Neural Network Interest Group

Título/*Title*:
Introduction to Statistical Learning Theory
PART II –Data Regression

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

Relatório Técnico/*Technical Report* No. 2 /2003

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

Título/*Title*:
Introduction to Statistical Learning Theory
PART II –Data Regression
Autor(es)/*Author(s)*:
F. Sereno, J.P. Marques de Sá
Relatório Técnico/*Technical Report* No. 2 /2003
Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Contents

8 The Data Regression Learning Problem 5
9 Diversity of a Set of Real Functions 6
10 Consistent Learning for Sets of Real Functions 8
11 Bounds on the Rate of Convergence 10
11.1 VC-Dimension of a Set of Real-Valued Functions 10
11.2 Distribution Independent Bounds for Convergence 11
12 Pseudo- and Fat-Shattering Dimensions 14
Appendix – Regression Solutions 21

8 The Data Regression Learning Problem

In data regression we are seeking a functional relation of one random variable y depending on a predictor variable x, which may or may not be random, as shown in Figure 8.1.

$$y = g(x).$$

We see that for every predictor value x_i, we must take into account the probability distribution of y as expressed by the density function $f_y(y)$.

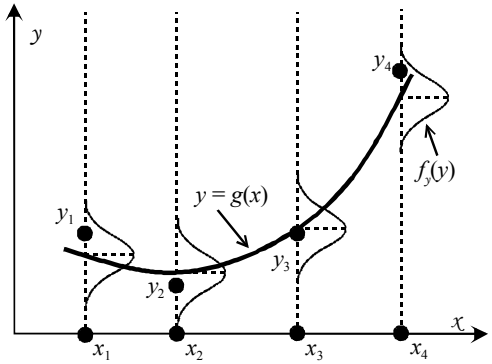


Figure 8.1

We can cast the problem of learning the functional dependence $y = g(x)$ in the same framework as minimizing a certain risk function, as given by formula 1.1:

$$R(\alpha) = \int Q(z,\alpha)dF(z), \quad \alpha \in \mathrm{A}. \tag{8.1}$$

As a matter of fact, classic regression consists in minimizing the above risk when the following loss function is used (the well-known *least mean square* method):

$$Q(z,\alpha) = Q((y,x),\alpha) = L(y, g(x,\alpha)) = (y - g(x,\alpha))^2. \tag{8.2}$$

For the above loss function the minimization leads to a particular α_0 such that (see Appendix):

$$g(x,\alpha_0) = \int_{-\infty}^{\infty} yf(y \mid x)dydx = \mathrm{E}[y \mid x]$$

Thus, for a quadratic loss function the sought for regression solution is the conditional mean of y given the predictor x, as depicted in Figure 8.1. This does not hold for other loss functions (see Appendix).

Figure 8.1 assumes a known conditional distribution of y given the predictor x, namely with normally distributed deviations (residuals) from $\mathrm{E}[y|x]$, with zero mean and equal variance: the *classical model*.

In the general case, the probability distribution of the data $F(z)$ is unknown (and $f(y|x)$ as well); we then attempt to minimize the following *empirical loss*:

$$R_{emp}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i,\alpha))^2 \ , \qquad 8.3$$

in a training sample $Z_n = \{(x_1,y_1),\ldots, (x_i,y_i),\ldots, (x_n,y_n)\}$ (using e.g. a neural network approach).

As we have already seen in section 3
, in order to have a consistent learning for the regression problem based on the ERM principle, the Theorem 3.1 has to hold true, since this Theorem applies to any risk functional.

9 Diversity of a Set of Real Functions

In section 6 the three milestones of learning theory for data classification were expressed in terms of an integer measure, $N^{\Lambda}(Z_n)$, that reflected the "expressiveness" of the family of clasifying functions.

Let us again consider the set of n-dimensional vectors for $\alpha\in\Lambda$:

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

For data classification the elements of $q(\alpha)$ were discrete (e.g. dichotomic). For data regression the vectors $q(\alpha)$, $\alpha\in\Lambda$, describe a subset of a continuous n-dimensional domain, depending on the training set Z_n and of the particular family of loss functions. For instance, in Example 4.1 the vectors $q(\alpha)$ for $n = 3$ objects were represented by vertices of a cube. Now, in data regression, the set of vectors $q(\alpha)$ contains an infinite number of elements. We assume that $Q(z,\alpha)$, $\alpha\in\Lambda$, is a family of uniformly bounded functions:

$$|Q(z,\alpha)| < C \ , \ \forall\,\alpha\in\Lambda.$$

We then have for $n = 3$ a set of infinite points inside a cube of edge 2C, as shown in Figure 9.1.

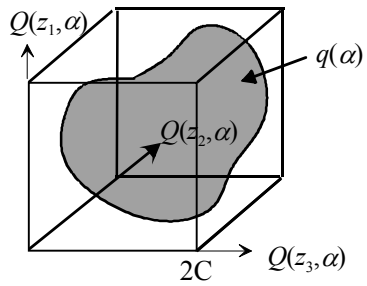


Figure 9.1

The generalization of $N^A(Z_n)$ for an infinite set is possible if the infinite set can be covered by a *finite ε-net*[1].

Definition
The set B of elements b in a metric space M (with a distance measure ρ) is called an *ε-net* of the set G, if any point $g \in G$ is distant from some point $b \in B$ by an amount not exceeding ε:

$$\rho(b, g) < \varepsilon$$

The set G admits a *finite ε-net* if for each ε there exists an *ε-net*, B_ε, with a finite number of elements. The B_ε^* set with minimal number of elements is the *minimal ε-net*, with a number of elements:

$$N(\varepsilon, z_1, \ldots, z_n) \equiv N(\varepsilon, Z_n) \qquad [2]$$

Example 9.1

Consider the following sample of 2 points in [0, 1]: $Z_2 = \{(0, 0.2), (1, 0.5)\}$. For $Q(z,\alpha) = L(y,g(x,\alpha))$ with $g(x,\alpha) = \{b;\ \alpha=b \in [0, 1]\}$ and loss function 8.2 determine the set $\{q(\alpha)\}$ and find a 0.1-net using Euclidian norm.
Figure 9.2 shows the solid curve corresponding to the set $G = \{q(\alpha)\}$. The open circles with center b have $\rho(b, g) < 0.1$, where ρ is the Euclidian norm. We see that the set B = {(0, 0.2), (0, 0.1), (0.1, 0), (0.23, 0), (0.4, 0.1), (0.5, 0.2), (0.6, 0.2)} constitutes a 0.1-net.

□

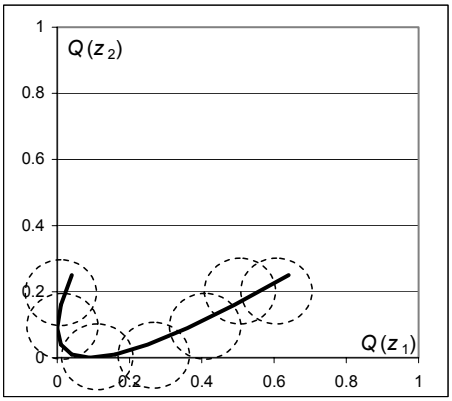


Figure 9.2

Even for such an easy configuration as in the previous example it may be a hard task to find the minimal *ε-net*. Next example illustrates simpler solutions for another type of loss function and distance measure.

Example 9.2

Consider the following loss function and Chebyshev distance measure:
$Q(z,\alpha) = L(y, g(x,\alpha)) = | y - g(x,\alpha) |$

[1] We follow the notation of Vapnik (1998). Also called *ε-cover* by other authors.
[2] For the parameter set A, i.e., $N^A(\varepsilon, Z_n)$; we omit A for uncluttered notation.

$\rho(b,g) = \max_i |b_i - g_i|$ (instead of balls we have hypercubes)

Also, consider the cover by closed hypercubes: $\rho(b,g) \le \varepsilon$.

Let Z_2 = {(0, 1), (1, 0.5)}and $g(x,\alpha)$ = {b; α=b $\in$ [0, 1]}. Then, $N(0.125; Z_2) = 4$ (Figure 9.3a).
Let Z_2 = {(0, 0.5), (1, 0.5)}and $g(x,\alpha)$ = {b; α=b $\in$ [0, 1]}. Then, $N(0.125; Z_2) = 2$ (Figure 9.3b).
In both cases $|Q(z,\alpha)| < 1$, $\forall \alpha \in$ A.
Let Z_2 = {(0, 1), (1, 0)}and $g(x,\alpha)$ = {$ax + b$; $\alpha = (a, b) \in$ A = $[-1, 1]^2$}. In this case $|Q(z,\alpha)| < 2$, $\forall \alpha \in$ A the{$q(\alpha)$} set is the dotted region in Figure 9.3c .Then, $N(0.125; Z_2) = (2/0.25)^2 - 12 = 52$.

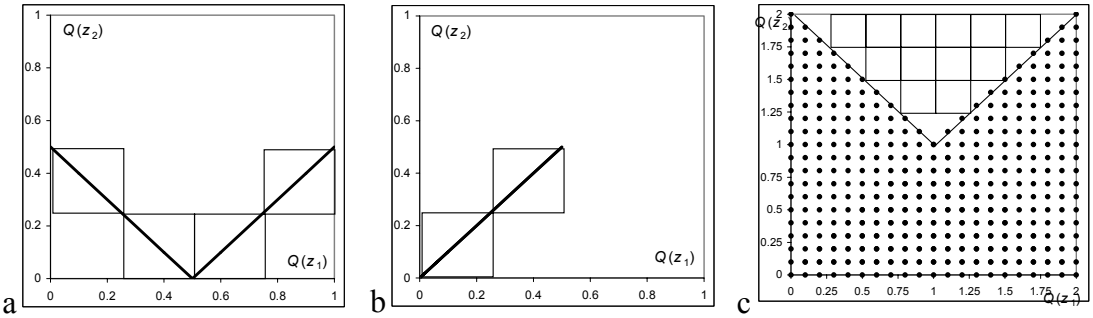


Figure 9.3

As in section 5 we define:

Random entropy of the set of bounded functions $Q(z,\alpha)$ on the sample Z_n (this is a r.v.):

$$H(\varepsilon; Z_n) = \ln N(\varepsilon; Z_n)$$

ε-entropy (or *VC-entropy*) of the set of bounded functions $Q(z,\alpha)$on the sample Z_n:

$$H(\varepsilon; n) = \mathrm{E}\left[H(\varepsilon; Z_n)\right] = \mathrm{E}\left[\ln N(\varepsilon; Z_n)\right]$$

10 Consistent Learning for Sets of Real Functions

Theorem for uniformly bounded functions (Vapnik)

In order that uniform convergence

$$P\left\{\sup_{\alpha \in \mathrm{A}} \left| \int Q(z,\alpha) dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \right| > \varepsilon \right\} \underset{n\to\infty}{\to} 0$$

over a set of uniformly bounded functions $Q(z,\alpha)$ be valid, it is necessary and sufficient that the following holds:

$$\frac{H(\varepsilon; n)}{n} \underset{n\to\infty}{\to} 0, \quad \forall\ \varepsilon > 0$$

A stronger Theorem also proves that the above condition guarantees almost sure convergence.

Similarly to section 6, the three milestones for learning sets of bounded real functions are:

1. **Sufficient condition for the consistency of the ERM principle:**

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

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

$$\lim_{n\to\infty} \frac{H_{\text{ann}}(\varepsilon;n)}{n} = 0, \quad \forall\ \varepsilon > 0, \text{ with } H_{\text{ann}}(\varepsilon;n) = \ln \mathrm{E}\left[N(\varepsilon;Z_n\right]$$

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

$$\lim_{n\to\infty} \frac{G(\varepsilon;n)}{n} = 0, \quad \forall\ \varepsilon > 0, \text{ with } G(\varepsilon;n) = \ln\left\{\sup_{Z_n} N(\varepsilon;Z_n)\right\}. \qquad 10.1$$

It is also a sufficient condition for fast convergence.

The computation of the entropy, the annealed entropy and the growth function as previously defined is usually very difficult (virtually impossible) in practical cases. In the following section more practical measures of expressiveness (capacity) of sets of real-valued functions are presented. The following is a naif example for illustrating the growth function concept.

Example 10.1

Consider the same conditions as in Example 9.2 with $g(x,\alpha) = \{b;\ \alpha=b \in [0, 1]\}$. Furthermore, consider that $Z = Y_X X = \{0,1\}_X [0,1]$, i.e., the observed y values to be approximated by $g(x,\alpha)$ in [0,1] only have two values, 0 or 1. Then, the $\{q(\alpha)\}$ set is always a main diagonal of the $[0,1]^n$ hypercube, with length $\sqrt{n}$. On the other hand, the ε-hypercube diagonal has length $2\varepsilon\sqrt{n}$. Thus:

$$G(\varepsilon;n) = \ln\left\{\sup_{Z_n} N(\varepsilon;Z_n)\right\} = \ln\frac{\sqrt{n}}{2\varepsilon\sqrt{n}} = -\ln(2\varepsilon),$$

and condition 10.1 is satisfied. □

11 Bounds on the Rate of Convergence

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

11.1 VC-Dimension of a Set of Real-Valued Functions

Definitions:

1 - The *set of indicators* for the real-valued function $Q(z,\alpha^*)$ is:

$$I\left(Q(z,\alpha^*)-\beta\right)=\theta\left(Q(z,\alpha^*)-\beta\right) \text{ with } \beta\in\left[\inf_z Q(z,\alpha^*), \sup_z Q(z,\alpha^*)\right]$$

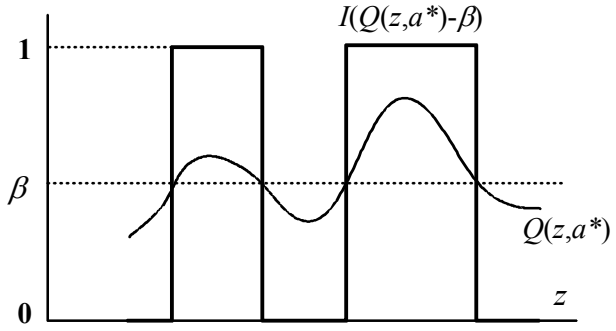


Figure 11.1

Figure 11.1 illustrates one of the indicator functions of the set of indicators of $Q(z,\alpha)$.

2 – The set of indicators of $Q(z,\alpha)$, for any $\alpha\in$A, is called the *complete set of indicators* of the family $Q(z,\alpha)$.

These two definitions allow to apply the same concepts already presented in Part I, as follows:

3 - Let $N(Z_n) \equiv N^{A,\beta}(z_1, \ldots, z_n)$ be the number of *different separations* of Z_n by a complete set of indicators of the family $Q(z,\alpha)$. As in Part I we similarly define:

4 – *Annealed entropy of the set of indicators of real-valued functions*:

$$H_{\text{ann}}^{A,\beta}(n) \equiv H_{\text{ann}}(n) = \ln \mathrm{E}\left[N(Z_n)\right]$$

5 – *Growth function of the set of indicators of real-valued functions*:

$$G^{A,\beta}(n) \equiv G(n) = \ln \max_{Z_n} N(Z_n)$$

As in the case of indicator functions presented in Part I, these two quantities are related as:

$$H_{\text{ann}}(n) \le G(n) \le h\left(\ln\frac{n}{h}+1\right),$$

where h is:

6 – *VC Dimension of a set of real-valued functions*: maximal number h of vectors z_1, …, z_n, that can be shattered by the complete set of indicators of $Q(z,\alpha)$.

Example 11.1

The VC dimension of a set of functions that are linear in their parameters:

$$f(z,\alpha) = \sum_{i=1}^{d} \alpha_i \phi_i(z) + \alpha_0 \ ,$$

equals d+1, the number of parameters. The proof is based on the result of Example 7.2.

Remark

Note that, as we saw already in section 7.1, the VC-dimension is defined in terms of a family of loss functions $Q(z,\alpha)$. In the case of 2-class data classification, the VC-dimension of the loss function equals the VC-dimension of the set of approximating functions $\phi(x,\alpha)$.

In data regression with a quadratic loss function, we have:

$$Q(z,\alpha) = L(y,\phi(x,\alpha)) = (y-\phi(x,\alpha))^2 \ .$$

Let h_f denote the VC-dimension of the set $\phi(x,\alpha)$. Then it can be shown (Vapnik, 1995) that the VC-dimension h of the set of real functions $Q(z,\alpha) = (y-\phi(x,\alpha))^2$ is bounded as:

$$h_f \le h \le \mathrm{c}\, h_f \,,$$

where c is some universal constant. According to Vapnik (cited in Cherkassky V, Mulier F, 1998), for practical applications one can use $h \approx h_f$.

11.2 Distribution Independent Bounds for Convergence

The following theorems are similar to the ones presented in section 7.4 and apply to a family of non-negative functions:

$$0 \le Q(z,\alpha) \le B, \quad \alpha \in \mathrm{A},\ B \in \Re^+$$

Theorem (Vapnik)

With probability at least 1-δ simultaneously for all functions in a set of non-negative real-valued functions the following inequality holds true:

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

$$\text{with} \quad \varepsilon(n) = 4\frac{h\left(\ln\frac{2n}{h}+1\right)-\ln(\delta/4)}{n}. \qquad 11.1$$

Corolary: with probability at least 1-2δ the following inequality holds true:

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

Example 11.2

Consider a set of linear functions $\phi(x,\alpha) = ax + b$, with VC-dimension $h = 2$ (see Example 10.1). Assume we use a significance level $\delta = 4/\sqrt{n}$, as recommended by (Vapnik, 1998) and also used by Cherkassky V, Mulier F, 1998). Assume further that the training error is $R_{\text{emp}}(\alpha) = 0.07$ and $B = 1$.
Figure 11.2 illustrates the behaviour of the $R - R_{\text{emp}}$ with n. (Compare with Figure 7.10.)

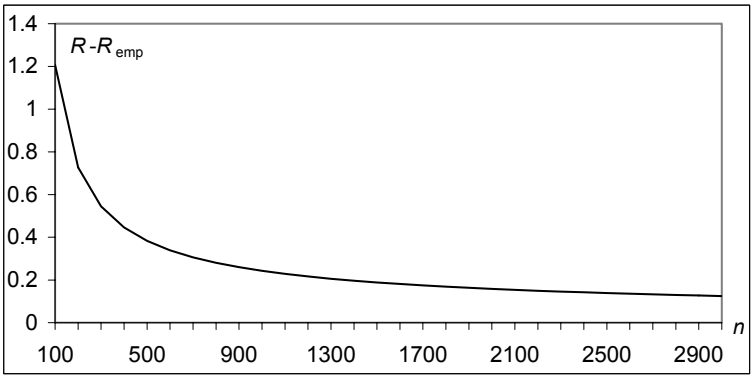


Figure 11.2

If now we fix $m = 10000$ and $\delta = 0.05$ and vary $h = 2,\ldots,18$ we can see that the expected risk increases with the VC-dimension (Figure 11.3) almost linearly.

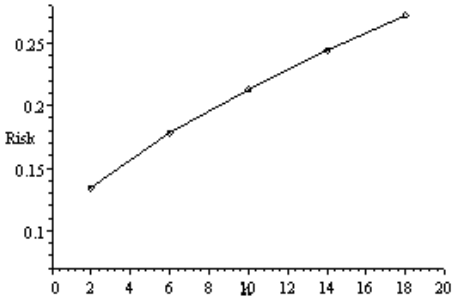


Figure 11.3

Tighter bounds can be derived for sets of *unbounded* nonnegative functions. As described in (Vapnik, 1998) for unbounded non negative real-valued loss functions, when $F(z)$ is a distribution with light tails, then with probability at least $1-\delta$, simultaneously for all loss functions in the set, the following distribution-independent bound holds:

$$R(\alpha) \le \left(\frac{R_{\text{emp}}(\alpha)}{1 - c\sqrt{\varepsilon(n)}} \right)_{\infty}^{3}. \qquad 11.2$$

In most practical problems, one may take $c = 1$.

Example 11.3

Figure 11.4 shows the values of the risk computed with formulas 11.1 (dotted curve) and 11.2 (solid curve) in the conditions of the prvious example. It is clear that formula 11.2 provides a much tighter bound.

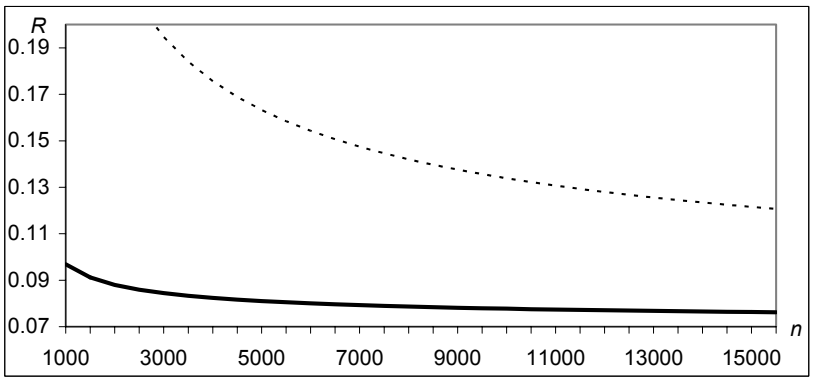


Figure 11.4

Note that in regression the loss functions should be considered as being unbounded and nonnegative, since usually we cannot provide finite bounds for mean squared error and the bounds on the true function or the additive noise are not known. There is

[3] The subscripted ∞ means that $R(\alpha) < \infty$ if the denominator turns out to be negative.

always a small probability of observing very large output values that can yield large values for the loss function as well. Formula 11.2 (Vapnik, 1998) considers distributions with so called "light tails", i.e., small probabilities of observing large values.

12 Pseudo- and Fat-Shattering Dimensions

Let us first consider the following example.

Example 12.1

Let $i = 1,\ldots,7$ and consider the following real vector $\mathbf{z} \in \{X_x Y\}^7 = \{[0,1]_x[0,1]\}^7$:

$$\mathbf{z}' = \left[(x_1, y_1)\ldots(x_7, y_7)\right]' = \begin{bmatrix} 0.167 & 0.093 \\ 0.278 & 0.317 \\ 0.389 & 0.330 \\ 0.500 & 0.517 \\ 0.611 & 0.633 \\ 0.722 & 0.697 \\ 0.833 & 0.767 \end{bmatrix}$$

Consider the following three linear functions (see Figure 12.1), belonging to a set F of linear functions:

$f_{\text{dash}} = 2.018\,x - 0.244$	dash line
$f_{\text{dot}} = 1.068\,x - 0.085$	dot line
$f_{\text{dash-dot}} = 0.117\,x + 0.284$	dash-dot line

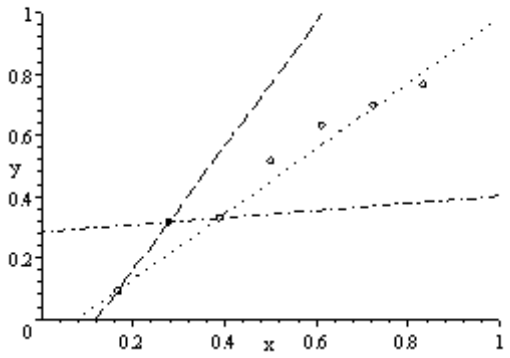


Figure 12.1

At each point $x_i \in \{0.167, 0.278, 0.389, 0.5, 0.611, 0.722, 0.833\}$ and for each function:

$$f \in \left\{ f_{dash},\ f_{dot},\ f_{dash-dot} \right\} \subset F$$

the graph of $f(x_i)$ can either pass above or through y_i, or else below y_i . We then can denote the set

$$\left\{ f_{dash},\ f_{dot},\ f_{dash-dot} \right\} \text{ by } \left\{ f_{[1100000]},\ f_{[1111110]},\ f_{[0111111]} \right\}$$

E.g., $f_{[1100000]}$ means that:

$f_{[1100000]}(x_1) \geq y_1$

$f_{[1100000]}(x_2) \geq y_2$

$f_{[1100000]}(x_i) < y_i,\ for\ i > 2$

For a given class F of linear functions it may be possible to obtain all 2^7 binary sequences (above/below).

□

We generalize the VC-dimensionof a class of functions $\mathcal{F}$, to the pseudo-dimension, denoted P-dim($\mathcal{F}$), in order to study the learnability of [0,1]-valued functions.

Definition:

Given $\mathcal{F}=\{X\rightarrow[0,1]\}$ *and a set* $S=\{x_1,\ldots,x_n\}$. *The set S is P-shattered by* $\mathcal{F}$, *with* $\mathbf{c}\in[0,1]^n$ *as the witness, if for every binary vector* $e\in\{0,1\}^n$ *there exists a function* $f_e\in\mathcal{F}$ *such that*:

$$f_e(x_i)\begin{cases} \geq c_i & if\ \ e_i = 1 \\ \leq c_i & if\ \ e_i = 0 \end{cases}$$

The P-dim($\mathcal{F}$) is the largest integer n for which there exists a set of cardinality n that is P-shattered by $\mathcal{F}$. Thus, the only extra feature of the P-dim($\mathcal{F}$) is the possibility of introducing the "*off-set*" vector $\mathbf{c}\in[0,1]^n$.

Example 12.2

(Function class with witness)

Given a class of (measurable) linear functions $\mathcal{F}=\{X\rightarrow[0,1]\}$ such that:

$f_e(x) = ax + b$

and a set S:

$S = \left\{ 0.167,\ \ 0.278,\ \ 0.389,\ \right\},$

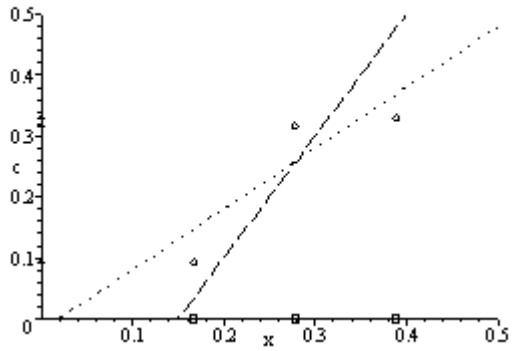
suppose we have the following pairs of real numbers, the first one, x_i being a member of the set S and the second one, c_i, being a *witness*, and that $(x_i, c_i) \in [0,1] \times [0,1]$, where $i = 1, \ldots, 3$:

i	x_i	c_i
1	0.167	0.093
2	0.278	0.317
3	0.389	0.330

We can say that the graph of the following instances of functions f_{dash} and f_{dot} that belong to the class $\mathcal{F}$, and are represented in Figure12.2, can either pass above or through c_i, or else below c_i

$f_{dash} = 2x - 0.3$ *dash line*

$f_{dot} = x - 0.02$ *dot line*



$f_{dash} = 2x - 0.3$ and $f_{dot} = x - 0.02$, points $x_i \in S$ are represented by boxes, and witness c_i by diamonds.

Figure 12.2

we can say that the graph of $f_e(x)$ can either pass *above or through* c_i, or else *below* c_i.
We can denote the set $\{f_{dash}, f_{dot}\}$ by:

$$\{ f_{[001]}, f_{[010]}, \}$$

where, e.g., $f_{[001]}$(the dash line in figure), $e = [e_1, e_2, e_3] = [001]$, and the meaning is that this f_e passes *above (or through)* c_3, therefore:

$e_3 = 1$, *because* $f_e(x_3) \geq c_3$

and f_e passes *below* c_1 and c_2 therefore:

$e_1 = 0$, *because* $f_e(x_1) < c_1$

$e_2 = 0$, *because* $f_e(x_2) < c_1$

Therefore as card(S) = 3 we can say that there are 2^3 different possible behaviors as f varies over the all possible linear functions $f_e \in F$, where $e \in \{0,1\}^3$

□

Lemma ([Vidyasagar03])

Given a collection of functions $\mathcal{F}$ mapping X onto [0,1] define an associated collection of functions F as follows: For each $f : X \to [0,1]$ define a corresponding $f : X \times [0,1] \to \{0,1\}$, using the Heaviside function $\theta(x)$ by:

$$f(x,c) = \theta(f(x) - c)$$

Let $F = \{f; f \in \mathcal{F}\}$ then:

$$P\text{-}dim(\mathcal{F}) = VC\text{-}dim(F).$$

Definition

We define the notion called *fat-shattering dimension*, denoted F-dim and usually referred to as a "*scale-sensitive*" version of the P-dim (see e.g. Vidyasagar, 2003):

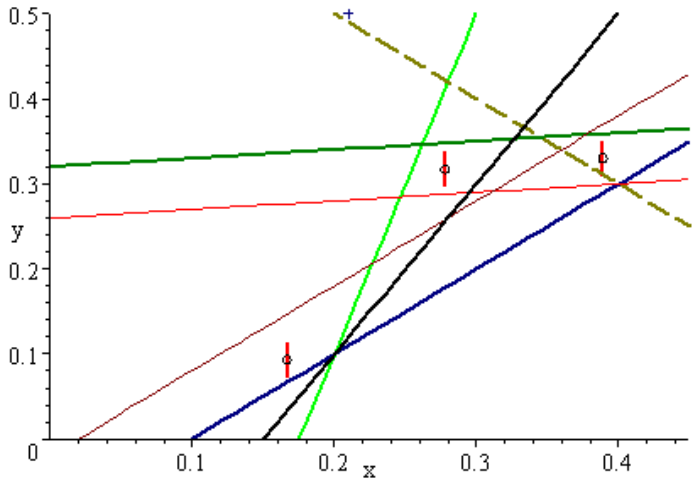
Given $\mathcal{F} = \{X \to [0,1]\}$ *and a set* $S = \{x_1, \ldots, x_n\}$. *The set* S *is fat-shattered by* $\mathcal{F}$ *to width* $\gamma > 0$ *with* $c \in [0,1]^n$ *as witness if, for every binary vector* $e \in \{0,1\}^n$, *there exists a function* $f_e \in \mathcal{F}$ *such that*

$$f_e(x_i) \begin{cases} \geq c_i + \gamma & \text{if } e_i = 1, \\ \leq c_i - \gamma & \text{if } e_i = 0. \end{cases} \qquad 12.1$$

The F-dim *of* $\mathcal{F}$ *to width* γ, *denoted by* F-dim($\mathcal{F}$,γ) *is the largest integer n for which there exists a set of cardinality n that is P-shattered by* $\mathcal{F}$.

Example 12.3

Given $\mathcal{F} = \{X \to [0,1]\}$ and a set S = {0.167, 0.278, 0.389} as stated in Example 12.2, we show how the set S is fat-shattered by 7 functions to width γ = 0.02 with the set {0.093, 0.317, 0.33} as witness, respectively. For every binary vector $e \in \{0,1\}^3$ with the exception of the vector [010] there exists a function $f_e \in \mathcal{F}$ such that expression 12.1 is verified.



The set $S = \{0.167, 0.278, 0.389\}$ is fat-shattered to width $\gamma > 0.02$ with witness $\{0.093, 0.317, 0.33\}$. Note f_{010} is missing.

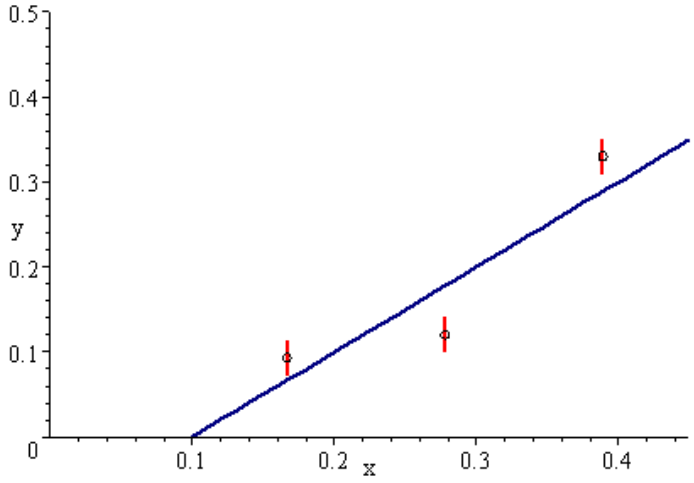
Figure 12.3

Figure 12.3 shows clockwise, and starting from the crossed end of the dash line, the following examples of linear functions $f_e \in F$:

$$F = \{f_{[110]}, f_{[011]}, f_{[001]}, f_{[101]}, f_{[111]}, f_{[000]}, f_{[100]}\}$$

Thus:

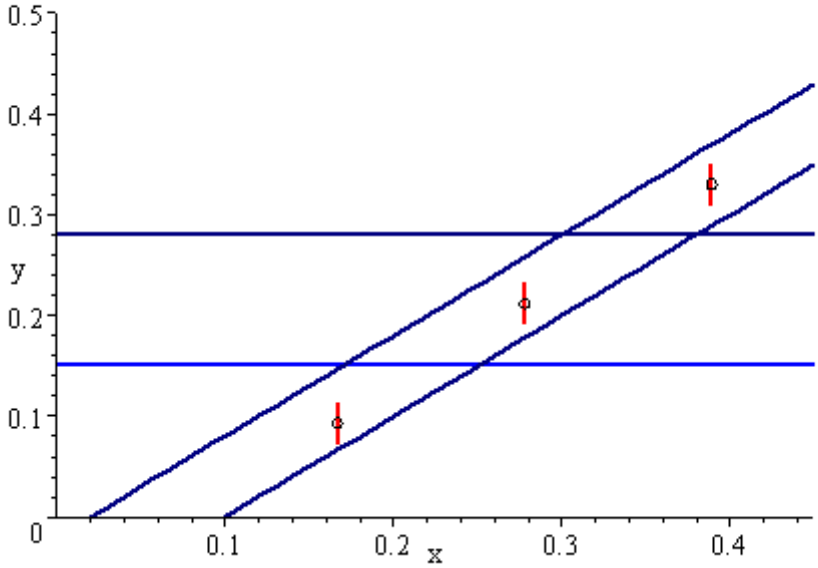
$$F\text{-}dim(\mathcal{F}) = VC\text{-}dim(F) \leq 2^7 = 128$$



The set $S = \{0.167, 0.278, 0.389\}$ is fat-shattered $\gamma > 0.02$ with witness $\{0.093, 0.12, 0.33\}$. Note f_{010} exists but f_{101} is missing.

Example 12.4

(The set S is fat-shattered by 4 functions)



The set $S = \{0.167, 0.278, 0.389\}$ is fat-shattered $\gamma > 0.02$ with with linearely dependent witeness $\{0.093, 0.2115, 0.33\}$ by only 4 functions.

Figure 12.4

Theorem (see Anthony, 1999)

Let $\mathcal{F} = \{[0,1] \rightarrow [0,1]\}$ *be the set of all functions mapping from the interval* [0,1] *to the interval* [0,1] *and having total variation at most* V. *Then:*

$$F\text{-}dim(\mathcal{F},\gamma) = 1 + \left\lfloor \frac{V}{2\gamma} \right\rfloor$$

where

$$V \geq \sum_{i=1}^{m} |f(y_{i+1}) - f(y_i)|$$

The following theorem (see Anthony, 1999), gives a lower bound on the sample complexity $m(\varepsilon,\delta,B)$ of any learning algorithm in terms of the fat-shattering dimension of a function class.

Theorem:

Suppose that $\mathcal{F} = \{X \rightarrow [0,1]\}$. *Then for* $B \geq 2$ *and* $0 < \varepsilon < 1$, $0 < \eta < 0.01$, *any learning algorithm for any function class* $\mathcal{F}$ *has sample complexity satisfying:*

$$m(\varepsilon,\delta,B)=\frac{F-\dim(\mathcal{F},\varepsilon/\alpha)}{16\alpha}$$

for any $0 < \alpha < 0.25$, w*here B is a bound on* $|\hat{y}-y|$ *in real prediction problem (p.233),* ε *is the estimation error of the algorithm,* η *is confidence level, and* α *is an unspecified parameter.*

Example 12.5

$$m = 7$$

$$z = [z_i]^T = [(x_1,y_1),\ldots,(x_7,y_7)]^T$$

$$= \begin{bmatrix} 0.167 & 0.093 \\ 0.278 & 0.317 \\ 0.389 & 0.330 \\ 0.500 & 0.517 \\ 0.611 & 0.633 \\ 0.722 & 0.697 \\ 0.833 & 0.767 \end{bmatrix}$$

Example 12.6

$$m = 280$$

$$\varepsilon = 0.005$$

$$\alpha = 0.010$$

$$\gamma = \frac{\varepsilon}{\alpha} = 0.500$$

V	$F\text{-}dim(\mathcal{F},\gamma)$	m_{LowB}
28	29	175
30	31	188
32	33	200
34	35	213
36	37	225

Appendix – Regression Solutions

1 - Let us first assume the quadratic loss function 8.2. We have:

$$R(\alpha) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y - g(x,\alpha))^2 dF(y,x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y - g(x,\alpha))^2 f(y,x)dydx \text{ ;}^4$$

i.e., $R(\alpha)$ is simply the expectation of the square deviations $\mathrm{E}[(y - g(x,\alpha))^2]$. But:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y - g(x,\alpha))^2 f(y,x)dydx = \int_{-\infty}^{\infty} f(x)\int_{-\infty}^{\infty}(y - g(x,\alpha))^2 f(y \mid x)dydx$$

The integrand above is nonnegative; therefore, minimizing $R(\alpha)$ amounts to minimizing the following:

$$R(\alpha) = \int_{-\infty}^{\infty}(y - g(x,\alpha))^2 f(y \mid x)dy = \mathrm{E}\left[y^2 \mid x\right] - 2g(x,\alpha)\mathrm{E}\left[y \mid x\right] + g(x,\alpha)^2$$

For every particular (x,α) the integral is a second-order moment relative to the *constant* $g(x,\alpha)$. It reaches a minimum for the particular value $g(x, \alpha_0)$ such that:

$$\frac{\partial R(g(x,\alpha))}{\partial g(x,\alpha)} = 0 \quad \Rightarrow \quad -2\mathrm{E}\left[y \mid x\right] + 2g(x,\alpha) = 0 \quad \Rightarrow$$

$$g(x,\alpha_0) = \mathrm{E}\left[y \mid x\right] = \int_{-\infty}^{\infty} y f(y \mid x)dydx$$

2 – Let us now use the following loss function:

$$Q(z,\alpha) = L(y, g(x,\alpha)) = \mid y - g(x,\alpha) \mid$$

we now have:

$$R(\alpha) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mid y - g(x,\alpha) \mid f(y,x)dydx = \int_{-\infty}^{\infty} f(x)\int_{-\infty}^{\infty} \mid y - g(x,\alpha) \mid f(y \mid x)dydx$$

Thus, we have to minimize:

$$R(\alpha) = \int_{-\infty}^{\infty} \mid y - g(x,\alpha) \mid f(y \mid x)dy =$$

$$\int_{-\infty}^{g(x,\alpha)} (g(x,\alpha) - y) f(y \mid x)dy + \int_{g(x,\alpha)}^{+\infty} (y - g(x,\alpha)) f(y \mid x)dy$$

For uncluttered notation let us denote the constant $g(x,\alpha)$ as a. The above expression is then developed as:

[4] We assume that the family of functions $g(x,\alpha)$, $\alpha \in \mathrm{A}$ are square integrable.

$$R(\alpha) = a\left[\int_{-\infty}^{a} f(y \mid x)dy - \int_{a}^{+\infty} f(y \mid x)dy\right] + \left[\int_{a}^{+\infty} yf(y \mid x)dy - \int_{-\infty}^{a} yf(y \mid x)dy\right]$$

Using the Fundamental Theorem of Calculus, we obtain:

$$\frac{\partial R(\alpha)}{\partial a} = 0 \quad \Rightarrow$$

$$\left[\int_{-\infty}^{a} f(y \mid x)dy - \int_{a}^{+\infty} f(y \mid x)dy\right] + a\left[f(a \mid x) - f(a \mid x)\right] + \left[af(a \mid x) - af(a \mid x)\right] = 0$$

Thus:

$$\frac{\partial R(\alpha)}{\partial a} = 0 \Rightarrow \left[\int_{-\infty}^{a} f(y \mid x)dy - \int_{a}^{+\infty} f(y \mid x)dy\right] = 0 \;\Rightarrow\; a_0 = g(x, \alpha_0) = \text{median}\left[f(y \mid x)\right]$$

References

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

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

Kecman V (2001). Learning and Soft Computing, Support Vector Machines, Neural Networks and Fuzzy Logic Models. MIT Press, Boston, 2001.

Scholkopf B, Smola A (2002). Learning with Kernels}. MIT Press, Cambridge MA.

Vapnik VN (1995) The Nature of Statistitical Learning. Springer Verlag.

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

Vapnik VN (1999). An overview of statistical learning theory. IEEE Trans. on Neural Networks}, 10 (5): 988-999, 1999.

Papoulis A (1965) Probability, Random Variables and Stochastic Processes, McGraw-Hill Book Co.