Information Science and Statistics

Series Editors:
M. Jordan
J. Kleinberg
B. Schölkopf

Information Science and Statistics

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 Vapnik

Estimation of Dependences Based on Empirical Data

Reprint of 1982 Edition

Empirical Inference Science

Afterword of 2006

Springer

Vladimir Vapnik
NEC Labs America
4 Independence Way
Princeton, NJ 08540
vlad@nec-labs.com

Samuel Kotz (*Translator*)
Department of Engineering Management
and Systems Engineering
The George Washington University
Washington, D.C. 20052

Library of Congress Control Number: 2005938355

ISBN-10: 0-387-30865-2
ISBN-13: 978-0387-30865-4

Printed on acid-free paper.

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springer.com

Vladimir Vapnik

Estimation of Dependences Based on Empirical Data

Translated by Samuel Kotz

With 22 illustrations

Springer

To the students of my students in memory of my violin teacher Ilia Shtein and PhD advisor Alexander Lerner, who taught me several important things that are very difficult to learn from books.

PREFACE

Twenty-five years have passed since the publication of the Russian version of the book *Estimation of Dependencies Based on Empirical Data* (*EDBED* for short). Twenty-five years is a long period of time. During these years many things have happened. Looking back, one can see how rapidly life and technology have changed, and how slow and difficult it is to change the theoretical foundation of the technology and its philosophy.

I pursued two goals writing this Afterword: to update the technical results presented in *EDBED* (the easy goal) and to describe a general picture of how the new ideas developed over these years (a much more difficult goal).

The picture which I would like to present is a very personal (and therefore very biased) account of the development of one particular branch of science, *Empirical Inference Science*.

Such accounts usually are not included in the content of technical publications. I have followed this rule in all of my previous books. But this time I would like to violate it for the following reasons. First of all, for me *EDBED* is the important milestone in the development of empirical inference theory and I would like to explain why. Second, during these years, there were a lot of discussions between supporters of the new paradigm (now it is called the VC theory[1]) and the old one (classical statistics). Being involved in these discussions from the very beginning I feel that it is my obligation to describe the main events.

The story related to the book, which I would like to tell, is the story of how it is difficult to overcome existing prejudices (both scientific and social), and how one should be careful when evaluating and interpreting new technical concepts.

This story can be split into three parts that reflect three main ideas in the development of empirical inference science: from the pure technical (mathematical) elements of the theory to a new paradigm in the philosophy of generalization.

[1] VC theory is an abbreviation for Vapnik–Chervonenkis theory. This name for the corresponding theory appeared in the 1990s after *EDBED* was published.

The first part of the story, which describes the main technical concepts behind the new mathematical and philosophical paradigm, can be titled

Realism and Instrumentalism: Classical Statistics and VC Theory

In this part I try to explain why between 1960 and 1980 a new approach to empirical inference science was developed in contrast to the existing classical statistics approach developed between 1930 and 1960.

The second part of the story is devoted to the rational justification of the new ideas of inference developed between 1980 and 2000. It can be titled

Falsifiability and Parsimony: VC Dimension and the Number of Entities

It describes why the concept of VC falsifiability is more relevant for predictive generalization problems than the classical concept of parsimony that is used both in classical philosophy and statistics.

The third part of the story, which started in the 2000s can be titled

Noninductive Methods of Inference: Direct Inference Instead of Generalization

It deals with the ongoing attempts to construct new predictive methods (direct inference) based on the new philosophy that is relevant to a complex world, in contrast to the existing methods that were developed based on the classical philosophy introduced for a simple world.

I wrote this Afterword with my students' students in mind, those who just began their careers in science. To be successful they should learn something very important that is not easy to find in academic publications.

In particular they should see the big picture: what is going on in the development of this science and in closely related branches of science in general (not only about some technical details). They also should know about the existence of very intense paradigm wars. They should understand that the remark of Cicero, "Among all features describing genius the most important is inner professional honesty", is not about ethics but about an intellectual imperative. They should know that Albert Einstein's observation about everyday scientific life that "Great spirits have always encountered violent opposition from mediocre minds," is still true. Knowledge of these things can help them to make the right decisions and avoid the wrong ones. Therefore I wrote a fourth part to this Afterword that can be titled

The Big Picture.

This, however, is an extremely difficult subject. That is why it is wise to avoid it in technical books, and risky to discuss it commenting on some more or less recent events in the development of the science.

Writing this Afterword was a difficult project for me and I was able to complete it in the way that it is written due to the strong support and help of my colleagues Mike Miller, David Waltz, Bernhard Schölkopf, Leon Bottou, and Ilya Muchnik.

I would like to express my deep gratitude to them.

Princeton, New Jersey, Vladimir Vapnik
November 2005

CONTENTS

Chapter 1

REALISM AND INSTRUMENTALISM: CLASSICAL STATISTICS AND VC THEORY (1960–1980)

1.1 THE BEGINNING

In the history of science two categories of intellectual giants played an important role:

(1) The giants that created the new models of nature such as Lavoisier, Dirac, and Pasteur;

(2) The giants that created a new vision, a new passion, and a new philosophy for dealing with nature such as Copernicus, Darwin, Tsiolkovsky, and Wiener.

In other words, there are giants who created new technical paradigms, and giants who created new conceptual (philosophical) paradigms. Among these, there are unique figures who did both, such as Isaac Newton and Albert Einstein.

Creating a new technical paradigm is always difficult. However, it is much more difficult to change a philosophical paradigm. To do this sometimes requires several generations of scientists.[1] Even now one can see the continuation of the old paradigm wars in articles discussing (in a negative way) the intellectual heritage of the great visionaries Charles Darwin, Albert Einstein, Norbert Wiener, and Isaac Newton.

My story is about attempts to shift one of the oldest philosophical paradigms related to the understanding of human intelligence. Let me start with the vision Wiener described in his book *Cybernetics*. The main message of this book was that there are no

[1] Fortunately scientific generations change reasonably fast, about every ten years.

big conceptual differences between solving intellectual problems by the brain or by a computer, and that it is possible to use computers to solve many intellectual problems.

Today every middle school student will agree with that (five scientific generations have passed since Wiener's time!). However, 50 years ago even such giants as Kolmogorov hesitated to accept this point of view.

1.1.1 The Perceptron

One of the first scientific realizations of Wiener's idea was a model of how the brain learns introduced by Rosenblatt. He created a computer program called the "Perceptron" and successfully checked it on the digit recognition problem. Very soon Novikoff proved that the Perceptron algorithm (inspired by pure neurophysiology) constructs a hyperplane in some high-dimensional feature space that separates the different categories of training vectors.

It should be mentioned that models of how the brain generalizes and different pattern recognition algorithms both existed at the time of the Perceptron. These algorithms demonstrated success in solving simple generalization problems (for example Selfridge's Pandemonium, or Steinbuch's Learning Matrix).

However, after Rosenblatt's Perceptron and Novikoff's theorem, it became clear that complex biological models can execute very simple mathematical ideas. Therefore it may be possible to understand the principles of the organization of the brain using abstract mathematical arguments applied to some general mathematical constructions (this was different from analysis of specific technical models suggested by physiologists).

1.1.2 Uniform Law of Large Numbers

The Novikoff theorem showed that a model of the brain described in standard physiological terms ("neurons," "reward and punishment," "stimulus") executes a very simple mathematical idea — it constructs a hyperplane that separates two different categories of data in some mathematical space. More generally, it minimizes in a given set of functions an empirical risk functional.

If it is true that by minimizing the empirical risk one can generalize, then one can construct more efficient minimization algorithms than the one that was used by the Perceptron. Therefore in the beginning of the 1960s many such algorithms were suggested. In particular Alexey Chervonenkis and I introduced the optimal separating hyperplane that was more efficient for solving practical problems than the Perceptron algorithm (especially for problems with a small sample size). In the 1990s this idea became a driving force for SVMs (we will discuss SVMs in Chapter 2, Section 2.3). However, just separation of the training data does not guarantee success on the test data. One can easily show that good separating of the training data is a necessary condition for the generalization. But what are the sufficient conditions?

This led to the main question of learning theory:

When does separation of the training data lead to generalization?

This question was not new. The problem, "How do humans generalize?" (What is the model of induction? Why is the rule that is correct for previous observations also correct for future observations?) was discussed in classical philosophy for many centuries. Now the same question — but posed for the simplest mathematical model of generalization, the pattern recognition problem — became the subject of interest.

In the beginning of the 1960s many researchers including Chervonenkis and I became involved in such discussions. We connected this question with the existence of uniform convergence of frequencies to their probabilities over a given set of events. To find the conditions that guarantee the generalization for the pattern recognition problem, it is sufficient to find the conditions for such convergence.

Very quickly we constructed a theory for uniform convergence over sets with a finite number of events (1964) and in four years we obtained the general answer, the necessary and sufficient conditions for uniform convergence for any (not necessarily finite) set of events. This path is described in *EDBED*.

What was not known at the time EDBED was written is that the uniform convergence describes not only sufficient conditions for generalization but also the necessary conditions:

Any algorithm that uses training data to choose a decision rule from the given admissible set of rules must satisfy it.

It took us another 20 years to prove this fact. In 1989 we proved the main theorem of VC theory that states:[2]

If the necessary and sufficient conditions for uniform convergence are not valid, that is, if the VC entropy over the number of observations does not converge to zero,

$$\frac{H_P^{\Lambda}(\ell)}{\ell} \longrightarrow c \neq 0,$$

then there exists a subspace X^* of the space R^n whose probability measure is equal to c,

$$P(X^*) = c,$$

such that almost any sample of vectors $x_1^*, \ldots, x_k^*$ of arbitrary size k from the subspace X^* can be separated in all 2^k possible ways by the functions from the admissible set of indicator functions $f(x, \alpha),\ \alpha \in \Lambda$. (See also *EDBED*, Chapter 6 Section 7 for the definition of VC entropy).

This means that if uniform convergence does not take place then any algorithm that does not use additional prior information and picks up one function from the set of admissible functions cannot generalize.[3]

[2]Below for the sake of simplicity we formulate the theorem for the pattern recognition case (sets of indicator functions), but the theorem has been proven for any set of real-valued functions [121;140]. Also to simplify formulation of the theorem we used the concept of "two-sided uniform convergence" discussed in *EDBED* instead of "one-sided" introduced in [121].

[3]This, however, leaves an opportunity to use averaging algorithms that possess a priori information about the set of admissible functions. In other words VC theory does not intersect with Bayesian theory.

If, however, the conditions for uniform convergence are valid then (as shown in Chapter 6 of *EDBED*) for any fixed number of observations one can obtain a bound that defines the guaranteed risk of error for the chosen function.

Using classical statistics terminology the uniform convergence of the frequencies to their probability over a given set of events can be called the *uniform law of large numbers* over the corresponding set of events. (The convergence of frequencies to their corresponding probability for a fixed event (the Bernoulli law) is called the law of large numbers.)

Analysis of Bernoulli's law of large numbers has been the subject of intensive research since the 1930s. Also in the 1930s it was shown that for one particular set of events the uniform law of large numbers always holds. This fact is the Glivenco–Cantelli theorem. The corresponding bound on the rate of convergence forms Kolmogorov's bound. Classical statistics took advantage of these results (the Glivenco–Cantelli theorem and Kolmogorov's bound are regarded as the foundation of theoretical statistics).

However, to analyze the problem of generalization for pattern recognition, one should have an answer to the more general question:

> *What is the demarcation line that describes whether the uniform law of large numbers holds?*

The obtaining of the existence conditions for the uniform law of large numbers and the corresponding bound on the rate of convergence was the turning point in the studies of empirical inference.

This was not recognized immediately, however. It took at least two decades to understand this fact in full detail. We will talk about this in what follows.

1.2 Realism and Instrumentalism in Statistics and the Philosophy of Science

1.2.1 The Curse of Dimensionality and Classical Statistics

The results of successfully training a Perceptron (which constructed decision rules for the ten-class digit classification problem in 400-dimensional space, using 512 training examples) immediately attracted the attention of the theorists.

In classical statistics a problem analogous to the pattern recognition problem was considered by Ronald Fisher in the 1930s, the so-called problem of discriminant analysis. Fisher considered the following problem. One knows the generating model of data for each class, the density function defined up to a fixed number of parameters (usually Gaussian functions). The problem was: given the generative models (the model how the data are generated known up to values of its parameters) estimate the discriminative rule. The proposed solution was:

> *First*, using the data, estimate the parameters of the statistical laws and

Second, construct the optimal decision rule using the estimated parameters.

To estimate the densities, Fisher suggested the maximum likelihood method.

This scheme later was generalized for the case when the unknown density belonged to a nonparametric family. To estimate these generative models the methods of nonparametric statistics were used (see example in Chapter 2 Section 2.3.5). However, the main principle of finding the desired rule remained the same: first estimate the generative models of data and then use these models to find the discriminative rule.

This idea of constructing a decision rule after finding the generative models was later named the *generative model of induction*. This model is based on understanding of how the data are generated. In a wide philosophical sense an understanding of how data are generated reflects an understanding of the corresponding law of nature.

By the time the Perceptron was introduced, classical discriminant analysis based on Gaussian distribution functions had been studied in great detail. One of the important results obtained for a particular model (two Gaussian distributions with the same covariance matrix) is the introduction of a concept called the Mahalanobis distance. A bound on the classification accuracy of the constructed linear discriminant rule depends on a value of the Mahalanobis distance.

However, to construct this model using classical methods requires the estimation of about $0.5n^2$ parameters where n is the dimensionality of the space. Roughly speaking, to estimate one parameter of the model requires C examples. Therefore to solve the ten-digit recognition problem using the classical technique one needs $\approx 10(400)^2C$ examples. The Perceptron used only 512.

This shocked theorists. It looked as if the classical statistical approach failed to overcome the curse of dimensionality in a situation where a heuristic method that minimized the empirical loss easily overcame this curse.

Later the methods based on the idea of minimizing different type of empirical losses were called the *predictive (discriminative) models of induction*, in contrast to the classical *generative models*. In a wide philosophical sense predictive models do not necessarily connect prediction of an event with understanding of the law that governs the event; they are just looking for a function that explains the data best.[4]

The VC theory was constructed to justify the empirical risk minimization induction principle: according to VC theory the generalization bounds for the methods that minimize the empirical loss do not depend directly on the dimension of the space. Instead they depend on the so-called capacity factors of the admissible set of functions — the VC entropy, the Growth function, or the VC dimension — that can be much smaller than the dimensionality. (In *EDBED* they are called *Entropy* and *Capacity*; the names VC entropy and VC dimension as well as VC theory appeared later due to R. Dudley.)

[4]It is interesting to note that Fisher suggested along with the classical generative models (which he was able to justify), the heuristic solution (that belongs to a discriminative model) now called Fisher's linear discriminant function. This function minimizes some empirical loss functional, whose construction is similar to the Mahalanobis distance. For a long time this heuristic of Fisher was not considered an important result (it was ignored in most classical statistics textbooks). Only recently (after computers appeared and statistical learning theory became a subject not only of theoretical but also of practical justification) did Fisher's suggestion become a subject of interest.

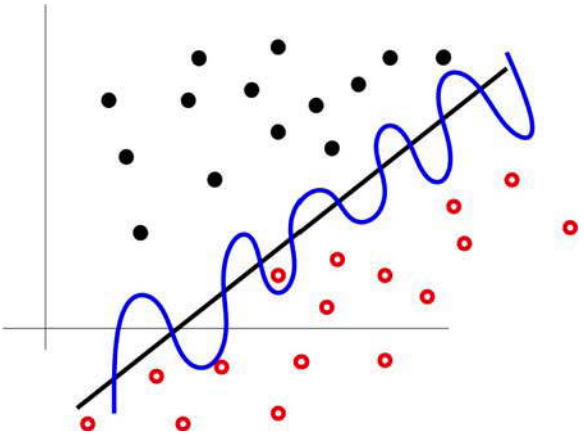

Figure 1.1: Two very different rules can make a similar classification.

Why do the generative and discriminative approaches lead to different results? There are two answers to this very important question which can be described from two different points of view: technical and philosophical (conceptual).

1.2.2 The Black Box Model

One can describe the pattern recognition problem as follows. There exists a black box BB that when given an input vector x_i returns an output y_i which can take only two values $y_i \in \{-1, +1\}$. The problem is: given the pairs $(y_i, x_i), i = 1, \ldots, \ell$ (the training data) find a function that approximates the rule that the black box uses.

Two different concepts of what is meant by a *good approximation* are possible:

(1) A good approximation of the BB rule is a function that is close (in a metric of functional space) to the function that the BB uses. (In the classical setting often we assume that the BB uses the Bayesian rule.)

(2) A good approximation of the BB rule is a function that provides approximately the same error rate as the one that the BB uses (provides the rule that predicts the outcomes of the BB well).

In other words, in the first case one uses a concept of closeness in the sense of being close to the *true function* used by the BB (closeness in a metric space of functions), while in the second case one uses a concept of closeness in the sense of being close to the accuracy of prediction (closeness in *functionals*). These definitions are very different.

In Figure 1.1 there are two different categories of data separated by two different rules. Suppose that the straight line is the function used by the black box. Then from the point of view of function estimation, the polynomial curve shown in Figure 1.1 is

very different from the line and therefore cannot be a good estimate of the *true BB* rule. From the other point of view, the polynomial rule separates the data well (and as we will show later can belong to a set with small VC dimension) and therefore can be a good *instrument* for prediction.

The lesson the Perceptron teaches us is that sometimes it is useful to give up the ambitious goal of estimating the rule the *BB* uses (the generative model of induction). Why?

Before discussing this question let me make the following remark. The problem of pattern recognition can be regarded as a generalization problem: using a set of data (observations) find a function[5] (theory). The same goals (but in more complicated situations) arise in the classical model of science: using observation of nature find the law. One can consider the pattern recognition problem as the simplest model of generalization where observations are just a set of i.i.d. vectors and the admissible laws are just a set of indicator functions. Therefore it is very useful to apply the ideas described in the general philosophy of induction to its simplest model and vice versa, to understand the ideas that appear in our particular model in the general terms of the classical philosophy. Later we will see that these interpretations are nontrivial.

1.2.3 Realism and Instrumentalism in the Philosophy of Science

The philosophy of science has two different points of view on the goals and the results of scientific activities.

(1) There is a group of philosophers who believe that the results of scientific discovery are the real laws that exist in nature. These philosophers are called the *realists*.

(2) There is another group of philosophers who believe the laws that are discovered by scientists are just an instrument to make a good prediction. The discovered laws can be very different from the ones that exist in Nature. These philosophers are called the *instrumentalists*.

The two types of approximations defined by classical discriminant analysis (using the generative model of data) and by statistical learning theory (using the function that explains the data best) reflect the positions of realists and instrumentalists in our simple model of the philosophy of generalization, the pattern recognition model. Later we will see that the position of philosophical instrumentalism played a crucial role in the success that pattern recognition technology has achieved.

However, to explain why this is so we must first discuss the theory of ill-posed problems, which in many respects describes the relationship between realism and instrumentalism in very clearly defined situations.

[5]The pattern recognition problem can be considered as the simplest generalization problem, since one has to find the function in a set of admissible *indicator* functions (that can take only two values, say 1 and -1.

1.3 Regularization and Structural Risk Minimization

1.3.1 Regularization of Ill-Posed Problems

In the beginning of the 1900s, Hadamard discovered a new mathematical phenomenon. He discovered that there are continuous operators A that map, in a one-to-one manner, elements of a space f to elements of a space F, but the inverse operator A^{-1} from the space F to the space f can be discontinuous. This means that there are operator equations

$$Af = F \tag{1.1}$$

whose solution in the set of functions $f \in \Phi$ exists, and is unique, but is unstable. (See Chapter 1 of EDED). That is, a small deviation $F + \Delta F$ of the (known) right-hand side of the equation can lead to a big deviation in the solution. Hadamard thought that this was just a mathematical phenomenon that could never appear in real-life problems. However, it was soon discovered that many important practical problems are described by such equations.

In particular, the problem of solving some types of linear operator equations (for example, Fredholm's integral equation of the second order) are ill-posed (see Chapter 1, Section 5 of *EDBED*). It was shown that many geophysical problems require solving (ill-posed) integral equations whose right-hand side is obtained from measurements (and therefore is not very accurate).

For us it is important that ill-posed problems can occur when one tries to estimate *unknown reasons from observed consequences*.

In 1943 an important step in understanding the structure of ill-posed problems was made. Tikhonov proved the so-called inverse operator lemma:

> *Let A be a continuous one-to-one operator from E_1 to E_2. Then the inverse operator A^{-1} defined on the images F of a compact set $f \in \Phi^*$ is stable.*

This means that if one possesses very strong prior knowledge about the solution (it belongs to a known compact set of functions), then it is possible to solve the equation. It took another 20 years before this lemma was transformed into specific approaches for solving ill-posed problems.

In 1962 Ivanov [21] suggested the following idea of solving operator equation (1.1). Consider the functional $\Omega(f) \geq 0$ that possesses the following two properties

(1) For any $c \geq 0$ the set of functions satisfying the constraint

$$\Omega(f) \leq c \tag{1.2}$$

is convex and compact.

(2) The solution f_0 of Equation (1.1) belongs to some compact set

$$\Omega(f_0) \leq c_0 \tag{1.3}$$

(where the constant $c_0 > 0$ may be unknown).

Under these conditions Ivanov proved that there exists a strategy for choosing $c = c(\varepsilon)$ depending on the accuracy of the right-hand side $||\Delta F||_{E_2} \leq \varepsilon$ such that the sequence of minima of the functional

$$R = ||Af - F||_{E_2} \tag{1.4}$$

subject to the constraints

$$\Omega(f) \leq c(\varepsilon) \tag{1.5}$$

converges to the solution of the ill-posed problem (1.1) as ε approaches zero.

In 1963 Tikhonov [55] proved the equivalent theorem that states: under conditions (1.2) and (1.3) defined on the functional $\Omega(f)$, there exists a function $\gamma_\varepsilon = \gamma(\varepsilon)$ such that the sequence of minima of the functionals

$$R_\gamma(f) = ||Af - F_\varepsilon||^2_{E_2} + \gamma_\varepsilon \Omega(f) \tag{1.6}$$

converges to the solution of the operator equation (1.1) as ε approaches zero.[6]

Both these results can be regarded as "*comforting ones*" since for any ε (even very small) one can guarantee nothing (the theorems guarantee only convergence of the sequence of solutions).

Therefore, one should try to avoid solving ill-posed problems by replacing them (if possible) with well-posed problems.

Keeping in mind the structure of ill-posed problems our problem of finding the *BB* solution can be split into two stages:

(1) Among a given set of admissible functions find a subset of functions that provides an expected loss that is close to the minimal one.

(2) Among functions that provide a small expected loss find one that is close to the *BB* function.

The first stage does not lead to an ill-posed problem, but the second stage might (if the corresponding operator is unstable).

The realist view requires solving both stages of the problem, while the instrumentalist view requires solving only the first stage and choosing for prediction any function that belongs to the set of functions obtained.

Technically, ill-posed problems appear in classical discriminant analysis as soon as one connects the construction of a discriminant function with the density estimation problem.

By definition, the density (if it exists) is a solution of the following equation

$$\int_a^x p(x')dx' = F(x), \tag{1.7}$$

[6]There is one more equivalent idea of how to solve ill-posed problems proposed in 1962 by Phillips [166]: minimize the functional $\Omega(f)$ satisfying the conditions defined above subject to the constraints

$$||Af - F||^2 \leq \varepsilon.$$

where $F(x)$ is a cumulative distribution function.

Therefore to estimate the density from the data

$$x_1, \ldots, x_\ell$$

means to solve Fredholm's equation (1.7) when the cumulative distribution function $F(x)$ is unknown but the data are given. One can construct an approximation to the unknown cumulative distribution function and use it as the right hand side of the equation. For example, one can construct the empirical distribution function

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i), \tag{1.8}$$

where

$$\theta(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{if } u < 0 \end{cases}.$$

It is known from Kolmogorov's bound for the Glivenco–Cantelli theorem that the empirical distribution function converges exponentially fast (not only asymptotically but for any set of fixed observations) to the desired cumulative distribution function. Using the empirical distribution function constructed from the data, one can try to solve this equation.

Note that this setting of the density estimation problem cannot be avoided since it reflects the definition of the density. Therefore in both parametric or nonparametric statistics, one has to solve this equation. The only difference is how the set of functions in which one is looking for the solution is defined: in a "narrow set of parametric functions" or in a "wide set of non-parametric functions".[7]

However, this point of view was not clearly developed in the framework of classical statistics, since both theories (parametric and nonparametric) of density estimation were constructed *before* the theory of solving ill-posed problems was introduced.

The general setting of the density estimation problem was described for the first time in *EDBED*. Later in Chapter 2, Section 2.3 when we discuss the SVM method, we will consider a pattern recognition problem, and show the difference between the solutions obtained by nonparametric statistics (based on the philosophy of realism) and by an SVM solution (based on the philosophy of instrumentalism).

Regularization Techniques

The regularization theory as introduced by Tikhonov suggests minimizing the equation

$$R_\gamma(f) = ||Af - F_\varepsilon||^2_{E_2} + \gamma_\varepsilon \Omega(f). \tag{1.9}$$

Under very specific requirements on the set of functions defined both by the functional $\Omega(f)$ and the value $c > 0$

$$\Omega(f) \leq c \tag{1.10}$$

[7]The maximum likelihood method suggested by Fisher is valid just for a very narrow admissible set of functions. It is already invalid, for example, for the set of densities defined by the sum of two Gaussians with unknown parameters (see example [139], Section 1.7.4.)

(for any $c > 0$ the set should be *convex and compact*), and under the condition that the desired solution *belongs to the set with some fixed* c_0, it is possible to define a strategy of choosing the values of the parameter γ that asymptotically lead to the solution.

1.3.2 Structural Risk Minimization

The Structural Risk Minimization (SRM) principle generalizes the Ivanov scheme in two ways:

(1) It considers a structure on any sets of functions (not necessarily defined by inequality (1.5)).

(2) It does not require compactness or convexity on the set of functions that define the element of the structure. It also does not require the desired solution belonging to one of the elements of the structure.

The only requirement is that every element of the nested sets possesses a finite VC dimension (or other capacity factor).

Under these general conditions the risks provided by functions that minimize the VC bound converge to the smallest possible risk (even if the desired function belongs to the closure of the elements). Also, for any fixed number of observations it defines the smallest guaranteed risk.

In the early 1970s Chervonenkis and I introduced SRM for sets of indicator functions (used in solving pattern recognition problems) [13]. In *EDBED* the SRM principle was generalized for sets of real-valued functions (used in solving regression estimation problems).

Therefore the difference between regularization and structural risk minimization can be described as follows.

Regularization was introduced for solving ill-posed problems. It requires strong knowledge about the problem to be solved (the solution has to belong to the compact (1.10) defined by some constant c) and (generally speaking) does not have guaranteed bounds for a finite number of observations.

Structural risk minimization was introduced for solving predictive problems. It is more general (does not require strong restrictions of admissible set of functions) and has a guaranteed bound for a finite number of observations.

Therefore if the regularization method is the main instrument for solving ill-posed problems using the *philosophical realism* approach, then the structural risk minimization method is the main instrument for solving problems using the *philosophical instrumentalism* approach.

Remark. In the late 1990s the concept of regularization started to be used in the general framework of minimizing the functionals (1.9) to solve predictive generalization problems. The idea was that under any definition of the functional $\Omega(f)$ there exists a parameter γ which leads to convergence to the desired result. This is, however,

incorrect: first, it depends on the concept of convergence; second, there are functionals (for which the set of functions (1.5) can violate finiteness of capacity conditions) that do not lead to convergence in any sense.

1.4 The Beginning of the Split Between Classical Statistics and Statistical Learning Theory

The philosophy described above was more or less clear by the end of the 1960s.[8] By that time there was no doubt that in analyzing the pattern recognition problem we came up with a new direction in the theory of generalization. The only question that remained was how to describe this new direction. Is this a new branch of science or is it a further development in classical statistics? This question was the subject of discussions in the seminars at the Institute of Control Sciences of the Academy of Sciences of USSR (Moscow).

The formal decision, however, was made when it came time to publish these results in the *Reports of Academy of Sciences of USSR* [143]. The problem was in which section of *Reports* it should be published — in "Control Sciences (Cybernetics)" or in "Statistics". It was published as a contribution in the "Control Sciences" section.

This is how one of the leading statisticians of the time, Boris Gnedenco, explained why it should not be published in the "Statistics" section:

> It is true that this theory came from the same roots and uses the same formal tools as statistics. However, to belong to the statistical branch of science this is not enough. Much more important is to share the same belief in the models and to share the same philosophy. Whatever you are suggesting is not in the spirit of what I am doing or what A. Kolmogorov is doing. It is not what our students are doing nor will it be what the students of our students do. Therefore, you must have your own students, develop your own philosophy, and create your own community.

More than 35 years have passed since this conversation. The more time passed, the more impressed I became with Gnedenco's judgment. The next three decades (1970s, 1980s, and 1990s) were crucial for developments in statistics. After the shocking discovery that the classical approach suffers from the curse of dimensionality, statisticians tried to find methods that could replace classical methods in solving real-life problems. During this time statistics was split into two very different parts: theoretical statistics that continued to develop the classical paradigm of generative models, and applied statistics that suggested a compromise between theoretical justification of the algorithms and heuristic approaches to solving real-life problems. They tried to justify such a position by inventing special names for such activities (exploratory data analysis), where in fact the superiority of common sense over theoretical justification was declared. However, they never tried to construct or justify new algorithms using VC

[8] It was the content of my first book *Pattern Recognition Problem* published in 1971 (in Russian).

theory. Only after SVM technology became a dominant force in data mining methods did they start to use its technical ideas (but not its philosophy) to modify classical algorithms.[9]

Statistical learning theory found its home in computer science. In particular, one of the most advanced institutions where SLT was developing in the 1970s and 1980s was the Institute of Control Sciences of the Academy of Sciences of USSR. Three different groups, each with different points of view on the generalization problem, became involved in such research: the Aizerman–Braverman–Rozonoer's group, the Tsypkin group, and the Vapnik–Chervonenkis group.

Of these groups ours was the youngest: I just got my PhD (candidate of science) thesis, and Chervonenkis got his several years later. Even so, our research direction was considered one of the most promising. In order to create a VC community I was granted permission from the Academy of Sciences to have my own PhD students.[10]

From this beginning we developed a statistical learning community. I had several very strong students including Tamara Glaskov, Anatoli Mikhalsky, Anatoli Stehanuyk, Alexander Sterin, Fellix Aidu, Sergey Kulikov, Natalia Markovich, Ada Sorin, and Alla Juravel who developed both machine learning theory and effective machine learning algorithms applied to geology and medicine.

By the end of the 1960s my department head, Alexander Lerner, made an extremely important advance in the application of machine learning: he convinced the high-level bureaucrats to create a laboratory for the application of machine learning techniques in medicine.

In 1970 such a laboratory was created in the State Oncology Centre. The director of the laboratory was my former PhD student, Tamara Glaskov.

It is hard to overestimate how much this laboratory accomplished during this time. Only recently have the most advanced oncology hospitals in the West created groups to analyze clinical data. This was routine in USSR decades earlier.

In beginning of the 1970s I prepared my doctoral thesis.

1.5 The Story Behind This Book

Government control under the Soviet Communist regime was total. One of its main modus operandi was to control who was promoted into more or less prominent positions. From the government bureaucrat's perspective a scientific degree (and especially a doctoral degree) holder possessed influence, and therefore they wanted to control who obtained this degree.

The execution of such control was one of the obligations of the institution called

[9] Statisticians did not recognise conceptual aspects of VC theory. Their criticism of this theory before SVM was that the VC bounds were too loose to be useful. Therefore the theory is not practical and to create new methods it is better to use common sense than the results of this theory.

[10] In the Russian system there were two academic degrees: *candidate of science* (which is equivalent to the PhD degree in the United States) and *doctor of science* (which is equivalent to the *Habilitation a Diriger des Recherches (HDR)* in France). Normally only doctors of science could have PhD students. I was granted this privilege and had to defend my doctoral thesis soon.

the Supreme Certifying Commission[11] (SCC) closely related to the KGB. The rule was that any decision on any thesis defense made by any Scientific Councils anywhere in the country must be approved by this commission. If the SCC disapproved several decisions by a particular Scientific Council it could be dismissed. Therefore the normal policy of academic institutions was not to enter into conflict with the SCC.

From the KGB's point of view I was a wrong person to obtain the doctoral level: I was not a member of the Communist Party, I was Jewish, my PhD adviser, Alexander Lerner, had applied for immigration to Israel and became a "refusenik," some of my friends were dissidents, and so on.

In this situation everybody understood that the Institute would be in conflict with the SCC's mandate. Nevertheless the feeling was that the support of the scientific community would be so strong that the SCC would not start the battle.

The SCC, however, reacted with a trick that to my knowledge was never used before: it requested that the Scientific Council change one of the reviewers to their trusted man who did his job: wrote a negative review.

I had a long conversation with the Chairman of the Scientific Council, Yakov Tsypkin, after he discussed the situation with the members of the Council. He told me that everyone on the Scientific Council understood what was going on and if I decided to defend my thesis the Scientific Council would unanimously support me. However, I had no chance of being approved by the SCC since they would have a formal reason to reject my thesis. Also they would have a formal reason to express distrust of the Scientific Council of the Institute. In this situation the best solution was to withdraw my thesis and publish it as a book. However, since the names of the authors of books were also under the KGB's control (the authors should also be "good guys") I would only be able to publish the book if my name did not attract too much attention. This would allow the editor, Vladimir Levantovsky (who was familiar with this story), to successfully carry out all necessary procedures to obtain permission (from the institution that controls the press) to publish the book.

So, I withdrew my thesis, rewrote it as a book, and due to the strong support of many scientists (especially Tsypkin), the editor Levantovsky was able to publish it (in Russian) in 1979.

In 1982 the well known American statistician, S. Kotz, translated it into English under the title *Estimation of Dependencies Based on Empirical Data* which was published by Springer. The first part of this volume is its reprint.

The main message that I tried to deliver in the book was that classical statistics could not overcome the curse of dimensionality but the new approach could. I devoted three chapters of the book to different classical approaches and demonstrated that none of them could overcome the curse of dimensionality. Only after that did I describe the new theory.

[11] The Russian abbreviation is VAK.

Chapter 2

FALSIFIABILITY AND PARSIMONY: VC DIMENSION AND THE NUMBER OF ENTITIES (1980–2000)

2.1 SIMPLIFICATION OF VC THEORY

For about ten years this book did not attract much attention either in Russia or in the West. It attracted attention later.

In the meantime, in 1984 (five years after the publication of the original version of this book and two years after its English translation) an important event happened. Leslie Valiant published a paper where he described his vision of how learning theory should be built [122].

Valiant proposed the model that later was called the *Probably Approximately Correct* (PAC) learning model. In this model, the goal of learning is to find a rule that reasonably well approximates the best possible rule. One has to construct algorithms which guarantee that such a rule will be found with some probability (not necessarily one). In fact, the PAC model is one of the major statistical models of convergence, called consistency. It has been widely used in statistics since at least Fisher's time.

Nevertheless Valiant's article was a big success. In the mid-1980s the general machine learning community was not very well connected to statistics. Valiant introduced to this community the concept of consistency and demonstrated its usefulness. The theory of consistency of learning processes as well as generalization bounds was the subject of our 1968 and 1971 articles [143, 11], and was described in detail in our 1974 book [12, 173] devoted to pattern recognition, and in a more general setting in *EDBED*. However, at that time these results were not well known in the West.[1]

[1] In 1989 I met Valiant in Santa Cruz, and he told me that he did not know of our results when he wrote

In the 20th century, and especially in the second half of it, mass culture began to play an important role. For us it is important to discuss the "scientific component" of mass culture.

With the increasing role of science in everyday life, the general public began to discuss scientific discoveries in different areas: physical science, computer science (cybernetics), cognitive science (pattern recognition), biology (genomics), and philosophy. The discussions were held using very simplified scientific models that could be understood by the masses. Also scientists tried to appeal to the general public by promoting their philosophy using simplified models (for example, as has been done by Wiener). There is nothing wrong with this.

However, when science becomes a mass profession, the elements of the scientific mass culture in some cases start to substitute for the real scientific culture: It is much easier to learn the slogans of the scientific mass culture than it is to learn many different concepts from the original scientific sources. Science and "scientific mass culture," however, are built on very different principles. In *Mathematical Discoveries*, Polya describes the principle of creating scientific mass culture observed by the remarkable mathematician Zermello. Here is the principle:

Gloss over the essentials and attract attention to the obvious.

Something that could remind this principle happened when (after appearance Valiant's article) the adaptation of ideas described in EDBED started. In the PAC adaptation the VC theory was significantly simplified by removing its essential parts.

In *EDBED* the main idea was the necessary and sufficient aspects of the theory based on three capacity concepts: the VC entropy, the Growth function, and the VC dimension. It stresses that the most accurate bounds can be obtained based on the VC entropy concept. This, however, requires information about the probability measure. One can construct less accurate bounds that are valid for all probability measures. To do this one has to calculate the Growth function which can have a different form for different sets of admissible functions. The Growth function can be upper bounded by the standard function that depends on only one integer parameter (the VC dimension). This also decreases accuracy, but makes the bounds simpler.

These three levels of the theory provide different possibilities for further developments in learning technology. For example, one can try to create theory for the case when the probability measure belongs to some specific sets of measures (say smooth ones), or one can try to find a better upper bound for the Growth function using a standard function that depends on say two (or more) parameters. This can lead to more accurate estimates and therefore to more advanced algorithms. The important component of the theory described in *EDBED* was the structural risk minimization principle. It was considered to be the main driving force behind predictive learning technology.

PAC theory started just from the definition of the VC dimension based on the combinatorial lemma used to estimate the bound for the Growth function (see *EDBED*, Chapter 6, Section A2). The main effort was placed on obtaining VC type bounds for

his article, and that he even visited a conference at Moscow University to explain this to me. Unfortunately we never met in Moscow. After his article was published Valiant tried to find the computer science aspects of machine learning research suggesting analyzing the computational complexity of learning problems. In 1990 he wrote [123]: "If the computational requirements is removed from the definition then we are left with the notion of non-parametric inference in sense of statistics as discussed in particular by Vapnik [*EDBED*]."

different classes of functions (say for neural networks), and on the generalizations of the theory for the set of nonindicator functions. In most cases these generalizations were based on extensions of the VC dimension concept for real-valued functions made in the style described in *EDBED*. The exception was the fat-shattering concept [141] related to VC entropy for real-valued functions described in Chapter 7.

In the early 1990s, some PAC researchers started to attack the VC theory. First, the VC theory was declared a "worst-case theory" since it is based on the uniform convergence concept. In contrast to this "worst-case-theory" the development of "real-case theory" was announced. However, this is impossible (see Section 1.2 of this Afterword) since the (one-sided) uniform convergence forms the necessary and sufficient conditions for consistency of learning (that is also true for PAC learning). Then in the mid-1990s an attempt was made to rename the Vapnik–Chervonenkis lemma (*EDBED*, Chapter 6, Sections 8 and A2) as the Sauer lemma. For the first time we published the formulation of this lemma in 1968 in the *Reports of the Academy of Sciences of USSR* [143]. In 1971, we published the corresponding proofs in the article devoted to the uniform law of large numbers [11]. In 1972, two mathematicians N. Sauer [130] and S. Shelah [131] independently proved this combinatorial lemma.

Researchers, who in the 1980s learned from *EDBED* (or from our articles) both the lemma and its role in statistical learning theory, renamed it in the 1990s.[2] Why?

My speculation is that renaming it was important for the dilution of VC theory and creating the following legend:

> In 1984 the PAC model was introduced. Early in statistics a concept called the VC dimension was developed. This concept plays an important role in the Sauer lemma, which is a key instrument in PAC theory.

Now, due to new developments in the VC theory and the interest in the advanced topics of statistical learning theory, this legend has died, and as a result interest in PAC theory has significantly decreased. This is, however, a shame because the computational complexity aspects of learning stressed by Valiant remain relevant.

2.2 Capacity Control

2.2.1 Bell Labs

In 1990 Larry Jackel, the head of the Adaptive Systems Research Department at AT&T Bell Labs, invited me to spend half a year with his group. It was a time of wide discussions on the VC dimension concept and its relationship to generalization problems. The obvious interpretation of the VC dimension was the number of free parameters that led to the curse of dimensionality. John Denker, a member of this department, showed,

[2] N. Sauer did not have in mind statistics proving this lemma. This is the content of the abstract of his article: "P. Erdös (oral communication) transmitted to me in Nice the following question: . . . (*the formulation of the lemma*). . . . .. In this paper we will answer this question in the affirmative by determining the exact upper bounds."

however, that the VC dimension is not necessarily the number of free parameters. He came up with the example

$$y = \theta\{\sin ax\},\ x \in R^1,\ a \in (0, \infty)$$

a set of indicator functions that has only one free parameter yet possesses an infinite VC dimension (see Section 2.7.5, footnote 7). In *EDBED* another situation was described: when the VC dimension was smaller than the number of free parameters. These intriguing facts could lead to new developments in learning theory.

Our department had twelve researchers. Six of them, L. Jackel, J. Denker, S. Solla, C. Burges, G. Nohl, and H.P. Graf were physicists, and six, Y. LeCun, L. Bottou, P. Simard, I. Guyon, B. Boser, and Y. Bengio were computer scientists. The main direction of research was to advance the understanding of pattern recognition phenomena. To do this they relied on the principles of research common in physics.

The main principle of research in physics can be thought of as the complete opposite of the Zermello principle for creating scientific mass culture. It can be formulated as follows:

Find the essential in the nonobvious.

The entire story of creating modern technology can be seen as an illustration of this principle. At the time when electricity, electromagnetic waves, annihilation, and other physical fundamentals were discovered they seemed to be insignificant elements of nature. It took a lot of joint efforts of theorists, experimental physicists and engineers to prove that these negligible artifacts are very important parts of nature and make it work.

The examples given by Denker and another one described in *EDBED* (see Chapter 10, Section 5) could be an indication that such a situation in machine learning is quite possible.

The goal of our department was to understand and advance new general principles of learning that are effective for solving real-life problems. As a model problem for ongoing experiments, the department focused on developing automatic systems that could read handwritten digits. This task was chosen for a number of reasons. First, it was known to be a difficult problem, with traditional machine vision approaches making only slow progress. Second, lots of data were available for training and testing. And third, accurate solutions to the problem would have significant commercial importance.

Initial success in our research department led to the creation of a development group supervised by Charlie Stenard. This group, which worked closely with us, had as a goal the construction of a machine for banks that could read handwritten checks from all over the world. Such a machine could not make too many errors (the number of errors should be comparable to the number made by humans). However, the machine could refuse to read some percentage of checks.

I spent ten years with this department. During this time check reading machines became an important instrument in the banking industry. About 10% of checks in US banks are read by technology developed at Bell Labs.

During these years the performance of digit recognition was significantly improved. However, it *never* happened that significant improvements in quality of classification were the results of smart engineering heuristics. All jumps in performance were results of advances in understanding fundamentals of the pattern recognition problem.

2.2.2 Neural Networks

When I joined the department, the main instrument for pattern recognition was neural networks constructed by Yann LeCun, one of the originators of neural networks. For the digit recognition problem he designed a series of convolutional networks called LeNet. In the early 1990s this was a revolutionary idea. The traditional scheme of applying pattern recognition techniques was the following: a researcher constructs several very carefully crafted features and uses them as inputs for a statistical parametric model. To construct the desired rule they estimated the parameters of this model. Therefore good rules in many respects reflected how smart the researcher was in constructing features.

LeNet uses as input a high-dimensional vector whose coordinates are the raw image pixels. This vector is processed using a multilayer convolutional network with many free parameters. Using the back propagation technique, LeNet tunes the parameters to minimize the training loss.[3]

For the digit recognition problem, the rules obtained by LeNet were significantly better than any rules obtained by the classical style algorithms. This taught a great lesson: one does not need to go into the details of the decision rule; it is enough to create an "appropriate architecture" and an "appropriate minimization method" to solve the problem.

2.2.3 Neural Networks: The Challenge

The success of neural nets in solving pattern recognition problems was a challenge for theorists. Here is why. When one is trying to understand how the brain is working two different questions arise:

(1) What happens? What are the principles of generalization that the brain executes?

(2) How does it happen? How does the brain execute these principles?

Neural networks attempt to answer the second question using an artificial brain model motivated by neurophysiologists.

According to the VC theory, however, this is not very important. VC theory declares that two and only two factors are responsible for generalization. They are the value of empirical loss, and the capacity of the admissible set of functions (the VC entropy, Growth function, or the VC dimension). The SRM principle states that any method that controls these two factors well (minimizing the right-hand side of the VC bounds) is strongly universally consistent.

It was clear that artificial neural networks executed the structural risk minimization principle. However, they seemed to do this rather inefficiently. Indeed, the loss function that artificial neural networks minimize has many local minima. One can guarantee convergence to one of these minima but cannot guarantee good generalization. Neural networks practitioners define some initial conditions that they believe will lead to a

[3] As computer power increased, LeCun constructed more powerful generations of LeNet.

"good" minimum. Also, the back-propagation method based on the gradient procedure of minimization in high-dimensional spaces requires a very subtle treatment of step values. The choice of these values does not have a good recommendation.

In order to control capacity the designer chooses an appropriate number of elements (neurons) for the networks. Therefore for different training data sizes one has to design different neural networks. All these factors make neural networks more of an art than a science.

Several ideas that tried to overcome the described shortcomings of neural networks were checked during 1991 and 1992 including measuring the VC dimension (capacity) of the learning machine [144, 142] and constructing local learning rules [145]. Now these ideas are developing in a new situation. However, in 1992 they were overshadowed by a new learning concept called Support Vector Machines (SVMs).

2.3 Support Vector Machines (SVMs)

The development of SVMs has a 30-year history, from 1965 until 1995. It was completed in three major steps.

2.3.1 Step One: The Optimal Separating Hyperplane

In 1964, Chervonenkis and I came up with an algorithm for constructing an optimal separating hyperplane called the generalized portrait method. Three chapters of our 1974 book *Theory of pattern recognition*, contain the detailed theory of this algorithm [12, 173]. In *EDBED* (Addendum I), a simplified version of this algorithm is given. Here are more details. The problem was: given the training data

$$(y_1, x_1), \ldots, (y_\ell, x_\ell), \tag{2.1}$$

construct the hyperplane

$$(w_0, x) + b_0 = 0 \tag{2.2}$$

that separates these data and has the largest margin. In our 1974 book and in *EDBED* we assumed that the data were separable. The generalization of this algorithm for constructing an optimal hyperplane in the nonseparable case was introduced in 1995 [132]. We will discuss it in a later section.

Thus, the goal was to maximize the functional

$$\rho_0 = \min_{\{i:\ y_i=1\}} \left[\left(\frac{w}{|w|}, x_i \right) + b \right] - \max_{\{j:\ y_j=-1\}} \left[\left(\frac{w}{|w|}, x_j \right) + b \right]$$

under the constraints

$$y_i((w, x_i) + b) \geq 1, \quad i = 1, \ldots, \ell. \tag{2.3}$$

It is easy to see that this problem is equivalent to finding the minimum of the quadratic form

$$R_1(w, b) = (w, w)$$

subject to the linear constraints (2.3). Let this minimum be achieved when $w = w_0$. Then

$$\rho_0 = \frac{2}{\sqrt{(w_0, w_0)}}.$$

To minimize the functional (w, w) subject to constraints (2.3) the standard Lagrange optimization technique was used. The Lagrangian

$$L(\alpha) = \frac{1}{2}(w, w) - \sum_{i=1}^{\ell} \alpha_i([y_i((w, x_i) + b) - 1]) \tag{2.4}$$

(where $\alpha_i \geq 0$ are the Lagrange multipliers) was constructed and its minimax (minimum over w and b and maximum over the multipliers $\alpha_i \geq 0$) was found. The solution of this quadratic optimization problem has the form

$$w_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i. \tag{2.5}$$

To find these coefficients one has to maximize the functional:

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (x_i, x_j) \tag{2.6}$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \ \alpha_i \geq 0, \ \ i = 1, \ldots, \ell.$$

Substituting (2.5) back into (2.2) we obtain the separating hyperplane expressed in terms of the Lagrange multipliers

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 (x, x_i) + b_0 = 0. \tag{2.7}$$

2.3.2 The VC Dimension of the Set of ρ-Margin Separating Hyperplanes

The following fact plays an important role in SVM theory. Let the vectors $x \in R^n$ belong to the sphere of radius $R = 1$. Then the VC dimension h of the set of hyperplanes with margin $\rho_0 = (w_0, w_0)^{-1}$ has the bound

$$h \leq \min\{(w_0, w_0), n\} + 1.$$

That is, the VC dimension is defined by the smallest of the two values: the dimensionality n of the vectors x and the value (w_0, w_0). In Hilbert (infinite dimensional) space,

the VC dimension of the set of separating hyperplanes with the margin ρ_0 depends just on the value (w_0, w_0).

In *EDBED* I gave a geometrical proof of the bound (See Chapter 10, Section 5). In 1997, Gurvits found an algebraic proof [124]. Therefore, the optimal separating hyperplane executes the SRM principle: it minimizes (to zero) the empirical loss, using the separating hyperplane that belongs to the set with the smallest VC dimension.

One can therefore introduce the following learning machine that executes the SRM principle:

> Map input vectors $x \in X$ into (a rich) Hilbert space $z \in Z$, and construct the maximal margin hyperplane in this space.

According to the VC theory the generalization bounds depend on the VC dimension. Therefore by controlling the margin of the separating hyperplane one controls the generalization ability.

2.3.3 Step Two: Capacity Control in Hilbert Space

The formal implementation of this idea requires one to specify the operator

$$z = \mathcal{F}x$$

which should be used for mapping. Then similar to (2.7) one constructs the separating hyperplane in image space

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 (z, z_i) + b_0 = 0,$$

where the coefficients $\alpha_i \geq 0$ are the ones that maximize the quadratic form

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (z_i, z_j) \tag{2.8}$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \ \alpha_i \geq 0, \quad i = 1, \ldots, \ell. \tag{2.9}$$

In 1992 Boser, Guyon and I found an effective way to construct the optimal separating hyperplane in Hilbert space without explicitly mapping the input vectors x into vectors z of the Hilbert space [125].

This was done using Mercer's theorem.

> *Let vectors $x \in X$ be mapped into vectors $z \in Z$ of some Hilbert space.*
>
> **1.** *Then there exists in X space a symmetric positive definite function $K(x_i, x_j)$ that defines the corresponding inner product in Z space:*
>
> $$(z_i, z_j) = K(x_i, x_j).$$

2. *Also, for any symmetric positive definite function $K(x_i, x_j)$ in X space there exists a mapping from X to Z such that this function defines an inner product in Z space.*

Therefore, according to Mercer's theorem, the separating hyperplane in image space has the form

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 K(x, x_i) + b_0 = 0,$$

where the coefficients α_i^0 are defined as the solution of the quadratic optimization problem: maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.10}$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \ \alpha_i \geq 0, \quad i = 1, \ldots, \ell. \tag{2.11}$$

Choosing specific kernel functions $K(x_i, x_j)$ one makes specific mappings from input vectors x into image vectors z.

The idea of using Mercer's theorem to map into Hilbert space was used in the mid-1960s by Aizerman, Braverman, and Rozonoer [2]. Thirty years later we used this idea in a wider context.

2.3.4 Step Three: Support Vector Machines

In 1995 Cortes and I generalized the maximal margin idea for constructing (in image space) the hyperplane

$$(w_0, z) + b_0 = 0$$

when the training data are nonseparable [132]. This technology became known as Support Vector Machines (SVMs). To construct such a hyperplane we follow the recommendations of the SRM principle.

Problem 1. Choose among the set hyperplanes with the predefined margin

$$\rho^2 = \frac{4}{(w_0, w_0)} \leq H = \frac{1}{h}$$

the one that separates the images of the training data with the smallest number of errors. That is, we minimize the functional

$$R = \sum_{i=1}^{\ell} \theta(\xi_i) \tag{2.12}$$

subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \;\; i = 1, \ldots, \ell \tag{2.13}$$

and the constraint

$$(w, w) \leq h, \tag{2.14}$$

where $\theta(u)$ is the step function:

$$\theta(u) = \begin{cases} 1, & \text{if } u \geq 0 \\ 0, & \text{if } u < 0. \end{cases}$$

For computational reasons, however, we approximate *Problem 1* with the following one.

Problem 2. Minimize the functional

$$R = \sum_{i=1}^{\ell} \xi_i \tag{2.15}$$

(instead of the functional (2.12)) subject to the constraints (2.13) and (2.14).

Using the Lagrange multiplier technique, one can show that the corresponding hyperplane has an expansion

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 (z_i, z) + b_0 = 0. \tag{2.16}$$

To find the multipliers one has to maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - h\sqrt{\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (z_i, z_j)} \tag{2.17}$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{2.18}$$

and the constraints

$$0 \leq \alpha_i \leq 1, \quad i = 1, \ldots, \ell.$$

Problem 3. Problem 2 is equivalent to the following (reparametrized) one: Minimize the functional

$$R = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} \xi_i \tag{2.19}$$

subject to constraints (2.13). This setting implies the following dual space solution: Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (z_i, z_j) \tag{2.20}$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{2.21}$$

and the constraints

$$0 \le \alpha_i \le C, \quad i = 1, \ldots, \ell.$$

One can show that for any h there exists a C such that the solutions of Problem 2 and Problem 3 coincide. From a computational point of view Problem 3 is simpler than Problem 2. However, in Problem 2 the parameter h estimates the VC dimension. Since the VC bound depends on the ratio h/ℓ one can choose the VC dimension to be some fraction of the training data, while in the reparametrized Problem 3 the corresponding parameter C cannot be specified; it can be any value depending on the VC dimension and the particular data.

Taking into account Mercer's theorem,

$$(z_i, z_j) = K(x_i, x_j),$$

we can rewrite the nonlinear separating rule in input space X as

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x) + b_0 = 0, \tag{2.22}$$

where the coefficients are the solution of the following problems:

Problem 1a. Minimize the functional

$$R = \sum_{i=1}^{\ell} \theta(\xi_i) \tag{2.23}$$

subject to the constraints

$$y_i \sum_{j=1}^{\ell} (y_j \alpha_j K(x_j, x_i) + b) \ge 1 - \xi_i, \;\; \xi_i \ge 0, \quad i = 1, \ldots, \ell \tag{2.24}$$

and the constraint

$$\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \le h. \tag{2.25}$$

Problem 2a. Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - h \sqrt{\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j)} \tag{2.26}$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{2.27}$$

and the constraints

$$0 \le \alpha_i \le 1, \quad i = 1, \ldots, \ell. \tag{2.28}$$

Problem 3a. Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{2.29}$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0$$

and the constraints

$$0 \le \alpha_i \le C, \quad i = 1, ..., \ell.$$

The solution of Problem 3a became the standard SVM method. In this solution only some of the coefficients α_i^0 are different from zero. The vectors x_i for which $\alpha_i^0 \neq 0$ in (2.22) are called the *support vectors*. Therefore, the separating rule (2.22) is the expansion on the support vectors.

To construct a support vector machine one can use any (conditionally) positive definite function $K(x_i, x_j)$ creating different types of SVMs. One can even use kernels in the situation when input vectors belong to nonvectorial spaces. For example, the inputs may be sequences of symbols of different size (as in problems of bioinformatics or text classification). Therefore SVMs form a universal generalization engine that can be used for different problems of interest.

Two examples of Mercer kernels are the polynomial kernel of degree d

$$K(x_i, x_j) = ((x_i, x_j) + c)^d, \; c \ge 0 \tag{2.30}$$

and the exponential kernel

$$K(x_i, x_j) = \exp\left\{ - \left(\frac{|x_i - x_j|}{\sigma} \right)^d \right\}, \; \sigma > 0, \; 0 \le d \le 2. \tag{2.31}$$

2.3.5 SVMs and Nonparametric Statistical Methods

SVMs execute the idea of the structural risk minimization principle, where the choice of the appropriate element of the structure is defined by the constant C (and a kernel parameters). Therefore, theoretically, for any appropriate kernel (say for (2.31) by controlling parameters (which depends on the training data) one guarantees asymptotic convergence of the SVM solutions to the best possible solution [167].

In 1980 Devroye and Wagner proved that classical nonparametric methods of density estimation are also universally consistent [134]. That is, by controlling the parameter $\sigma_\ell = \sigma(\ell) > 0$ depending on the size ℓ of the training data, the following approximation of the density function

$$\bar{p}(x) = \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{1}{(2\pi)^{n/2}\sigma_\ell^n} \exp\left\{-\left(\frac{|x_i - x|}{\sigma_\ell}\right)^2\right\} \tag{2.32}$$

converges (in the uniform metric) to the desired density *with increasing* ℓ.

However, by choosing an appropriate parameter C of SVM, one controls the VC bound for any finite number of observations. One can also control these bounds by choosing the parameters of the kernels.

This section illustrates the practical advantage of this fact.

Let us use the nonparametric density estimation method to approximate the optimal (generative) decision rule for binary classification

$$p_1(x) - p_2(x) = 0, \tag{2.33}$$

where $p_1(x)$ is the density function of the vectors belonging to the first class and $p_2(x)$ is the density function of the vectors belonging to the second class. Here for notational simplicity we assume that the two classes are equally likely and that the number of training samples from the first and second class is the same. Using (2.32) the approximation (2.33) can be rewritten as follows.

$$\sum_{\{i:\ y_i=1\}} \exp\left\{-\left(\frac{|x - x_i|}{\sigma}\right)^2\right\} - \sum_{\{j:\ y_j=-1\}} \exp\left\{-\left(\frac{|x - x_j|}{\sigma}\right)^2\right\} = 0.$$

The SVM solution using the same kernel has the form

$$\sum_{\{i:\ y_i=1\}} \alpha_i \exp\left\{-\left(\frac{|x - x_i|}{\sigma}\right)^2\right\} - \sum_{\{j:\ y_j=-1\}} \alpha_j \exp\left\{-\left(\frac{|x - x_j|}{\sigma}\right)^2\right\} = 0.$$

Since our kernel is a positive definite function there exists a space Z where it defines an inner product (by the second part of Mercer's theorem). In Z space both solutions define separating hyperplanes

$$\sum_{\{i:\ y_i=1\}} (z_i, z) - \sum_{\{j:\ y_j=-1\}} (z_j, z) = 0$$

(the classical non-parametric solution) [152] and

$$\sum_{\{i:\ y_i=1\}} \alpha_i(z_i, z) - \sum_{\{j:\ y_j=-1\}} \alpha_j(z_j, z) = 0.$$

(the SVM solution). Figure 2.1 shows these solutions in Z space. The separating hyperplane obtained by nonparametric statistics is defined by the hyperplane orthogonal

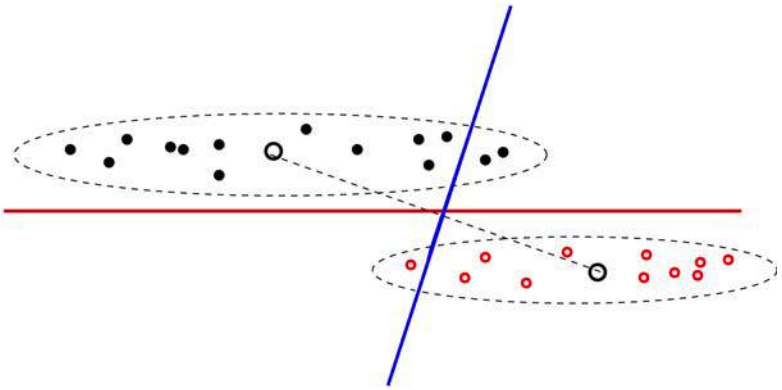

Figure 2.1: Classifications given by the classical nonparametric method and the SVM are very different.

to the line connecting the center of mass of two different classes. The SVM produces the optimal separating hyperplane.

In spite of the fact that both solutions converge asymptotically to the best one[4] they are very different for a fixed number of training data since the SVM solution is optimal (for any number of observations it guarantees the smallest predictive loss), while the non-parametric technique is not.

This makes SVM a state-of-the-art technology in solving real-life problems.

2.4 An Extension of SVMs: SVM+

In this section we consider a new algorithm called SVM+, which is an extension of SVM. SVM+ takes into account a known structure of the given data.

2.4.1 Basic Extension of SVMs

Suppose that our data are the union of $t \geq 1$ groups:

$$(X, Y)_r = (x_{r_1}, y_{r_1}), \ldots, (x_{r_{n_r}}, y_{r_{n_r}}), \quad r = 1, \ldots, t.$$

Let us denote indices from the group r by

$$T_r = \{i_{n_1}, \ldots, i_{n_r}\}, \quad r = 1, \ldots, t.$$

[4]Note that nonparametric density estimate (2.32) requires dependence of σ from ℓ. Therefore, it uses different Z spaces for different ℓ.

Let inside one group the slacks be defined by some correcting function that belongs to a given set of functions

$$\xi_i = \xi_r(x_i) = \phi_r(x_i, w_r), \;\; w_r \in W_r, \; i \in T_r, \; r = 1, \ldots, t. \tag{2.34}$$

The goal is to define the decision function for a situation when sets of admissible correcting functions are restricted (when sets of admissible correcting functions are not restricted we are back to conventional SVM). By introducing groups of data and different sets of correcting functions for different groups one introduces additional information about the problem to be solved.

To define the correcting function $\xi(x) = \phi_r(x, w_r)$ for group T_r we map the input vectors $x_i, \; i \in T_r$ simultaneously into two different Hilbert spaces: into the space $z_i \in Z$ which defines the decision function (as we did for the conventional SVM) and into correcting function space $z_i^r \in Z_r$ which defines the set of correcting functions for a given group r. (Note that vectors of different groups are mapped into the same decision space Z but different correcting spaces Z_r.)

Let the inner products in the corresponding spaces be defined by the kernels

$$(z_i, z_j) = K(x_i, x_j), \; \forall i, j$$

and

$$(z_i^r, z_j^r) = K_r(x_i, x_j), \; i, j \in T_r, \; r = 1, \ldots, t. \tag{2.35}$$

Let the set of admissible correcting functions $\xi_r(x) = \phi_r(x, w_r), \; w_r \in W_r$, be linear in each Z_r space

$$\xi(x_i) = \phi_r(x, w_r) = [(w_r, z_i^r) + d_r] \geq 0, \; i \in T_r, \; r = 1, \ldots, t. \tag{2.36}$$

As before our goal is to find the separating hyperplane in decision space Z,

$$(w_0, z) + b_0 = 0$$

whose parameters w_0 and b_0 minimize the functional

$$R(w, w_1, \ldots, w_t) = \frac{1}{2}(w, w) + C \sum_{r=1}^{t} \sum_{i \in T_r} ((w_r, z_i^r) + d_r), \tag{2.37}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - ((z_i^r, w_r) + d_r), \; i \in T_r, \; r = 1, \ldots, t \tag{2.38}$$

and the constraints

$$(w_r, z_i^r) + d_r \geq 0, \; i \in T_r, \; r = 1, \ldots, t. \tag{2.39}$$

Note that for set (2.36) the solution of this optimization problem does exist.

The corresponding Lagrangian is

$$L(w, w_1, \ldots, w_t; \alpha, \mu) = \frac{1}{2}(w, w) + C \sum_{r=1}^{t} \sum_{i \in T_r} ((w_r, z_i^r) + d_r) \tag{2.40}$$

$$-\sum_{i=1}^{\ell}\alpha_i[y_i((w,z_i)+b)-1+d_r+(w_r,z_i^r)]-\sum_{i=1}^{\ell}\mu_i((w_r,z_i^r)+d_r).$$

Using the same dual optimization technique as above one can show that the optimal separating hyperplane in Z space has the form

$$\sum_{i=1}^{\ell}\alpha_i^0 y_i(z_i,z)+b_0=0,$$

where the coefficients $\alpha_i^0\geq 0$ minimize the same quadratic form as before

$$W(\alpha)=\sum_{i=1}^{\ell}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{\ell}y_iy_j\alpha_i\alpha_jK(x_i,x_j) \tag{2.41}$$

subject to the conventional constraint

$$\sum_{i=1}^{\ell}y_i\alpha_i=0 \tag{2.42}$$

and the new constraints

$$\sum_{i\in T_r}(\alpha_i+\mu_i)=|T_r|C,\ r=1,\ldots,t \tag{2.43}$$

($|T_r|$ is the number of elements in T_r),

$$\sum_{i\in T_r}(\alpha_i+\mu_i)K_r(x_i,x_j)=C\sum_{i\in T_r}K_r(x_i,x_j),\ j\in T_r,\ r=1,\ldots,t. \tag{2.44}$$

and constraints

$$\alpha_i\geq 0,\ \mu_i\geq 0,\quad i=1,\ldots,\ell.$$

When either

(1) There is no structure in the data: any vector belongs to its own group,

or

(2) There is no correlation between slacks inside all groups: $K_r(x_i,x_j)$ is an identity matrix for all r

$$K_r(x_i,x_j)=\begin{cases}1, & \text{if } i=j\\ 0, & \text{if } i\neq j\end{cases} \tag{2.45}$$

then Equation (2.44) defines the box constraints as in conventional SVMs (in case (2) Equations (2.43) are satisfied automatically). Therefore the SVM+ model contains the classical SVM model as a particular case.

The advantage of the SVM+ is the ability to consider the global structure of the problem that the conventional SVM ignores (see Section 2.4.3 for details).

This, however, requires solving a more general quadratic optimization problem to minimize in the space of 2ℓ nonnegative variables the same objective function subject to $(\ell+t+1)$ linear constraints (instead of one minimizing this objective function in the space of ℓ variable subjects of one linear constraint and ℓ box constraints in the conventional SVM).

2.4.2 Another Extension of SVM: SVM$_\gamma$+

Consider another extension of SVM, the so-called SVM$_\gamma$+, which directly controls the capacity of sets of correcting functions.

Let us instead of objective function (2.37) consider the function

$$R(w, w_1, \ldots, w_t) = \frac{1}{2}(w, w) + \frac{\gamma}{2}\sum_{r=1}^{t}(w_r, w_r) + C\sum_{r=1}^{t}\sum_{i\in T_r}((w_r, z_i^r) + d_r), \quad (2.46)$$

where $\gamma > 0$ is some value. When γ approaches zero (2.46) and (2.37) coincide.

The SVM$_\gamma$+ solution minimizes functional (2.46) subject to the constraints (2.38) and (2.39). To solve this problem we construct the Lagrangian. Comparing it to (2.40), this Lagrangian has one extra term $\gamma/2\sum(w_r, w_r)$. Repeating almost the same algebra as in the previous section we obtain that for the modified Lagrangian the dual space solution that defines the coefficients α_i^0 must maximize the functional

$$W(\alpha, \mu) = \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) +$$

$$\frac{C}{\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \mu_i)K_r(x_i, x_j) - \frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \mu_i)(\alpha_j + \mu_j)K_r(x_i, x_j)$$

subject to the constraints (2.42) and the constraints

$$\sum_{i\in T_r}(\alpha_i + \mu_i) = |T_r|C,\ r = 1, \ldots, t,$$

$$\alpha_i \geq 0,\ \mu_i \geq 0,\ i = 1, \ldots, \ell.$$

Note that when either:

(1) There is no structure (every training vector belongs to its own group),

or

(2) There is no correlation inside groups ((2.45) holds for all r) and $\gamma \longrightarrow 0$

then the SVM$_\gamma$+ solution coincides with the conventional SVM solution.

This solution requires maximizing the quadratic objective function in the space of 2ℓ nonnegative variables subject to $t + 1$ equality constraints.

One can simplify the computation when using models of correcting functions (2.36) with $d_r = 0,\ r = 1, ..., t$. In this case one has to maximize the functional $W(\alpha, \mu)$ over non-negative variables $\alpha_i\ \mu_i,\ i = 1, ..., \ell$ subject to one equality constraint (2.42).

2.4.3 Learning Hidden Information

SVM+ is an instrument for a new inference technology which can be called *Learning Hidden Information* (LHI). It allows one to extract additional information in situations where conventional technologies cannot be used.

WHAT INFORMATION CAN BE HIDDEN?

Consider the pattern recognition problem. Let one be given the training set

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

Suppose that one can add to this set additional information from two sources:

(1) information that exists in *hidden classifications* of the training set and

(2) information that exists in *hidden variables* of the training set.

The next two examples describe such situations.

EXAMPLE 1 (Information given in hidden classifications).
Suppose that one's goal is to find a rule that separates cancer patients from non cancer patients. One collects training data and assigns class $y_i = 1$, or $y_i = -1$, to patient x_i depending on the result of analysis of tissue taken during surgery. Analyzing the tissue, a doctor composes a report which not only concludes that the patient has a cancer (+1) or benign diagnosis (-1) but also that the patient belongs to a particular group (has a specific type of cancer or has a specific type of cell and so on). That is, the doctor's classification of the training data y_i^* is more detailed than the desired classification y_i. When constructing a classification rule $y = f(x)$, one can take into account information about y_i^*. This information can be used, for example, to create appropriate groups.

EXAMPLE 2 (Information given in hidden variables).
Suppose that one's goal is to construct a rule $y = f(x)$. However, for the training data along with the nonhidden variables x_i, one can determine the hidden variables x_i^*. The problem is using the data

$$(x_1, x_1^*, y_1), \ldots, (x_\ell, x_\ell^*, y_\ell)$$

which contain both nonhidden and hidden variables and their classifications y_i, to construct a rule $y = f(x)$ (rather than a rule $y = f(x, x^*)$) that makes a prediction based on nonhidden variables. By using variables x for a decision space and variables x, x^* for a correcting space one can solve this problem.

EXAMPLE 3 (Special rule for selected features).
A particular case of the problem described in Example 2 is constructing a decision rule for selected features, using information about the whole set of features. In this problem, the selected features are considered as non hidden variables while the rest of the features are hidden variables.

THE GENERAL PROBLEM
How should one construct (a more accurate than conventional) rule $y = f(x)$ using the data

$$(x_1, x_1^*, y_1, y_1^*), \ldots, (x_\ell, x_\ell^*, y_\ell, y_\ell^*)$$

instead of the data

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

To do this one can use the SVM+ method. Constructing the desired decision rule in the solution space, SVM+ uses two new ideas:

(1) It uses structure on training data and

(2) It uses several different spaces: (a) the solution space of nonhidden variables and (b) the correcting spaces of joint hidden and nonhidden variables.

SVM+ allows one to effectively use additional hidden information. The success of SVM+ depends on the quality of recovered hidden information.

The LHI technology using SVM+ requires the following three steps:

1. Use the data (x_i, x_i^*, y_i, y_i^*) for constructing a structure on the training set.

2. Use the kernel $K(x_i, x_j)$ for constructing a rule in the decision space, and

3. Use the kernels $K_r(x_i, x_i^* y_i^*; x_j, x_j^*, y_j^*)$ in the correcting spaces.

Note that in the SVM+ method the idea of creating a structure on the training set differs from the classical idea of clustering of the training set.

2.5 Generalization for Regression Estimation Problem

In this section we use the ε-insensitive loss function introduced in [140],

$$u_\varepsilon = \begin{cases} |u| - \varepsilon, & \text{if } |u| \geq \varepsilon \\ 0, & \text{if } |u| < \varepsilon. \end{cases}$$

This function allows one to transfer some properties of the SVM for pattern recognition (the accuracy and the sparsity) to the regression problem.

2.5.1 SVM Regression

Consider the regression problem: given iid data

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

where $x \in X$ is a vector and $y \in (-\infty, \infty)$ is a real value, estimate the function in a given set of real-valued functions.

As before using kernel techniques we map input vectors x into the space of image vectors $z \in Z$ and approximate the regression by a linear function

$$y = (w, z) + b, \tag{2.47}$$

where w and b have to be defined. Our goal is to minimize the following loss,

$$R = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} |y_i - (w, z) - b|_\varepsilon. \tag{2.48}$$

To minimize the functional (2.48) we solve the following equivalent problem [140]: Minimize the functional

$$R = \frac{1}{2}(w,w) + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) \tag{2.49}$$

subject to the constraints

$$y_i - (w,z_i) - b \le \varepsilon + \xi_i^*,\ \ \xi_i^* \ge 0, \quad i = 1,\dots,\ell, \tag{2.50}$$

$$(w,z_i) + b - y_i \le \varepsilon + \xi_i,\ \ \xi_i \ge 0, \quad i = 1,\dots,\ell. \tag{2.51}$$

To solve this problem one constructs the Lagrangian

$$L = \frac{1}{2}(w,w) + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) - \sum_{i=1}^{\ell}\alpha_i[y_i - (w,z_i) - b + \varepsilon + \xi_i] \tag{2.52}$$

$$-\sum_{i=1}^{\ell}\alpha_i^*[(w,z_i) + b - y_i + \varepsilon + \xi_i^*] - \sum_{i=1}^{\ell}(\beta_i\xi_i + \beta_i^*\xi_i^*)$$

whose minimum over w, b, and $\xi,\ \xi_i^*$ leads to the equations

$$w = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)z_i, \tag{2.53}$$

$$\sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0, \tag{2.54}$$

and

$$\alpha_i^* + \beta_i^* = C,\ \ \alpha_i + \beta_i = C, \tag{2.55}$$

where $\alpha,\ \alpha^*, \beta, \beta^* \ge 0$ are the Lagrange multipliers. Putting (2.53) into (2.47) we obtain that in X space the desired function has the kernel form

$$y = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)K(x_i,x) + b. \tag{2.56}$$

To find the Lagrange multipliers one has to put the obtained equation back into the Lagrangian and maximize the obtained expression.

Putting (2.53), (2.54), and (2.55) back into (2.52) we obtain

$$W = -\sum_{i=1}^{\ell}\varepsilon(\alpha_i^*+\alpha_i) + \sum_{i=1}^{\ell}y_i(\alpha_i^*-\alpha_i) - \frac{1}{2}\sum_{i,j}^{\ell}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j). \tag{2.57}$$

To find $\alpha_i,\ \alpha_i^*$ for the approximation (2.56) one has to maximize this functional subject to the constraints

$$\sum_{i=1}^{\ell}\alpha_i^* = \sum_{i=1}^{\ell}\alpha_i,$$

$$0 \le \alpha_i \le C, \quad 0 \le \alpha_i^* \le C, \quad i = 1,\dots,\ell.$$

2.5.2 SVM+ Regression

Now let us solve the same regression problem of minimizing the functional (2.49) subject to the constraints (2.50) and (2.51) in the situation when the slacks ξ_i and ξ_i^* are defined by functions from the set described in Section 2.4:

$$\xi_i = \phi_r(x_i, w_r) = (w_r, z_i) - d_r \geq 0,\ i \in T_r,\ \ r = 1, \ldots, t \tag{2.58}$$

$$\xi_i^* = \phi_r^*(x_i, w_r^*) = (w_r^*, z_i) - d_r^* \geq 0,\ i \in T_r,\ \ r = 1, \ldots, t. \tag{2.59}$$

To find the regression we construct the Lagrangian similar to (2.52) where instead of slacks ξ_i and ξ_i^* we use their expressions (2.58) and (2.59).

Minimizing this Lagrangian over w, b (as before) and over $w_r,\ d_r, w_r^*,\ d_r^*,\ r = 1, ..., t$ (instead of slacks ξ_i, and ξ_i^*) we obtain Equations (2.53) and (2.54) and the equations

$$\sum_{i \in T_r} (\alpha_i + \beta_i) z_i^r = C \sum_{i \in T_r} z_i^r,\ \sum_{i \in T_r} (\alpha_i^* + \beta_i^*) z_i^r = C \sum_{i \in T_r} z_i^r,\ r = 1, \ldots, t, \tag{2.60}$$

$$\sum_{i \in T_r} (\alpha_i^* + \beta_i^*) = C|T_r|,\ \sum_{i \in T_r} (\alpha_i + \beta_i) = C|T_r|,\ r = 1, \ldots, t \tag{2.61}$$

Putting these equations back into the Lagrangian we obtain

$$W = -\sum_{i=1}^{\ell} \varepsilon(\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i,j}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j). \tag{2.62}$$

From (2.60) and (2.61) we obtain

$$\sum_{i \in T_r} (\alpha_i + \beta_i) K_r(x_j, x_j) = C \sum_{i \in T_r} K_r(x_i, x_j),\ r = 1, \ldots, t,\ j \in T_r, \tag{2.63}$$

$$\sum_{i \in T_r} (\alpha_i^* + \beta_i^*) K_r(x_i, x_j) = C \sum_{i \in T_r} K_r(x_i, x_j), \quad r = 1, ..., t\ j \in T_r, \tag{2.64}$$

$$\alpha_i \geq 0,\ \ \beta_i \geq 0, \quad i = 1, \ldots, \ell.$$

Therefore to estimate the SVM+ regression function (2.56) one has to maximize the functional (2.62) subject to the constraints (2.54), (2.61), (2.63), (2.64).

2.5.3 SVM$_\gamma$+ Regression

Consider SVM$_\gamma$+ extension of regression estimation problem: Minimize the functional

$$R = \frac{1}{2}(w, w) + \frac{\gamma}{2} \left(\sum_{r=1}^{t} (w_r, w_r) + \sum_{r=1}^{t} (w_r^*, w_r^*) \right) + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \tag{2.65}$$

(instead of functional (2.49)) subject to constraints (2.50) and (2.51), where slacks ξ_i and ξ_i^* are defined by the correcting functions (2.58) and (2.59). The new objective function approaches (2.49) when γ approaches zero.

The same algebra of the Lagrange multiplier technique that was used above now implies that to find the coefficients $\alpha_i,\ \alpha_i^*$ for approximation (2.56) one has to maximize the functional

$$W = -\sum_{i=1}^{\ell} \varepsilon(\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) +$$

$$\frac{C}{\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \beta_i)K_r(x_i, x_j) - \frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \beta_i)(\alpha_j + \beta_j)K_r(x_i, x_j) +$$

$$\frac{C}{\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i^* + \beta_i^*)K_r(x_i, x_j) - \frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i^* + \beta_i^*)(\alpha_j^* + \beta_j^*)K_r(x_i, x_j)$$

subject to the constraints

$$\sum_{i=1}^{\ell}\alpha_i^* = \sum_{i=1}^{\ell}\alpha_i,$$

$$\sum_{i\in T_r}(\alpha_i + \beta_i) = |T_r|C,\ r = 1, \ldots, t,$$

$$\sum_{i\in T_r}(\alpha_i^* + \beta_i^*) = |T_r|C,\ r = 1, \ldots, t,$$

$$\alpha_i \geq 0,\ \ \alpha_i^* \geq 0,\ \ \beta_i \geq 0,\ \ \beta_i^* \geq 0,\quad i = 1, ..., \ell.$$

When either (1) there is no structure ($t = \ell$) or (2) there are no correlations ($K_r(x_i, x_j)$ has the form (2.45)) and $\gamma \to 0$ the solutions defined by SVM+ or SVM$_\gamma$+ regression coincide with the conventional SVM solution for regression.

2.6 The Third Generation

In the mid-1990s the third generation of statistical learning theory (SLT) researchers appeared. They were well-educated, strongly motivated, and hard working PhD students from Europe. Many European universities allow their PhD students to work on their theses anywhere in the world, and several such students joined our department in order to work on their thesis. First came Bernhard Schölkopf, Volker Blanz, and Alex Smola from Germany, then Jason Weston from England, followed by Olivier Chapelle, Olivier Bousquet, and Andre Eliseeff from France, Pascal Vincent from Canada, and Corina Cortes (PhD student from Rochester university). At that time support vector technology had just started to develop. Later many talented young people followed this direction but these were the first from the third generation of researchers.

I would like to add to this group two young AT&T researchers of that time: Yoav Freund and Robert Schapire, who did not directly follow the line of statistical learning theory and developed *boosting* technology that is close to the one discussed here [135, 136].

The third generation transformed both the area of machine learning research and the style of research. During a short period of time (less than ten years) they created a new direction in statistical learning theory: SVM and kernel methods. The format of this Afterword does not allow me to go into details of their work (there are hundreds of first-class articles devoted to this subject and it is very difficult to choose from them). I will just quote some of their textbooks [152–158], collective monographs and workshop materials [159–164]. Also I would like to mention the tutorial by Burges [165] which demonstrated the unity of theoretical and algorithmical parts of VC theory in a simple and convincing way.

The important achievement of the third generation was creating a large international SVM (kernel) community. They did it by accomplishing three things:

(1) Constructing and supporting a special Website called Kernel Machine (www.kernel-machines.org).

(2) Organizing eight machine learning workshops and five Summer Schools, where advanced topics relevant to empirical inference research were taught. These topics included:
— Statistical learning theory,
— Theory of empirical processes,
— Functional analysis,
— Theory of approximation,
— Optimization theory, and
— Machine learning algorithms.
In fact they created the curriculum for a new discipline: *Empirical Inference Science*.

(3) Developing high-quality professional software for empirical inference problems that can be downloaded and used by anyone in the world.[5]

This generation took advantage of computer technology to change forever the style and atmosphere of data mining research: from the very hierarchical group structure of the 1970–1980s lead by old statistical gurus (with their *know-how* and dominating opinion) to an open new society (with widely available information, free technical tools, and open professional discussions).

Many of the third generation researchers of SLT became university professors. This Afterword is dedicated to their students.

[5]The three most popular software are:
(1) SVM-$Light$ developed by Thorsten Joachims (Germany) http://svmlight.joachims.org/,
(2) Lib-SVM developed by Chin-Chang Chang and Chih-Jen Lin (Taiwan) http://www.csie.ntu.edu.tw/∼cjlin/libsvm/, and
(3) SVM-$Torch$ developed by Ronan Collobert (Switzerland) http://www.torch.ch/

2.7 Relation to the Philosophy of Science

By the end of the 1990s it became clear that there were strong ties between machine learning research and research conducted in the classical philosophy of induction. The problem of generalization (induction) always was one of the central problems in philosophy. Pattern recognition can be considered as the simplest problem of generalization (its drosophila fly: any idea of generalization has its reflection in this model). It forms a very good object for analysis and verification of a general inductive principle. Such analysis includes not only speculations but also experiments on computers.

Two main principles of induction were introduced in classical philosophy: the principle of simplicity (parsimony) formulated by the 14th century English monk Occam (Ocham), and the principle of falsifiability, formulated by the Austrian philosopher of the 20th century Karl Popper. Both of them have a direct reflection in statistical learning theory.

2.7.1 Occam's Razor Principle

The Occam's Razor (or parsimony) principle was formulated as follows:

Entities are not to be multiplied beyond necessity.

Such a formulation leaves two open questions:

(1) What are the *entities*?

(2) What does *beyond necessity* mean?

According to *The Concise Oxford Dictionary of Current English* [172] the word *entity* means

> A thing's existence, as opposite to its qualities or relations; thing that has real existence.

So the number of entities is commonly understood to be the number of different parameters related to different physical (that which can be measured) features. The predictive rule is a function defined by these features.

The expression *not to be multiplied beyond necessity* has the following meaning: *not more than one needs to explain the observed facts*.

In accordance with such an interpretations the Occam's Razor principle can be reformulated as follows:

> *Find the function from the set with the smallest number of free parameters that explains the observed facts.*[6]

[6]There exist wide interpretation of Occam's Razor principle as a request to minimize some functional (without specifying which). Such interpretation is too general to be useful since it depends on the definition of the functional. The original Occam formulation (assuming that entities are free parameters) is unambiguous and in many cases is a useful instrument of inference.

2.7.2 Principles of Falsifiability

To introduce the principles of falsifiability we need some definitions.

Suppose we are given a set of indicator functions $f(x, \alpha), \alpha \in \Lambda$. We say that the set of vectors

$$x_1, \ldots, x_\ell, \ x_i \in X \tag{2.66}$$

cannot falsify the set of indicator functions $f(x, \alpha), \alpha \in \Lambda$ if all 2^ℓ possible separation of vectors (2.66) into two categories can be accomplished using functions from this set.

This means that on the data (2.66) one can obtain any classification (using functions from the admissible set). In other words, from these vectors one can obtain any possible law (given appropriate $y_i, \ i = 1, ...\ell$): the vectors themselves do not forbid (do not falsify) any possible law.

We say that the set of vectors (2.66) *falsifies* the set $f(x, \alpha), \alpha \in \Lambda$ if there exists such separation of the set (2.66) into two categories that cannot be obtained using an indicator function from the set $f(x, \alpha), \alpha \in \Lambda$.

Using the concept of falsifiability of a given set of functions by the given set of vectors, two different combinatorial definitions of the dimension of a given set of indicator functions were suggested: the VC dimension and the Popper dimension. These definitions lead to different concepts of falsifiability.

The Definition of the VC Dimension and VC Falsifiability

The VC dimension is defined as follows (in *EDBED* it is called capacity. See Chapter 6, Sections 8 and A2:)

A set of functions $f(x, \alpha), \alpha \in \Lambda$ has VC dimension h if:

(1) **there exist** h vectors that cannot falsify this set and

(2) **any** $h + 1$ vectors falsify it.

The set of functions $f(x, \alpha), \alpha \in \Lambda$ is *VC falsifiable* if its VC dimension is finite and *VC nonfalsifiable* if its VC dimension is infinite.

The VC dimension of the set of hyperplanes in R^n is $n + 1$ (the number of free parameters of a hyperplane in R^n) since there exist $n + 1$ vectors that cannot falsify this set but any $n + 2$ vectors falsify it.

The Definition of the Popper Dimension and Popper Falsifiability

The Popper dimension is defined as follows [137, Section 38]

A set of functions $f(x, \alpha), \alpha \in \Lambda$ has the Popper dimension h if:

(1) **any** h vectors cannot falsify it and

(2) **there exist** $h + 1$ vectors that can falsify this set.

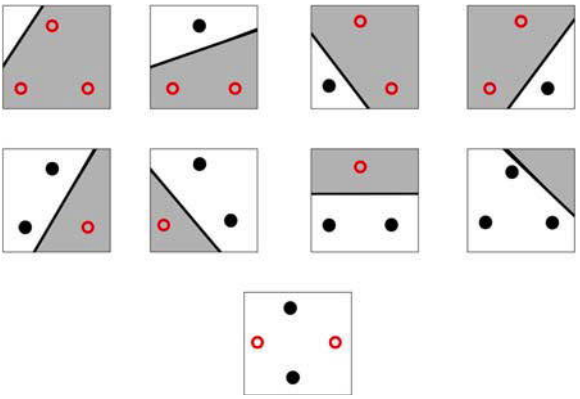

Figure 2.2: The VC dimension of the set of oriented lines in the plane is three since there exist three vectors that cannot falsify this set and any four vectors falsify it.

Popper called value h the degree of falsifiability or the dimension.

The set of functions $f(x, \alpha), \alpha \in \Lambda$ is *Popper falsifiable* if its Popper dimension is finite and *Popper nonfalsifiable* if its Popper dimension is infinite.

Popper's dimension of the set of hyperplanes in R^n is at most two (independent of the dimensionality of the space n) since only two vectors that belong to the one-dimensional linear manifold can not falsify the set of hyperplanes in R^n and three vectors from this manifold falsify this set.

2.7.3 Popper's Mistakes

In contrast to the VC dimension, the Popper concept of dimensionality does not lead to useful theoretical results for the pattern recognition model of generalization. The requirements of nonfalsifiability for any h vectors include, for example, the nonfalsifiability of vectors belonging to the line (one-dimensional manifold). Therefore, Popper's dimension will be defined by combinatorial properties restricted at most by the one-dimensional situation.

Discussing the concept of simplicity, Popper made several incorrect mathematical claims. This is the most crucial:

> In an algebraic representation, the dimension of a set of curves depends upon the number of parameters whose value we can freely choose. We can therefore say that the number of freely determinable parameters of a set of curves by which a theory is represented is characteristic of the degree of falsifiability. [137, Section 43]

This is wrong for the Popper dimension. The claim is correct only in a restricted situation for the VC dimension, namely when the set of functions in $R^n, n > 2$ *linearly depends on the parameters*.

In other (more interesting) situations as in Denker's example with a set of $\theta(\{\sin ax\})$ functions (Section 2.2.1 and Section 2.7.5 below) and in the example of a separating hyperplane with the margin given in *EDBED* (Chapter 10, Section 5) that led to SVM technology, the considered set of functions depends nonlinearly upon the free parameters.

Popper did not distinguish the type of dependency on the parameters. Therefore he claimed that the set $\{\theta(\sin ax)\}$ (with only one free parameter a) is a simple set of functions [137, Section 44]. However, the VC dimension of this set is infinite[7] and therefore generalization using this set of functions is impossible.

It is surprising that the mathematical correctness of Popper's claims has never been discussed in the literature.[8]

2.7.4 Principle of VC Falsifiability

In terms of the philosophy of science, the structural risk minimization principle for the structure organized by the nested set with increasing VC dimension can be reformulated as follows:

Explain the facts using the function from the set that is easiest to falsify.

The mathematical consistency of SRM therefore can have the following philosophical interpretation:

> Since one was able to find the function that separates the training data well, in the set of functions that is easy to falsify, these data are very special and the function which one chooses reflects the intrinsic properties of these data.[9].

It is possible, however, to organize the structure of nested elements on which capacity is defined by a more advanced measure than VC dimension (say, the Growth

[7]Since for any ℓ the set of values $x_1 = 2^{-1}, \ldots, x_\ell = 2^{-\ell}$ cannot falsify $\{\theta(\sin ax)\}$. The desired classifications $y_1, \ldots, y_\ell,\ y_i \in \{1, -1\}$ of this set provide the function $y = \theta(\sin a^* x)$ where the coefficient a^* is

$$a^* = \left(\pi \sum_{i=1}^{\ell} \frac{(1 - y_i)}{2} 2^i + 1 \right).$$

[8]Karl Popper's books were forbidden in the Soviet Union because of his criticism of communism. Therefore, I had no chance to learn about his philosophy until Gorbachev's time. In 1987 I attended a lecture on Popper's philosophy of science and learned about the falsifiability concept. After this lecture I became convinced that Popper described the VC dimension. (It was hard to imagine such a mistake.) Therefore in my 1995 and 1998 books I wrongly referred to Popper falsifiability as VC falsifiability. Only in the Spring of 2005 in the process of writing a philosophical article (see Corfield, Schölkopf, and Vapnik: "Popper, falsification and the VC dimension." Technical Report # 145, Max Planck Institute for Biological Cybernetics, Tübingen, 2005) did we check Popper's statements and realize my mistake.

[9]The *Minimum Message Length (MML)–Minimum Description Length* (MDL) principle [127, 128] that takes Kolmogorov's *algorithmic complexity* [129] into account can have the same interpretation. It is remarkable that even though the concepts of VC dimension and algorithmic complexity are very different, the MML-MDL principle leads to the same generalization bound for the pattern recognition problem that is given in *EDBED*. (See [139], Chapter 4, Section 4.6.)

function, or even better the VC entropy). This can lead to more advanced inference techniques (see Section 2.8 of this chapter).

Therefore the falsifiability principle is closely related to the VC dimension concept and can be improved by more refined capacity concepts.

2.7.5 Principle of Parsimony and VC Falsifiability

The principle of simplicity was introduced as a principle of parsimony or a principle of economy of thought.

The definition of simplicity, however, is crucial since it can be very different. Here is an example. Which set of functions is simpler:

(1) One that has the parametric form

$$f(x, \alpha),\ \alpha \in \Lambda,\ \text{ or}$$

(2) One that has the parametric form

$$f(x, \alpha),\ \alpha \in \Lambda$$

and satisfies the constraint

$$\Omega(f) \leq C,$$

where $\Omega(f) \geq 0$ is some functional?

From a computational point of view, finding the desired function in situation 1 can be much simpler than in situation 2 (especially if the $\Omega(f) \leq C$ is a nonconvex set).

From an information theory point of view, however, to find the solution in situation 2 is simpler, since one is looking for the solution in a more restricted set of functions.

Therefore the inductive principle based on the (intuitive) idea of simplicity can lead to a contradiction. That is why Popper used the "degree of falsifiability" concept (Popper dimension) to characterize the simplicity:

> The epistemological question which arise in connection with the concept of simplicity can all be answered if we equate this concept with degree of falsifiability. ([137], Section 43)

In the Occam's Razor principle, the number of "entities" defines the simplicity. Popper incorrectly claimed the equality of Popper dimension to be the number of free parameters (entities), and considered the falsifiability principle to be a justification of the parsimony (Occam's Razor) principle.

The principle of VC falsifiability does not coincide with the Occam's Razor principle of induction, and this principle (but not Occam's Razor) guarantee the generalization. VC dimension describes diversity of the set of functions. It does not refer either to the number of free parameters nor to our intuition of simplicity. Recall once again that Popper (and many other philosophers) had the intuition that $\{\theta(\sin ax)\}$ is the simple set of functions,[10] while the VC dimension of this set is infinite.

[10] In the beginning of Section 44 [137] Popper wrote: "According to common opinion the sine-function is a simple one . . .."

The principle of VC falsifiability forms the necessary and sufficient conditions of consistency for the pattern recognition problem while there are pattern recognition algorithms that contradict the parsimony principle.[11]

2.8 Inductive Inference Based on Contradictions

In my 1998 book, I discussed an idea of inference through contradictions [140, p.707]. In this Afterword, I introduce this idea as an algorithm for SVM. Sections 2.8.1 and 3.1.5 give the details of the algorithm. This section presents a simplified description of the general concept (see remark in Section 3.1.3 for details) of inductive inference through contradictions.

Suppose we are given a set of admissible indicator functions $f(x, \alpha), \alpha \in \Lambda$ and the training data. The vectors x from the training data split our admissible set of functions into a finite number of equivalence classes $F_1, \ldots, F_N$. The equivalence class contains functions that have the same values on the training vectors x (separate them in the same way).

Suppose we would like to make a structure on the set of equivalence classes to perform SRM principle. That is, we would like to collect some equivalence classes in the first element of the structure, then add to them some other equivalence classes, constructing the second element, and so on. To do this we need to characterize every equivalence class by some value that describes our preference for it. Using such a measure, one can create the desired structure on the equivalence classes. When we constructed SVMs, we characterized the equivalence class by the size of the largest margin defined by the hyperplane belonging to this class.

Now let us consider a different characteristic. Suppose along with the training data we possess a set of vectors called *the Universum* or *the Virtual Universum*

$$x_1^*, \ldots, x_k^*, \quad x^* \in X. \tag{2.67}$$

The Universum plays the role of prior information in Bayesian inference. It describes our knowledge of the problem we are solving. However, there are important differences between the prior information in Bayesian inference and the prior information given by the Universum. In Bayesian inference, prior information is information about the relationship of the functions in the set of admissible functions to the desired one. With the Universum, prior information is information related to possible training and test vectors. For example, in the digit recognition problem it can be some vectors whose

[11] The example of a machine learning algorithm that contradicts the parsimony principle is *boosting*. This algorithm constructs so-called weak features (entities) which it linearly combines in a decision rule. Often this algorithm constructs some set of weak features and the corresponding decision rule that separates the training data with no mistakes but continues to add new weak features (new entities) to construct a better rule. With an increasing number of (unnecessary, i.e., those that have no effect on separating the training data) weak features, the algorithm improves its performance on the test data. One can show that with an increasing number of entities this algorithm increases the margin (as the SVM). The idea of this algorithm is *to increase the number of entities (number of free parameters) in order to decrease the VC dimension* [136].

images resemble a particular digit (say some artificial characters). It defines the style of the digit recognition task, and geometrically belongs to the same part of input space to which the training data belong.

We use the Universum to characterize the equivalence class. We say that a vector x^* is contradictive for the equivalence class F_s if there exists a function $f_1(x^*) \in F_s$ such that

$$f_1(x^*) > 0$$

and there also exists a function $f_2(x^*) \in F_s$ such that

$$f_2(x^*) < 0.$$

We will characterize our preference for an equivalence class by the number of contradictions that occur on the Universum: the more contradictions, the more preferable the equivalence class.[12] We construct structure on equivalence classes using these numbers.

When using the Universum to solve a classification problem based on SRM principle, we choose the function (say one that has the maximal margin) from the equivalence class that makes no (or a small number of) training mistakes and has the maximal number of contradictions on the Universum. In other words, for inductive inference, when constructing the structure for SRM, we replace the *maximal margin* score with the *maximal contradiction on Universum* (MCU) score and select maximal margin function from the choosen equivalence class.

The main problem with MCU inference is, how does one create the appropriate Universum? Note that since one uses Universum only for evaluation of sizes of equivalence classes, its elements do not need to have the same distribution as the training vectors.

2.8.1 SVMs in the Universum Environment

The inference through contradictions can be implemented using SVM techniques as follows. Let us map both the training data and the Universum into Hilbert space

$$(y_1, z_1), \ldots, (y_\ell, z_\ell) \tag{2.68}$$

$$z_1^*, \ldots, z_u^*. \tag{2.69}$$

Quadratic Optimization Framework

In the quadratic optimization framework for an SVM, to conduct inference through contradictions means finding the hyperplane

$$(w^0, z) + b_0 = 0 \tag{2.70}$$

[12] A more interesting characteristic of an equivalence class would be the value of the VC entropy of the set of functions belonging to this equivalence class calculated on the Universum. This, however, leads to difficult computational problems. The number of contradictions can be seen as a characteristic of the entropy.

that minimizes the functional

$$R = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=1}^{u} \theta(\xi_j^*),\ C_1, C_2 > 0 \tag{2.71}$$

subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0,\ i = 1, \ldots, \ell \tag{2.72}$$

(related to the training data) and the constraints

$$|(w, z_j^*) + b| \leq a + \xi_j^*,\ \ \xi_j^* \geq 0,\ j = 1, \ldots, u \tag{2.73}$$

(related to the Universum) where $a \geq 0$.

As before, for computational reasons we approximate the target function (2.71) by the function[13]

$$R = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{s=1}^{u} \xi_s^*,\ C_1, C_2 > 0. \tag{2.74}$$

Using the Lagrange multipliers technique we determine that our hyperplane in feature space has the form

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i (z_i, z) + \sum_{s=1}^{u} (\mu_s^0 - \nu_s^0)(z_s^*, z) + b = 0, \tag{2.75}$$

where the coefficients $\alpha_i^0 \geq 0,\ \mu_s^0 \geq 0$, and $\nu_s^0 \geq 0$ are the solution of the following optimization problem: Maximize the functional

$$W(\alpha, \mu, \nu) = \sum_{i=1}^{\ell} \alpha_i - a \sum_{s=1}^{u} (\mu_s + \nu_s) - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (z_i, z_j) \tag{2.76}$$

$$- \sum_{i=1}^{\ell} \sum_{s=1}^{u} \alpha_i y_i (\mu_s - \nu_s)(z_i, z_s^*) - \frac{1}{2} \sum_{s,t=1}^{u} (\mu_s - \nu_s)(\mu_t - \nu_t)(z_s^*, z_t^*)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i + \sum_{s=1}^{u} (\mu_s - \nu_s) = 0 \tag{2.77}$$

and the constraints

$$0 \leq \alpha_i \leq C_1 \tag{2.78}$$

$$0 \leq \mu_s, \nu_s \leq C_2. \tag{2.79}$$

[13] One also can use a least squares technique by choosing ξ_i^2 and $(\xi_i^*)^2$ instead of ξ_i and ξ_i^* in objective function (2.74).

Taking into account Mercer's theorem, one can rewrite our separating function in input space as

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x) + \sum_{s=1}^{u} (\mu_s^0 - \nu_s^0) K(x_s^*, x) + b_0 = 0, \tag{2.80}$$

where the coefficients $\alpha_i^0 \geq 0,\ \mu_s^0 \geq 0$, and $\nu_s^0 \geq 0$ are the solution of the following optimization problem: Maximize the functional

$$W(\alpha, \mu, \nu) = \sum_{i=1}^{\ell} \alpha_i - a \sum_{s=1}^{u} (\mu_s + \nu_s) - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.81}$$

$$- \sum_{i=1}^{\ell} \sum_{s=1}^{u} \alpha_i y_i (\mu_s - \nu_s) K(x_i, x_s^*) - \frac{1}{2} \sum_{s,t=1}^{u} (\mu_s - \nu_s)(\mu_t - \nu_t) K(x_s^*, x_t^*)$$

subject to the constraints (2.77), (2.78), (2.79).

Linear Optimization Framework

To conduct inference based only on contradictions arguments (taking some function from the choosen equivalence class, not necessarily one with the largest margin) one has to find the coefficients $\alpha^0,\ \mu^0,\ \nu^0$ in (2.80) using the following linear programming technique: Minimize the functional

$$W(\alpha, \mu, \nu) = \gamma \sum_{i=1}^{\ell} \alpha_i + \gamma \sum_{s=1}^{u} (\mu_s + \nu_s) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{t=1}^{u} \xi_t^*,\ \gamma \geq 0 \tag{2.82}$$

subject to the constraints

$$y_i \left[\sum_{j=1}^{\ell} \alpha_j y_j K(x_i, x_j) + \sum_{s=1}^{u} (\mu_s - \nu_s) K(x_i, x_s^*) + b \right] \geq 1 - \xi_i,\ i = 1, \ldots, \ell \tag{2.83}$$

and the constraints

$$\sum_{j=1}^{\ell} \alpha_j y_j K(x_t^*, x_j) + \sum_{s=1}^{u} (\mu_s - \nu_s) K(x_t^*, x_s^*) + b \leq a + \xi_t^*,\ t = 1, \ldots, k, \tag{2.84}$$

$$\sum_{j=1}^{\ell} \alpha_j y_j K(x_t^*, x_j) + \sum_{s=1}^{u} (\mu_s - \nu_s) K(x_t^*, x_s^*) + b \geq -a - \xi_t^*,\ t = 1, \ldots, u, \tag{2.85}$$

where $a \geq 0$. In the functional (2.82) the parameter $\gamma \geq 0$ controls the sparsity of the solution.

2.8.2 The First Experiments and General Speculations

In the summer of 2005, Ronan Collobert and Jason Weston conducted the first experiments on training SVM with Universum using the algorithm described in Section 2.8.1. They discriminated digit 8 from digit 5 from the MNIST database, using a conventional SVM and an SVM trained in three different Universum environments.

The following table shows for different sizes of training sets the performance of a conventional SVM and the SVMs trained using Universums U_1, U_2, U_3 (each containing 5000 examples). In all cases the parameter $a = .01$, the parameters C_1, C_2, and the parameter of the Gaussian kernel were tuned using the tenfold cross-validation technique.

The Universums were constructed as follows:

U_1: Selects random digits from the other classes (0,1,2,3,4,6,7,9).

U_2: Creates an artificial image by first selecting a random 5 and a random 8, (from pool of 3,000 non-test examples) and then for each pixel of the artificial image choosing with probability 1/2 the corresponding pixel from the image 5 or from the image 8.

U_3: Creates an artificial image by first selecting a random 5 and a random 8, (from pool of 3,000 non-test examples) and then constructing the mean of these two digits.

No. of train. examples	250	500	1000	2000	3000
Test Err. SVM (%)	2.83	1.92	1.37	0.99	0.83
Test Err. SVM+U_1 (%)	2.43	1.58	1.11	0.75	0.63
Test Err. SVM+U_2 (%)	1.51	1.12	0.89	0.68	0.60
Test Err. SVM+U_3 (%)	1.33	0.89	0.72	0.60	0.58

The table shows that:
(a) The Universum can significantly improve the performance of SVMs.
(b) The role of knowledge provided by the Universum becomes more important with decreasing training size. However, even when the training size is large, the Universum still has a significant effect on performance.

We expect that advancing the understanding of the idea how to create a good Universum for the problem of interest will further boost the performance. This fact opens a new dimension in machine learning technology: How does one create a Virtual Universum for the problem of interest?

In trying to find an interpretation of the role of the Universum in machine learning, it is natural to compare it to the role of culture in the learning of humans, where knowledge about real life is concentrated not only in examples of reality but also in artificial images that reflect this reality. To classify well, one uses inspiration from both sources.

Chapter 3

NONINDUCTIVE METHODS OF INFERENCE: DIRECT INFERENCE INSTEAD OF GENERALIZATION (2000–· · ·)

3.1 INDUCTIVE AND TRANSDUCTIVE INFERENCE

Chapter 10 of *EDBED* distinguishes between two different problems of estimation: estimation of the function and estimation of the values of the function at given points of interest.

(1) *Estimation of the function*. Given training data

$$(y_1, x_1), \ldots, (y_\ell, x_\ell), \tag{3.1}$$

find in the set of admissible functions $f(x, \alpha), \alpha \in \Lambda$ the one which guarantees that its expected loss is close to the smallest loss.

(2) *Estimation of the value of the function at the points of interest*. Given a set of training data (3.1) and a sequence of k test vectors

$$x_{\ell+1}, \ldots, x_{\ell+k}, \tag{3.2}$$

find among an admissible set of binary vectors

$$\{Y_* = (y^*_{\ell+1}, \ldots, y^*_{\ell+k})\}$$

the one that classifies the test vectors with the smallest number of errors. Here we consider

$$x_1, \ldots, x_{\ell+k} \tag{3.3}$$

to be random i.i.d. vectors drawn according to the same (unknown) distribution $P(x)$. The classifications y of the vectors x are defined by some (unknown) conditional probability function $P(y|x)$.

This setting is quite general. In the book we considered a particular setting where the set of admissible vectors is defined by the admissible set of indicator functions $f(x, \alpha),\ \alpha \in \Lambda$. In other words, every admissible vector of classification Y_* is defined as follows

$$Y_* = (f(x_1, \alpha_*), \ldots, f(x_k, \alpha_*)) .$$

In the mid-1990s (after understanding the relationship between the pattern recognition problem and the philosophy of inference), I changed the technical terminology [139]. That is, I called the problem of function estimation that requires one to find a function given particular data *inductive inference*. I called the problem of estimating the values of the function at particular points of interest given the observations *transductive inference*.

These two different ideas of inference reflect two different philosophies, which we will discuss next.

3.1.1 Transductive Inference and the Symmetrization Lemma

The mechanism that provides an advantage to the transductive mode of inference over the inductive mode was clear from the very beginning of statistical learning theory. It can be seen in the proof of the very basic theorems on uniform convergence. This proof is based on the following inequality which is the content of the so-called symmetrization lemma (see Basic lemma in *EDBED* Chapter 6, Section A3):

$$P\left\{\sup_{\alpha} |R(\alpha) - R_{emp}(\alpha)| \geq \varepsilon\right\} \leq 2P\left\{\sup_{\alpha} \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| \geq \frac{\varepsilon}{2}\right\}, \quad (3.4)$$

where

$$R^{(1)}_{emp}(\alpha) = \frac{1}{\ell}\sum_{i=1}^{\ell} |y_i - f(x_i, \alpha)| \quad (3.5)$$

and

$$R^{(2)}_{emp}(\alpha) = \frac{1}{\ell}\sum_{i=\ell+1}^{2\ell} |y_i - f(x_i, \alpha)| \quad (3.6)$$

are the empirical risks constructed using two different samples.

The bound for uniform convergence was obtained as an upper bound of the right-hand side of (3.4).

Therefore the symmetrization lemma implies that to obtain a bound for inductive inference we first obtain a bound for transductive inference (for the right-hand side of (3.4)) and then obtain an upper bound for it.

It should be noted that since the bound on uniform convergence was introduced in 1968, many efforts were made to improve it. However, all attempts maintain some form of the symmetrization lemma. That is, in the proofs of the bounds for uniform convergence the first (and most difficult) step was to obtain the bound for transductive inference. The trivial upper bound of this bound gives the desired result for inductive inference.

This means that transductive inference is a fundamental step in machine learning.

3.1.2 Structural Risk Minimization for Transductive Inference

The proof of the symmetrization lemma is based on the following observation: The following two models are equivalent (see Chapter 10, Section 1 of EDBED):

(a) one chooses two i.i.d. sets[1]

$$x_1, \ldots, x_\ell, \quad \text{and} \quad x_{\ell+1}, \ldots, x_{2\ell};$$

(b) one chooses an i.i.d. set of size 2ℓ and then randomly splits it into two subsets of size ℓ.

Using model (b) one can rewrite the right-hand side of (3.4) as follows

$$P\left\{\sup_\alpha \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \frac{\varepsilon}{2}\right\} =$$

$$E_{\{x_1,\ldots,x_{2\ell}\}} P\left\{\sup_\alpha |R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)| > \frac{\varepsilon}{2} \ |\{x_1, \ldots, x_{2\ell}\}\right\}. \tag{3.7}$$

To obtain the bound we first bound the conditional probability

$$P\left\{\sup_\alpha |R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)| > \frac{\varepsilon}{2} \ |\{x_1, \ldots, x_{2\ell}\}\right\} \leq$$

$$\Delta^\Lambda(x_1, \ldots, x_{2\ell}) \exp\left\{-\varepsilon^2 \ell\right\} \tag{3.8}$$

where $\Delta^\Lambda(x_1, ..., x_{2\ell})$ is the number of equivalence classes on the set (3.3). The probability is obtained with respect to the random split data into two parts (training and testing). Then we take the expectation over working sets of size 2ℓ. As a result, we obtain

$$E_{\{x_1,\ldots,x_{2\ell}\}} P\left\{\sup_\alpha \left|R^{(1)}_{emp}(\alpha) - R^{(2)}_{emp}(\alpha)\right| > \frac{\varepsilon}{2} |\{x_1, \ldots, x_{2\ell}\}\right\} \leq$$

$$\Delta^\Lambda_P(2\ell) \exp\left\{-\varepsilon^2 \ell\right\}. \tag{3.9}$$

Note that for the transductive model of inference we do not even need to take the expectation over sets of size 2ℓ. We can just use the bounds (3.8).

[1] For simplicity of the formulas we choose two sets of equal size.

Let us consider both models of inference, transductive and inductive from one unified point of view: In both cases we are given a set of functions defined on some space R. We randomly choose the training examples from this space. In the inductive case we choose by sampling from the space, and in the transductive case we choose by splitting the working set into the training and testing parts. We define the values of the function of interest over the domain of definition of the function: In the inductive case in the whole space, and in the transductive case on the working set.

The difference is that in transductive inference the space of interest is discrete (defined on $\ell + k$ elements of the working set (3.3)), while in inductive inference it is R^n.

One can conduct a nontrivial analysis of the discrete space but not the continuous space R^n. This is the key advantage of transductive inference.

3.1.3 Large Margin Transductive Inference

Let $F_1, \ldots, F_N$ be the set of equivalence classes defined by the working set (3.3). Our goal is to construct an appropriate structure on this set of equivalence classes.

In Chapter 2, Section 2.6 we constructed a similar structure on the set of equivalence classes for inductive inference. However, we violated one of the important requirements of the theory: The structure must be constructed *before* the training data appear. In fact we constructed it *after* (in the inductive mode of inference the set of equivalence classes was defined by the training data), creating a *data-dependent structure*. There are technical means to justify such an approach. However, the bound for a data-dependent structure will be worse [138].

In transductive inference we construct the set of equivalence classes using a joint working set of vectors that contain both the training and test sets. Since in constructing the equivalence class we do not use information about how our space will be split into training and test sets we do not violate the statistical requirements.

Let us define the size of an equivalence class F_i by the value of the corresponding margin: We find, among the functions belonging to the equivalence class, the one that has the largest margin[2] and use the value of the margin $\mu(F_i)$ as the size of the equivalence class F_i.

Using this concept of the size of an equivalence class, SVM transductive inference suggests:

Classify the test vectors (3.2) by the equivalence class (defined on the working set (3.3)) that classifies the training data well and has the largest value of the (soft) margin.

Formally, this requires us to classify the test data using the rule

$$y_i = \text{sgn}((w_0, z_i) + b_0), \; i = \ell + 1, \ldots, \ell + k,$$

[2]We consider the hard margin setting just for the sake of simplicity. One can easily generalize this setting to the soft margin situation as described in Section 2.3.4.

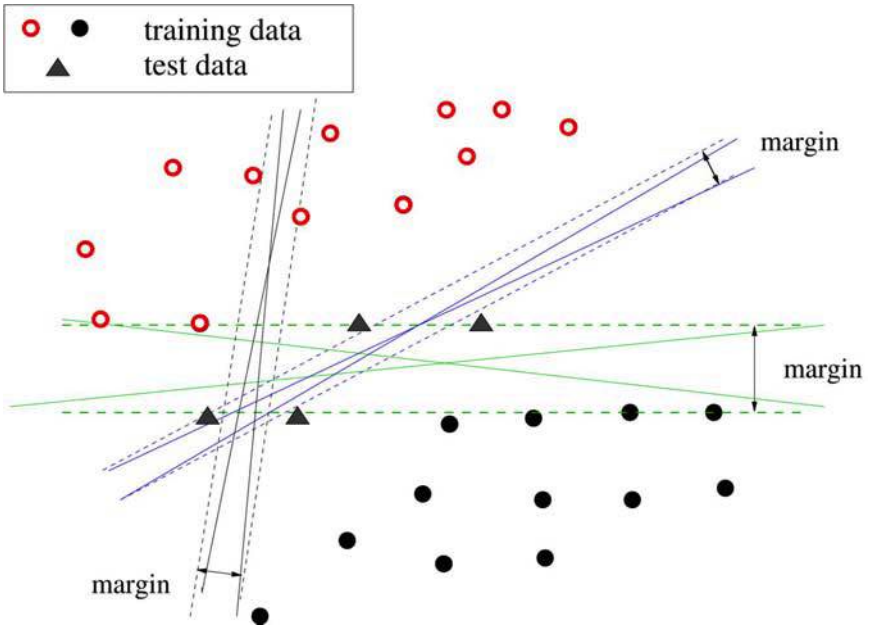

Figure 3.1: Large margin defines a large equivalence class.

where the parameters w_0, b_0 are the ones that minimize the functional

$$R(w) = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=\ell+1}^{\ell+k} \theta(\xi_j),\ C_1, C_2 \geq 0 \tag{3.10}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0,\ i = 1, \ldots, \ell \tag{3.11}$$

(defined by the images of the training data (3.1)) and the constraints

$$y_j^*((z_j, w) + b) \geq 1 - \xi_j, \quad \xi_j \geq 0,\ j = \ell + 1, \ldots, \ell + k \tag{3.12}$$

(defined by the set (3.2) and its desired classification $Y_* = (y_{\ell+1}^*, \ldots, y_{\ell+k}^*)$.

One more constraint. To avoid unbalanced solution Chapelle and Zien [174], following ideas of Joachims [154], suggested the following constraint:

$$\frac{1}{k} \sum_{j=\ell+1}^{\ell+k} ((w, z_j) + b) \approx \frac{1}{\ell} \sum_{i=1}^{\ell} y_i. \tag{3.13}$$

This constraint requires that the proportion of test vectors in the first and second categories be similar to the proportion observed in the training vectors.

For computational reasons we will replace the objective function (3.10) with the function

$$R(w) = \frac{1}{2}(w,w) + C_1 \sum_{i=1}^{\ell} \xi_i, + C_2 \sum_{s=\ell+1}^{\ell+k} \xi_s^*,\ C_1, C_2 \geq 0 \tag{3.14}$$

Therefore (taking into account kernelization based on Mercer's theorem) we can obtain the following solution of this problem (in the dual space).

The classification rules for the test data in the dual space have the form

$$y_\tau = \text{sgn}(\sum_{i=1}^{\ell} \alpha_i^0 K(x_i, x_\tau) + \sum_{s=\ell+1}^{\ell+k} \beta_s y_s^* K(x_s, x) + b_0),\ \tau = \ell+1, \ldots, \ell+k,$$

where the coefficients $\alpha_i^0, \beta_s^0, b_0$ and desired classifications of test data are the solution of the following problem: Maximize (over $\alpha,\ \beta,\ y^*$) the functional

$$W(\alpha, \beta, y^*) = \sum_{i=1}^{\ell} \alpha_i + \sum_{s=\ell+1}^{\ell+k} \beta_s - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$- \sum_{i=1}^{\ell} \sum_{s=\ell+1}^{\ell+k} \alpha_i y_i \beta_s y_s^* K(x_i, x_s) - \frac{1}{2} \sum_{s,t=\ell+1}^{\ell+k} \beta_s y_s^* \beta_t y_t^* K(x_s, x_t)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i + \sum_{s=\ell+1}^{\ell+k} y_s^* \beta_s = 0,$$

the constraints

$$0 \leq \alpha_i \leq C_1,\ i = 1, \ldots, \ell$$

$$0 \leq \beta_s \leq C_2,\ s = \ell+1, \ldots, \ell+k,$$

and the constraint (3.13):

$$\frac{1}{k} \sum_{j=\ell+1}^{\ell+k} \left(\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x_j) + \sum_{t=\ell+1}^{\ell+k} \beta_t y_t^* K(x_t, x_j b_0 \right) \approx \frac{1}{\ell} \sum_{i=1}^{\ell} y_i.$$

Note that this problem does not have a unique solution. This makes transductive inference difficult. However, whenever one can maximize the functional well, one obtains an improvement over inductive SVMs.

3.1.4 Examples of Transductive Inference

Here are examples of real-life problems solved using transductive inference.

1. Prediction of molecular bioactivity for drug discovery [146]. The KDD CUP-2001 competition on data analysis methods required the construction

of a rule for predicting molecular bioactivity using data provided by the DuPont Pharmaceutical company. The data belonged to a binary space of dimension 139,351, which contained a training set of 1909 vectors, and a test set of 634 vectors.

The results are given here for the winner of the competition (among the 119 competitors who used traditional approaches), SVM inductive inference and SVM transductive inference.

Winner's accuracy	68.1 %
SVM inductive mode accuracy	74.5 %
SVM transductive mode accuracy	82.3 %

It is remarkable that the jump in performance obtained due to a new philosophy of inference (transductive instead of inductive) was larger than the jump resulting from the reinforcement of the technology in the construction of inductive predictive rules.

2. TEXT CATEGORIZATION [138]. In a text categorization problem, using transductive inference instead of inductive inference reduced the error rate from 30% to 15%.

REMARK. The discovery of transductive inference and its advantages over inductive inference is not just a technical achievement, but a breakthrough in the philosophy of generalization.

Until now, the traditional method of inference was the *inductive–deductive* method, where one first defines a general rule using the available information, and then deduces the answer using this rule. That is, one goes from *particular to general* and then from *general to particular*.

In transductive mode one provides direct inference from *particular to particular*, avoiding the ill-posed part of the inference problem (inference from particular to general).

3.1.5 TRANSDUCTIVE INFERENCE THROUGH CONTRADICTIONS

Replacing the maximal margin generalization principle with the maximal contradiction on the Universum (MCU) princple leads to the following algorithm: Using the working set (3.3) create a set of equivalence classes of functions, then using the Universum (2.67) calculate the size of the equivalence classes by the number of contradictions.

The recommendation of SRM for such a structure would be:

> *To classify test vectors (3.2), choose the equivalence class (defined on the working set (3.3)) that classifies the training data (3.1) well and has the largest number of contradictions on the Universum.*

The idea of maximizing the number of contradictions on the Universum can have the following interpretation:

> *When classifying the test vectors, be very specific; try to avoid extra generalizations on the Universum (2.67).*

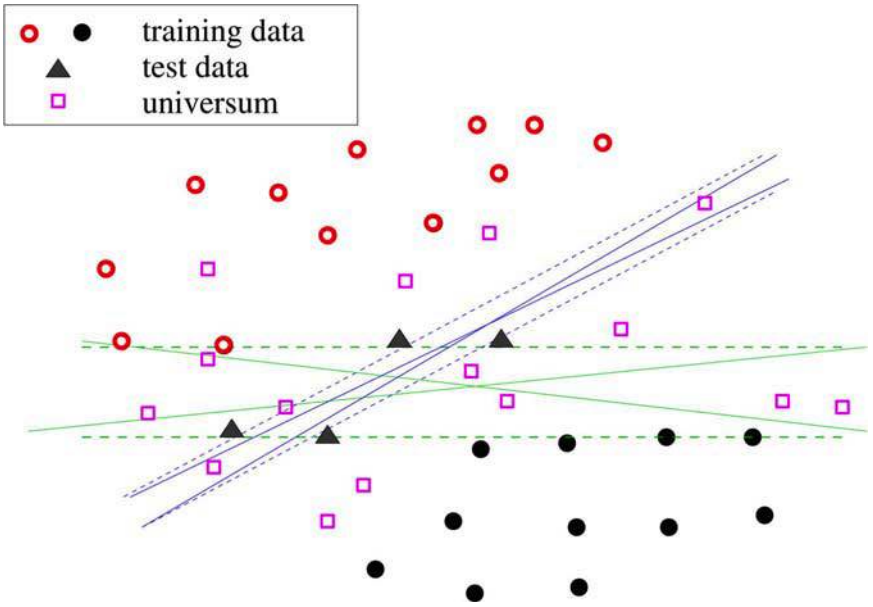

Figure 3.2: Large number of contradictions on Universum (boxes inside the margin) defines a large equivalence class.

From a technical point of view, the number of contradictions takes into account the inhomogeneity of image space, especially when the input vectors are nonlinearly mapped into feature space.

Technically, to implement transductive inference through contradictions one has to solve the following problem.

Given the images of the training data (3.1), the images of the test data (3.2), and the images of the Universum (2.67), construct the linear decision rule

$$I(x) = \theta[(w_0, z) + b_0],$$

where the vector w_0 and threshold b_0 are the solution of the following optimization problem: Minimize the functional

$$R(w) = \frac{1}{2}(w, w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=\ell+1}^{\ell+k} \theta(\xi_j) + C_3 \sum_{s=1}^{u} \theta(\xi_s^*), \quad C_1, C_2, C_3 \geq 0 \tag{3.15}$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \; i = 1, \ldots, \ell \tag{3.16}$$

(defined by the images of the training data (3.1)), the constraints

$$y_j^*((z_j, w) + b) \geq 1 - \xi_j, \quad \xi_j \geq 0, \; j = \ell + 1, \ldots, \ell + k \tag{3.17}$$

(defined by the set (3.2) and the desired vector $(y^*_{\ell+1}, \ldots, y^*_{\ell+k})$), and the constraints

$$|(z^*_s, w) + b| \le a + \xi^*_s, \quad \xi^*_s \ge 0, \ s = 1, \ldots, u, \ \ a \ge 0 \tag{3.18}$$

(defined by the images of the Universum (2.67)).

As before (for computational reasons), we replace $\theta(\xi)$ in the objective function with ξ. Therefore we minimize the functional

$$R(w) = \frac{1}{2}(w, w) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{j=\ell+1}^{\ell+k} \xi_j + C_3 \sum_{s=1}^{u} \xi^*_s, \quad C_1, C_2, C_3 \ge 0 \tag{3.19}$$

subject to the constraints (3.16), (3.17), and (3.18).

Dual Form Solution

The solutions to all of the above problems in the dual space of Lagrange multipliers can be unified as follows. Find the function

$$f(x) = \sum_{i=1}^{\ell} \alpha^0_i y_i K(x, x_i) + \sum_{t=\ell+1}^{\ell+k} \beta^0_t y^*_t K(x, x_t) + \sum_{m=1}^{u} (\mu^0_m - \nu^0_m) K(x, x^*_m) + b_0 \tag{3.20}$$

whose test classifications y^*_j and coefficients $\alpha^0, \beta^0, \mu^0, \nu^0, b_0$ maximise the functional

$$W(\alpha, \beta, \gamma, \mu, \nu, y^*) = \sum_{i=1}^{\ell} \alpha_i + \sum_{t=\ell+1}^{\ell+k} \beta_t - a \sum_{n=1}^{u} (\mu_n + \nu_n) \tag{3.21}$$

$$-\frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \frac{1}{2} \sum_{s,t=\ell+1}^{\ell+k} \beta_t y^*_t \beta_s y^*_s K(x_t, x_s)$$

$$-\frac{1}{2} \sum_{m,n=1}^{u} (\mu_m - \nu_m)(\mu_n - \nu_n) K(x^*_m, x^*_n) - \sum_{i=1}^{\ell} \sum_{t=\ell+1}^{\ell+k} \alpha_i y_i \beta_t y^*_t K(x_i, x_t)$$

$$-\sum_{i=1}^{\ell} \sum_{m=1}^{u} \alpha_i y_i (\mu_m - \nu_m) K(x_i, x^*_m) - \sum_{m=1}^{u} \sum_{t=\ell+1}^{\ell+k} (\mu_m - \nu_m) \beta_t y^*_t K(x^*_m, x_t)$$

subject to the constraints

$$0 \le \alpha_i \le C_1, \tag{3.22}$$

$$0 \le \beta_t, \le C_2, \tag{3.23}$$

$$0 \le \mu_m, \nu_m \le C_3, \tag{3.24}$$

and the constraint

$$\sum_{i=1}^{\ell} \alpha_i y_i + \sum_{t=\ell+1}^{\ell+k} \beta_t y^*_t + \sum_{m=1}^{u} (\mu_m - \nu_m) = 0. \tag{3.25}$$

In particular, when $C_2 = C_3 = 0$ we obtain the solution for the conventional SVM, when $C_2 = 0$ we obtain the solution for inductive SVMs with the Universum, and when $C_3 = 0$ we obtain the solution for transductive SVMs.

Note that just taking into account the Universum ($C_2 = 0$) does not change the convexity of the optimization task. The problem becomes nonconvex (and therefore can have a nonunique solution) only for transductive mode.

It is good to use hint (3.13) when solving transductive problems.

3.2 Beyond Transduction: The Transductive Selection Problem

The transductive selection problem was not discussed in the original Russian edition of *EDBED*. It was written at the last moment for the English translation. In *EDBED* the corresponding section (Chapter 10, Section 13) has a very technical title "The Problem of Finding the Best Point of a Given Set." Here we call this type of inference transductive selection.

3.2.1 Formulation of Transductive Selection Problem

The transductive selection problem is the following: Given the training examples (pairs $(x_i, y_i), x \in R^n,\ y \in \{-1, +1\},\ i = 1, \ldots, \ell$) and given a working set ($x_j^* \in R^*,\ j = 1, \ldots, m$), find in the working set the k elements that belong to the first class ($y = +1$) with the highest probability.

Here are some examples of the selection problem:

- *Drug discovery.* In this problem, we are given examples of effective drugs $(x_i, +1)$ and examples of ineffective drugs $(x_s, -1)$. The goal is to find among the given candidates $(x_1^*, \ldots, x_m^*)$ the k candidates with the highest probability of being effective drugs.

- *National security.* In this problem, we are given examples (descriptions) of terrorists $(x_i, +1)$ and examples of non-terrorists $(x_s, -1)$. The goal is to find among the given candidates $(x_1^*, \ldots, x_m^*)$ the k most likely terrorists.

Note that in contrast to general transductive inference, this setting does not require the classification of all candidates[3]. The key to solving the selective inference problem is to create an appropriate factorization of a given set of functions that contains fewer equivalence classes than the factorization for transductive inference. The transductive selective models are the main instrument for solving decision-making problems in high-dimensional spaces. However, this instrument has not yet been developed.

[3]In such problems, the most difficult cases are "border candidates." In transductive selection problems, we exclude this most difficult part of the task (classification of border candidates). Here again we obtain the same advantage that we obtained by replacing the model identification scheme by the prediction scheme and replacing the predictive scheme by the transductive scheme: we replaced a not very well-posed problem by a better-posed problem.

3.3 Directed Ad Hoc Inference (DAHI)

3.3.1 The Idea Behind DAHI

This section discusses *directed ad hoc inference*, inference that occupies an intermediate position between *inductive–deductive* inference and *transductive* inference.

The main idea of DAHI is a reconsideration of the roles of the training and testing stages during the inference process. The classical *inductive–deductive* model of inference contains two different stages:

(1) The training (inductive) stage, where one constructs a rule for classification using the training data, and

(2) The testing (deductive) stage where one classifies the test data using the constructed rule.

The *transductive* model of inference solves the classification problem in one step:

– Given a set of training data and a set of test data, it finds the labels for the test data directly.

DAHI works differently. During the training stage, DAHI looks for a principal direction (concept) used to construct different rules for future inferences. This is different from the inductive stage of inference where the goal is to find one fixed rule. During the test stage DAHI uses this principal direction to construct a specific rule for each given test vector (the ad hoc rule). Therefore, DAHI contains elements of both inductive and transductive inference:

(1) It constructs one general direction of inference (as in inductive inference).

(2) It constructs an individual (ad hoc) rule for each given test example (as in transductive inference).

The idea of DAHI is: *To construct a linear (in feature space) decision rule that has fixed homogeneous terms and individual (for different test vectors) thresholds.*

The problem is how to find thresholds that make inferences more accurate than ones based on one fixed threshold (as in SVM).

From a technical point of view DAHI is a combination of ideas from statistical learning theory (in particular, support vector machines), and from nonparametric statistics (methods for conditional probability estimation).

3.3.2 Local and Semi-Local Rules

To discuss the details of DAHI let us consider the idea of *local algorithms* suggested by nonparametric statistics and in particular the k-nearest neighbors method.

k-NEAREST NEIGHBORS METHOD

According to the k-nearest neighbours method for any point of interest x_0 one chooses from the training data the k-nearest (in a given metric) vectors $x_i,\ i = 1, \ldots, k$ and classifies the point of interest x_0 depending on which class dominates among these k chosen vectors.

The k-nearest neighbors method can be described as a *local* estimating method. Consider the set of constant-valued functions. For a set of indicator functions it contains only two functions: one takes the value -1; another takes the value 1. Consider the following local algorithm: define the spherical vicinity of the point of interest x_0 based on the given metric and a value for the radius (defined by the distance from a point of interest x_0 to its k nearest neighbors). Then choose from the admissible set of functions the function that minimizes the empirical loss on the training vectors belonging to the vicinity of the point of interest x_0. Finally use this function to classify the point of interest.

This description of the k-nearest neighbors method as a local algorithm immediately allows one to generalize it in two respects:

(1) One can use a richer set of admissible functions (for example, the set of large margin linear decision rules, see Section 2.3)

(2) One can use different rules to specify the value of the radius that defines the locality (not just the distance to the kth nearest neighbor).

In 1992 the idea of local algorithms for pattern recognition was used where (local) linear rules (instead of local constant rules) and VC bounds (instead of the distance to the kth nearest neighbor) were utilized [145]. The local linear rules demonstrated a significant improvement in performance (3.2% error rate instead of 4.1% for digit recognition on the US Postal Service database).

For the regression estimation problem a similar idea was used in the Nadaraya–Watson estimator [147, 148] with a slightly different concept of locality. Nadaraya and Watson suggested considering "soft locality": they introduced a weight function (e.g., a monotonically decreasing nonnegative function from the distance between a point of interest x_0 and elements x_i of training data $f(||x_0 - x_i||),\ \ i = 1, \ldots, \ell$), and used this function for estimating the value of interest

$$y_0 = \sum_{i=1}^{\ell} \tau_i(x_0) y_i, \tag{3.26}$$

where coefficients $\tau_i(x_0)$ were defined as follows,

$$\tau_i(x_0) = \frac{f(||x_0 - x_i||)}{\sum_{i=1}^{\ell} f(||x_0 - x_i||)}. \tag{3.27}$$

This concept is a generalization of the hard locality concept. We will use this construction later.

However in all of these methods the concept of locality is the same: it is a sphere (a "soft sphere" in the Nadaraya–Watson method) defined by a given metric with the center at the point of interest.

SEMI-LOCAL RULE

In DAHI we use a new concept of vicinity. We map input vectors x into a feature space z where we specify the vicinity. We consider a cylinder (or more generally a "soft cylinder"; see Section 3.3.4 below) whose axis passes through the image z_0 of the point of interest x_0. The defined vicinity is unbounded in one direction (defined by the axis of the cylinder) and bounded in all other directions. We call such a vicinity a *semi-local* vicinity.

The difference between the local and semi-local concepts of vicinity is the following. In a sphere with a fixed center there are no preferable directions in a feature space, while a cylinder has one preferable direction (along the axis of the cylinder). DAHI uses this direction to define vicinities for all points of interest.

During the training stage DAHI looks for the direction of the cylinder that defines the axis (in feature space) for all possible vicinities (cylinders). To find this direction one can use the methods of statistical learning theory (e.g., SVMs).

During the test stage DAHI uses only data from the (semi-local) vicinity of the point of interest z_0 and constructs a one-dimensional conditional probability function defined on the axis of the cylinder passing in the specified direction w_0 through the point of interest z_0. DAHI then uses this conditional probability $P(y_0 = 1|z_0)$ to classify z_0, where z_0 is the image of the point of interest x_0 in feature space.

Note that DAHI generalizes the SVM idea. In SVM one chooses both the direction w_0 and the threshold b_0 for the decision rule. In DAHI one chooses only the direction w_0, and for any test vector constructs an individual decision rule (threshold).

3.3.3 ESTIMATION OF CONDITIONAL PROBABILITY ALONG THE LINE

To solve the classification part of the problem we estimate the conditional probability $P(y(t) = 1|t)$ that the point t on the axis of a cylinder (passing through the point of interest t_0) belongs to the first class. To do this we have to solve the integral equation

$$\int_a^t P(y = 1|t')dF(t') = F(y = 1, t), \tag{3.28}$$

where both the cumulative distribution function of the point on the line $F(t)$ and the probability function $F(y = 1, t)$ of that point on the line with $t' \leq t$ belong to the first class are unknown, but data (inside cylinder) are given.

Note that when the density function $p(t)$ exists for $F(t)$, the conditional probability

$$P(y = 1|t) = \frac{p(y = 1, t)}{p(t)}$$

defines the solution of Equation (3.28).

To solve this problem given data one must first estimate the cumulative distribution functions along the line and then use these estimates $F_{est}(t),\ F_{est}(1,t)$ in Equation (3.28) instead of the actual functions $F(\xi)$ and $F(y=1,\xi)$.

$$\int_a^t P(y=1|t')dF_{mp}(t') = F_{emp}(y=1,t). \tag{3.29}$$

This Equation forms an ill-posed problem where not only the right-hand side of the equation is an approximation of the real right-hand side but also the operator is an approximation of the real operator (since we use $F_{emp}(t)$ instead of $F(t)$).

In [140] it is shown that if the approximations $F_{emp}(t)$, and $F_{emp}(y=1,t)$ are consistent then there exists a law $\gamma_\ell = \gamma(\ell)$ such that the Tikhonov regularization method

$$R(P) = \left|\left| \int_a^t P(y=1|t')dF_{emp}(t') - F_{emp}(y=1,t) \right|\right|^2 + \gamma_\ell \Omega(P) \tag{3.30}$$

provides the solutions that converge to the solution of Equation (3.28) as $\ell \to \infty$.

3.3.4 Estimation of Cumulative Distribution Functions

A consistent method of estimating cumulative distribution functions along a line was first suggested by Stute in 1986 [149]. He considered a cylinder of radius r whose axis coincides with the line, projected on this line the vectors z of the training data that were inside the cylinder (suppose that there are $r(\ell)$ such vectors), and constructed a one-dimensional empirical distribution function using these projections:

$$F^*_{r(\ell)}(x) = \frac{1}{r(\ell)} \sum_{i=1}^{r(\ell)} \theta(t-t_i). \tag{3.31}$$

Stute showed that under some general law of choosing the radius of the cylinder (which depends on the number of observations ℓ) with an increasing number of observations, this empirical cumulative distribution function converges with probability one to the desired function. To estimate conditional probability one can use in (3.30) the approximation (3.31) and the approximation

$$F_{r(\ell)}(1,t) = \frac{1}{2r(\ell)} \sum_{i=1}^{r(\ell)} (1+y_i)\theta(t-t_i). \tag{3.32}$$

Also one can estimate a cumulative distribution function along the line in the Nadaraya-Watson style using the distances between images of training vectors and the line passing through the point of interest z_0 in direction w_0 ,

$$d_i(z_0) = \sqrt{|z_i - z_0|^2 - t_0^2}, \tag{3.33}$$

where $t_0 = (z_0, w_0)$ is the projection of the vector z_0 on the direction w_0. Using $d_i(z_0)$ instead of $||x_0 - x_i||$ in (3.27) one obtains the Nadaraya–Watson type approximations of the elements of Equation (3.29):

$$F_{emp}(t) = \sum_{i=1}^{\ell} \tau_i(z_0)\theta(t - t_i), \tag{3.34}$$

$$F_{emp}(y = 1, t) = \frac{1}{2}\sum_{i=1}^{\ell} (1 + y_i)\tau_i(z_0)\theta(t - t_i). \tag{3.35}$$

Both the Stute estimate and modified Nadaraya–Watson estimate are step functions. The difference is that in Stute's estimate there are $r(\ell)$ steps where all values of the step are equal to $1/r(\ell)$ while in the Nadarya–Watson estimate there are ℓ steps but the step values $\tau_i(z_0)$ are different, and depend on the distance between the vector z_i and the line passing through the point z_0 in the direction w_0.

3.3.5 Synergy Between Inductive and Ad Hoc Rules

In DAHI we combine two consistent methods: the SVM method for estimating the direction in feature space, and the method for estimating the conditional probability along the line passing through the point of interest.

However, when the number of training data is not large (and this is always the case in a high-dimensional problem) one needs to provide both methods with additional information: In order to choose a good SVM solution one has to map the input vectors into a "good" Hilbert space (to choose a "good" kernel). In order to obtain a good solution for solving the ill-posed problem of estimating a conditional probability function along the line one has to use a priori information about the admissible set of functions that contain the desired conditional probability function.

By combining the above two methods, one tries to construct a robust classification method that reduces the dependency on a priori information.

This is because:

(1) When one chooses a direction that is "reasonably close" to the one that defines a "good" separating hyperplane, the corresponding conditional probability function belongs to the set of *monotonic* functions (the larger the SVM score is, the larger is the probability of the positive class). Finding a direction that maintains the monotonicity property for the conditional probabilities requires fewer training examples than finding a direction that provides a good classification.

(2) The problem of finding a conditional probability function from the set of monotonic nondecreasing functions is much better posed than the more general problem of finding a solution from the set of continuous nonnegative functions.[4]

Therefore, in the set of monotonic functions one can solve this problem well, using a restricted (small) number of observations.

[4] A set of monotonically increasing (or monotonically decreasing) functions has VC dimension one while a set of continuous nonnegative functions has an infinite VC dimension.

(3) Using the leave-one-out technique one can use the same training data for constructing the main direction and later for constructing conditional probability functions.

The minimization of functional (3.30) in a set of monotonic functions is not too difficult a computational problem. The idea behind DAHI is to use this possible synergy.

Figure 3.3 shows two examples of the binary classification problem: separating digit 3 from digit 5. Two examples of conditional probabilities $P(3|t)$ estimated along the line are presented in Figure 3.3. For each example the figure shows the image of interest, the functions $F_{emp}(t)$ and $F_{emp}(3,t)$, and the solution of Equation (3.29). The position of the point of interest on the line corresponds to an ordinate value of 0. Part (a) of the figure shows the probability that the image is a 3 is 0.34, but in part (b) the probability that the image is a 3 is 0.

3.3.6 DAHI AND THE PROBLEM OF EXPLAINABILITY

The idea of DAHI is appealing from a philosophical point of view since it addresses the question of *explainability* of complex rules [169]. DAHI divides the model of explainability for complex rules into two parts: the "main direction" and the "ad hoc" parts where only the "main direction" part of the rule has to be explained (described by the formal model).

One speculation on the DAHI model of explainability can be given by the example of how medical doctors distinguish between cancer and benign cases. They use principal rules to evaluate the cancer and if the corresponding score exceeds a threshold value, they decide the case is cancer.

The threshold, however, is very individual: it depends on the family history of the patient, and many other factors. The success of a doctor depends on his experience in determining the individual threshold. The threshold can make all the difference in diagnostics. Nevertheless the explainability is mostly related to the "main direction" part of the rule.

3.4 PHILOSOPHY OF SCIENCE FOR A COMPLEX WORLD

3.4.1 EXISTENCE OF DIFFERENT MODELS OF SCIENCE

The limitations of the classical model of science when dealing with the real-life complex world have been discussed for quite some time. For example, according to Einstein, the classical model of science is relevant for a simple world. For a complex world it is inapplicable.

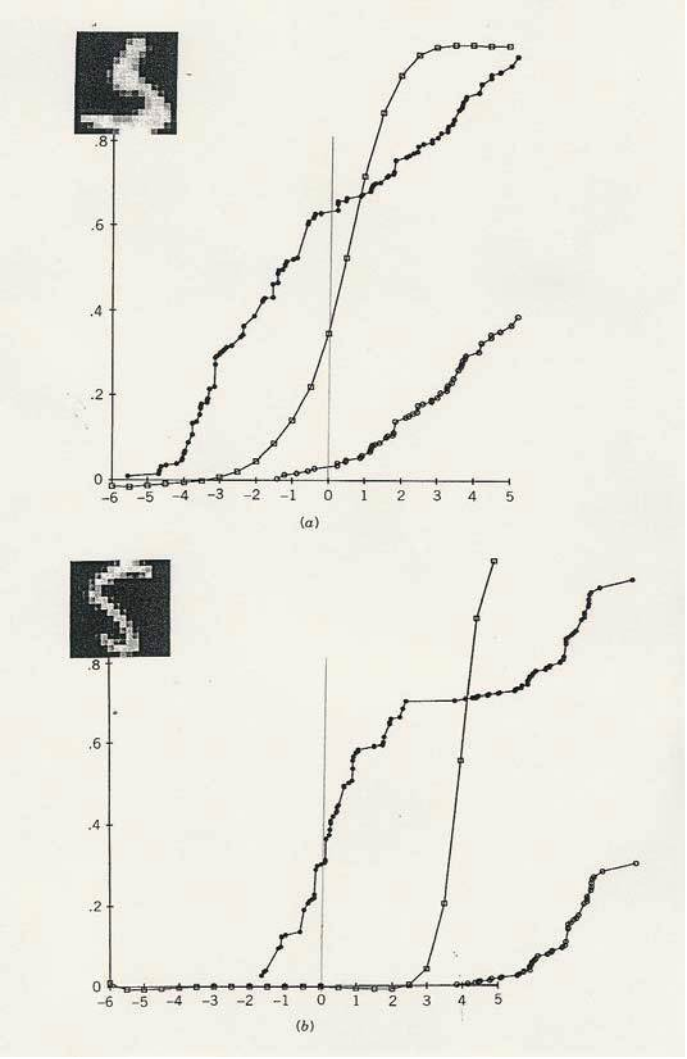

Figure 3.3: Solutions of the integral equation for different data.

– Einstein on the simple world:

When the solution is simple, God[5] *is answering.*

– Einstein on the complex world:

When the number of factors coming into play in a phenomenological complex is too large, scientific methods in most cases fail.[6]

One can see the idea of limitation of scientific models and existence of non-scientific ones in the following Richard Feynman's remark (*Lectures on physics*):

> *If something is said not to be a science, it does not mean that there is something wrong with it ... it just means that it is not a science.*

In other words there was an understanding that:

> *Classical science is an instrument for a simple world. When a world is complex, in most cases classical science fails. For a complex world there are methods that do not belong to classical science.*

Nevertheless, the success of the physical sciences strongly influenced the methodology used to analyze the phenomena of a complex world (one based on many factors). In particular, such a methodology was adopted in the biological, behavioural, and social sciences where researchers tried to construct low-dimensional models to explain complex phenomena.

The development of machine learning technology challenged the research in the methodology of science.

3.4.2 Imperative for a Complex World

Statistical learning theory stresses that the main difficulties of solving generalization problems arise because, in most cases, they are ill-posed.

To be successful in such situations, it suggests to give up attempts of solving ill-posed problems of interest replacing them by less demanding but better posed problems. In many cases this leads to renunciation of explainability of obtained solutions (which is one of the main goals declared by the classical science). Therefore, a science for a complex world has different goals (may be it should be called differently).

For solving specific ill-posed problems the regularization technique was suggested [20, 21, 54, 55]. However, to advance high-dimensional problems of inference just applying classical regularization ideas is not enough. The SRM principle of inference is another way to control the capacity of admissible sets of functions. Recently a new general idea of capacity control was suggested in the form of the following imperative [139]:

[5]Here and below Einstein uses the word God as a metaphor for nature.

[6]Great theoretical physicist Lev Landau did not trust physical theories that combine more than a few factors. This is how he explained why: "With four free parameters one can draw an elephant, with five one can draw an elephant rotating its tail."

IMPERATIVE

When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one.

According to this imperative:
— Do not estimate a density if you need to estimate a function.
(Do not use the classical statistics paradigm for prediction in a high-dimensional world: Do not use generative models for prediction.)
— Do not estimate a function if you only need to estimate its values at given points. *(Try to perform direct inference rather than induction.)*
— Do not estimate predictive values if your goal is to act well.
(A good strategy of action does not necessary rely on good predictive ability.)

3.4.3 RESTRICTIONS ON THE FREEDOM OF CHOICE IN INFERENCE MODELS

In this Afterword we have discussed three levels of restrictions on the freedom of choice in the inference problem:

(1) *Regularization*, which controls the smoothness properties of the admissible set of functions (it forbids choosing an approximation to the desired function from not a "not smooth enough set of functions").

(2) *Structural risk minimization*, which controls the diversity of the set of admissible functions (it forbids choosing an approximation to the desired function from too diverse a set of functions, that is, from the set of functions which can be falsified only using a large number of examples).

(3) *Imperatives*, which control the goals of possible inferences in order to consider a better-posed problem. In our case it means creating the concept of equivalence classes of functions and making an inference using a large equivalence class (it forbids an inference obtained using a "small" equivalence class).

It should be noted that an understanding of the role of a general theory as an instrument to restrict directions of inference has existed in philosophy for a long time. However, the specific formulations of the restrictions as described above were developed only recently. The idea of using regularization to solve ill-posed problems was introduced in the mid-1960s [21, 55]. Structural risk minimization was introduced in the early 1970s [*EDBED*], and the imperative was introduced in the mid-1990s [139].

In order to develop the philosophy of science for a complex world it is important to consider different forms of restriction on the freedom of choice in inference problems and then analyze their roles in obtaining accurate predictive rules for the pattern recognition problem.

One of the main goals of research in the methodology of analysis of a complex world is to introduce new imperatives and for each of them establish interpretations in the corresponding branches of science.

3.4.4 Metaphors for Simple and Complex Worlds

I would like to finish this part of the Afterword with metaphors that stress the difference in the philosophy for simple and complex worlds. As such metaphors let me again use quotes from Albert Einstein.

Two Metaphors for a Simple World

1. *I want to know God's thoughts.* (*A. Einstein*)
2. *When the solution is simple, God is answering.* (*A. Einstein*)

Interpretation

Nature is a realization of the simplest conceivable mathematical ideas. I am convinced that we can discover, by means of purely mathematical constructions, concepts and laws, connect them to each other, which furnish the key to understanding of natural phenomena. (*A. Einstein.*)

Three Metaphors for a Complex World

First Metaphor

Subtle is the Lord, but malicious He is not. (*A. Einstein*)

Interpretation[7]

Subtle is the Lord — one can not understand His thoughts.

But malicious He is not — one can act well without understanding them.

Second Metaphor

The devil imitates God.[8] (*Medieval concept of the devil.)*

Interpretation

Actions based on your understanding of God's thoughts can bring you to catastrophe.

Third Metaphor

If God does exist then many things must be forbidden. (*F. Dostoevsky*)

Interpretation

If a subtle and nonmalicious God exists, then many ways of generalization must be forbidden. The subject of the complex world philosophy of inference is to define corresponding imperatives (to define what should be forbidden). These imperatives are the basis for generalization in real-life high-dimensional problems.

The imperative described in Section 3.4.2 is an example of the general principle that forbids certain ways of generalization.

[7] Surely what Einstein meant is that the laws of nature may be elusive and difficult to discover, but not because the Lord is trying to trick us or defeat our attempts to discover them. Discovering the laws of nature may be difficult, but *it is not impossible*. Einstein considered comprehensibility of the physical world as a "mystery of the world". My interpretation of his metaphor for a *complex world* given below is different.

[8] This includes the claim that for humans the problem of distinguishing imitating ideas of the devil from thoughts of God is ill-posed.

Chapter 4

THE BIG PICTURE

4.1 RETROSPECTIVE OF RECENT HISTORY

The recent history of empirical inference science can be described by Kuhn's model of the development of science which distinguishes between periods with fast development of ideas (development of the new paradigms) and periods with slow developments (incremental research) [168].

In empirical inference science one can clearly see three fast periods: in the 1930s, 1960s, and 1990s.

4.1.1 THE GREAT 1930S: INTRODUCTION OF THE MAIN MODELS

The modern development of empirical inference science started in the early 1930s. Three important theoretical results indicated this beginning:

(1) The foundation of the theory of probability and statistics based on Andrei Kolmogorov's axiomatization and the beginning of the development of the classical theory of statistics.[1]

(2) The development of a basis of applied statistics by Sir Ronald Fisher.[2]

(3) The development of the falsifiability principle of induction by Sir Karl Popper[3].

[1] Andrei Kolmogorov was a leading figure in mathematics in the 20th century. He was the recipient of many of the highest prizes and international awards.

[2] Ronald Fisher was a creator of applied statistics. He was knighted by Queen Elizabeth II in 1952 for his works in statistics and genetics.

[3] Karl Popper was a creator of the modern philosophy of science. He was knighted by Queen Elizabeth II in 1965 for his works in philosophy.

AXIOMATIZATION OF STATISTICS AND PROBABILITY THEORY AND THE PROBLEM OF EMPIRICAL INFERENCE

At the beginning of the 20th century there was a great interest in the philosophy of probabilistic analysis. It was the time of wide discussions about the nature of randomness. These discussions, however, contained a lot of wide speculations. Such speculations were not very useful for development of the formal mathematical theory of random events. In order to separate formal mathematical development of the theory from its interpretation, mathematicians discussed the opportunity to axiomatize the theory of random events. In particular, the problem of the axiomatization of probability theory was mentioned by David Hilbert at the Second Mathematical Congress in Paris in 1900 as one of the important problems of the 20th century.

It took more than 30 years until in 1933 Kolmogorov introduced simple axioms for probability theory and statistics.

Kolmogorov began with a set Ω which is called the *sample space* or set of elementary events A (outcomes of all possible experiments). On the set of possible elementary events a system $\mathcal{F} = \{F\}$ of subsets $F \subset \Omega$, which are called *events*, is defined. He considered that the set $F_0 = \Omega \subset \mathcal{F}$ determines a situation corresponding to an event that always occurs. It is also assumed that the set of events contains the empty set $\emptyset = F_\emptyset \subset \mathcal{F}$, the event that never occurs. Let $\mathcal{F}$ be an algebra of sets. (When $\mathcal{F}$ is also closed under countably infinite intersections and unions, it is called a σ-algebra.) The pair $(\Omega, \mathcal{F})$ defines the *qualitative* aspects of random experiments.

To define the quantitative aspects he introduced the concept called a *probability measure* $P(F)$ defined on the elements F of the set $\mathcal{F}$. The function has the following properties,

$$P(\emptyset) = 0, \quad P(\Omega) = 1, \quad 0 \leq P(F) \leq 1,$$

$$P\left(\cup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i), \quad \text{if } F_i, F_j \in \mathcal{F} \text{ and } F_i \cap F_j = \emptyset \ \ \forall i, j.$$

He then introduced the idea of conditional probability of events

$$P(F|E) = \frac{P(F \cap E)}{P(E)}, \quad P(E) > 0$$

and defined mutual independence of events $F_1, \ldots, F_n$ as the situation when

$$P(F_1 \cap ... \cap F_n) = \Pi_{i=1}^{n} P(F_i).$$

By introducing this axiomatization, Kolmogorov made the theory of probabilities a pure mathematical (deductive) discipline with the following basic problem.

THE BASIC PROBLEM OF PROBABILITY THEORY

> Given the triplet $(\Omega, \mathcal{F}, P)$ and an event B, calculate the probability measure $P(B)$.

The axiomatization led to the definition of statistics as the inverse problem.

THE BASIC PROBLEM OF STATISTICS

Given the pair $(\Omega, \mathcal{F})$ and a finite number of i.i.d. data

$$A_1, \ldots, A_\ell$$

estimate the probability measure $P(F)$ defined on all subsets $F \in \mathcal{F}$.

This inverse problem reflects the inductive idea of inference. The general theoretical analysis of inductive inference started with the particular instance of this problem described in the Glivenco–Cantelli theorem (1933) and later was extended to a general theory for uniform convergence (VC theory) in 1968 and 1971.

FISHER'S APPLIED STATISTICS

At about the same time that theoretical statistics was introduced, Fisher suggested the basics of applied statistics. The key element of his simplification of statistical theory was his suggestion of *the existence of a density function* $p(\xi)$ that defines the randomness (noise) for a problem of interest.

Using the density function Fisher introduced the model of observed data

$$(y_1, x_1), \ldots, (y_\ell, x_\ell) \tag{4.1}$$

as measurements of an unknown function of interest $f(x, \alpha_0)$ that belongs to some parametric family $f(x, \alpha), \alpha \in \Lambda$ contaminated by uncorrelated noise defined by the known density $p(\xi)$

$$y_i = f(x_i, \alpha_0) + \xi_i, \quad E\xi_i x_i = 0. \tag{4.2}$$

He developed the maximum likelihood method for estimating the density function given the data (4.1) and the parametric family

$$p_\alpha(\xi) = p(y - f(x, \alpha))$$

that contains the density function of interest. Fisher suggested choosing the density function with parameter α that maximizes the functional

$$R(\alpha) = \sum_{i=1}^{\ell} \ln p(y_i - f(x_i, \alpha)).$$

It took 20 years before LeCam, *using uniform convergence arguments*, proved in 1953 the consistency of the maximum likelihood method for specific sets of parametric families.

Since this time many efforts were made to generalize Fisher's scheme for a wide set of densities (to remove Fisher's requirement to explicitly define the model of noise). In particular, Huber suggested the model of robust estimation that is based on a wide class of density functions. Later, nonparametric techniques also generalized Fisher's model for wide sets of admissible functions.

However, the key element of applied statistics remained the *philosophical realism* based on the generative model (4.2) of the observed data (4.1).

POPPER'S CONCEPT OF FALSIFIABILITY

In the early 1930s Popper suggested his idea of falsifiability. It was considered both as the demarcation line between metaphysics and natural science as well as a justification of Occam's Razor principle. Popper developed the falsifiability idea over his entire lifetime: his first publication appeared in German in 1934 his last addition was in an English edition in 1972. This idea is considered as one of the most important achievements in the philosophy of science of the 20th century.

Fisher's philosophy of applied statistics and Popper's justification of the dependence of generalization ability on the number of entities formed the classical paradigm of philosophical realism for induction.

The continuation of the Kolmogorov–Glivenco–Cantelli line of theoretical statistics led to the development of the VC theory that reflected the philosophical instrumentalism paradigm.

4.1.2 THE GREAT 1960S: INTRODUCTION OF THE NEW CONCEPTS

In the 1960s several revolutionary ideas for empirical inference science were introduced. In particular

1. Tikhonov, Ivanov, and Phillips developed the main elements of the theory of ill-posed problems.
2. Kolmogorov and Tikhomirov introduced the capacity concepts (ε-entropy, covering numbers, width) for sets of functions.
3. Solomonoff, Kolmogorov, and Chaitin developed the concept of algorithmic complexity and used it to justify inductive inference.
4. Vapnik and Chervonenkis developed the basics of empirical inference science.
5. The empirical inference problem became a problem of natural science.

ILL-POSED PROBLEMS

I consider the philosophy of ill-posed problems as the turning point in the understanding of inference. It can have the following interpretation:

(1) *The general problem of inference — obtaining the unknown reasons from the known consequences — is ill-posed.*

(2) *To solve it one has to use very rich prior information about the desired solution. However, even if one has this information it is impossible to guarantee that the number of observations that one has is enough to obtain a reasonable approximation to the solution.*

Therefore one should try to avoid solving an ill-posed problem if possible. The development of the VC theory is just an illustration of this thesis.

THE BASIS FOR AN ALTERNATIVE

The 1960s marked the beginning of the mathematical development of the instrumentalist point of view. In the late 1950s and early 1960s Kolmogorov and Tikhomirov introduced the idea of the capacity of a set of functions and demonstrated its usefulness in function approximation theory.

The VC entropy, Growth function, and VC dimension concepts of capacity (which are different from the one suggested by Kolmogorov and Tikhomirov) became the main concepts that define the generalization ability in the instrumentalist framework.

In the 1960s Solomonoff, Kolmogorov, and Chaitin introduced the concept of algorithmic complexity. Solomonoff introduced this concept in order to understand the inductive principle, while Kolmogorov tried to address issues about nature of randomness,[4] which was a subject of discussion at the beginning of the 20th century.

The book [139] shows that for the pattern recognition problem the idea of Kolmogorov complexity leads to essentially the same bound for the pattern recognition problem that the VC theory gives. The VC theory, however, defines the necessary and sufficient conditions for consistency. It is unclear if algorithmic complexity also provides the necessary and sufficient conditions.

By the end of 1960s we constructed the theory for the uniform law of large numbers and connected it to the pattern recognition problem. By doing this, we had developed the mathematical foundation of predictive learning.

An extremely important fact is that by the end of the 1960s the methodology of solving the inference problem adopted the methodology of the natural sciences. Any results on the generalization problem must be confirmed by computer experiments on a variety of problems.

This forever changed the approach to both empirical inference science and to the philosophy of inference.

4.1.3 THE GREAT 1990S: INTRODUCTION OF THE NEW TECHNOLOGY

In the 1990s the following events took place.

(1) Vapnik and Chervonenkis proved that the existence of the uniform law of large numbers is the necessary and sufficient condition for consistency of the empirical risk minimization principle. This means that if one chooses the function from an admissible set of functions one can not avoid VC type arguments.

(2) Estimation of high-dimensional functions became a practical problem.

(3) Large-margin methods based on the VC theory of generalization (SVM, boosting, neural networks) proved advantageous over classical statistical methods.

These results have led to the development of new learning technologies.

[4] In 1933 when Kolmogorov introduced his axiomatization of probability theory, he effectively stopped these discussions. Thirty years later he came back to this question connecting randomness with algorithmic complexity: Random events are ones that have high algorithmic complexity.

4.1.4 The Great 2000s: Connection to the Philosophy of Science

In the early 2000s the following important developments took place. These all contributed to a philosophy of science for a complex world.

(1) The development of the theory of empirical inference based on VC falsifiability as opposed to Occam's Razor principle.

(2) The development of noninductive methods of inference.

This can lead to a reconsideration of the psychological and behavioral sciences based on noninductive inference. Also it will lead to reconsideration of the goals and methods of pedagogical science: Teaching not just inductive inference but also direct inference that can use the (cultural) Universum.

4.1.5 Philosophical Retrospective

This is how the philosophy of inference has developed.

– At the end of 1930s
the basics of two different paradigms of empirical inference (generative and predictive) were introduced.

– At the end of the 1960s it became clear that
the classical statistical paradigm is too restrictive:
It cannot be applied to high-dimensional problems.

– At the end of the 1990s it became clear that
the Occam's Razor principle of induction is too restrictive:
Experiments with SVMs, boosting, and neural nets provided counter-examples.

– In the beginning of the 2000s it became clear that
the classical model of science is too restrictive:
It does not include noninductive (transductive and ad hoc) types of inference which, in high-dimensional situations, can be more accurate than inductive inference.

– In the beginning of the 2000s it also became clear that
in creating a philosophy of science for a complex world the machine learning problem will play the same role that physics played in creating the philosophy of science for a simple world.

4.2 Large Scale Retrospective

Since ancient times, philosophy has distinguished among three branches: natural science, metaphysics, and mathematics. Over many centuries, there have been ongoing

discussions about the demarcation among these branches, and this has changed many times. Even the existence of these three categories is still under discussion.

Some scholars consider only two categories, including mathematics in the category of natural science or in the category of metaphysics. However, it is convenient for our discussion to consider three categories.

4.2.1 Natural Science

The goal of natural science is to understand and describe the real world. It can be characterized by two features:

> The subject of analysis is defined by the real World.
>
> The methodology is to construct a theory (model) based on the results of both passive (observation-based), and active (query-based) experiments in the real world.

Examples of natural sciences include astronomy, biology, physics, and chemistry where scholars can observe facts of the physical world (conduct passive experiments), ask particular questions about the real world (perform active experiments), and (based on analyzing results of these experiments) create models of the world. In these activities, experiments play a crucial role. From experiments scientists obtain facts that reflect the relationships existing in the world; also, experiments are used to verify the correctness of the theory (models) that are suggested as a result of analysis[5].

Because of this, sometimes natural science is called empirical science. This stresses that the subject of natural science is the real world and its method is the analysis of results of (passive and active) experiments.

4.2.2 Metaphysics

In contrast to empirical (natural) science, metaphysics does not require analysis of experimental facts, or the verification of results of inference. It tries to develop a general way of reasoning with which truth can be found for any imaginary models. Metaphysics stresses the power of pure reasoning.

Examples of metaphysical problems are the following:

> What is the essence of the devil? Here the devil is not necessarily a personalized concept. It can be a metaphor. The metaphors on a complex world used in Section 3.4.4: "The devil imitates God" is one of the concepts of the devil given in the Middle Ages. This was formulated following very wide discussions.
>
> Another example: What is freedom of will?

[5]The relationship between the number of active and passive experiments differs in different sciences. For example, in astronomy there are more passive observations and fewer active experiments. However, in chemistry there are more active experiments and fewer passive experiments.

According to Kant the question

What are the principles of induction?

defines the demarcation line between empirical science (that can be applied only for a particular world) and metaphysics (that can be applied to any possible world). It is commonly accepted that Popper gave the answer to Kant's question by introducing the falsifiability idea. That is, empirical science must be falsifiable.

4.2.3 MATHEMATICS

Mathematics contains elements of both natural science and metaphysics.
There is the following (not quite) jocular definition of pure mathematics (Eugene Wigner):

> *Mathematics is the science of skillful operation with concepts and rules invented just for this purpose.*

From this view, pure mathematics (rather than applied) is a part of metaphysics since mathematics invents concepts and rules for analysis, constructs objects of analysis, and analyzes these objects. It does not rely on experiments either to construct theories or to verify their correctness.

The ideal scheme of pure mathematics is the system that has been used since Euclid introduced his geometry: Define a system of definitions and axioms, and from these deduce a theory.[6] Some scholars, however, consider mathematics as a part of the natural sciences because (according to these scholars) systems of definitions and axioms (concepts and rules) used in mathematics are inspired by the real world.

Another view makes a bridge among mathematics, natural science, and metaphysics by declaring that:

> Everything which is a law in the real world has a description in mathematical terms and everything which is true in mathematics has a manifestation in the real world.

Many scholars consider mathematics as a language that one uses to describe the laws of nature.

We, however, will not discuss the demarcation lines between mathematics and metaphysics, or mathematics and natural science and will instead consider all three as different branches of knowledge.

The duality of the position of mathematics with respect to natural science and metaphysics has important consequences in the history of the development of natural science.

[6]The real picture is much more complicated. According to Israel Gelfand one can distinguish among three periods of mathematical development in the 20th century:

(1) Axiomatization (constructing axioms for different branches of mathematics)

(2) Structurization (finding similar structures in different branches of mathematics)

(3) Renaissance (discovering new facts in different branches of mathematics).

4.3 SHOULDERS OF GIANTS

4.3.1 THREE ELEMENTS OF SCIENTIFIC THEORY

According to Kant, any scientific theory contains three elements:

(1) The setting of the problem,

(2) The resolution of the problem, and

(3) Proofs.

At first glance this remark seems obvious. However, it has a deep meaning. The crux of this remark is the idea that these three elements of the theory in some sense are independent and equally important.

(1) The precise setting of the problem provides a general point of view of the problem and its relation to other problems.

(2) The resolution of the problem comes not from deep theoretical analysis of the setting of the problem but rather precedes this analysis.

(3) Proofs are constructed not to search for the resolution of the problem, but to justify the solution that has already been suggested.

The first two elements of the theory reflect the understanding of the essence of the problem, its philosophy. The proofs make a general (philosophical) model a scientific theory. Mathematics mostly deals with one of these three elements, namely proofs, and much less with setting and resolution.

One interpretation of the Einstein remark:

> *Do not worry about your problems with mathematics, I assure you mine are far greater.*

could be the following:

> The solution of a problem in natural science contains three elements. Proofs are just one of them. There are two other elements: the setting of the problem, and its resolution, which make basis for a theory.

For the empirical inference problem, these three elements are clearly defined:

SETTING.

> The setting of the Inference problems is based on the risk minimization model: Minimize the risk functional
>
> $$R(\alpha) = \int Q(y, f(x, \alpha))dP(y, x), \quad \alpha \in \Lambda$$

in the situation when the probability measure is unknown, but i.i.d. data $(y_1, , x_1), \ldots, (y_\ell, x_\ell)$ are given.

It is very difficult to say who suggested this setting for the first time.[7] I learned about this setting from seminars at the Moscow Institute of Control Science in the mid-1960s.

RESOLUTION.

Two different resolutions of solving this problem were suggested:

(1) Aizerman, Braverman, Rozonoer, and Tsypkin from Russia and Amari from Japan suggested using methods based on gradient-type procedures

$$\alpha_n = \alpha_{n-1} + \gamma_n \text{grad}_\alpha Q(y_n, f(x_n, \alpha_{n-1})).$$

(2) Vapnik and Chervonenkis suggested methods that use the empirical risk minimization principle under the condition of uniform convergence.

PROOFS.

Proofs that justify these resolutions are based on:

(1) The theory of stochastic approximation for gradient based procedures, and

(2) The theory of uniform convergence for the empirical risk minimization principle (1968, 1971). In 1989, Vapnik and Chervonenkis proved that (one-sided) uniform convergence defines the necessary condition for consistency not only for the empirical risk minimization method, but for any method that selects one function from a given set of admissible functions.

4.3.2 BETWEEN TRIVIAL AND INACCESSIBLE

According to Kolmogorov, in the space of problems suggested by the real world there is a huge subspace where one can find trivial solutions. There is also a huge subspace where solutions are inaccessible. Between these two subspaces there is a tiny subspace where one can find non-trivial solutions. Mathematics operates inside this subspace.

It is therefore a big achievement when one can suggest a problem setting and a resolution to this setting and also invent concepts and rules that make proofs both nontrivial and accessible (this is interesting for mathematicians).

In order to transform a problem from an inaccessible one to one that has a mathematical solution very often one must simplify the setting of the problem, perform mathematical analysis, and then apply the result of this analysis to the nonsimplified real-life problem.[8]

[7] I believe that it was formulated by Tsypkin.

[8] In our discussions, the main simplification that made the analysis of induction possible was i.i.d. data in the training and test sets.

Einstein's remark:

> As far as the laws of Mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.

describes this situation.

Nevertheless mathematics forms a universal language for describing the laws of nature that (as we believe) does not contain inner contradictions. There is an understanding that

> The more science uses mathematics, the more truth it contains.

The language, however, is not always equivalent to thought.

4.3.3 Three Types of Answers

Analyzing the real world, mathematics gives three types of answers:

(1) Direct answers,

(2) Comforting answers, and

(3) Tautologies.

(1) A *direct answer* means a direct answer to the posed question. These answers are not necessarily accurate but they are answers to your questions. For example, the answer to the question "How many examples are sufficient to find the ε-approximation to the best possible solution?" is the VC bound. It can be possiby improved, but this is a direct answer to the question.

Few questions have direct answers.

(2) A *comforting answer* does not answer the question of interest since the direct answer is impossible or inaccessible. Instead, it answers another *accessible* question that is somehow related to the question of interest.

For example, one might wonder whether there are enough data to solve a specific ill-posed problem of interest. There is no answer to this question (since it is impossible). Instead mathematics suggests considering a resolution (regularization techniques) for which under some circumstances, in an imaginary (asymptotic) situation, an answer is attainable.

Many more questions have *comforting* answers.

(3) Lastly, mathematics is an instrument that can easily produce many trivial tautologies.[9] As soon as one has a good setting, a decent resolution to this setting, and examples of proofs, one can easily repeat the same construction under slightly different conditions.

It will produce

[9] Many good theorems can be considered as nontrivial tautologies.

formulas, formulas,. . . , formulas

that can be regarded in the same way as Hamlet regarded

words, words,. . . ., words.

Just words, nothing more.

There are many trivial tautologies among the results of mathematical analyses of natural phenomena.

4.3.4 The Two-Thousand-Year-Old War Between Natural Science and Metaphysics

Therefore there is a complicated relationship between metaphysics and natural science which has its reflection in discussions of the role of pure mathematics in natural science.

On the one hand the more mathematics a science uses, the more truth it contains (because we believe that its language does not contain contradictions).

On the other hand, there is a two-thousand-year-old war between metaphysics and natural science.

To discuss the nature of this war, let me start with some well-known quotes that describe it (there are hundreds of similar quotes but these are from intellectual giants):

- I am not a mathematician. I am a natural scientist. *(Kolmogorov, 1973)*
- Theoretical physics is too difficult for physicists. *(Poincare, 1910)*
- A mathematician may say anything he pleases, but a physicist must be at least partially sane. *(Gibbs, 1889)*
- I have hardly ever known a mathematician who was able to reason. *(Plato, 370 BC)*

Why did Kolmogorov not like to call himself a mathematician?[10] Why didn't Plato take mathematicians as seriously as the natural philosophers (people involved in discussion about fundamental principles of nature)?

[10]Kolmogorov did not play with formulas. The concepts and rules that he introduced in different branches of mathematics (probability theory, information theory, theory of approximation, functional analysis, logic, differential equations) helped to advance philosophy in natural science. These are some examples of his ideas related to the subject of this book:

He obtained the bound whose generalization is the bound on the uniform law of large numbers.

He introduced the concept of ε-entropy which provided the opportunity to consider capacity concepts of learning theory and in particular the VC entropy.

His idea of algorithmic complexity was used in the minimum message (minimum description) length principle leading to learning methods with the same generalization bound as the VC bound [139].

Kolmogorov did not work on pattern recognition problems but he developed the concepts and rules that were very similar to the one behind the main philosophy of learning theory.

This could be the answer. Natural science is not only about proofs but more about the setting and resolution of problems. Mathematics is just a language that is useful for the setting, the resolutions, and especially for the proofs. To use this language well requires a high level of professionalism. This is probably what Poincare had in mind when he made the above-quoted remark.[11]

However, to find a good setting and a good resolution requires another sort of professionalism. I believe the tendency to underestimate the role of this sort of professionalism and overestimate the role of technical (mathematical) professionalism in analysis of nature was the reason for Plato's remark.

The research in empirical inference science requires searching for new models of inference (different from inductive inference, such as inference in Universum environment, transductive, selective, ad hoc inferences, and so on). They are currently not under the scope of interest of mathematicians since they do not yet have clear settings and clear resolutions (this is the main subject of research). Mathematicians will become interested in this subject later when new settings, new resolutions, and new ideas for proofs are found.

The goal of the empirical inference discipline is to find these elements of the theory.

4.4 To My Students' Students

4.4.1 Three Components of Success

To be successful in creativity, and in particular in scientific creativity, one has to possess the following three gifts:

(1) Talent and strong motivation,

(2) Ability to work hard, and

(3) Aspiration for perfection and uncompromising honesty to one's inner truth.

Most discussions about the components of success concentrate on the first two gifts. One can easily recognize them observing the work of an individual (how bright the individual is in solving problems, how fast he understands new concepts, how many hours he works, and so on). These two components form the *necessary* conditions for success.

The third (maybe the most important) component that provides the spirit of creativity, the inner quality control for creation, a concept of high standards, and the willingness to pay any price for this high standard is more delicate. It can not be seen as easily as the first two. Nevertheless, it is the demarcation line between individuals whose lifetime achievements are above the expectations of their colleagues and individuals whose lifetime achievements are below the expectations of their colleagues.

In the next section I will try to describe this gift and to show that when creating something new one encounters two problems:

[11] By the way, the main part of theoretical physics was done by theoretical physicists who sometimes used "dirty mathematics." Later, mathematicians cleaned up the mathematics.

(1) to develop one's ideas in the way one desires, and

(2) to develop them perfectly.

The second of these two problems requires the most effort.

4.4.2 The Misleading Legend About Mozart

There is a highest standard of genius: Mozart, the greatest wunderkind, the greatest musician, and the greatest composer of all time. The legend gives the impression that everything Mozart touched achieved perfection automatically, without much effort. In many languages there is an expression "Mozart's lightness."

Legend admits, however, that when he was very young he worked extremely hard (he did not have a normal childhood; he was under very strong pressure from his father, who forced him to practice a lot).

Then legend tells us about Mozart, a merry young man, visiting a variety of Vienna cafes, who had admirers in everyday life, yet created the greatest music. He wrote it down with no draft.

That is true, Mozart did not write any drafts. He possessed a phenomenal professional memory, created his compositions in his mind, and could work simultaneously on several different compositions. Because of this, it would seem that his creativity also came easily. This was not the case. Legend tells us stories that he was almost always late in finishing the masterpieces which he committed to create.

The work which no one saw that he did in his mind was so exhausting that Mozart sometimes was not able to speak; he barked like a dog and behaved inappropriately sometimes like an idiot. He badly needed relaxation from his inner work, therefore he visited Vienna cafes (the simpler, the better) where he looked for a break from his exhausting concentration. He almost killed himself by such work that no one could see. By the end of his life (he was only 35 years old when he died) he was a very sick person who had used up his life: he had no time to properly build his family life (he married, almost by chance, the daughter of his landlord) and he had no time for friendship. He gave up everything for his genius.

One can say that this is just speculation; no one can tell you what was going on inside Mozart. That is true. But fortunately there is a recording made for Deutsche Gramophon called the "Magic of Horowitz." In this album there are 2 CDs and one bonus DVD. The DVD documents the recording of Mozart's 23rd piano concerto, played by the pianist of the century, Vladimir Horowitz. He is accompanied by one of the world's best orchestras, the orchestra of Teatro alla Scallo, under the direction of maestro Carlo Maria Giulini.

I believe that Horowitz's interpretation of Mozart's work and his uncompromised demand for excellence reflects Mozart's spirit for perfection.

4.4.3 Horowitz's Recording of Mozart's Piano Concerto

The record was made in 1987 in Horowitz's final years (he was 84). You see an old man who can hardly walk, and who probably has different health problems but who does not forget for a second about the necessity to play perfectly. In spite of all of his past achievements, he is not sure he will succeed. He asks his assistant (who turns the pages of his score), "Are my fingers good?" When one of the visitors (who came from England to Italy just to see this recording) tells him that she likes his tie he immediately reacts, "Do you like my tie more than my playing?" and repeats this several times throughout the session.

Probably the best part of this DVD is when, after recording the glamorous second movement, Horowitz, Giulini (who was selected by Horowitz for this record), and their producer evaluate the recording. You see the pianist's striving for perfection, the conductor's uncertainty, and how long it takes them to relax and agree that the record is good.

At the start of the third movement, Horowitz and the orchestra did not play together perfectly. The producer immediately stopped recording and suggested repeating it. You see how the great Horowitz without any doubt accepts his part of the fault and then how deeply he concentrated and how wonderfully he performed on the last movement in the next recording attempt.

Then he chats with musicians that came from all over the World just to see this session. The very last words of Horowitz on the DVD were his recollection of excellent reviews on the previous performance. However, he immediately added, "But this makes no difference."

That is, it does not matter how good you were yesterday; it makes no difference for today's results. Today a new challenge starts. That is the way of all great intellectual leaders.

4.4.4 Three Stories

Since ancient times people saw a very specific relationship between a genius and his professional work. Cicero formulated this as follows:

> *Among all features describing genius the most important is inner professional honesty.*

There are a lot of examples where the moral quality of a genius in everyday life does not meet high standards, but they never lose these high standards in their professional lives.

A person who plays games with professional honesty loses his demand for inner truth and compromises with himself. This leads to a decrease in his ability to look for the truth. Let me give examples of actions of my heros Kolmogorov, and Einstein.

Kolmogorov. There is a legend that Kolmogorov read everything. Nobody knows when and how he accomplished it, but somehow he did. In the beginning of the 1960s an unknown young researcher Ray Solomonoff, working for a small private company

in Boston released a report titled, " A Preliminary Report on a General Theory of Inductive Inference" [171]. This report contained ideas on inference and algorithmic complexity. Probably very few researchers read this report but Kolmogorov did. In 1965 Kolmogorov published his seminal paper where he introduced the concept of Kolmogorov complexity to answer the question "What is the nature of randomness?"

In this article he wrote that he was familiar with Solomonoff's work and that Solomonoff was the first to suggest the idea of algorithmic complexity.

In 2002 Solomonoff became the first person awarded the Kolmogorov medal established by the Royal Holloway University of London.

Einstein. In 1924 Einstein received a letter from the Indian researcher Bose which contained the handwritten manuscript "Planck's Law and the Hypothesis of Light Quanta," written in English. Einstein translated this letter into German and submitted it to the *Zeitschrift für Physik* with a strong recommendation. This was the main work of Bose. Later Einstein significantly extended the ideas of Bose and published several papers on this subject.

One can say that this just reflects human decency. Not only that. They also did these things to keep themselves honest in order not to betray their own individuality. The smallest compromise here leads to a compromised demand of yourself, which leads to a decline in creativity. They also did it because they had responsibility before their talents.

Galois. The last story is about Evariste Galois which I have been trying to understand ever since I read about it. This story is about the responsibility of the great talent with respect to results of his work.

Galois was a very talented mathematician and squabbling young man who during almost all of his short life produced nothing but trouble. As a result, this kid entered into a duel and was killed when he was just 20. The night before this duel he wrote down the mathematical theory that is now called "Galois Theory."

Why? He was not a stupid kid. He understood the consequences of not sleeping the night before a duel. Why didn't he think like this: "I must sleep to perform well tomorrow. This, is the most important thing. If I do not die I will write down the theory later, if I do die who cares?"

Why did this kid, who looked like just a troublemaker, accepted such a big responsibility? It was something bigger than himself. As with all great people he had a burden of responsibility that came along with his talent and Galois paid the full price of his life for this. He belonged not only to himself.

He was very different from most people.

4.4.5 Destructive Socialist Values

People are not born equal in their potential (they are not identical). Among those who are born there are future beautiful women, and future gentlemen, future musicians, and future scientists, there are very warm family people (this is also a great talent), and there are misanthropes. Among these who are born are future Mozarts and Einsteins. They

are all very different, and they bring a variety of human talents into society. People are not equal in their abilities.

This inequality of individuals inspired strong negative feelings among people who identify themselves as socialists. Socialist discussions on equality of individuals have continued for thousands of years. The main subject of these discussions has not been how to distribute wealth justly (this is a common misconception) but how to make people equal (essentially identical). One can find reflections of these discussions in philosophy, literature, social movements, and state systems. The practical implications of socialist ideas always were attempts to make people identical *by force*, not allowing them to be too different from an accepted standard.

A deep analysis of this phenomena was done by Igor Shafarevich in his book *The socialist phenomenon* [151].[12] The main results of his analysis is that the impulse to seek utopia (identity of individuals) is deeply rooted in the human psyche. This impulse, however, is very dangerous, because it leads to suppression of individuals in favor of unificated community and this in turn leads to societal decline.

Over thousands of years (ranging from ancient Mesopotamia and medieval Inca Empire to recent Russia and Cambodia) the socialist ideas has always brought humans to catastrophic consequences, but it has nonetheless always resurfaced, despite all past experience. For reasons that are unclear it can be attractive even to high-level intellectuals (Plato supported it in his *Republic*).

But despite the miserable failure of *all* classical socialist systems, its ideas remain attractive for many at the level of *socialist values*. This is how *The Concise Oxford Dictionary of Current English* [172] defines the essence of these values:

> Socialism is a principle that individual freedom should be completely subordinated to interest of community, with any deductions that may be correctly or incorrectly drawn from it.[13]

Renaming the Vapnik–Chervonenkis lemma as the Sauer lemma and the attempt to create the PAC legend described in Section 2.1 were beneficial to no one personally. These were the actions that execute the main slogan of a socialist community:

Expropriate extras and split them equally.

Socialist values have a very negative effect in science since they lead to a strong resistance to original ideas in favor of ones shared by a community.[14]

[12] I believe that one cannot consider his/her education complete without reading this book. It shows that the socialist phenomenon based on the idea of fundamental identity of individuals (udenticalness of individuals) existed from very ancient times and constitutes one of the basic forces of history. It is the instinct of *self-destruction* of a society that has a strong attraction for some of its parts (similar, for example, to a strong attraction for some individuals to jump down being on an edge of a very high place).

[13] I have had experience of the strong pressure of this principle both in Soviet Russia and in the United States. However below I refer only to my limited experience in the United States.

[14] As an illustration of this statement let me quote from three reviews on three different proposals that suggested to develop the ideas described in this Afterword and rejected by the US National Science Foundation (NSF) based on socialist arguments (I emphasize them in bold font).

REVIEW 1. Release date 07/21/2004. (On Science of Learning Center.)
The described intellectual merit is substantial and impressive. This proposal purports to break fundamentally new ground in our approach to scientific reasoning and developing powerful new algorithms. The claimed

This creates (using Popper's words [150, VIII]) *"the fundamental subjectivist position which can conceive knowledge only as a special kind of mental state, or as a disposition, or as a special kind of belief.*[15]

My understanding of the reason why:

> *Great spirits have always encountered violent opposition from mediocre minds* (A. Einstein)

is because the great spirits are *unique*, and this contradicts the socialist belief of mediocracy, the fundamental identity of individuals.

---

impact would be significant both in practice and in our fundamental view of science.

This is an extremely well written, argued, and engineered proposal for a center with a strong theme. It is a pleasure to read, and highly convincing. There is a level of excitement and importance that is communicated. This list of practical implications, in itself, is well worth the effort, but the theoretical ones are rather intriguing. These are highly qualified scientists asking for funding to accelerate development toward a new kind of reasoning. One does not routinely read a proposal like this.

Unfortunately the case for "centerness" is weak inasmuch as the work is rather narrowly structured, and poorly argued for in some of the critical characteristics. **The intellectual atmosphere might be enriched for the students if there were a great diversity of viewpoints, and it would enhance the case for forming a center.**

FROM REVIEW 2. Release date 06/23/2005. (On Empirical Inference in Social and Behavioural Science.) Dr. Vapnik proposes to develop a philosophy of science for the complex world using ideas from statistical learning theory, support vector machines, and transductive reasoning. The proposal was actually fascinating to read, and I thought the background information on SLT was especially clear. However, the scope of the proposal is enormous if the aim is to construct such a comprehensive philosophy and relate it to the social science. **One thing that strikes me is that this philosophy of science and ideas related to generalizability would all be centred around Dr. Vapnik's personal theoretical contributions to SLT.**

FROM REVIEW 3. Release date 09/17/2005. (On Directed Ad-Hoc Inference.)
Although the idea of developing a new type of inference is very interesting, **the proposed methodology heavily relies on the work previously done by the PI on support vector machines, and non-parametric estimation of conditional probabilities.**

The projects were rejected not because there were doubts about their scientific significance but because they are based on ideas that do not represent a widespread community.

[15] As an example of what is *to conceive knowledge as a belief* let me quote from another NSF review (on another proposal) that contains no arguments but agressive attack from a position of blind corporate belief (the quote from the review I made in bold font and my comments in parentheses).

FROM REVIEW 4. Release date 01/31/06. (On Relation to the Philosophy of Science.)
The discussion of philosophy of science is breathtakingly naive including an over-simplistic account of simplicity and complexity (this is about advanced complexity concepts, the VC dimension and VC entropy; is there something better? *V.V.*), **a completely implausible characterization of key methodological difference between physics and social science** (this is about Einstein's remark that the methodology of classical science cannot be used when one must consider too many factors and an attempt to create methodology for such situations *V.V.*), **and promise to give that dead-horse Popperian concept of corroboration another whipping.** (Sir Karl Popper was the first who tried to justify induction using the idea of capacity of set (falsifiability). He made mistakes. However, another capacity concept (the VC falsfiability) does define the necessary and sufficient conditions for predictive induction. *V.V.*). **Moreover, there is virtually no reference to pertinent philosophical discussion such as . . .** (two irrelevant traditional works *V.V.*).

It is sad to see the same collective socialist logic in all above reviews: *ideas of individuals should be completely subordinated to interest of community, with any deductions that may be correctly or incorrectly drawn from it*.

This is why there exist rude attack against great spirits demonstrating disrespect to them and in particular against Einstein as the brightest figure in scientific originality ("Many of his ideas were suggested by his wife." "He was not original since similar ideas were suggested by Poincare, Lorenz, and Minkowski." "He was a terrible family man," and so on). The same sort of criticism exists even against Sir Isaac Newton ("He spent most of his life working on stupid things and was a terrible man"). The main message is:

" Look at them, they are almost no different from us."

Maybe it is true that sometimes they behave like us (or maybe even worse than us). But they are very different in their vision of the truth, their devotion to this truth, and their honesty in pursuing the truth.

4.4.6 Theoretical Science Is Not Only a Profession — It Is a Way of Life

Being a natural science theorist is not only a profession, but it is also a difficult way of life:

— You come into this world with your individual seed of truth.

— You work hard to make your truth clear.

— You push yourself to be unconditionally honest with respect to your truth.

— You have a life-long fight protecting your truth from old paradigms.

— You resist strong pressure to betray your truth and become part of a socialist community.

If you have a talent, the character to bear such a life, and a little luck, then you have a chance to succeed: To come into this world bringing your own seed of truth, to work hard to make your truth clear, and to add it to the Grand Truth.

There is a warm feeling of deep satisfaction for those who have made it. And even if one cannot call this genuine happiness, it can be a very good substitute.

BIBLIOGRAPHY

For items with numbers 1 – 120 see pages 391 – 396.

121 Vapnik, V., and Chervonenkis, A.: The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Yearbook of the Academia of Sciences of USSR on Recognition, Classification, Forecasting* Vol 2, Moscow, Nauka, 207–249 (in Russian), 1989.

English translation: Vapnik, V., and Chervonenkis, A.:The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recogn. Image Anal.* **1** (3), 284-305, 1991.

122 Valiant, L.: Theory of the learnability. *Commun. ACM* **27**, 1134–1142, 1984.

123 Valiant,L.: A view of computational learning theory. In *Computation & Cognition*, Proceeding of the First NEC Research Symposium, Society for Industrial and Applied Mathematics, Philadelphia, 32–51, 1990.

124 Gurvits, L.: A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In Lee and Maruoka (Eds), *Algorithmic Learning Theory* ALT-97, LNAI-1316, Berlin, Springer, 352–363, 1997.

125 Boser, B., Guyon, I., and Vapnik, V.: A training algorithm for an optimal margin classifier. *Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh AMC, 144–152, 1992.

126 Aizerman, M., Braverman, I., and Rosonoer, L.: Theoretical foundations of the potential functions method in pattern recognition learning. *Automation and Remote Control* **25**, 821–837.

127 Wallace, C., and Boulton, D.: An information measure for classification. *Comput. J.* **11**, 85–95, 1968.

128 Rissanen, J.: Modelling by shortest data description. *Automatica* **14**, 465–471, 1978.

129 Kolmogorov, A.: Three approaches to the quantitative definition of information. *Prob. Inf. Transm.* **1** (1), 1–7, 1965.

130 Sauer, N.: On the density of families of sets. *J. Comb. Theor. (A)* **13**, 145–147, 1972.

131 Shelah, S.: A computational problem: Stability and order of models and theory of infinitary languages. *Pacific J. Math.* **41**, 247–261, 1972.

132 Cortes, C., and Vapnik, V.: Support vector networks. *Mach. Learn.* **20**, 1–25, 1995.

133 Lugosi, G., and Zeger, K.: Concept of learning using complexity regularization. *IEEE Trans. on Information Theory* **41**, 677–678, 1994.

134 Devroye, L., and Wagner, T.: Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics* **8**, 231–239, 1982.

135 Freund, Y. and Schapire, R.: Experiments with new boosting algorithms. In *Proceedings of the 13th International Conference on Machine Learning*, San Francisco: Morgan-Kaufmann, 148–156, 1996.

136 Schapire, R., Freund, Y., Bartlett, P., and Lee, W.: Boosting the margin: A new explanation for effectiveness of voting methods. *Ann. Stat.* **26**, 1651–1686, 1998.

137 Popper, K.: *The Logic of scientific discovery*(2nd ed.). New York: Harper Torch, 1968.

138 Shawe-Taylor, J., Bartlett, P., Williamson, C., and Anthony, M.: Structural risk minimization over data-dependent hierarchies. IEEE Trans. Info. Theor. **44** (5), 1926–1940, 1998.

139 Vapnik, V.: *The nature of statistical learning theory*. New York: Springer, 1995.

140 Vapnik, V.: *Statistical learning theory*. New York: Wiley, 1998.

141 Alon, N., Ben-David, S., Cesa-Bianch, N., and Haussler, D.: Scale-sensitive dimension, uniform convergence, and learnability. *J. ACM* **44**(4), 615–631, 1997.

142 Bottou, L., Cortes, C., and Vapnik, V.: On the effective VC dimension. Tech. Rep. Neuroprose, (ftp://archive.cis.ohio–state.edu/pub/neuroprose), 1994.

143 Vapnik, V., and Chervonenkis, A.: Uniform convergence of frequencies of occurrence of events to their probabilities. *Dokl. Akad. Nauk USSR* **181**, 915–918, 1968. (English translation: *Soviet Math. Dokl.* **9**, 4, (1968).)

144 Vapnik, V., Levin, E., and LeCun, Y.: Measuring the VC dimension of a learning machine. *Neural Computation* **10** (5), 1994.

145 Bottou, L., and Vapnik, V.: Local learning algorithms. *Neural Computation* **6** (6), 888–901, 1992.

146 Weston, J., Perez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., and Shölkopf, B.: KDD cup 2001 data analysis: prediction of molecular bioactivity for drug design—binding to thrombin. *Bioinformatics*, 2003.

147 Nadaraya, E.: On estimating regression. *Theor. Prob. Appl.* **9**, 1964.

148 Watson,G.: Smooth regression analysis, *Shankhaya*, Seria A **26**, 1964.

149 Stute, W.: On almost sure convergence of conditional empirical distribution function, *Ann. Probab.* **14**(3), 891–901, 1986.

150 Popper, K.: *Conjectures and Refutations*. New York: Routledge, 2000.

151 Shafarevich, I.: *The Socialist phenomenon*. New York: Harper and Row, 1980. (Currently this book is out of print. One can find it at http://www.robertlstephens.com/essays/shafarevich/001SocialistPhenomenon.html.)

152 Schölkopf, B., and Smola, A.: *Learning with kernels*, Cambridge MA: MIT Press, 2002.

153 Cristianini, N., and Shawe-Taylor, J.: *An introduction to Support Vector Machines*. Cambridge: Cambridge University Press, 2000.

154 Joachims, T.: *Learning to classify text using support vector machines*, Hingham, MA: Kluwer Academic Publishers, 2002.

155 Herbrich, R.: *Learning kernel classifiers: theory and algorithms*. Cambridge MA: MIT Press, 2002.

156 Shawe-Taylor, J., and Cristianini, N.: *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press, 2004.

157 Abe, S.: *Support vector machines for pattern classification*. New York: Springer, 2005

158 Schölkopf, B., Burges, C., and Smola, A. (Eds.): *Advances in Kernel Methods. Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

159 Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D. (Eds.) *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 2000.

160 Schölkopf, B., Tsuda, K., and Vert, J. (Eds.): Kernel Methods in Computational Biology. Cambridge, MA: MIT Press, 2004.

161 Chapelle, O., Schölkopf, B., and Zien, A. (Eds.): *Semi-supervised Learning*, Cambridge MA: MIT Press, 2006.

162 Wang, L. (Ed.): *Support vector machines: theory and applications*, New York: Springer, 2005.

163 Lee, S.-W, and Verri, A.(Eds.): *Pattern recognition with support vector machines*, New York: Springer, 2002.

164 Schölkopf, B., and Warmuth, M. (Eds.): *Learning theory and kernel machines*. New York: Springer, 2003.

165 Burges, C.: A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, **2**(2), 121–167, 1995.

166 Phillips, D.: A technique for numerical solution of certain integral equation of first kind. *J. Assoc. Comput. Math.* **9**, 84–96, 1962.

167 Steinwart, I.: Consistency of Support Vector Machines and other regularized kernel machines. *IEEE Trans. Info. Theor.* , **51**, 128–142, 2005.

168 Kuhn, T.: *The Structure of Scientific Revolutions* (3rd ed.). Chicago: Chicago University Press, 1996.

169 Hempel, K.: *The philosophy of K. Hempel: Studies in Science, Explanation, and Rationality*. Oxford, UK: Oxford University Press, 2001.

170 Chaitin, G.: On the length of programs for computing finite binary sequences. *J. Assoc. Comput. Math.* **4**, 547–569, 1996.

171 Solomonov, R.: A preliminary report on general theory of inductive inference, Technical Report ZTB–138, Cambridge, MA: Zator Company, 1960.

172 Fowler, H.W. (Ed.): The Concise Oxford Dictionary of Current English, Oxford, UK: Oxford University Press, 1956.

173 Wapnik, W.N., Tscherwonenkis, A. Ya: *Theorie der Zeichenerkennung*. Berlin: Akademia, 1979. (German translation of the book [12]).

174 Chapelle O. and Zien A.: Semi-supervised classification of low density separation. Proc. of the Thenth International Workshop on Artificial Intelligence and Statistics, pp 57 – 64, 2005

INDEX