathcal{F}_k) = \{b(f_1(.),...,f_k(.)); \quad f_i \in \mathcal{F}_i,\ i = 1,...,k\}$$

Then

$$VCD_{\mathcal{B}} < 2k\log_2(ek)\max_{i=1,...,k}\{VCD_{\mathcal{F}_i}\}. \qquad \textbf{4.8}$$

Proof:
Assume that $S \subseteq X$ is shattered by $\mathcal{B}$, with $|S| = n$, and consider the restrictions:

$$\mathcal{F}_{i|S} : S \rightarrow \{0,1\}; \qquad \mathcal{F} = \mathcal{F}_{1|S} \times ... \times \mathcal{F}_{k|S} : S \rightarrow \{0,1\}^k$$

Now, the mapping $\mathcal{F} \rightarrow \mathcal{B}$ is onto, since to any k-tuple of functions (f_1, …, f_k) corresponds a Boolean composition $b \circ$ (f_1, …, f_k). Therefore:

$$|\mathcal{B}| \le |\mathcal{F}| = \prod_i |\mathcal{F}_i|$$

Let $h_i = VCD_{\mathcal{F}_i} < \infty$ (otherwise there is nothing to prove). By Lemma 4.2,

$$|\mathcal{F}_i| \le \left(\frac{en}{h_i}\right)^{h_i}$$

Thus, with $h = \max_{i=1,..,k} h_i$,

$$|\mathcal{B}| \le \left(\frac{en}{h}\right)^{kh}$$

As S is shattered by $\mathcal{B}$, we have

$$2^n \le \left(\frac{en}{h}\right)^{kh} \quad \Rightarrow \quad \frac{en}{h} \le ke\log_2\left(\frac{en}{h}\right) \quad \Rightarrow \quad n < 2dk\log_2(ek)$$

This last result follows from the calculus argument: $m \le q\log_2(m) \Rightarrow m < 2q\log_2(q)$, when $q > 4$ and $m \ge 1$, which is certainly true in our case.

■

4.1.4 Growth Function of Perceptrons with Linear Thresholds

Theorem 4.3 (Cover, 1965)

Let $$\mathcal{G} = \left\{\sum_{i=1}^{d} a_i\psi_i; \quad a_i \in \Re\right\}$$

be the linear space of functions spanned by a set of d fixed functions ψ_i, i =1,…,d: $\Re^k \rightarrow \Re$. Define $\Psi(\mathbf{x}) = (\psi_1(\mathbf{x}), \ldots, \psi_d(\mathbf{x}))$ (a point in $\Re^d$) and assume that every r-element subset of n points $\{\Psi(x_1), \ldots, \Psi(x_n)\}$ is *linearly independent* (equivalently, the n points are in *general posotion*). Then, the n-th shatter coefficient of the class of sets $\mathcal{A} = \{\{x; g(x) \ge 0\}; g \in \mathcal{G}\}$ is:

$$S_{\mathcal{G}}(n) = 2\sum_{i=0}^{d-1}\binom{n-1}{i}. \qquad \textbf{4.9}$$

■

This result shows that $VCD_{\mathcal{G}} = d$. As a matter of fact, using the previous formula, we have $S_{\mathcal{G}}(d) = 2^d$ and $S_{\mathcal{G}}(d+1) = 2.2^d - 2 < 2^{d+1}$.

Corolary 4.1

Let $\mathcal{A}$ be the class of half-spaces in $\Re^d$, of the form $\{\mathbf{x};\ a\ \mathbf{x} - b \geq 0\}$. Then $VCD_{\mathcal{A}} = d+1$ and

$$S_{\mathcal{A}}(n) = 2\sum_{i=0}^{d}\binom{n-1}{i} \leq 2(n-1)^d + 2$$

The proof is based on Theorem 4.3 by choosing $\psi_i = x_i$ for $1 \leq i \leq d$ and $\psi_{d+1} = 1$

■

Corolary 4.2

The number of linearly separable dichotomies of n points in general position or (*regularly distributed* [9]) in $\Re^d$, is:

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

■

We may also write:

$$D(n,d) = 2\sum_{i=0}^{d}\binom{n-1}{i}. \qquad \textbf{4.10}$$

with the convention $\binom{a}{b} = 0$ if $b > a$. As a matter of fact, we then obtain for $n < d+1$:

$$D(n,d) = 2\left(\sum_{i=0}^{n-1}\binom{n-1}{i} + \underbrace{\sum_{i=n}^{d}\binom{n-1}{i}}_{0}\right) = 2\ 2^{n-1} = 2^n$$

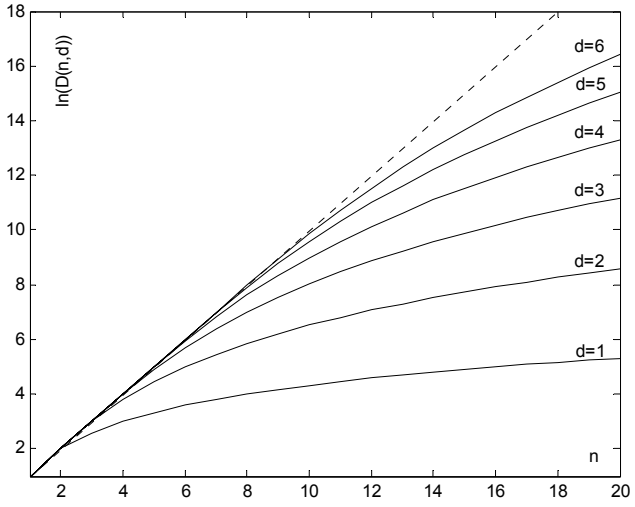

Figure 4.5

Other results can be obtained along theselines ([10]).

---

[9] A set of n points is *regularly distributed* or in *general position* in $\Re^d$ if no d+1 points lie on a linear variety of $\Re^d$. Equivalently, as we did in Theorem 4.4, every r-element subset of the vectors defined by the n points, $r \leq d$, is linearly independent.

Corolary 4.3

Let $\mathcal{H}$ be the class of functions implemented by a simple perceptron with d real inputs and hard-limiting activation function.We have:

$$S_{\mathcal{H}}(n) = D(n,d)\,.$$

Therefore, $VCD_{\mathcal{H}} = d + 1$.

■

Corolaries 4.1 and 4.3 didn't take into account the "general position" restriction of Theorem 4.4. We now show that this restriction is not relevant, since if it is not verified it can only decrease the number of dichotomies. We will show it for the perceptron.

Consider the set of dichotomies implemented by a perceptron $\mathcal{H}$ in a set $D = \{ \mathbf{x}_1, \ldots, \mathbf{x}_n\}$, represented as the restriction of $\mathcal{H}$ to D:

$$\mathcal{H}_{|D} = \{f_1,...,f_k\}$$

Each f_i represents a dichotomy: f_i: $D \rightarrow \{0,1\}$. We don't know the value of $k = |\mathcal{H}_{|D}|$.

Now, consider a set of perceptron weights $\{(\mathbf{w}_j, w_0)\}$ that correspond to each f_j in $\mathcal{H}_{|D}$, i.e.

$$f_j(\mathbf{x}_i) = \begin{cases} 1 & \text{if} \quad \mathbf{w}_j{}'\mathbf{x}_i + w_{0j} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The distance of any point $\mathbf{x}_i$ to the hyperplane represented by $(\mathbf{w}_j, w_0)$ is given by:

$$\Delta_{ji} = \left|\mathbf{w}_j{}'\mathbf{x}_i + w_{0j}\right| / \| \mathbf{w}_j \|$$

Now, some of these distances can be zero; but let us consider:

$$\delta_j = \min\left\{\left|\mathbf{w}_j{}'\mathbf{x}_i + w_{0j}\right|; \quad 1 \leq i \leq n, \ \ \mathbf{w}_j{}'\mathbf{x}_i + w_{0j} \neq 0\right\},$$

and $\delta = \min_j \delta_j$. If we replace w_{0j} by $w'_{0j} = w_{0j} + \delta/2$, we obtain a new set of parameters $\{(\mathbf{w}_j, \ w'_{0j})\}$, corresponding to the same dichotomies ($\mathbf{w}_j{}'\mathbf{x}_i + w_{0j} \geq 0 \Rightarrow \mathbf{w}_j{}'\mathbf{x}_i + w_{0j} + \delta/2 \geq 0$), with the additional separation property that $\left|\mathbf{w}_j{}'\mathbf{x}_i + w'_{0j}\right| \geq \delta/2 > 0$ for all i and j (therefore, all distances are $\neq 0$).

If there are points that are not in general position, let us perturb these points in a small ball around them, by a distance less than:

$$\delta/(2w) \ \text{ with } \ w = \max_j \left\|\mathbf{w}_j\right\|$$

This perturbation will create a new distribution of points $\widetilde{D}$, such that: $\left|\mathcal{H}_{|D}\right| \leq \left|\mathcal{H}_{|\widetilde{D}}\right|$. Now, the set of n-tuples of points in $\Re^d$ that are not in general position has Lebesgue measure

[10] The following can also be proved: Let $\mathcal{A}$ be the class of all closed balls in $\Re^d$. Then, $VCD_{\mathcal{A}} \leq d+2$.

zero; therefore, one will always find an n-tuple lying inside the balls that is in general position. Thus, one always has $\left|H_{|D}\right| \le D(n,d)$.

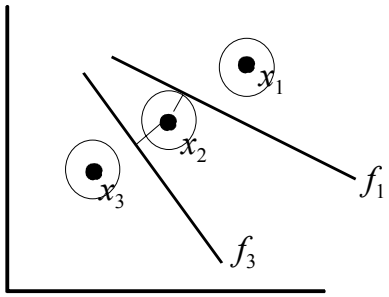

Figura 4.1

Theorem 4.4 (Baum and Haussler, 1989)

Let $C^{(k)}$ be the NN class with h hidden nodes and thresholds as activation functions. Then:

$$S_{C^{(k)}}(n) \le \left(\sum_{i=0}^{d+1}\binom{n}{i}\right)^{h}\left(\sum_{i=0}^{h+1}\binom{n}{i}\right) \le \left(\frac{ne}{d+1}\right)^{h(d+1)}\left(\frac{ne}{h+1}\right)^{h+1} \le (ne)^{hd+2h+1}. \qquad \textbf{4.11}$$

Hence: $V_{C^{(k)}} \le (2hd+4h+2)\log_2(e(hd+2h+1))$.

Sharper bounds are obtained with:

Theorem 4.5 (Baum and Haussler, 1989)

Let $\mathcal{H}$ be the class of functions computed by a feed-forward linear threshold network with a total of w weights and thresholds, and k computation units. Then, for $n \ge w$ we have

$$S_{\mathcal{H}}(n) \le \left(\frac{enk}{w}\right)^{w}, \qquad \textbf{4.12}$$

and hence $VCD_{\mathcal{H}} < 2w\log_2(2k/\ln 2)$.

■

In order to find the VC dimension notice that if $n > w\log_2(enk/w)$, then $2^n > S_{\mathcal{H}}(n) \Rightarrow VCD_{\mathcal{H}} < n$. An inequality in the Appendix shows the stated bound.

This formula yields pessimistic upper bounds. Example for $d = 2$:

h	1	2	3	4	5	6
$VCD_{\max}$	10	57	92	131	173	217

We know that for h=1,2 the true values are 3 and 6.

Theorem 4.6 (Sakurai, 1993)

Let $\mathcal{H}$ be the class of functions computed by a two-layer feed-forward linear threshold network, fully connected, with a total of w weights and thresholds, h computation units in the first layer and $d \ge 3$ inputs, such that $h \le 2^{d/2-2}$. Then

$$VCD_{\mathcal{H}} \ge \frac{dh}{8}\log_2\left(\frac{h}{4}\right) \ge \frac{w}{32}\log_2\left(\frac{h}{4}\right). \qquad \textbf{4.13}$$

Note that $w = dh + 2h + 1$. ■

The constraint on h can be written as: $d \ge 2(\log_2 h+2)$.

Example 4.9

VCD bounds for two layers MLP with $d = 10$ (setting $VCDmin \geq 1$)

h	1	2	3	4	5	6
$VCDmax$	34	156	261	377	502	633
$VCDmin$	1	1	1	1	2	4

□

Definition 4.5

Let $\mathcal{H}$ be a class of $\{0,1\}$-valued functions defined on X, and $\mathcal{F}$ a class of real-valued functions defined on $\Re^d \times X$. $\mathcal{H}$ is a *k-combination of sgn($\mathcal{F}$)* if there is a boolean function g: $\{0,1\}^k \rightarrow \{0,1\}$ and functions $f_1,\ldots,f_k$ in $\mathcal{F}$ such that for all h in $\mathcal{H}$ there is a parameter vector $a \in \Re^d$ satisfying

$$h(x) = g(\text{sgn}(f_1(a,\mathbf{x})),\ldots,\text{sgn}(f_k(a,\mathbf{x}))) \;\forall\; \mathbf{x} \in X$$

A function f in $\mathcal{F}$ is continuous in its parameters (continuous in its p derivatives, C^p) if $\forall$ $\mathbf{x} \in X$ $f(.,\mathbf{x})$ is continuous (respectively, C^p).

Definition 4.6

A set $\{f_1,\ldots,f_k\}$ of differentiable functions mapping from $\Re^d$ to $\Re$ is said to have *regular zero-set intersections* if, for all nonempty subsets $\{i_1,\ldots,i_l\} \subseteq \{1,\ldots,k\}$, the Jacobian of $(f_{i_1},\ldots,f_{i_l})$ has rank l at every point a of the solution set

$$\left\{a \in \Re^d;\, f_{i_1}(a) = \cdots = f_{i_l}(a) = 0\right\}$$

■

This definition forbids degenerate intersections of the zero-sets of the functions, i.e., they must have "true" intersections

Definition 4.7

Let G be a set of real-valued functions defined on $\Re^d$. We say that G has *solution set components bound B* if for any $1 \leq k \leq d$ and any $\{f_1,\ldots,f_k\} \subseteq G$ that has regular zero-set intersections, we have

$$\text{CC}\left(\bigcap_{i=1}^{k}\left\{a \in \Re^d;\quad f_i(a) = 0\right\}\right) \leq B$$

where CC denotes the number of connected components of a set.

■

Example 4.10

Figure 4.5a shows straight lines in $\Re^2$, as could be obtained by a simple perceptron; i.e., in this case we have:

$$f_i : \Re^2 \rightarrow \Re$$
$$(x_1, x_2) \rightarrow f_i(x_1, x_2) = ax_1 + bx_2 + c$$

and the set $\{(x_1,x_2)\in\Re;\ f_i(x_1,x_2) = 0\}$ is a straight line. Thus, we obtain 1 connected component for either $k = 1$ or $k = 2$. Hence, $B = 1$. (Notice that the intersection is always empty for $k > d$ regular zero-set intersections.)
In Figure 4.5b the zero-sets of the functions are parabolas. f_1 and f_3 have no regular intersections. For k=1 we have 1 CC; for k=2 we have 2 CC. Hence B=2.

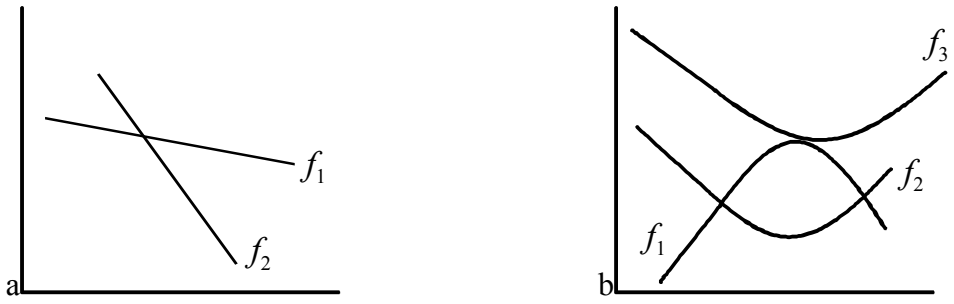

Figure 4.6

□

Theorem 4.7

Suppose that $\mathcal{F}$ is a class of real-valued functions defined on $\Re^d\times X$, and $\mathcal{H}$ is a k-combination of sgn($\mathcal{F}$). If $\mathcal{F}$ is closed under addition of constants[11], has solution set components bound B, and the functions in $\mathcal{F}$ are C^d in their parameters, then

$$S_{\mathcal{H}}(n) \le B\sum_{i=0}^{d}\binom{nk}{i} \le B\left(\frac{enk}{d}\right)^d$$

for $n \ge d/k$.

■

This theorem is useful for deriving results for other types of activation functions (e.g. sigmoids).
For the perceptron (a 1-combination of sgn($\mathcal{F}$) with B=1) one obtains:

$$S_{\mathcal{H}}(n) \le \sum_{i=0}^{d+1}\binom{n}{i} = 2\sum_{i=0}^{d}\binom{n-1}{i} + \binom{n-1}{d+1}$$

larger than the correct value by $\binom{n-1}{d+1}$.

Theorem 4.8 (Goldberg and Jerrum, 1993)

Suppose that $\mathcal{F}$ is a class of real-valued functions defined on $\Re^d\times X$, so that, for all $x\in X$ and $f\in\mathcal{F}$ the function $f(a,x)$ is a polynomial on $\Re^d$ of degree no more than l. Suppose that $\mathcal{H}$ is a k-combination of sgn($\mathcal{F}$). Then if $n \ge d/k$

$$S_{\mathcal{H}}(n) \le 2\left(\frac{2enkl}{d}\right)^d$$

and hence $VCD_{\mathcal{H}} \le 2d\log_2(12kl)$.

■

[11] $f\in\mathcal{F} \Rightarrow f + c\in\mathcal{F},\quad \forall c\in\Re$

4.1.5 Growth Function of Perceptrons with Sigmoids

Theorem 4.9

Suppose $\sigma:\Re\to\Re$ satisfies $\lim_{x\to\infty}\sigma(x)=1$ and $\lim_{x\to-\infty}\sigma(x)=0$. Let N be a feed-forward linear threshold network and N' a network with the same structure as N but with the threshold activation functions replaced by σ in all non-output neurons. Suppose that S is any finite set of inputs. Then, any function computable by N on S is also computed by N' and $\mathrm{VCD}_{N''} \geq \mathrm{VCD}_N$.

■

Lemma 4.4

The class $\mathcal{F}=\{x\to \mathrm{sgn}(\sin(ax);\ x\in\mathcal{N},\ a\in\Re^+\}$ has $VCD=\infty$.

Proof:

Is based on showing that for any $d\in\mathcal{N}$ and a set of points $x_i = 2^{i-1}$, $i = 1,\ldots,d$, one can always shatter this set with functions of $\mathcal{F}$.

■

With this Lemma one can prove:

Theorem 4.10 (Anthony and Bartlett)

Define

$$\sigma(x)=\frac{1}{1+e^{-x}}+cx^3e^{-x^2}\sin(x)$$

for $c > 0$. Then $\sigma(x)$ is analytic (continuous derivatives), and for sufficiently small $c > 0$, we have

$$\lim_{x\to\infty}\sigma(x)=1;\qquad \lim_{x\to-\infty}\sigma(x)=0;\qquad \frac{d^2\sigma(x)}{dx^2}\begin{cases}<0 & x>0\\ >0 & x<0\end{cases}$$

Let N be a two-layer network with one real input, two first-layer neurons using $\sigma(x)$, and one output neuron with threshold function, such that the set of functions of N is

$$\mathcal{H}_N=\{x\to \mathrm{sgn}(w_0+w_1\sigma(a_1x)+w_2\sigma(a_2x));\quad x,w_0,w_1,a_1,a_2\in\Re\}$$

Then $VCD_{\mathcal{H}_N}=\infty$.

■

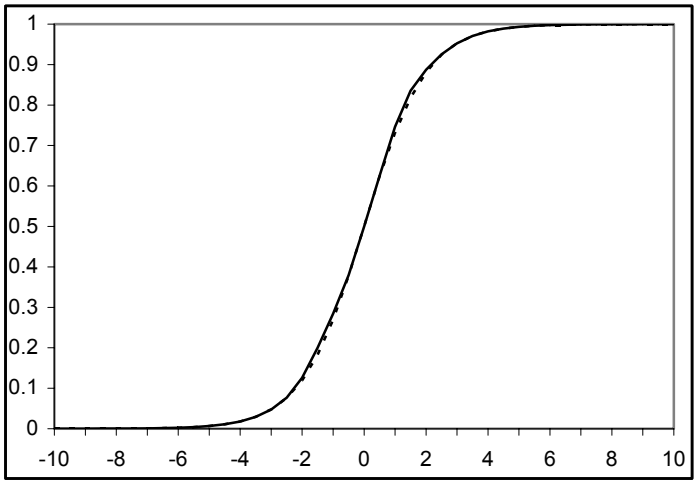

Figure 4.7

Theorem 4.11 (Bartlett, Maiorov and Meir, 1998)

Suppose $\sigma:\Re\to\Re$ satisfies $\lim_{x\to\infty}\sigma(x)=1$ and $\lim_{x\to-\infty}\sigma(x)=0$ and is differentiable at some point α_0, with $\sigma'(\alpha_0)\neq 0$. For any $k \geq 1$ and $w \geq 10k - 14$, there is a feed-forward network with l layers and a total of w parameters, where every computation unit but the output has activation function σ, the output being a linear threshold unit, and for which the set $\mathcal{H}$ of functions computed by the network has

$$VCD_{\mathcal{H}} \geq \left\lfloor \frac{l}{2} \right\rfloor \left\lfloor \frac{w}{2} \right\rfloor$$

■

Theorem 4.12 (Karpinsky and Macintyre, 1997)

Let $\mathcal{H}$ be the set of functions computed by a feed-forward network with w parameters and k computatipn units, in which each computation unit other than the output unit has the standard sigmoid activation function (the output unit being a linear threshold unit. Then

$$S_{\mathcal{H}}(n) \leq 2^{(wk)^2/2}(18wk^2)^{5wk}\left(\frac{en}{w}\right)^w$$

provided $n \geq w$ and

$$VCD_H \leq (wk)^2 + 11wk\log_2(18wk^2)$$

■

Example 4.11

VCD bounds for two layers with d=10 and sigmoid

h	1	2	3	4	5	6
VCDmax	1044	15512	43686	98910	195435	350913
VCDmin	5	12	18	24	30	36

□

4.2 Learning Bounds for Infinite Classes of Classifiers

4.2.1 Upper Bounds

Theorem 4.13 (Vapnik and Chervonenkis, 1971)

Suppose $C=\{\phi\}$ is a class of classifiers defined on a set X. Then, for $n > 0$ and $1 > \varepsilon > 0$,

$$P\left(\left|R(\phi)-\hat{R}_n(\phi)\right| \geq \varepsilon\right) \leq 4S_C(2n)e^{-n\varepsilon^2/8} \qquad \textbf{4.14}$$

The proof is based on the Glivenko-Cantelli Theorem.

■

Corolary 4.4

Suppose C has finite VC-dimension $h = VCD_C \geq 1$ and L is the ERM algorithm. Then L is a learning algorithm for C and if $n \geq h/2$, we have:

$$\varepsilon_L(n,\delta) = \left(\frac{32}{n}\left(h\ln\left(\frac{2em}{h}\right) + \ln\left(\frac{4}{\delta}\right)\right)\right)^{1/2}. \quad \textbf{4.15}$$

$$n_L(\varepsilon,\delta) = \frac{64}{\varepsilon^2}\left(2h\ln\left(\frac{12}{\varepsilon}\right) + \ln\left(\frac{4}{\delta}\right)\right). \quad \textbf{4.16}$$

The proof is based on the previous Theorem and Lemma 3.1. Thus,

$$P\left(\ R(\phi_n^*) - \inf_{\phi\in C} R(\phi) \geq \varepsilon\right) \leq 4S_C(2n)e^{-n\varepsilon^2/32}$$

■

Sharper bounds are obtained with the folowing theorem.

Theorem 4.14 (Vapnik and Chervonenkis, 1998)

Suppose $C=\{\phi\}$ is a class of classifiers defined on a set X. Then, for $n > 0$ and $1 > \varepsilon > 0$,

$$P\left(\ \sup_{\phi\in C}\left|R(\phi) - \hat{R}_n(\phi)\right| > \varepsilon\right) < 4S_C(2n)e^{-n(\varepsilon-1/n)^2} \quad \textbf{4.17}$$

$$P\left(\ \sup_{\phi\in C}\frac{R(\phi) - \hat{R}_n(\phi)}{\sqrt{R(\phi)}} > \varepsilon\right) < 4S_C(2n)e^{-n\varepsilon^2/4} \quad \textbf{4.18}$$

The proof is based on the Glivenko-Cantelli Theorem.

■

Formula 4.6 establishes a bound of two-sided uniform convergence; formula 4.7 establishes a bound of relative uniform convergence.

Corolary 4.5

Using 4.6 and Lemma 4.2 ($P\left(R(\phi_n^*) - \inf_{\phi\in C} R(\phi) > \varepsilon\right) = P\left(\sup_{\phi\in C}\left|R_n(\phi) - r(\phi)\right| > \varepsilon/2\right)$) we have

$$P\left(\ R(\phi_n^{'}) - \inf_{\phi\in C} R(\phi) > \varepsilon\right) < 4S_C(2n)e^{-n(\varepsilon/2-1/n)^2} \quad \textbf{4.19}$$

and

$$\varepsilon(n,\delta) = 2\left(1/n + \sqrt{\left(h(1+\ln(2n/h)) - \ln(\delta/4)\right)/n}\right). \quad \textbf{4.20}$$

■

Formula 4.7 can be rewritten as

$$P\left\{\sup_{\phi\in C}\frac{R(\phi) - \hat{R}_n(\phi)}{\sqrt{R(\phi)}} \geq \varepsilon\right\} \leq 4\exp\left\{\left(\frac{G(2n)}{n} - \frac{\varepsilon^2}{4}\right)n\right\}, \quad \textbf{4.21}$$

where $G(2n) \leq h(1+\ln(2n/h)$ is the growth function for $2n$ and VC-dimension h. (The slightly tighter combinatorial bound leads to computational problems for large n.)
Note that 4.7 holds for any $\phi \in C$; therefore it also holds for ϕ_n^*, the ERM function.

Let $\phi_0 = \arg\inf_{\phi \in C} R(\phi)$. The additive Chernoff bound allows us to state that with probability at least 1-δ the following inequality holds true:

$$R(\phi_0) = \inf_{\phi \in C} R(\phi) \geq \hat{R}_n(\phi_0) - \sqrt{\frac{-\ln \delta}{2n}}. \qquad \textbf{4.22}$$

Using this result together with formula 4.9, one can conclude that with probability at least 1-2δ the following inequality holds true:

$$R(\phi_n^*) - R(\phi_0) \leq \sqrt{\frac{-\ln \delta}{2n}} + \frac{\varepsilon^2}{2}\left(1 + \sqrt{1 + \frac{4\hat{R}_n(\phi_n^*)}{\varepsilon^2}}\right). \qquad \textbf{4.23}$$

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

The result 4.11 is obtained by solving 4.9 in terms of $\hat{R}_n(\phi_n^*)$ and using the bound on $G(2n)$. In order to obtain bounds in terms of $R(\phi_n^*)$ we use:

Theorem 4.15 (*)

Let C be a class of classifiers with VC dimension h. Then for any P, n and δ the following holds:

$$P\left(R(\phi_n^*) - R(\phi_0) \geq \varepsilon(n, \delta, h)\right) \leq \delta$$

with

$$\varepsilon(n, \delta, h) = \sqrt{\frac{-\ln(\delta/2)}{2n}} + 2\sqrt{\frac{R(\phi_n^*)\left(h(1 + \ln(2n/h)) - \ln(\delta/8)\right)}{n}}. \qquad \textbf{4.24}$$

Equivalently, the ERM algorithm is a learning algorithm with estimation error $\varepsilon(n,\delta,h)$.

Proof:
From 4.9 we have for the ERM classifier ϕ_n^* and any given δ:

$$P\left\{\sup_{\phi \in C}(R(\phi_n^*) - \hat{R}_n(\phi_n^*)) \geq \varepsilon\right\} \leq \delta$$

by choosing: $\varepsilon^2 = \dfrac{4R(\phi_n^*)}{n}\left(G(2n) - \ln(\delta/4)\right)$.

Using 4.11 and expressing the probability in terms of δ we obtain the above result.

■

Example 4.12

$VCD = 3$, $\delta = 0.05$

ε	0.01	0.02	0.03	0.04	0.05
4.15	30030392	6842177	2867969	1544189	954006
4.23 (R = 0.1)	312722	72344	30613	16599	10312
4.23 (R = 0.25)	676236	155819	65772	35596	22082

□

4.2.2 Lower Bounds

Theorem 4.16 (Devroye and Lugosi, 1995)

Let C be a class of classifiers with VC-dimension $V \geq 2$ and such that $R = \inf_{\phi \in C} R(\phi) \in \left]0,1/4\right]$. Then, for any classifier ϕ_n based upon D_n, and any $\varepsilon \leq R$

$$\sup_{Z} P(R(\phi_n) - R \geq \varepsilon) \geq \frac{1}{4} e^{-4n\varepsilon^2 / R}. \qquad \textbf{4.25}$$

■

The theorem applies to any ϕ_n (not only the ERM-based ϕ_n).

From 4.16 we obtain the bound:

$$n_L(\varepsilon, \delta) \geq \frac{R \ln(1/4\delta)}{4\varepsilon^2}. \qquad \textbf{4.26}$$

Theorem 4.17 (Simon, 1996)

Suppose C is a class of classifiers with VC-dimension h. For any learning algorithm L the sample complexity $n_L(\varepsilon,\delta)$ satisfies

$$n_L(\varepsilon, \delta) \geq \frac{h}{320\varepsilon^2}$$

for all $0 < \varepsilon, \delta < 1/64$. Furthermore, if C contains at least two functions, we have

$$n_L(\varepsilon, \delta) \geq 2\left\lfloor \frac{1-\varepsilon^2}{\varepsilon^2} \ln\left(\frac{1}{8\delta(1-2\delta)}\right)\right\rfloor. \qquad \textbf{4.27}$$

for all $0 < \varepsilon < 1$ and $0 < \delta < 1/64$.

■

Example 4.13

$VCD = 3$, $\delta = 0.05$

ε	0.01	0.02	0.0.3	0.04	0.05
4.26	10214	2552	1134	636	406
4.25 (R = 0.1)	402	101	45	25	16
4.25 (R = 0.25)	1006	251	112	63	40

□

5 Restricted Learning Model

5.1 Basic Definitions

In the restricted learning model we have the values of t defined in terms of the values of x, i.e., there is a "correct" classification of any $\mathbf{x} \in X$ represented by some *target function* $t(\mathbf{x})$. Therefore, there is only one probability distribution $\mu(A)$ defined on X (instead of Z as before) and $R^* = 0$.

Definitions:

Concept: $C = \{ \mathbf{x} \in X ; t(\mathbf{x}) = 1\}$

Concept class: nonempty $\{C\} \subseteq 2^X$

Training sample *corresponding to t*:

$$D_n = \{(\mathbf{x}_1, t(\mathbf{x}_1)),\ldots,(\mathbf{x}_n, t(\mathbf{x}_n))\} \in Z^n .$$

Hypothesis (classifier): $h : X \to Y$

Class of hypotheses:

$$\mathcal{H} = \{h : X \to Y\}; \quad t \in \mathcal{H}$$

Error of a hypothesis, h:

$$R(h) = R(h,t) = P_\mu\big(h(x) \neq t(x)\big) \equiv P_\mu\big(h(X) \Delta t(X)\big)$$

Definition 5.1

Given $\mathcal{H} = \{h : X \to Y\}$ a *learning algorithm L* for $\mathcal{H}$. is a function

$$L : \bigcup_{n=1}^{\infty} \{D^n\} \to \mathcal{H}$$

such that given $\varepsilon, \delta \in]0, 1[$, there is an integer $n_0(\varepsilon, \delta)$ such that for $n \geq n_0(\varepsilon, \delta)$ and every training sample D_n (as above), then $h_n = L(D_n)$ satisfies

$$P\big(R(h_n) \leq \varepsilon\big) > 1 - \delta$$

for any probability distribution μ on X.

■

The restricted model is a special case of the general model, since given $t \in \mathcal{H}$ and a distibution μ on X, there is a corresponding distribution P on Z. As a matter of fact, for any measurable subset $A \subseteq \Re^d$:

$$P((\mathbf{x}, t(\mathbf{x})); \mathbf{x} \in A;) = \mu(A)$$
$$P((\mathbf{x}, y); \mathbf{x} \in A, y \neq t(\mathbf{x})) = 0$$

Furthermore, $R_P(h) = R_\mu(h,t)$. Thus the restricted model corresponds to considering only a subset of all distributions on Z.

Theorem 5.1 (Vapnik-Chervonenkis, 1974)

Assume $|C| < \infty$ and $\min_{\phi \in C} R(\phi) = 0$. Then, for every $\varepsilon > 0$,

$$P\left(R(\phi_n^*) > \varepsilon\right) \le |C| e^{-n\varepsilon} \quad \text{and} \quad \mathrm{E}\left[R(\phi_n^*)\right] \le \frac{1 + \ln|C|}{n}. \qquad \textbf{5.1}$$

■

In this case the ERM principle converges to $\min_{\phi \in C} R(\phi) = 0$ with n of $O(1/\varepsilon, 1/\delta)$.

Definition 5.2

A learning algorithm for the restricted model is said to be *PAC* (*Probably Approximately Correct*) if it satisfies the following conditions: $\varepsilon, \delta \in]0, ½]$; the learning algorithm time (thus $n_0(\varepsilon, \delta)$) is polynomial in d, $1/\varepsilon$, $1/\delta$ and size(c), where size(c) is the number of parameters needed to represent a concept.

■

5.2 Consistent Learning

A *consistent learning algorithm* for the restricted model is one that outputs a hypothesis that perfectly fits the training data:

$$\forall x_i \in D_n, \quad h(x_i) = t(x_i)$$

As far as we use consistent learning, we can relax the condition $|C| < \infty$ in Theorem 3.1, as shown in

Theorem 5.2

Assume a consistent learning algorithm L that outputs $h = L(D_n)$. Then, the sample complexity is:

$$m_L(\varepsilon, \delta) = \frac{1}{\varepsilon} \ln\left(\frac{|\mathcal{H}|}{\delta}\right). \qquad \textbf{5.2}$$

■

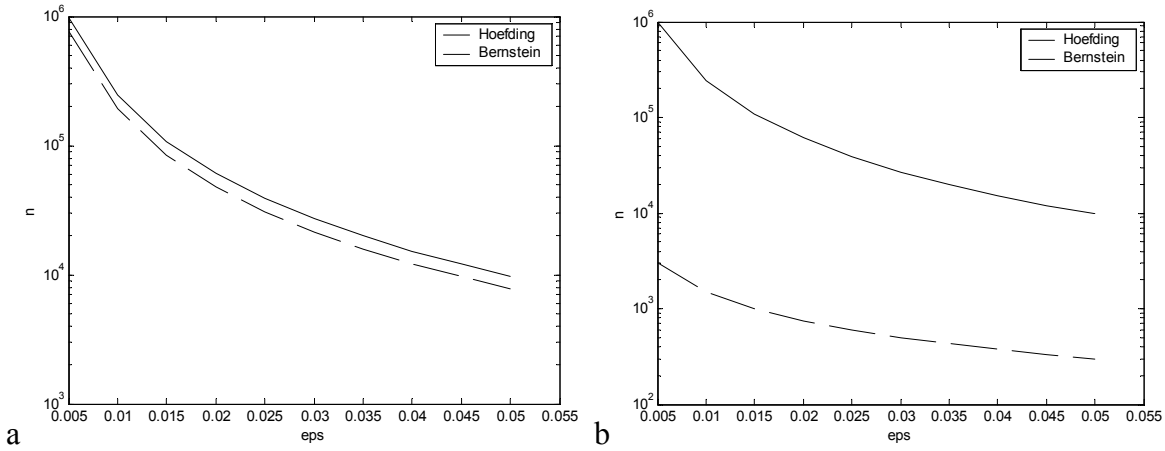

Figure 5.1. Bounds for $d = 2$, $\delta = 0.05$, $k = 8$ in logarithmic scale: a) $R(\phi) = 0.4$; b) $R(\phi) = 0$.

5.3 Learning Bounds

Theorem 5.3 (Blumer et al., 1989)

Let C be a class of classifiers. Suppose that $\inf_{\phi \in C} R(\phi) = 0$ (i.e., the Bayes classifier is in C). Let ϕ_n^* denote the classifier that minimizes the empirical error, with $R^* = 0$. Then:

$$P\left(R(\phi_n^*) > \varepsilon\right) \leq 2S_C(2n)2^{-n\varepsilon/2}. \quad \textbf{5.3}$$

■

Theorem 5.4 (Blumer *et al.*, 1989)

Let C be a class of concepts and H a hypothesis space. Then:

i. C is PAC-learnable iff VCD_C is finite.

ii. If VCD_C is finite, then:

(a) For $0 < \varepsilon < 1$ and sample size at least

$$n_u = \max\left[\frac{4}{\varepsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8VCD_C}{\varepsilon}\log_2\left(\frac{13}{\varepsilon}\right)\right], \quad \textbf{5.4}$$

any consistent algorithm is of PAC learning for C.

(b) For $0 < \varepsilon < ½$ and sample size less than

$$n_l = \max\left[\frac{1-\varepsilon}{\varepsilon}\ln\left(\frac{1}{\delta}\right), DVC_C\left(1-2\left(\varepsilon\left(1-\delta\right)+\delta\right)\right)\right], \quad \textbf{5.5}$$

no learning algorithm, for any hypothesis space H, is of PAC learning for C.

A sharper upper bound is obtained with:

Theorem 5.5 (Shawe-Taylor et al., 1993)

$$P\left(R(\phi_n^*) > \varepsilon\right) \leq 2S_C(n^2)2^{-n\varepsilon}$$

Thus:
$$\varepsilon(n,\delta) \geq \frac{h(1+\ln(n^2/h)) - \ln(\delta/2)}{n\ln 2}. \quad \textbf{5.6}$$

where h is the VC-dimension.

■

Example 5.1

$VCD = 3$, $\delta = 0.05$

ε	0.05	0.1	0.15	0.2	0.25
5.4	3851	1685	1030	723	547
5.7	1346	604	375	267	204

□

Theorem 5.6 (Devroye e Lugosi, 1995)

Let C be a class of classifiers with VC-dimension $V \geq 2$. Suppose that $\inf_{\phi \in C} R(\phi) = 0$ (i.e., the Bayes classifier is in C). Let ϕ_n^* denote the classifier that minimizes the empirical error, with $R^* = 0$. Then for any classifier based on D_n, with $n \geq V - 1$ and for any $\varepsilon \leq 1/4$

$$\sup_{(x,y)\in\aleph} P(R_n \geq \varepsilon) \geq \frac{1}{e\sqrt{\pi V}} \left(\frac{2ne\varepsilon}{V-1} \right)^{(V-1)/2} e^{-4n\varepsilon/(1-4\varepsilon)}. \qquad \textbf{5.7}$$

Furthermore, if $15 \leq n$ and $n \leq (V-1)/(12\varepsilon)$

$$\sup_{(x,y)\in\aleph} P(R_n \geq \varepsilon) \geq \frac{1}{10}$$

■

Example 5.2

$VCD = 3$, $\delta = 0.05$

ε	0.01	0.02	0.0.3	0.04	0.05
5.5	297	147	97	72	57

□

6 Appendix

6.1 The Glivenko-Cantelli Theorem

Let $z_1, \ldots, z_n$ be i.i.d. real-valued r.v. with distribution function $F(z) = \mathrm{P}(z_1 \leq z)$. Denote the empirical distribution function by

$$F_n(z) = \frac{1}{n} \sum_{i=1}^{n} I_{z_i \leq z}$$

Then

$$P\left(\sup_{z\in\Re} |F(z) - F_n(z)| > \varepsilon \right) \leq 8(n+1)e^{-n\varepsilon^2/32}$$

and, in particular, by the Borel-Cantelli lemma,

$$\lim_{n\to\infty} \sup_{z\in\Re} |F(z) - F_n(z)| = 0 \text{ with probability one.}$$

■

This theorem states a.s. convergence of the empirical distribution to the true one and is sometimes referred to as the fundamental theorem of mathematical statistics.

6.2 Useful Formulas

6.2.1 Markov's inequality

If a r.v. x is almost surely nonnegative, then

$$P(x \geq a) \leq \frac{\mathrm{E}[x]}{a} \quad \forall a > 0$$

To see this, notice that $\mathrm{E}[x] \geq \mathrm{E}[x \mid x \geq a] P(x \geq a) \geq aP(x \geq a)$

6.2.2 Logarithms

$2\ln x < x$

$$\ln x = 2\left[\frac{x-1}{x+1} + \frac{(x-1)^3}{3(x+1)^3} + \frac{(x-1)^5}{5(x+1)^5} + \ldots + \frac{(x-1)^{2n+1}}{(2n+1)(x+1)^{2n+1}} + \ldots\right] \qquad x > 0$$

$$\ln x = \frac{x-1}{x} + \frac{(x-1)^2}{2x^2} + \frac{(x-1)^3}{3x^3} + \ldots + \frac{(x-1)^n}{nx^n} + .. \qquad x > \frac{1}{2}$$

$$x \ln x \geq x - 1 + \frac{(x-1)^2}{2x} \qquad x > \frac{1}{2}$$

$$x \ln x \leq x - 1 + \frac{(x-1)^2}{2} \qquad x > 1 \qquad \left(x \ln x = \frac{x-1}{1!} + \frac{(x-1)^2}{2!} - \frac{(x-1)^3}{3!} \ldots\right)$$

Suppose $q > 4$, $m \geq 1$. Then

$$m \geq 2q \log_2(q) \Rightarrow m > q \log_2(m)$$

Equivalently,

$$m \leq q\log_2(m) \Rightarrow m < 2q\log_2(q)$$

For any $\alpha, x > 0$

$$\ln x \leq \alpha x - \ln\alpha - 1$$

with equality only if $\alpha x = 1$.

6.2.3 Binomial Formulas

Newton binomial formula: $(1+x)^p = \sum_{i=0}^{p} \binom{p}{i} x^i$

$$\sum_{i=0}^{p} \binom{p}{i} = 2^p$$

$$\sum_{i=0}^{h} \binom{n}{i} = \sum_{i=0}^{h} \binom{n-1}{i} + \sum_{i=0}^{h-1} \binom{n-1}{i} \qquad \left(\text{porque } \binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i}\right)$$

$$\sum_{i=0}^{h} \binom{n}{i} < 2\left(\frac{n^h}{h!}\right) \leq \left(\frac{en}{h}\right)^h \qquad \forall n > h$$

$$\sum_{i=0}^{h} \binom{n}{i} \leq n^h + 1 \qquad \forall h, \forall n > 2h; \qquad \sum_{i=0}^{h} \binom{n}{i} \leq n^h \qquad \forall h > 2, \forall n > 2h$$

6.2.4 Exponentials

$$1 + x \le e^x \quad \forall x \in \Re$$

$$\left(1 + \frac{1}{x}\right)^x < e \qquad ;\left(1 - \frac{1}{x}\right)^x < e^{-1} \qquad \forall x > 0$$

Euler's inequality:

$$\left(1 + \frac{a}{x}\right)^x < e^a \qquad \forall x > 0, a \in \Re, a \neq 0$$

6.2.5 Stirling Formula

$$n! = n^n e^{-n} \sqrt{2\pi n}\left(1 + \varepsilon_n\right) \text{ com } \varepsilon_n = \frac{1}{12n} + \frac{1 + \theta_n}{288n^2} \underset{n \to \infty}{\to} 0$$

$$n^n e^{-n} \sqrt{2\pi n}\, e^{1/(12n+1)} < n! < n^n e^{-n} \sqrt{2\pi n}\, e^{1/(12n)}$$

References

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

Antos A, Lugosi G (1998) Strong Minimax Lower Bounds for Learning. Machine Learning, 30:31-56.

Baum EB (1988) On the Capabilities of Multilayer Perceprons. Journal of Complexity, 4:193-215.

Baum EB, Haussler D (1989) What Size Net Gives Valid Generalization? Neural Computation, vol. 1(1):151-160.

Baum EB, Haussler D (1989) What Size Net Gives Valid Generalization? Neural Computation, 1:151-160.

Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK (1989) Learnability and the Vapnik-Chervonenkis Dimension. Journal ACM, 36: 929-965.

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

Devroye L, Györfi L, Lugosi G (1996) A Probabilistic Theory of Pattern Recognition. Springer-Verlag

Ehrenfeucht A, Kearns M, Valiant L (1989) A General Lower Bound on the Number of Examples Needed for Learning. Information and Computation, 82: 247-261.

Haussler D, Kearns M, Schapire R (1994) Bounds on the Sample Complexity of

Kearns MJ, Vazirani UV (1994) An Introduction to Computational Learning Theory. MIT Press.

Kearns M, Seung HS (1995) Learning from a Population of Hypotheses. Machine Learning, 18:255-276.

Linial N, Mansour Y, Rivest RL (1991) Results on Learnability and the Vapnik-Chervonenkis Dimension. Information and Computation, 90: 33-49.

Petrov VV (1995) Limit Theorems of Probability Theory. Clarendon Press. Oxford.

Sontag ED (1999) VC Simension of Neural Networks. In Neural Networks and Machine Learning (Nato ASI Series. Series F, Computer and Systems Sciences, vol. 168). Ed. CM Bishop.

Vapnik VN (1974) Theorie der Zeichnerkennung. Academie Verlag, Berlin.

Vapnik VN (1998) Statistical Learning Theory. Wiley, New York.

Vapnik VN (1999) The Nature of Statistical Learning Theory. Springer-Berlag.

Vidyasagar M (2003) Learning and Generalization. With Applications to Neural Networks. Springer-Verlag.

Wenocur RS, Dudley RM (1981). Some Special Vapnik-Chervonenkis Classes. Discrete Mathematics, 33: 313-318.