Statistics for Engineering and Information Science

Akaike and Kitagawa: The Practice of Time Series Analysis.
Cowell, Dawid, Lauritzen, and Spiegelhalter: Probabilistic Networks and Expert Systems.
Doucet, de Freitas, and Gordon: Sequential Monte Carlo Methods in Practice.
Fine: Feedforward Neural Network Methodology.
Hawkins and Olwell: Cumulative Sum Charts and Charting for Quality Improvement.
Jensen: Bayesian Networks and Decision Graphs.
Marchette: Computer Intrusion Detection and Network Monitoring: A Statistical Viewpoint.
Rubinstein and Kroese: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation, and Machine Learning.
Studený: Probabilistic Conditional Independence Structures.
Vapnik: The Nature of Statistical Learning Theory, Second Edition.
Wallace: Statistical and Inductive Inference by Minimum Massage Length.

Vladimir N. Vapnik

The Nature of Statistical Learning Theory

Second Edition

With 50 Illustrations

Springer

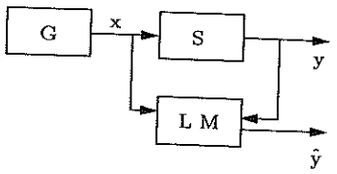

FIGURE 1.1. A model of learning from examples. During the learning process, the learning machine observes the pairs (x, y) (the training set). After training, the machine must on any given x return a value $\bar{y}$. The goal is to return a value $\bar{y}$ that is close to the supervisor's response y.

The selection of the desired function is based on a training set of ℓ independent and identically distributed (i.i.d.) observations drawn according to $F(x, y) = F(x)F(y|x)$:

$$(x_1, y_1), \ldots, (x_\ell, y_\ell). \tag{1.1}$$

1.2 THE PROBLEM OF RISK MINIMIZATION

In order to choose the best available approximation to the supervisor's response, one measures the *loss*, or discrepancy, $L(y, f(x, \alpha))$ between the response y of the supervisor to a given input x and the response $f(x, \alpha)$ provided by the learning machine. Consider the expected value of the loss, given by the *risk functional*

$$R(\alpha) = \int L(y, f(x, \alpha)) dF(x, y). \tag{1.2}$$

The goal is to find the function $f(x, \alpha_0)$ that minimizes the risk functional $R(\alpha)$ (over the class of functions $f(x, \alpha)$, $\alpha \in \Lambda$) in the situation where the joint probability distribution function $F(x, y)$ is unknown and the only available information is contained in the training set (1.1).

1.3 THREE MAIN LEARNING PROBLEMS

This formulation of the learning problem is rather broad. It encompasses many specific problems. Consider the main ones: the problems of pattern recognition, regression estimation, and density estimation.

1.3.1 Pattern Recognition

Let the supervisor's output y take only two values $y = \{0, 1\}$ and let $f(x, \alpha)$, $\alpha \in \Lambda$, be a set of *indicator* functions (functions which take only two values: zero and one). Consider the following loss function:

$$L(y, f(x, \alpha)) = \begin{cases} 0 & \text{if } y = f(x, \alpha), \\ 1 & \text{if } y \neq f(x, \alpha). \end{cases} \tag{1.3}$$

For this loss function, the functional (1.2) determines the probability of different answers given by the supervisor and by the indicator function $f(x, \alpha)$. We call the case of different answers a *classification error*.

The problem, therefore, is to find a function that minimizes the probability of classification error when the probability measure $F(x, y)$ is unknown, but the data (1.1) are given.

1.3.2 Regression Estimation

Let the supervisor's answer y be a real value, and let $f(x, \alpha), \alpha \in \Lambda$, be a set of real functions that contains the *regression function*

$$f(x, \alpha_0) = \int y \, dF(y|x).$$

It is known that the regression function is the one that minimizes the functional (1.2) with the following loss function:[3]

$$L(y, f(x, \alpha)) = (y - f(x, \alpha))^2. \tag{1.4}$$

Thus the problem of regression estimation is the problem of minimizing the risk functional (1.2) with the loss function (1.4) in the situation where the probability measure $F(x, y)$ is unknown but the data (1.1) are given.

1.3.3 Density Estimation (Fisher–Wald Setting)

Finally, consider the problem of density estimation from the set of densities $p(x, \alpha), \alpha \in \Lambda$. For this problem we consider the following loss function:

$$L(p(x, \alpha)) = -\log p(x, \alpha). \tag{1.5}$$

[3] If the regression function $f(x)$ does not belong to $f(x, \alpha), \alpha \in \Lambda$, then the function $f(x, \alpha_0)$ minimizing the functional (1.2) with loss function (1.4) is the closest to the regression in the metric $L_2(F)$:

$$\rho(f(x), f(x, \alpha_0)) = \sqrt{\int (f(x) - f(x, \alpha_0))^2 dF(x)}.$$

It is known that the desired density minimizes the risk functional (1.2) with the loss function (1.5). Thus, again, to estimate the density from the data one has to minimize the risk functional under the condition that the corresponding probability measure $F(x)$ is unknown, but i.i.d. data

$$x_1, \ldots, x_n$$

are given.

1.4 THE GENERAL SETTING OF THE LEARNING PROBLEM

The general setting of the learning problem can be described as follows. Let the probability measure $F(z)$ be defined on the space Z. Consider the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$. The goal is to minimize the risk functional

$$R(\alpha) = \int Q(z, \alpha) dF(z), \quad \alpha \in \Lambda, \tag{1.6}$$

where the probability measure $F(z)$ is unknown, but an i.i.d. sample

$$z_1, \ldots, z_\ell \tag{1.7}$$

is given.

The learning problems considered above are particular cases of this general problem of *minimizing the risk functional* (1.6) *on the basis of empirical data* (1.7), where z describes a pair (x, y) and $Q(z, \alpha)$ is the specific loss function (e.g., one of (1.3), (1.4), or (1.5)). In the following we will describe the results obtained for the general statement of the problem. To apply them to specific problems, one has to substitute the corresponding loss functions in the formulas obtained.

1.5 THE EMPIRICAL RISK MINIMIZATION (ERM) INDUCTIVE PRINCIPLE

In order to minimize the risk functional (1.6) with an unknown distribution function $F(z)$, the following inductive principle can be applied:

(i) The risk functional $R(\alpha)$ is replaced by the so-called *empirical risk functional*

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \tag{1.8}$$

constructed on the basis of the training set (1.7).

(ii) One approximates the function $Q(z, \alpha_0)$ that minimizes risk (1.6) by the function $Q(z, \alpha_\ell)$ minimizing the empirical risk (1.8).

This principle is called the *empirical risk minimization* inductive principle (ERM principle).

We say that an inductive principle defines a *learning process* if for any given set of observations the learning machine chooses the approximation using this inductive principle. In learning theory the ERM principle plays a crucial role.

The ERM principle is quite general. The classical methods for the solution of a specific learning problem, such as the least-squares method in the problem of regression estimation or the maximum likelihood (ML) method in the problem of density estimation, are realizations of the ERM principle for the specific loss functions considered above.

Indeed, by substituting the specific loss function (1.4) in (1.8) one obtains the functional to be minimized

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

which forms the least-squares method, while by substituting the specific loss function (1.5) in (1.8) one obtains the functional to be minimized

$$R_{\text{emp}}(\alpha) = -\frac{1}{\ell} \sum_{i=1}^{\ell} \ln p(x_i, \alpha).$$

Minimizing this functional is equivalent to the ML method (the latter uses a plus sign on the right-hand side).

1.6 THE FOUR PARTS OF LEARNING THEORY

Learning theory has to address the following four questions:

(i) *What are (necessary and sufficient) conditions for consistency of a learning process based on the ERM principle?*

(ii) *How fast is the rate of convergence of the learning process?*

(iii) *How can one control the rate of convergence (the generalization ability) of the learning process?*

(iv) *How can one construct algorithms that can control the generalization ability?*

The answers to these questions form the four parts of learning theory:

(i) Theory of consistency of learning processes.

(ii) Nonasymptotic theory of the rate of convergence of learning processes.

(iii) Theory of controlling the generalization ability of learning processes.

(iv) Theory of constructing learning algorithms.

Each of these four parts will be discussed in the following chapters.

Informal Reasoning and Comments — 1

The setting of learning problems given in Chapter 1 reflects two major requirements:

(i) To estimate the desired function from a wide set of functions.

(ii) To estimate the desired function on the basis of a limited number of examples.

The methods developed in the framework of the classical paradigm (created in the 1920s and 1930s) did not take into account these requirements. Therefore, in the 1960s considerable effort was put into both the generalization of classical results for wider sets of functions and the improvement of existing techniques of statistical inference for small sample sizes. In the following we will describe some of these efforts.

1.7 THE CLASSICAL PARADIGM OF SOLVING LEARNING PROBLEMS

In the framework of the classical paradigm all models of function estimation are based on the maximum likelihood method. It forms an inductive engine in the classical paradigm.

Chapter 4

Controlling the Generalization Ability of Learning Processes

The theory for controlling the generalization ability of learning machines is devoted to constructing an inductive principle for minimizing the risk functional using a *small sample* of training instances.

The sample size ℓ is considered to be small if the ratio ℓ/h (ratio of the number of training patterns to the VC dimension of functions of a learning machine) is small, say $\ell/h < 20$.

To construct small sample size methods we use both the bounds for the generalization ability of learning machines with sets of totally bounded nonnegative functions,

$$R(\alpha_\ell) \le R_{\text{emp}}(\alpha_\ell) + \frac{B\mathcal{E}}{2}\left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{B\mathcal{E}}}\right), \tag{4.1}$$

and the bounds for the generalization ability of learning machines with sets of unbounded functions,

$$R(\alpha_\ell) \le \frac{R_{\text{emp}}(\alpha_\ell)}{\left(1 - a(p)\tau\sqrt{\mathcal{E}}\right)_+}, \tag{4.2}$$

$$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}},$$

where

$$\mathcal{E} = 2\frac{\ln N - \ln \eta}{\ell}$$

if the set of functions $Q(z, \alpha_i)$, $1, \ldots, N$, contains N elements, and

$$\mathcal{E} = 4\frac{h\left(\ln\frac{2\ell}{h} + 1\right) - \ln(\eta/4)}{\ell} \qquad (3.2.4)$$

if the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, contains an infinite number of elements and has a finite VC dimension h. Each bound is valid with probability at least $1 - \eta$.

4.1 STRUCTURAL RISK MINIMIZATION (SRM) INDUCTIVE PRINCIPLE

The ERM principle is intended for dealing with large sample sizes. It can be justified by considering the inequality (4.1) or the inequality (4.2).

When ℓ/h is large, $\mathcal{E}$ is small. Therefore, the second summand on the right-hand side of inequality (4.1) (the second summand in the denominator of (4.2)) becomes small. The actual risk is then close to the value of the empirical risk. In this case, a small value of the empirical risk guarantees a small value of the (expected) risk.

However, if ℓ/h is small, a small $R_{\text{emp}}(\alpha_\ell)$ does not guarantee a small value of the actual risk. In this case, to minimize the actual risk $R(\alpha)$ one has to minimize the right-hand side of inequality (4.1) (or (4.2)) simultaneously over both terms. Note, however, that the first term in inequality (4.1) depends on a specific function of the set of functions, while the second term depends on the VC dimension of the whole set of functions. To minimize the right-hand side of the bound of risk, (4.1) (or (4.2)), simultaneously over both terms, one has to make the VC dimension a *controlling variable.*

The following general principle, which is called the *structural risk minimization* (SRM) inductive principle, is intended to minimize the risk functional with respect to both terms, the empirical risk, and the confidence interval (Vapnik and Chervonenkis, 1974).

Let the set S of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, be provided with a *structure* consisting of nested subsets of functions $S_k = \{Q(z, \alpha),\ \alpha \in \Lambda_k\}$, such that (Fig. 4.1)

$$S_1 \subset S_2 \subset \cdots \subset S_n \cdots, \qquad (4.3)$$

where the elements of the structure satisfy the following two properties:

(i) The VC dimension h_k of each set S_k of functions is finite.[1] Therefore,

$$h_1 \leq h_2 \ldots \leq h_n \ldots .$$

[1]However, the VC dimension of the set S can be infinite.

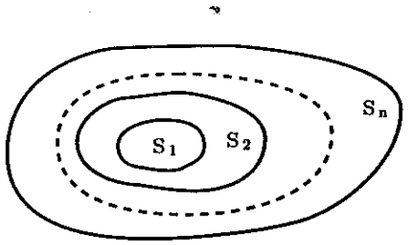

FIGURE 4.1. A structure on the set of functions is determined by the nested subsets of functions.

(ii) Any element S_k of the structure contains either

a set of totally bounded functions,

$$0 \leq Q(z, \alpha) \leq B_k,\ \alpha \in \Lambda_k,$$

or a set of functions satisfying the inequality

$$\sup_{\alpha \in \Lambda_k} \frac{\left(\int Q^p(z, \alpha) dF(z)\right)^{\frac{1}{p}}}{\int Q(z, \alpha) dF(z)} \leq \tau_k, \quad p > 2, \qquad (4.4)$$

for some pair (p, τ_k).

We call this structure an *admissible structure.*

For a given set of observations $z_1, \ldots, z_\ell$ the SRM principle chooses the function $Q(z, \alpha_\ell^k)$ minimizing the empirical risk in the subset S_k for which the guaranteed risk (determined by the right-hand side of inequality (4.1) or by the right-hand side of inequality (4.2) depending on the circumstances) is minimal.

The SRM principle defines a *trade-off between the quality of the approximation of the given data and the complexity of the approximating function.* As the subset index n increases, the minima of the empirical risks decrease. However, the term responsible for the confidence interval (the second summand in inequality (4.1) or the multiplier in inequality (4.2) (Fig. 4.2)) increases. The SRM principle takes both factors into account by choosing the subset S_n for which minimizing the empirical risk yields the best bound on the actual risk.

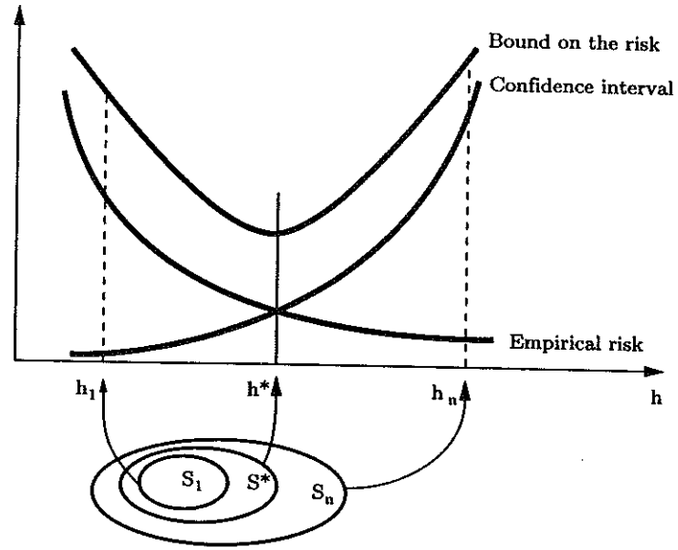

FIGURE 4.2. The bound on the risk is the sum of the empirical risk and the confidence interval. The empirical risk decreases with the index of the element of the structure, while the confidence interval increases. The smallest bound of the risk is achieved on some appropriate element of the structure.

4.2 ASYMPTOTIC ANALYSIS OF THE RATE OF CONVERGENCE

Denote by S^* the set of functions

$$S^* = \bigcup_{k=1}^{\infty} S_k.$$

Suppose that the set of functions S^* is everywhere dense[2] in S (recall $S = \{Q(z,\alpha),\ \alpha \in \Lambda\}$) with respect to the metric

$$\rho(Q(z,\alpha_1), Q(z,\alpha_2)) = \int |Q(z,\alpha_1) - Q(z,\alpha_2)| dF(z).$$

For asymptotic analysis of the SRM principle one considers a law determining, for any given ℓ, the number

$$n = n(\ell) \tag{4.5}$$

of the element S_n of the structure (4.3) in which we will minimize the empirical risk. The following theorem holds true.

Theorem 4.1. *The SRM method provides approximations $Q(z, \alpha_\ell^{n(\ell)})$ for which the sequence of risks $R(\alpha_\ell^{n(\ell)})$ converges to the smallest risk*

$$R(\alpha_0) = \inf_{\alpha \in \Lambda} \int Q(z,\alpha) dF(z)$$

with asymptotic rate of convergence[3]

$$V(\ell) = r_{n(\ell)} + T_{n(\ell)} \sqrt{\frac{h_{n(\ell)} \ln \ell}{\ell}} \tag{4.6}$$

[2]The set of functions $R(z,\beta)$, $\beta \in \mathcal{B}$, is everywhere dense in the set $Q(z,\alpha)$, $\alpha \in \Lambda$, in the metric $\rho(Q,R)$ if for any $\varepsilon > 0$ and for any $Q(z,\alpha^*)$ one can find a function $R(z,\beta^*)$ such that the inequality

$$\rho(Q(z,\alpha^*), R(z,\beta^*)) \le \varepsilon$$

holds true.

[3]We say that the random variables ξ_ℓ, $\ell = 1, 2, \ldots$, converge to the value ξ_0 with asymptotic rate $V(\ell)$ if there exists a constant C such that

$$V^{-1}(\ell)\,|\xi_\ell - \xi_0| \xrightarrow[\ell \to \infty]{P} C.$$

if the law $n = n(\ell)$ is such that

$$\lim_{\ell \to \infty} \frac{T^2_{n(\ell)} h_{n(\ell)} \ln \ell}{\ell} = 0, \tag{4.7}$$

where

(i) $T_n = B_n$ if one considers a structure with totally bounded functions $Q(z, \alpha) \le B_n$ in subsets S_n, and

(ii) $T_n = \tau_n$ if one considers a structure with elements satisfying the equality (4.4);

$r_{n(\ell)}$ is the rate of approximation

$$r_n = \inf_{\alpha \in \Lambda_n} \int Q(z, \alpha) dF(z) - \inf_{\alpha \in \Lambda} \int Q(z, \alpha) dF(z). \tag{4.8}$$

To provide the best rate of convergence one has to know the *rate of approximation r_n for the chosen structure.* The problem of estimating r_n for different structures on sets of functions is the subject of classical function approximation theory. We will discuss this problem in the next section. If one knows the rate of approximation r_n one can *a priori* find the law $n = n(\ell)$ that provides the best asymptotic rate of convergence by minimizing the right-hand side of equality (4.6).

Example. Let $Q(z, \alpha), \alpha \in \Lambda$, be a set of functions satisfying the inequality (4.4) for $p > 2$ with $\tau_k < \tau^* < \infty$. Consider a structure for which $n = h_n$. Let the asymptotic rate of approximation be described by the law

$$r_n = \left(\frac{1}{n}\right)^c.$$

(This law describes the main classical results in approximation theory; see the next section.) Then the asymptotic rate of convergence reaches its maximum value if

$$n(\ell) = \left[\frac{\ell}{\ln \ell}\right]^{\frac{1}{2c+1}},$$

where $[a]$ is the integer part of a. The asymptotic rate of convergence is

$$V(\ell) = \left(\frac{\ln \ell}{\ell}\right)^{\frac{c}{2c+1}}. \tag{4.9}$$

4.3 THE PROBLEM OF FUNCTION APPROXIMATION IN LEARNING THEORY

The attractive properties of the asymptotic theory of the rate of convergence described in Theorem 4.1 are that one can *a priori* (before the learning process begins) find the law $n = n(\ell)$ that provides the best (asymptotic) rate of convergence, and that one can *a priori* estimate the value of the asymptotic rate of convergence.[4] The rate depends on the construction of the admissible structure (on the sequence of pairs $(h_n\ T_n),\ n = 1, 2, ...$) and also depends on the rate of approximation $r_n,\ n = 1, 2, \ldots$.

On the basis on this information one can evaluate the rate of convergence by minimizing (4.6). Note that in equation (4.6), the second term, which is responsible for the stochastic behavior of the learning processes, is determined by nonasymptotic bounds on the risk (see (4.1) and (4.2)). The first term (which describes the deterministic component of the learning processes) usually only has an asymptotic bound, however.

Classical approximation theory studies connections between the smoothness properties of functions and the rate of approximation of the function by the structure with elements S_n containing polynomials (algebraic or trigonometric) of degree n, or expansions in other series with n terms. Usually, smoothness of an unknown function is characterized by the number s of existing derivatives. Typical results of the asymptotic rate of approximation have the form

$$r_n = n^{-\frac{s}{N}}, \tag{4.10}$$

where N is the dimensionality of the input space (Lorentz, 1966). Note that this implies that a high asymptotic rate of convergence[5] in high-dimensional spaces can be guaranteed only for very smooth functions.

In learning theory we would like to find the rate of approximation in the following case:

(i) $Q(z, \alpha),\ \alpha \in \Lambda$, is a set of high-dimensional functions.

(ii) The elements S_k of the structure are not necessarily linear manifolds. (They can be any set of functions with finite VC dimension.)

Furthermore, we are interested in the cases where the rate of approximation is high.

Therefore, in learning theory we face the problem of describing the cases for which a high rate of approximation is possible. This requires describing different sets of "smooth" functions and structures for these sets that provide the bound $O(\frac{1}{\sqrt{n}})$ for r_n (i.e., fast rate of convergence).

[4] Note, however, that a high asymptotic rate of convergence does not necessarily reflect a high rate of convergence on a limited sample size.

[5] Let the rate of convergence be considered high if $r_n \le n^{-1/2}$.

In 1989 Cybenko proved that using a superposition of sigmoid functions (neurons) one can approximate any smooth function (Cybenko, 1989).

In 1992-1993 Jones, Barron, and Breiman described a structure on different sets of functions that has a fast rate of approximation (Jones, 1992), (Barron, 1993), and (Breiman, 1993).

They considered the following concept of smooth functions. Let $\{f(x)\}$ be a set of functions and let $\{\bar{f}(\omega)\}$ be the set of their Fourier transforms.

Let us characterize the smoothness of the function $f(x)$ by the quantity

$$\int |\omega|^d |\bar{f}(\omega)| d\omega = C_d(f) < \infty, \qquad d \geq 0. \tag{4.11}$$

In terms of this concept the following theorem for the rate of approximation r_n holds true:

Theorem 4.2. (Jones, Barron, and Breiman) *Let the set of functions $f(x)$ satisfy (4.11). Then the rate of approximation of the desired functions by the best function of the elements of the structure is bounded by $O(\frac{1}{\sqrt{n}})$ if one of the following holds:*

(i) *The set of functions $\{f(x)\}$ is determined by (4.11) with $d = 0$, and the elements S_n of the structure contain the functions*

$$f(x, \alpha, w, v) = \sum_{i=1}^{n} \alpha_i \sin\left[(x \cdot w_i) + v_i\right], \tag{4.12}$$

where α_i and v_i are arbitrary values and w_i are arbitrary vectors (Jones, 1992).

(ii) *The set of functions $\{f(x)\}$ is determined by equation (4.11) with $d = 1$, and the elements S_n of the structure contain the functions*

$$f(x, \alpha, w, v) = \sum_{i=1}^{n} \alpha_i S\left[(x \cdot w_i) + v_i\right], \tag{4.13}$$

where α_i and v_i are arbitrary values, w_i are arbitrary vectors, and $S(u)$ is a sigmoid function (a monotonically increasing function such that $\lim_{u \to -\infty} S(u) = -1$, $\lim_{u \to \infty} S(u) = 1$) (Barron, 1993).

(iii) *The set of functions $\{f(x)\}$ is determined by (4.11) with $d = 2$, and the elements S_n of the structure contain the functions*

$$f(x, \alpha, w, v) = \sum_{i=1}^{n} \alpha_i |(x \cdot w_i) + v_i|_+, \quad |u|_+ = \max(0,\ u), \tag{4.14}$$

where α_i and v_i are arbitrary values and w_i are arbitrary vectors (Breiman, 1993).

In spite of the fact that in this theorem the concept of smoothness is different from the number of bounded derivatives, one can observe a similar phenomenon here as in the classical case: To keep a high rate of convergence for a space with increasing dimensionality, one has to increase the smoothness of the functions simultaneously as the dimensionality of the space is increased. Using constraint (4.11) one attains it automatically. Girosi and Anzellotti (Girosi and Anzellotti, 1993) observed that the set of functions satisfying (4.11) with $d = 1$ and $d = 2$ can be rewritten as

$$f(x) = \frac{1}{|x|^{n-1}} * \lambda(x), \quad f(x) = \frac{1}{|x|^{n-2}} * \lambda(x),$$

where $\lambda(x)$ is any function whose Fourier transform is integrable, and $*$ stands for the convolution operator. In these forms it becomes more apparent that due to more rapid fall-off of the terms $1/|x|^{n-1}$, functions satisfying (4.11) become more and more constrained as the dimensionality increases.

The same phenomenon is also clear in the results of Mhaskar (Mhaskar, 1992), who proved that the rate of convergence of approximation of functions with s continuous derivatives by the structure (4.13) is $O(n^{-s/N})$.

Therefore, if the desired function is not *very smooth*, one cannot guarantee a high asymptotic rate of convergence of the functions to the unknown function.

In Section 4.5 we describe a new model of learning that is based on the idea of local approximation of the desired function (instead of global, as considered above). We consider the approximation of the desired function in some neighborhood of the point of interest, where the radius of the neighborhood can decrease with increasing number of observations.

The rate of local approximation can be higher than the rate of global approximation, and this effect provides a better generalization ability of the learning machine.

4.4 EXAMPLES OF STRUCTURES FOR NEURAL NETS

The general principle of SRM can be implemented in many different ways. Here we consider three different examples of structures built for the set of functions implemented by a neural network.

1. A structure given by the architecture of the neural network

Consider an ensemble of fully connected feed-forward neural networks in which the number of units in one of the hidden layers is monotonically increased. The sets of implementable functions define a structure as the

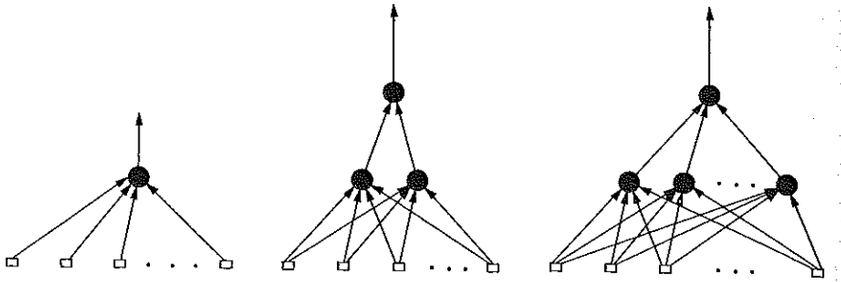

FIGURE 4.3. A structure determined by the number of hidden units.

number of hidden units is increased (Fig. 4.3).

2. A structure given by the learning procedure

Consider the set of functions $S = \{f(x, w),\ w \in W\}$, implementable by a neural net of fixed architecture. The parameters $\{w\}$ are the weights of the neural network. A structure is introduced through $S_p = \{f(x, w),\ ||w|| \leq C_p\}$ and $C_1 < C_2 < \cdots < C_n$. Under very general conditions on the set of loss functions, the minimization of the empirical risk within the element S_p of the structure is achieved through the minimization of

$$E(w, \gamma_p) = \frac{1}{\ell}\sum_{i=1}^{\ell} L(y_i, f(x_i, w)) + \gamma_p ||w||^2$$

with appropriately chosen Lagrange multipliers $\gamma_1 > \gamma_2 > \cdots > \gamma_n$. The well-known "weight decay" procedure refers to the minimization of this functional.

3. A structure given by preprocessing

Consider a neural net with fixed architecture. The input representation is modified by a transformation $z = K(x, \beta)$, where the parameter β controls the degree of degeneracy introduced by this transformation (β could, for instance, be the width of a smoothing kernel).

A structure is introduced in the set of functions $S = \{f(K(x, \beta), w),\ w \in W\}$ through $\beta \geq C_p$, and $C_1 > C_2 > \ldots > C_n$.

To implement the SRM principle using these structures, one has to know (estimate) the VC dimension of any element S_k of the structure, and has to be able for any S_k to find the function that minimizes the empirical risk.

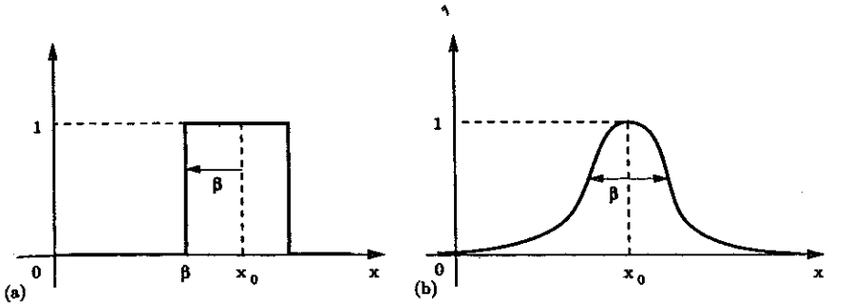

FIGURE 4.4. Examples of vicinity functions: (a) shows a hard-threshold vicinity function and (b) shows a soft-threshold vicinity function.

4.5 THE PROBLEM OF LOCAL FUNCTION ESTIMATION

Let us consider a model of local risk minimization (in the neighborhood of a given point x_0) on the basis of empirical data. Consider a nonnegative function $K(x, x_0; \beta)$ that embodies the concept of neighborhood. This function depends on the point x_0 and a "locality" parameter $\beta \in (0, \infty)$ and satisfies two conditions:

$$0 \leq K(x, x_0; \beta) \leq 1,$$

$$K(x_0, x_0; \beta) = 1. \tag{4.15}$$

For example, both the "hard threshold" vicinity function (Fig. 4.4(a))

$$K_1(x, x_0; \beta) = \begin{cases} 1 & \text{if } ||x - x_0|| < \frac{\beta}{2}, \\ 0 & \text{otherwise,} \end{cases} \tag{4.16}$$

and the "soft threshold" vicinity function (Fig. 4.4(b))

$$K_2(x, x_0; \beta) = \exp\left\{-\frac{(x - x_0)^2}{\beta^2}\right\} \tag{4.17}$$

meet these conditions.

Let us define a value

$$\mathcal{K}(x_0, \beta) = \int K(x, x_0; \beta) dF(x). \tag{4.18}$$

For the set of functions $f(x, \alpha),\ \alpha \in \Lambda$, let us consider the set of loss functions $Q(z, \alpha) = L(y, f(x, \alpha)),\ \alpha \in \Lambda$. Our goal is to minimize the *local*

risk functional

$$R(\alpha,\beta;x_0)=\int L(y,f(x,\alpha))\frac{K(x,x_0;\beta)}{\mathcal{K}(x_0;\beta)}dF(x,y) \tag{4.19}$$

over both the set of functions $f(x,\alpha)$, $\alpha\in\Lambda$, and different vicinities of the point x_0 (defined by parameter β) in situations where the probability measure $F(x,y)$ is unknown, but we are given the independent identically distributed examples

$$(x_1,y_1),\ldots,(x_\ell,y_\ell).$$

Note that the problem of local risk minimization on the basis of empirical data is a generalization of the problem of global risk minimization. (In the last problem we have to minimize the functional (4.19) with $K(x,x_0;\beta)=1$.)

For the problem of local risk minimization one can generalize the bound obtained for the problem of global risk minimization: With probability $1-\eta$ simultaneously for all bounded functions $A\le L(y,f(x,\alpha))\le B$, $\alpha\in\Lambda$, and all functions $0\le K(x,x_0,\beta)\le 1$, $\beta\in(0,\infty)$, the inequality

$$R(\alpha,\beta;x_0)\le\frac{\frac{1}{\ell}\sum_{i=1}^{\ell}L(y_i,f(x_i,\alpha))K(x_i,x_0;\beta)+(B-A)\mathcal{E}(\ell,h_\Sigma)}{\left(\frac{1}{\ell}\sum_{i=1}^{\ell}K(x_i,x_0;\beta)-\mathcal{E}(\ell,h_\beta)\right)_+},$$

$$\mathcal{E}(\ell,h)=\sqrt{\frac{h(\ln(2\ell/h+1)-\ln\eta/2}{\ell}},$$

holds true, where h_Σ is the VC dimension of the set of functions

$$L(y,f(x,\alpha))K(x,x_0;\beta),\ \alpha\in\Lambda,\ \beta\in(0,\infty)$$

and h_β is the VC dimension of the set of functions $K(x,x_0,\beta)$ (Vapnik and Bottou, 1993).

Now using the SRM principle one can minimize the right-hand side of the inequality over three parameters: the value of empirical risk, the VC dimension h_Σ, and the value of the vicinity β (VC dimension h_β).

The local risk minimization approach has an advantage when on the basis of the given structure on the set of functions it is impossible to approximate well the desired function using a given number of observations. However, it may be possible to provide a reasonable *local approximation* to the desired function at any point of interest (Fig. 4.5).

4.6 THE MINIMUM DESCRIPTION LENGTH (MDL) AND SRM PRINCIPLES

Along with the SRM inductive principle, which is based on the statistical analysis of the rate of convergence of empirical processes, there ex-

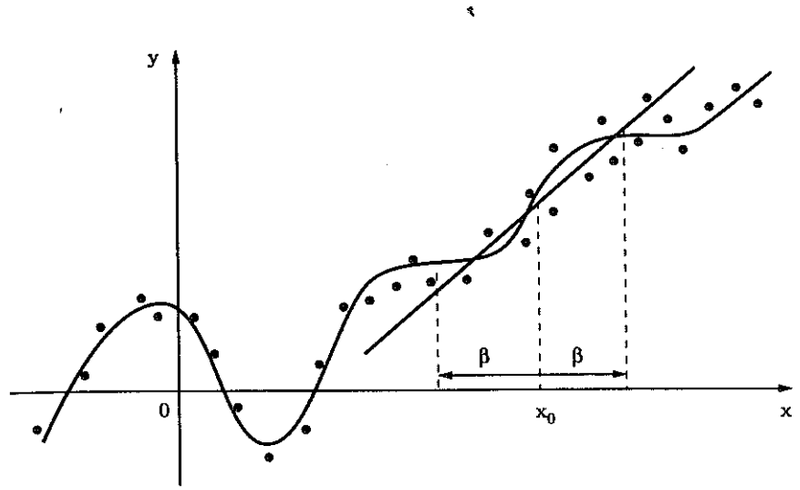

FIGURE 4.5. Using linear functions one can estimate an unknown smooth function in the vicinity of any point of interest.

ists another principle of inductive inference for small sample sizes, the so-called minimum description length (MDL) principle, which is based on an information-theoretic analysis of the randomness concept. In this section we consider the MDL principle and point out the connections between the SRM and the MDL principles for the pattern recognition problem.

In 1965 Kolmogorov defined a random string using the concept of algorithmic complexity.

He defined the algorithmic complexity of an object to be the length of the shortest binary computer program that describes this object, and he proved that the value of the algorithmic complexity, up to an additive constant, does not depend on the type of computer. Therefore, it is a universal characteristic of the object.

The main idea of Kolmogorov is this:

Consider the string describing an object to be random if the algorithmic complexity of the object is high — that is, if the string that describes the object cannot be compressed significantly.

Ten years after the concept of algorithmic complexity was introduced, Rissanen suggested using Kolmogorov's concept as the main tool of inductive inference of learning machines; he suggested the so-called MDL principle[6] (Rissanen, 1978]).

[6]The use of the algorithmic complexity as a general inductive principle

4.6.1 The MDL Principle

Suppose that we are given a training set of pairs

$$(\omega_1, x_1), \ldots, (\omega_\ell, x_\ell)$$

(pairs drawn randomly and independently according to some unknown probability measure). Consider two strings: the binary string

$$\omega_1, \ldots, \omega_\ell \tag{4.20}$$

and the string of vectors

$$x_1, \ldots, x_\ell. \tag{4.21}$$

The question is,

Given (4.21) is the string (4.20) a random object?

To answer this question let us analyze the algorithmic complexity of the string (4.20) in the spirit of Solomonoff–Kolmogorov's ideas. Since the $\omega_1, \ldots, \omega_\ell$ are binary valued, the string (4.20) is described by ℓ bits.

To determine the complexity of this string let us try to compress its description. Since training pairs were drawn randomly and independently, the value ω_i may depend only on vector x_i but not on vector x_j, $i \neq j$ (of course, only if the dependency exists).

Consider the following model: Suppose that we are given some fixed codebook C_b with $N \ll 2^\ell$ different tables T_i, $i = 1, \ldots, N$. Any table T_i describes some function[7] from x to ω.

Let us try to find the table T in the codebook C_b that describes the string (4.20) in the best possible way, namely, the table that on the given string (4.21) returns the binary string

$$\omega_1^*, \ldots, \omega_\ell^* \tag{4.22}$$

for which the Hamming distance between string (4.20) and string (4.22) is minimal (i.e., the number of errors in decoding string (4.20) by this table T is minimal).

Suppose we found a perfect table T_o for which the Hamming distance between the generated string (4.22) and string (4.20) is zero. This table decodes the string (4.20).

---

was considered by Solomonoff even before Kolmogorov suggested his model of randomness. Therefore, the principle of descriptive complexity is called the Solomonoff–Kolmogorov principle. However, only starting with Rissanen's work was this principle considered as a tool for inference in learning theory.

[7] Formally speakng, to get tables of finite length in codebook, the input vector x has to be discrete. However, as we will see, the number of levels in quantization will not affect the bounds on generalization ability. Therefore, one can consider any degree of quantization, even giving tables with an infinite number of entries.

Since the codebook C_b is fixed, to describe the string (4.20) it is sufficient to give the number o of table T_o in the codebook. The minimal number of bits to describe the number of any one of the N tables is $\lceil \lg_2 N \rceil$, where $\lceil A \rceil$ is the minimal integer that is not smaller than A. Therefore, in this case to describe string (4.20) we need $\lceil \lg_2 N \rceil$ (rather than ℓ) bits. Thus using a codebook with a perfect decoding table, we can compress the description length of string (4.20) by a factor

$$K(T_o) = \frac{\lceil \lg_2 N \rceil}{\ell}. \tag{4.23}$$

Let us call $K(T)$ the *coefficient of compression* for the string (4.20).

Consider now the general case: The codebook C_b does not contain the perfect table. Let the smallest Hamming distance between the strings (generated string (4.22) and desired string (4.20)) be $d \geq 0$. Without loss of generality we can assume that $d \leq \ell/2$. (Otherwise, instead of the smallest distance one could look for the largest Hamming distance and during decoding change one to zero and vice versa. This will cost one extra bit in the coding scheme). This means that to describe the string one has to make d corrections to the results given by the chosen table in the codebook.

For fixed d there are C_ℓ^d different possible corrections to the string of length ℓ. To specify one of them (i.e., to specify one of the C_ℓ^d variants) one needs $\lceil \lg_2 C_\ell^d \rceil$ bits.

Therefore, to describe the string (4.20) we need $\lceil \lg_2 N \rceil$ bits to define the number of the table, and $\lceil \lg_2 C_\ell^d \rceil$ bits to describe the corrections. We also need $\lceil \lg_2 d \rceil + \Delta_d$ bits to specify the number of corrections d, where $\Delta_d < 2 \lg_2 \lg_2 d$, $d > 2$. Altogether, we need $\lceil \lg_2 N \rceil + \lceil \lg_2 C_\ell^d \rceil + \lceil \lg_2 d \rceil + \Delta_d$ bits for describing the string (4.20). This number should be compared to ℓ, the number of bits needed to describe the arbitrary binary string (4.20). Therefore, the coefficient of compression is

$$K(T) = \frac{\lceil \lg_2 N \rceil + \lceil \lg_2 C_\ell^d \rceil + \lceil \lceil \lg_2 d \rceil + \Delta_d}{\ell}. \tag{4.24}$$

If the coefficient of compression $K(T)$ is small, then according to the Solomonoff–Kolmogorov idea, the string is not random and somehow depends on the input vectors x. In this case, the decoding table T somehow approximates the unknown functional relation between x and ω.

4.6.2 Bounds for the MDL Principle

The important question is the following:

Does the compression coefficient $K(T)$ determine the probability of test error in classification (decoding) vectors x by the table T?

The answer is yes.

To prove this, let us compare the result obtained for the MDL principle to that obtained for the ERM principle in the simplest model (the learning machine with a finite set of functions).

In the beginning of this section we considered the bound (4.1) for the generalization ability of a learning machine for the pattern recognition problem. For the particular case where the learning machine has a finite number N of functions, we obtained that with probability at least $1-\eta$, the inequality

$$R(T_i) \le R_{\text{emp}}(T_i) + \frac{\ln N - \ln \eta}{\ell}\left(1 + \sqrt{1 + \frac{2R_{\text{emp}}(T_i)\ell}{\ln N - \ln \eta}}\right) \tag{4.25}$$

holds true simultaneously for all N functions in the given set of functions (for all N tables in the given codebook). Let us transform the right-hand side of this inequality using the concept of the compression coefficient, and the fact that

$$R_{\text{emp}}(T_i) = \frac{d}{\ell}.$$

Note that for $d \le \ell/2$ and $\ell > 6$ the inequality

$$R(T_i) \le \frac{d}{\ell} + \frac{\ln N - \ln \eta}{\ell}\left(1 + \sqrt{1 + \frac{2d}{\ln N - \ln \eta}}\right)$$

$$< 2\left(\frac{\lceil \ln N \rceil + \lceil \ln C_\ell^d \rceil + \lceil \lg_2 d \rceil + \Delta_d}{\ell} - \frac{\ln \eta}{\ell}\right) \tag{4.26}$$

is valid (one can easily check it). Now let us rewrite the right-hand side of inequality (4.26) in terms of the compression coefficient (4.24):

$$2\left(\ln 2 \frac{\lceil \lg_2 \lceil N \rceil + \lceil \lg_2 C_\ell^d \rceil}{\ell} + \frac{\lceil \lg_2 d \rceil + \Delta_d}{\ell} - \frac{\ln \eta}{\ell}\right) \le 2\left(K \ln 2 - \frac{\ln \eta}{\ell}\right).$$

Since inequality (4.25) holds true with probability at least $1-\eta$ and inequality (4.26) holds with probability 1, the inequality

$$R(T_i) < 2\left(K(T_i)\ln 2 - \frac{\ln \eta}{\ell}\right) \tag{4.27}$$

holds with probability at least $1-\eta$.

4.6.3 The SRM and MDL Principles

Now suppose that we are given M codebooks that have the following structure: Codebook 1 contains a small number of tables, codebook 2 contains these tables and some more tables, and so on.

In this case one can use a more sophisticated decoding scheme to describe string (4.20): First, describe the number m of the codebook (this requires $\lceil \lg_2 m \rceil + \Delta_m$, $\Delta_m < 2\lceil \lg_2 \lg_2 m \rceil$ bits) and then, using this codebook, describe the string (which as shown above takes $\lceil \lg_2 N \rceil + \lceil \lg_2 C_\ell^d \rceil + \lceil \lg_2 d \rceil + \Delta_d$ bits).

The total length of the description in this case is not less than $\lceil \ln_2 N \rceil + \lceil \ln_2 C_\ell^d \rceil + \lceil \lg_2 d \rceil + \Delta_d + \lceil \lg_2 m \rceil + \Delta_m$, and the compression coefficient is

$$K(T) = \frac{\lceil \lg_2 N \rceil + \lceil \lg_2 C_\ell^d \rceil + \lceil \lg_2 d \rceil + \Delta_d + \lceil \lg_2 m \rceil + \Delta_m}{\ell}.$$

For this case an inequality analogous to inequality (4.27) holds. Therefore, the probability of error for the table that was used for compressing the description of string (4.20) is bounded by inequality (4.27).

Thus, for $d < \ell/2$ and $\ell > 6$ we have proved the following theorem:

Theorem 4.3. *If on a given structure of codebooks one compresses by a factor $K(T)$ the description of string (4.20) using a table T, then with probability at least $1-\eta$ one can assert that the probability committing an error by the table T is bounded by*

$$R(T) < 2\left(K(T)\ln 2 - \frac{\ln \eta}{\ell}\right), \quad \ell > 6. \tag{4.28}$$

Note how powerful the concept of the compression coefficient is: To obtain a bound on the probability of error, we actually need only information about this coefficient.[8] We do not need such details as

(i) How many examples we used,

(ii) how the structure of the codebooks was organized,

(iii) which codebook was used,

(iv) how many tables were in the codebook,

(v) how many training errors were made using this table.

Nevertheless, the bound (4.28) is not much worse than the bound on the risk (4.25) obtained on the basis of the theory of uniform convergence. The latter has a more sophisticated structure and uses information about the number of functions (tables) in the sets, the number of errors on the training set, and the number of elements of the training set.

[8] The second term, $-\ln \eta/\ell$, on the right-hand side is actually foolproof: For reasonable η and ℓ it is negligible compared to the first term, but it prevents one from considering too small η and/or too small ℓ.

Note also that the bound (4.28) cannot be improved more than by factor 2: It is easy to show that in the case where there exists a perfect table in the codebook, the equality can be achieved with factor 1.

This theorem justifies the MDL principle: To minimize the probability of error one has to minimize the coefficient of compression.

4.6.4 A Weak Point of the MDL Principle

There exists, however, a weak point in the MDL principle.

Recall that the MDL principle uses a codebook with *a finite number* of tables. Therefore, to deal with a set of functions determined by a continuous range of parameters, one must make a finite number of tables.

This can be done in many ways. The problem is this:

What is a "smart" codebook for the given set of functions?

In other words, how, for a given set of functions, can one construct a codebook with a small number of tables, but with good approximation ability?

A "smart" quantization could significantly reduce the number of tables in the codebook. This affects the compression coefficient. Unfortunately, finding a "smart" quantization is an extremely hard problem. This is the weak point of the MDL principle.

In the next chapter we will consider a normalized set of linear functions in a very high dimensional space (in our experiments we use linear functions in $N \approx 10^{13}$ dimensional space). We will show that the VC dimension h of the subset of functions with bounded norm depends on the value of the bound. It can be a small (in our experiments $h \approx 10^2$ to 10^3). One can guarantee that if a function from this set separates a training set of size ℓ without error, then the probability of test error is proportional to $h \ln \ell / \ell$.

The problem for the MDL approach to this set of indicator functions is how to construct a codebook with $\approx \ell^h$ tables (but not with $\approx \ell^N$ tables) that approximates this set of linear functions well.

The MDL principle works well when the problem of constructing reasonable codebooks has an obvious solution. But even in this case, it is not better than the SRM principle. Recall that the bound for the MDL principle (which cannot be improved using only the concept of the compression coefficient) was obtained by roughening the bound for the SRM principle.

Informal Reasoning and Comments — 4

Attempts to improve performance in various areas of computational mathematics and statistics have essentially led to the same idea that we call the structural risk minimization inductive principle.

First this idea appeared in the methods for solving ill-posed problems:

(i) Methods of quasi-solutions (Ivanov, 1962),

(ii) methods of regularization (Tikhonov, 1963)).

It then appeared in the method for nonparametric density estimation:

(i) Parzen windows (Parzen, 1962),

(ii) projection methods (Chentsov, 1963),

(iii) conditional maximum likelihood method (the method of sieves (Grenander, 1981)),

(iv) maximum penalized likelihood method (Tapia and Thompson, 1978)), etc.

The idea then appeared in methods for regression estimation:

(i) Ridge regression (Hoerl and Kennard, 1970),

(ii) model selection (see review in (Miller, 1990)).

Finally, it appeared in regularization techniques for both pattern recognition and regression estimation algorithms (Poggio and Girosi, 1990).

Of course, there were a number of attempts to justify the idea of searching for a solution using a structure on the admissible set of functions. However, in the framework of the classical approach justifications were obtained only for specific problems and only for the asymptotic case.

In the model of risk minimization from empirical data, the SRM principle provides capacity (VC dimension) control, and it can be justified for a finite number of observations.

4.7 METHODS FOR SOLVING ILL-POSED PROBLEMS

In 1962 Ivanov suggested an idea for finding a quasi-solution of the linear operator equation

$$Af = F, \quad f \in M, \tag{4.29}$$

in order to solve ill-posed problems. (The linear operator A maps elements of the metric space $M \subset E_1$ with metric ρ_{E_1} to elements of the metric space $N \subset E_2$ with metric ρ_{E_2}.) He suggested considering a set of nested convex compact subsets

$$M_1 \subset M_2 \subset \cdots \subset M_k, \cdots, \tag{4.30}$$

$$\overline{\bigcup_{i=1}^{\infty} M_i} = M, \tag{4.31}$$

and for any subset M_i to find a function $f_i^* \in M_i$ minimizing the distance

$$\rho = \rho_{E_2}(Af, F).$$

Ivanov proved that under some general conditions the sequence of solutions

$$f_1^*, \ldots, f_k^*, \ldots$$

converges to the desired one.

The quasi-solution method was suggested at the same time as Tikhonov proposed his regularization technique; in fact, the two are equivalent. In the regularization technique, one introduces a nonnegative semicontinuous (from below) functional $\Omega(f)$ that possesses the following properties:

(i) The domain of the functional coincides with M (the domain to which the solution of (4.29) belongs).

(ii) The region for which the inequality

$$M_j = \{f : \Omega(f) \le d_j\}, \quad d_j > 0,$$

holds forms a compactum in the metric of space E_1.

(iii) The solution of (4.29) belongs to some M_i^*:

$$\Omega(f) \le d^* < \infty.$$

Tikhonov suggested finding a sequence of functions f_γ minimizing the functionals

$$\Phi_\gamma(f) = \rho_{E_2}^2(Af, F) + \gamma\Omega(f)$$

for different γ. He proved that f_γ converges to the desired solution as γ converges to 0.

Tikhonov also suggested using the regularization technique even in the case where the right-hand side of the operator equation is given only within some δ-accuracy:

$$\rho_{E_2}(F, F_\delta) \le \delta.$$

In this case, in minimizing the functionals

$$\Phi^*(f) = \rho_{E_2}^2(Af, F_\delta) + \gamma(\delta)\Omega(f) \tag{4.32}$$

one obtains a sequence f_δ of solutions converging (in the metric of E_1) to the desired one f_0 as $\delta \to 0$ if

$$\lim_{\delta \to 0} \gamma(\delta) = 0,$$

$$\lim_{\delta \to 0} \frac{\delta^2}{\gamma(\delta)} = 0.$$

In both methods the formal convergence proofs do not explicitly contain "capacity control." Essential, however, was the fact that any subset M_i in Ivanov's scheme and any subset $M = \{f : \Omega(f) \le c\}$ in Tikhonov's scheme is compact. That means it has a bounded capacity (a metric ε-entropy).

Therefore, both schemes implement an SRM principle: First define a structure on the set of admissible functions such that any element of the structure has a finite capacity, increasing with the number of the element. Then, on any element of the structure, the function providing the best approximation of the right-hand side of the equation is found. The sequence of the obtained solutions converges to the desired one.

4.8 STOCHASTIC ILL-POSED PROBLEMS AND THE PROBLEM OF DENSITY ESTIMATION

In 1978 we generalized the theory of regularization to stochastic ill-posed problems (Vapnik and Stefanyuk, 1978). We considered a problem of solving the operator equation (4.29) in the case where the right-hand side is unknown, but we are given a sequence of approximations F_δ possessing the following properties:

(i) Each of these approximations F_δ is *a random function.*[9]

(ii) The sequence of approximations converges in probability (in the metric of the space E_2) to the unknown function F as δ converges to zero.

In other words, the sequence of random functions F_δ has the property

$$P\{\rho_{E_2}(F, F_\delta) > \varepsilon\} \underset{\delta \to 0}{\longrightarrow} 0, \quad \forall \varepsilon > 0.$$

Using Tikhonov's regularization technique one can obtain, on the basis of random functions F_δ, a sequence of approximations f_δ to the solution of (4.29).

We proved that for any $\varepsilon > 0$ there exists $\gamma_0 = \gamma_0(\varepsilon)$ such that for any $\gamma(\delta) \leq \gamma_0$ the functions minimizing functional (4.32) satisfy the inequality

$$P\left\{\rho_{E_1}(f, f_\delta) > \varepsilon\right\} \leq 2P\left\{\rho^2_{E_2}(F, F_\delta) > \gamma(\delta)\varepsilon\right\}. \tag{4.33}$$

In other words, we connected the distribution of the random deviation of the approximations from the exact right-hand side (in the E_2 metric) with the distribution of the deviations of the solutions obtained by the regularization method from the desired one (in the E_1 metric).

In particular, this theorem gave us an opportunity to find a general method for constructing various density estimation methods.

As mentioned in Section 1.8, density estimation requires us to solve the integral equation

$$\int_{-\infty}^{x} p(t)dt = F(x),$$

where $F(x)$ is an unknown probability distribution function, using i.i.d. data $x_1, \ldots, x_\ell, \ldots$.

Let us construct the empirical distribution function

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i),$$

which is a random approximation to $F(x)$, since it was constructed using random data $x_1, \ldots, x_\ell$.

In Section 3.9 we found that the differences $\sup_x |F(x) - F_\ell(x)|$ are described by the Kolmogorov–Smirnov bound. Using this bound we obtain

$$P\left\{\sup_x |F(x) - F_\ell(x)| > \varepsilon\right\} < 2e^{-2\varepsilon^2 \ell}.$$

[9]A random function is one that is defined by a realization of some random event. For a definition of random functions see any advanced textbook in probability theory, for example, A.N. Schiryaev, *Probability*, Springer, New York.

Therefore, if one minimizes the regularized functional

$$R(p) = \rho^2_{E_2}\left(\int_{\infty}^{x} p(t)dt,\ F_\ell(x)\right) + \gamma_\ell \Omega(p), \tag{4.34}$$

then according to inequality (4.33) one obtains the estimates $p_\ell(t)$, whose deviation from the desired solution can be described as follows:

$$P\left\{\rho_{E_1}(p, p_\ell) > \varepsilon\right\} \leq 2\exp\{-2\varepsilon\ell\gamma_\ell\}.$$

Therefore, the conditions for consistency of the obtained estimators are

$$\gamma_\ell \underset{\ell \to \infty}{\longrightarrow} 0,$$

$$\ell\gamma_\ell \underset{\ell \to \infty}{\longrightarrow} \infty. \tag{4.35}$$

Thus, minimizing functionals of type (4.34) under the constraint (4.35) gives consistent estimators. Using various norms E_2 and various functionals $\Omega(p)$ one can obtain various types of density estimators (including all classical estimators[10]). For our reasoning it is important that all nonparametric density estimators implement the SRM principle. By choosing the functional $\Omega(p)$, one defines a structure on the set of admissible solutions (the nested set of functions $M_c = \{p : \Omega(p) \leq c\}$ determined by constant c); using the law γ_ℓ one determines the appropriate element of the structure.

In Chapter 7 using this approach we will construct direct method of the density, the conditional density, and the conditional probability estimation.

4.9 THE PROBLEM OF POLYNOMIAL APPROXIMATION OF THE REGRESSION

The problem of constructing a polynomial approximation of regression, which was very popular in the 1970s, played an important role in understanding the problems that arose in small sample size statistics.

[10]By the way, one can obtain all classical estimators if one approximates an unknown distribution function $F(x)$ by the the empirical distribution function $F_\ell(x)$. The empirical distribution function, however, is not the best approximation to the distribution function, since, according to definition, the distribution function should be an absolutely continuous one, while the empirical distribution function is discontinuous. Using absolutely continuous approximations (e.g., a polygon in the one-dimensional case) one can obtain estimators that in addition to nice asymptotic properties (shared by the classical estimators) possess some useful properties from the point of view of limited numbers of observations (Vapnik, 1988).

Consider for simplicity the problem of estimating a one-dimensional regression by polynomials. Let the regression $f(x)$ be a smooth function. Suppose that we are given a finite number of measurements of this function corrupted with additive noise

$$y_i = f(x_i) + \xi_i, \quad i = 1, \ldots, \ell,$$

(in different settings of the problem, different types of information about the unknown noise are used; in this model of measuring with noise we suppose that the value of noise ξ_i does not depend on x_i, and that the point of measurement x_i is chosen randomly according to an unknown probability distribution $F(x)$).

The problem is to find the polynomial that is the closest (say in the $L_2(F)$ metric) to the unknown regression function $f(x)$. In contrast to the classical regression problem described in Section 1.7.3, the set of functions in which one has to approximate the regression is now rather wide (polynomial of any degree), and the number of observations is fixed.

Solving this problem taught statisticians a lesson in understanding the nature of the small sample size problem. First the simplified version of this problem was considered: The case where the regression itself is a polynomial (but the degree of the polynomial is unknown) and the model of noise is described by a normal density with zero mean. For this particular problem the classical asymptotic approach was used: On the basis of the technique of testing hypotheses, the degree of the regression polynomial was estimated and then the coefficients of the polynomial were estimated. Experiments, however, showed that for small sample sizes this idea was wrong: Even if one knows the actual degree of the regression polynomial, one often has to choose a smaller degree for the approximation, depending on the available number of observations.

Therefore, several ideas for estimating the degree of the approximating polynomial were suggested, including (Akaike, 1970), and (Schwartz, 1978) (see (Miller, 1990)). These methods, however, were justified only in asymptotic cases.

4.10 THE PROBLEM OF CAPACITY CONTROL

4.10.1 Choosing the Degree of the Polynomial

Choosing the appropriate degree p of the polynomial in the regression problem can be considered on the basis of the SRM principle, where the set of polynomials is provided with the simplest structure: The first element of the structure contains polynomials of degree one:

$$f_1(x, \alpha) = \alpha_1 x + \alpha_0, \quad \alpha = (\alpha_1, \alpha_0) \in R^2;$$

the second element contains polynomials of degree two:

$$f_2(x, \alpha) = \alpha_2 x^2 + \alpha_1 x + \alpha_0, \quad \alpha = (\alpha_2, \alpha_1, \alpha_0) \in R^3;$$

and so on.

To choose the polynomial of the best degree, one can minimize the following functional (the righthand side of bound (3.30)):

$$R(\alpha, m) = \frac{\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - f_m(x_i, \alpha))^2}{\left(1 - c\sqrt{\mathcal{E}_\ell}\right)_+}, \tag{4.36}$$

$$\mathcal{E}_\ell = 4\frac{h_m(\ln\frac{2\ell}{h_m} + 1) - \ln\eta/4}{\ell},$$

where h_m is the VC dimension of the set of the loss functions

$$Q(z, \alpha) = (y - f_m(x, \alpha))^2, \quad \alpha \in \Lambda,$$

and c is a constant determining the "tails of distributions" (see Sections 3.4 and 3.7).

One can show that the VC dimension h of the set of real functions

$$Q(z, \alpha) = F(|g(z, \alpha)|), \quad \alpha \in \Lambda,$$

where $F(u)$ is any fixed monotonic function, does not exceed eh^*, where $e < 9.34$ and h^* is the VC dimension of the set of indicators

$$I(z, \alpha, \beta) = \theta(g(x, \alpha) - \beta), \quad \alpha \in \Lambda, \ \beta \in R^1.$$

Therefore, for our loss functions the VC dimension is bounded as follows:

$$h_m \le e(m + 1).$$

To find the best approximating polynomial, one has to choose both the degree m of the polynomial and the coefficients α minimizing functional[11] (4.36).

4.10.2 Choosing the Best Sparse Algebraic Polynomial

Let us now introduce another structure on the set of algebraic polynomials: Let the first element of the structure contain polynomials $P_1(x, \alpha) = \alpha_1 x^d$, $\alpha \in R^1$ (of arbitrary degree d), with one nonzero term; let the second element contain polynomials $P_2(x, \alpha) = \alpha_1 x^{d_1} + \alpha_2 x^{d_2}$, $\alpha \in R^2$, with

[11] We used this functional (with constant $c = 1$, and $\mathcal{E}_\ell = [m(\ln \ell/m + 1) - \ln\eta]/\ell$, where $\eta = \ell^{-1/2}$) in several benchmark studies for choosing the degree of the best approximating polynomial. For small sample sizes the results obtained were often better than ones based on the classical suggestions.

two nonzero terms; and so on. The problem is to choose the best sparse polynomial $P_m(x)$ to approximate a smooth regression function.

To do this, one has to estimate the VC dimension of the set of loss functions

$$Q(z, \alpha) = (y - P_m(x, \alpha))^2,$$

where $P_m(x, \alpha)$, $\alpha \in R^m$, is a set of polynomials of arbitrary degree that contain m terms. Consider the case of one variable x.

The VC dimension h for this set of loss functions can be bounded by $2h^*$, where h^* is the VC dimension of the indicators

$$I(y, x) = \theta(y - P_m(x, \alpha) - \beta), \quad \alpha \in R^m, \ \beta \in R^1.$$

Karpinski and Werther showed that the VC dimension h^* of this set of indicators is bounded as follows:

$$3m \le h^* \le 4m + 3$$

(Karpinski and Werther, 1989). Therefore, our set of loss functions has VC dimension less than $e(4m + 3)$. This estimate can be used for finding the sparse algebraic polynomial that minimizes the functional (4.36).

4.10.3 Structures on the Set of Trigonometric Polynomials

Consider now structures on the set of trigonometric polynomials. First we consider a structure that is determined by the degree of the polynomials.[12] The VC dimension of the set of our loss function with trigonometric polynomials of degree m is less than $h = 4m + 2$. Therefore, to choose the best trigonometric approximation one can minimize the functional (4.36). For this structure there is no difference between algebraic and trigonometric polynomials.

The difference appears when one constructs a structure of sparse trigonometric polynomials. In contrast to the sparse algebraic polynomials, where any element of the structure has finite VC dimension, the VC dimension of *any* element of the structure on the sparse trigonometric polynomials is infinite.

This follows from the fact that the VC dimension of the set of indicator functions

$$f(x, \alpha) = \theta(\sin \alpha x), \quad \alpha \in R^1, \quad x \in (0, 1),$$

is infinite (see Example 2, Section 3.6).

[12] Trigonometric polynomials of degree m have the form

$$f_p(x) = \sum_{k=1}^{m} (a_k \sin kx + b_k \cos kx) + a_0.$$

4.10.4 The Problem of Feature Selection

The problem of choosing sparse polynomials plays an extremely important role in learning theory, since the generalization of this problem is a problem of feature selection (feature construction) using empirical data.

As was demonstrated in the examples, the above problem of feature selection (the terms in the sparse polynomials can be considered as the features) is quite delicate. To avoid the effect encountered for sparse trigonometric polynomials, one needs to construct *a priori* a structure containing elements with *bounded VC dimension* and then choose decision rules from the functions of this structure.

Constructing a structure for learning algorithms that select (construct) features and control capacity is usually a hard combinatorial problem.

In the 1980s in applied statistics, several attempts were made to find reliable methods of selecting *nonlinear functions* that control capacity. In particular, statisticians started to study the problem of function estimation in the following sets of the functions:

$$y = \sum_{j=1}^{m} \alpha_j K(x, w_j) + \alpha_0,$$

where $K(x, w)$ is a symmetric function with respect to vectors x and w, $w_1, ..., w_m$ are unknown vectors, and $\alpha_1, ..., \alpha_m$ are unknown scalars (Friedman and Stuetzle, 1981), (Breiman, Friedman, Olshen, and Stone, 1984) (in contrast to approaches developed in the 1970s for estimating *linear in parameters functions* (Miller, 1990)). In these classes of functions choosing the functions $K(x, w_j)$, $j = 1, ..., m$, can be interpreted as feature selection.

As we will see in the next chapter, for the sets of functions of this type, it is possible to effectively control both factors responsible for generalization ability — the value of the empirical risk and the VC dimension.

4.11 THE PROBLEM OF CAPACITY CONTROL AND BAYESIAN INFERENCE

4.11.1 The Bayesian Approach in Learning Theory

In the classical paradigm of function estimation, an important place belongs to the Bayesian approach (Berger, 1985).

According to Bayes's formula two events A and B are connected by the equality

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

One uses this formula to modify the ML models of function estimation discussed in the comments on Chapter 1.

Consider, for simplicity, the problem of regression estimation from measurements corrupted by additive noise

$$y_i = f(x, \alpha_0) + \xi_i.$$

In order to estimate the regression by the ML method, one has to know a parametric set of functions $f(x, \alpha), \alpha \in \Lambda \subset R^n$, that contain the regression $f(x, \alpha_0)$, and one has to know a model of noise $P(\xi)$.

In the Bayesian approach, one has to possess additional information: One has to know the *a priori* density function $P(\alpha)$ that for any function from the parametric set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, defines the probability for it to be the regression. If $f(x, \alpha_0)$ is the regression function, then the probability of the training data

$$[Y, X] = (y_1, x_1), \ldots, (y_\ell, x_\ell)$$

equals

$$P([Y, X]|\alpha_0) = \prod_{i=1}^{\ell} P(y_i - f(x_i, \alpha_0)).$$

Having seen the data, one can *a posteriori* estimate the probability that parameter α defines the regression:

$$P(\alpha|[Y, X]) = \frac{P([Y, X]|\alpha)P(\alpha)}{P([Y, X])}. \tag{4.37}$$

One can use this expression to choose an approximation to the regression function.

Let us consider the simplest way: We choose the approximation $f(x, \alpha^*)$ such that it yields the maximum conditional probability.[13] Finding α^* that maximizes this probability is equivalent to maximizing the following functional:

$$\Phi(\alpha) = \sum_{i=1}^{\ell} \ln P(y_i - f(x_i, \alpha)) + \ln P(\alpha). \tag{4.38}$$

[13] Another estimator constructed on the basis of the *a posteriori* probability

$$\phi_0(x|[Y, X]) = \int f(x, \alpha) P(\alpha|[Y, X]) d\alpha$$

possesses the following remarkable property: It minimizes the average quadratic deviation from the admissible regression functions

$$R(\phi) = \int (f(x, \alpha) - \phi(x|[Y, X]))^2 P([Y, X]|\alpha) P(\alpha) dx\, d([Y, X])\, d\alpha.$$

To find this estimator in explicit form one has to conduct integration analytically (numerical integration is impossible due to the high dimensionality of α). Unfortunately, analytic integration of this expression is mostly an unsolvable problem.

Let us for simplicity consider the case where the noise is distributed according to the normal law

$$P(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}.$$

Then from (4.37) one obtains the functional

$$\Phi^*(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i, \alpha))^2 - \frac{2\sigma^2}{\ell} \ln P(\alpha), \tag{4.39}$$

which has to be minimized with respect to α in order to find the approximation function. The first term of this functional is the value of the empirical risk, and the second term can be interpreted as a regularization term with the explicit form of the regularization parameter.

Therefore, the Bayesian approach brings us to the same scheme that is used in SRM or MDL inference.

The goal of these comments is, however, to describe a difference between the Bayesian approach and SRM or MDL.

4.11.2 Discussion of the Bayesian Approach and Capacity Control Methods

The only (but significant) shortcoming of the Bayesian approach is that it is restricted to the case where the set of functions of the learning machine coincides with the set of problems that the machine has to solve. Strictly speaking, it cannot be applied in a situation where the set of admissible problems differs from the set of admissible functions of the learning machine. For example, it cannot be applied to the problem of approximation of the regression function by polynomials if the regression function is not polynomial, since the *a priori* probability $P(\alpha)$ for any function from the admissible set of polynomials to be the regression is equal to zero. Therefore, the *a posteriori* probability (4.37) for any admissible function of the learning machine is zero. To use the Bayesian approach one must possess the following strong *a priori* information:

(i) The given set of functions of the learning machine coincides with the set of problems to be solved.

(ii) The *a priori* distribution on the set of problems is described by the given expression $P(\alpha)$.[14]

[14] This part of the *a priori* information is not as important as the first one. One can prove that with increasing numbers of observations the influence of an inaccurate description of $P(\alpha)$ is decreased.

In contrast to the Bayesian method, the capacity (complexity) control methods SRM or MDL use weak (qualitative) *a priori* information about reality: They use a structure on the admissible set of functions (the set of functions is ordered according to an idea of usefulness of the functions); this *a priori* information does not include any quantitative description of reality. Therefore, using these approaches, one can approximate a set of functions that is different from the admissible set of functions of the learning machine.

Thus, inductive inference in the Bayesian approach is based (along with training data) on given *strong* (quantitative) *a priori* information about reality, while inductive inference in the SRM or MDL approaches is based (along with training data) on *weak* (qualitative) *a priori* information about reality, but uses capacity (complexity) control.

In discussions with advocates of the Bayesian formalism, who use this formalism in the case where the set of problems to be solved and the set of admissible functions of the machine do not coincide, one hears the following claim:

The Bayesian approach also works in general situations.

The fact that the Bayesian formalism sometimes works in general situations (where the functions implemented by the machine do not necessarily coincide with those being approximated) has the following explanation. Bayesian inference has an outward form of capacity control. It has two stages: an informal stage, where one chooses a function describing (quantitative) *a priori* information $P(\alpha)$ for the problem at hand, and a formal stage, where one finds the solution by minimizing the functional (4.38). By choosing the distribution $P(\alpha)$ one controls capacity.

Therefore, in the general situation the Bayesian formalism realizes a human-machine procedure for solving the problem at hand, where capacity control is implemented by a human choice of the regularizer $\ln P(\alpha)$.

In contrast to Bayesian inference, SRM and MDL inference are pure machine methods for solving problems. For *any* ℓ they use the same structure on the set of admissible functions and the same formal mechanisms for capacity control.

Chapter 5

Methods of Pattern Recognition

To implement the SRM inductive principle in learning algorithms one has to minimize the risk in a given set of functions by controlling two factors: the value of the empirical risk and the value of the confidence interval.

Developing such methods is the goal of the theory of constructing learning algorithms.

In this chapter we describe learning algorithms for pattern recognition and consider their generalizations for the regression estimation problem.

5.1 WHY CAN LEARNING MACHINES GENERALIZE?

The generalization ability of learning machines is based on the factors described in the theory for controlling the generalization ability of learning processes. According to this theory, to guarantee a high level of generalization ability of the learning process one has to construct a structure

$$S_1 \subset S_2 \subset \cdots \subset S$$

on the set of loss functions $S = \{Q(z, \alpha), \alpha \in \Lambda\}$ and then choose both an appropriate element S_k of the structure and a function $Q(z, \alpha_\ell^k) \in S_k$ in this element that minimizes the corresponding bounds, for example, bound (4.1). The bound (4.1) can be rewritten in the simple form

$$R(\alpha_\ell^k) \le R_{\text{emp}}(\alpha_\ell^k) + \Phi(\frac{\ell}{h_k}), \tag{5.1}$$