

Vladimir N. Vapnik

The Nature of Statistical Learning Theory

Second Edition

With 50 Illustrations



Springer

Vladimir N. Vapnik
AT&T Labs—Research
Room 3-130
100 Schultz Drive
Red Bank, NJ 07701
USA
vlad@research.att.com

Series Editors

Michael Jordan
Department of Computer Science
University of California, Berkeley
Berkeley, CA 94720
USA

Jerald F. Lawless
Department of Statistics
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

Steffen L. Lauritzen
Department of Mathematical Sciences
Aalborg University
DK-9220 Aalborg
Denmark

Vijay Nair
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
USA

Library of Congress Cataloging-in-Publication Data
Vapnik, Vladimir Naumovich.

The nature of statistical learning theory/Vladimir N. Vapnik.

— 2nd ed.

p. cm. — (Statistics for engineering and information
science)

Includes bibliographical references and index.

ISBN 978-1-4419-3160-3 ISBN 978-1-4757-3264-1 (eBook)

DOI 10.1007/978-1-4757-3264-1

1. Computational learning theory. 2. Reasoning. I. Title.

II. Series.

Q325.7.V37 1999

006.3'1'015195—dc21

99-39803

Printed on acid-free paper.

© 2000, 1995 Springer Science+Business Media New York
Originally published by Springer-Verlag New York, Inc. in 2000
Softcover reprint of the hardcover 2nd edition 2000

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher Springer Science+Business Media, LLC, except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Frank McGuckin; manufacturing supervised by Erica Bresler.
Photocomposed copy prepared from the author's L^AT_EX files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4419-3160-3

SPIN 10713304

In memory of my mother

Preface to the Second Edition

Four years have passed since the first edition of this book. These years were “fast time” in the development of new approaches in statistical inference inspired by learning theory.

During this time, new function estimation methods have been created where a high dimensionality of the unknown function does not always require a large number of observations in order to obtain a good estimate. The new methods control generalization using capacity factors that do not necessarily depend on dimensionality of the space.

These factors were known in the VC theory for many years. However, the practical significance of capacity control has become clear only recently after the appearance of support vector machines (SVM). In contrast to classical methods of statistics where in order to control performance one decreases the dimensionality of a feature space, the SVM dramatically increases dimensionality and relies on the so-called large margin factor.

In the first edition of this book general learning theory including SVM methods was introduced. At that time SVM methods of learning were brand new, some of them were introduced for a first time. Now SVM margin control methods represents one of the most important directions both in theory and application of learning.

In the second edition of the book three new chapters devoted to the SVM methods were added. They include generalization of SVM method for estimating real-valued functions, direct methods of learning based on solving (using SVM) multidimensional integral equations, and extension of the empirical risk minimization principle and its application to SVM.

The years since the first edition of the book have also changed the general

philosophy in our understanding the of nature of the induction problem. After many successful experiments with SVM, researchers became more determined in criticism of the classical philosophy of generalization based on the principle of Occam's razor.

This intellectual determination also is a very important part of scientific achievement. Note that the creation of the new methods of inference could have happened in the early 1970: All the necessary elements of the theory and the SVM algorithm were known. It took twenty-five years to reach this intellectual determination.

Now the analysis of generalization from the pure theoretical issues become a very practical subject, and this fact adds important details to a general picture of the developing computer learning problem described in the first edition of the book.

Red Bank, New Jersey
August 1999

Vladimir N. Vapnik

Preface to the First Edition

Between 1960 and 1980 a revolution in statistics occurred: Fisher's paradigm, introduced in the 1920s and 1930s was replaced by a new one. This paradigm reflects a new answer to the fundamental question:

*What must one know *a priori* about an unknown functional dependency in order to estimate it on the basis of observations?*

In Fisher's paradigm the answer was very restrictive—one must know almost everything. Namely, one must know the desired dependency up to the values of a finite number of parameters. Estimating the values of these parameters was considered to be the problem of dependency estimation.

The new paradigm overcame the restriction of the old one. It was shown that in order to estimate dependency from the data, it is sufficient to know some general properties of the set of functions to which the unknown dependency belongs.

Determining general conditions under which estimating the unknown dependency is possible, describing the (inductive) principles that allow one to find the best approximation to the unknown dependency, and finally developing effective algorithms for implementing these principles are the subjects of the new theory.

Four discoveries made in the 1960s led to the revolution:

- (i) Discovery of regularization principles for solving ill-posed problems by Tikhonov, Ivanov, and Phillips.
- (ii) Discovery of nonparametric statistics by Parzen, Rosenblatt, and Chentsov.

- (iii) Discovery of the law of large numbers in functional space and its relation to the learning processes by Vapnik and Chervonenkis.
- (iv) Discovery of algorithmic complexity and its relation to inductive inference by Kolmogorov, Solomonoff, and Chaitin.

These four discoveries also form a basis for any progress in studies of learning processes.

The problem of learning is so general that almost any question that has been discussed in statistical science has its analog in learning theory. Furthermore, some very important general results were first found in the framework of learning theory and then reformulated in the terms of statistics.

In particular, learning theory for the first time stressed the problem of *small sample statistics*. It was shown that by taking into account the size of the sample one can obtain better solutions to many problems of function estimation than by using the methods based on classical statistical techniques.

Small sample statistics in the framework of the new paradigm constitutes an advanced subject of research both in statistical learning theory and in theoretical and applied statistics. The rules of statistical inference developed in the framework of the new paradigm should not only satisfy the existing asymptotic requirements but also guarantee that one does one's best in using the available restricted information. The result of this theory is new methods of inference for various statistical problems.

To develop these methods (which often contradict intuition), a comprehensive theory was built that includes:

- (i) Concepts describing the necessary and sufficient conditions for consistency of inference.
- (ii) Bounds describing the generalization ability of learning machines based on these concepts.
- (iii) Inductive inference for small sample sizes, based on these bounds.
- (iv) Methods for implementing this new type of inference.

Two difficulties arise when one tries to study statistical learning theory: a technical one and a conceptual one—to understand the proofs and to understand the nature of the problem, its philosophy.

To overcome the technical difficulties one has to be patient and persistent in following the details of the formal inferences.

To understand the nature of the problem, its spirit, and its philosophy, one has to see the theory as a whole, not only as a collection of its different parts. Understanding the nature of the problem is extremely important

because it leads to searching in the right direction for results and prevents searching in wrong directions.

The goal of this book is to describe the nature of statistical learning theory. I would like to show how abstract reasoning implies new algorithms. To make the reasoning easier to follow, I made the book short.

I tried to describe things as simply as possible but without conceptual simplifications. Therefore, the book contains neither details of the theory nor proofs of the theorems (both details of the theory and proofs of the theorems can be found (partly) in my 1982 book *Estimation of Dependencies Based on Empirical Data* (Springer) and (in full) in my book *Statistical Learning Theory* (J. Wiley, 1998)). However, to describe the ideas without simplifications I needed to introduce new concepts (new mathematical constructions) some of which are nontrivial.

The book contains an introduction, five chapters, informal reasoning and comments on the chapters, and a conclusion.

The introduction describes the history of the study of the learning problem which is not as straightforward as one might think from reading the main chapters.

Chapter 1 is devoted to the setting of the learning problem. Here the general model of minimizing the risk functional from empirical data is introduced.

Chapter 2 is probably both the most important one for understanding the new philosophy and the most difficult one for reading. In this chapter, the conceptual theory of learning processes is described. This includes the concepts that allow construction of the necessary and sufficient conditions for consistency of the learning processes.

Chapter 3 describes the nonasymptotic theory of bounds on the convergence rate of the learning processes. The theory of bounds is based on the concepts obtained from the conceptual model of learning.

Chapter 4 is devoted to a theory of small sample sizes. Here we introduce inductive principles for small sample sizes that can control the generalization ability.

Chapter 5 describes, along with classical neural networks, a new type of universal learning machine that is constructed on the basis of small sample sizes theory.

Comments on the chapters are devoted to describing the relations between classical research in mathematical statistics and research in learning theory.

In the conclusion some open problems of learning theory are discussed.

The book is intended for a wide range of readers: students, engineers, and scientists of different backgrounds (statisticians, mathematicians, physicists, computer scientists). Its understanding does not require knowledge of special branches of mathematics. Nevertheless, it is not easy reading, since the book does describe a (conceptual) forest even if it does not con-

sider the (mathematical) trees.

In writing this book I had one more goal in mind: I wanted to stress the practical power of abstract reasoning. The point is that during the last few years at different computer science conferences, I heard reiteration of the following claim:

Complex theories do not work, simple algorithms do.

One of the goals of this book is to show that, at least in the problems of statistical inference, this is not true. I would like to demonstrate that in this area of science a good old principle is valid:

Nothing is more practical than a good theory.

The book is not a survey of the standard theory. It is an attempt to promote a certain point of view not only on the problem of learning and generalization but on theoretical and applied statistics as a whole.

It is my hope that the reader will find the book interesting and useful.

AKNOWLEDGMENTS

This book became possible due to the support of Larry Jackel, the head of the Adaptive System Research Department, AT&T Bell Laboratories.

It was inspired by collaboration with my colleagues Jim Alvich, Jan Ben, Yoshua Bengio, Bernhard Boser, Léon Bottou, Jane Bromley, Chris Burges, Corinna Cortes, Eric Cosatto, Joanne DeMarco, John Denker, Harris Drucker, Hans Peter Graf, Isabelle Guyon, Patrick Haffner, Donnie Henderson, Larry Jackel, Yann LeCun, Robert Lyons, Nada Matic, Urs Mueller, Craig Nohl, Edwin Pednault, Eduard Säckinger, Bernhard Schölkopf, Patrice Simard, Sara Solla, Sandi von Pier, and Chris Watkins.

Chris Burges, Edwin Pednault, and Bernhard Schölkopf read various versions of the manuscript and improved and simplified the exposition.

When the manuscript was ready I gave it to Andrew Barron, Yoshua Bengio, Robert Berwick, John Denker, Federico Girosi, Ilia Izmailov, Larry Jackel, Yakov Kogan, Esther Levin, Vincent Mirelly, Tomaso Poggio, Edward Reitman, Alexander Shustorovich, and Chris Watkins for remarks. These remarks also improved the exposition.

I would like to express my deep gratitude to everyone who helped make this book.

Red Bank, New Jersey
March 1995

Vladimir N. Vapnik

Contents

Preface to the Second Edition	vii
Preface to the First Edition	ix
Introduction: Four Periods in the Research of the Learning Problem	1
Rosenblatt's Perceptron (The 1960s)	1
Construction of the Fundamentals of Learning Theory (The 1960s–1970s)	7
Neural Networks (The 1980s)	11
Returning to the Origin (The 1990s)	14
Chapter 1 Setting of the Learning Problem	17
1.1 Function Estimation Model	17
1.2 The Problem of Risk Minimization	18
1.3 Three Main Learning Problems	18
1.3.1 Pattern Recognition	19
1.3.2 Regression Estimation	19
1.3.3 Density Estimation (Fisher–Wald Setting)	19
1.4 The General Setting of the Learning Problem	20
1.5 The Empirical Risk Minimization (ERM) Inductive Principle	20
1.6 The Four Parts of Learning Theory	21
Informal Reasoning and Comments — 1	23

1.7	The Classical Paradigm of Solving Learning Problems	23
1.7.1	Density Estimation Problem (Maximum Likelihood Method)	24
1.7.2	Pattern Recognition (Discriminant Analysis) Problem	24
1.7.3	Regression Estimation Model	25
1.7.4	Narrowness of the ML Method	26
1.8	Nonparametric Methods of Density Estimation	27
1.8.1	Parzen's Windows	27
1.8.2	The Problem of Density Estimation Is Ill-Posed	28
1.9	Main Principle for Solving Problems Using a Restricted Amount of Information	30
1.10	Model Minimization of the Risk Based on Empirical Data	31
1.10.1	Pattern Recognition	31
1.10.2	Regression Estimation	31
1.10.3	Density Estimation	32
1.11	Stochastic Approximation Inference	33
Chapter 2	Consistency of Learning Processes	35
2.1	The Classical Definition of Consistency and the Concept of Nontrivial Consistency	36
2.2	The Key Theorem of Learning Theory	38
2.2.1	Remark on the ML Method	39
2.3	Necessary and Sufficient Conditions for Uniform Two-Sided Convergence	40
2.3.1	Remark on Law of Large Numbers and Its Generalization	41
2.3.2	Entropy of the Set of Indicator Functions	42
2.3.3	Entropy of the Set of Real Functions	43
2.3.4	Conditions for Uniform Two-Sided Convergence	45
2.4	Necessary and Sufficient Conditions for Uniform One-Sided Convergence	45
2.5	Theory of Nonfalsifiability	47
2.5.1	Kant's Problem of Demarcation and Popper's Theory of Nonfalsifiability	47
2.6	Theorems on Nonfalsifiability	49
2.6.1	Case of Complete (Popper's) Nonfalsifiability	50
2.6.2	Theorem on Partial Nonfalsifiability	50
2.6.3	Theorem on Potential Nonfalsifiability	52
2.7	Three Milestones in Learning Theory	55
Informal Reasoning and Comments — 2		59
2.8	The Basic Problems of Probability Theory and Statistics	60
2.8.1	Axioms of Probability Theory	60
2.9	Two Modes of Estimating a Probability Measure	63

2.10 Strong Mode Estimation of Probability Measures and the Density Estimation Problem	65
2.11 The Glivenko–Cantelli Theorem and its Generalization	66
2.12 Mathematical Theory of Induction	67
Chapter 3 Bounds on the Rate of Convergence of Learning Processes	69
3.1 The Basic Inequalities	70
3.2 Generalization for the Set of Real Functions	72
3.3 The Main Distribution–Independent Bounds	75
3.4 Bounds on the Generalization Ability of Learning Machines	76
3.5 The Structure of the Growth Function	78
3.6 The VC Dimension of a Set of Functions	80
3.7 Constructive Distribution–Independent Bounds	83
3.8 The Problem of Constructing Rigorous (Distribution–Dependent) Bounds	85
Informal Reasoning and Comments — 3	87
3.9 Kolmogorov–Smirnov Distributions	87
3.10 Racing for the Constant	89
3.11 Bounds on Empirical Processes	90
Chapter 4 Controlling the Generalization Ability of Learning Processes	93
4.1 Structural Risk Minimization (SRM) Inductive Principle	94
4.2 Asymptotic Analysis of the Rate of Convergence	97
4.3 The Problem of Function Approximation in Learning Theory	99
4.4 Examples of Structures for Neural Nets	101
4.5 The Problem of Local Function Estimation	103
4.6 The Minimum Description Length (MDL) and SRM Principles	104
4.6.1 The MDL Principle	106
4.6.2 Bounds for the MDL Principle	107
4.6.3 The SRM and MDL Principles	108
4.6.4 A Weak Point of the MDL Principle	110
Informal Reasoning and Comments — 4	111
4.7 Methods for Solving Ill-Posed Problems	112
4.8 Stochastic Ill-Posed Problems and the Problem of Density Estimation	113
4.9 The Problem of Polynomial Approximation of the Regression	115
4.10 The Problem of Capacity Control	116
4.10.1 Choosing the Degree of the Polynomial	116
4.10.2 Choosing the Best Sparse Algebraic Polynomial	117
4.10.3 Structures on the Set of Trigonometric Polynomials	118

4.10.4	The Problem of Features Selection	119
4.11	The Problem of Capacity Control and Bayesian Inference .	119
4.11.1	The Bayesian Approach in Learning Theory	119
4.11.2	Discussion of the Bayesian Approach and Capacity Control Methods	121
Chapter 5 Methods of Pattern Recognition		123
5.1	Why Can Learning Machines Generalize?	123
5.2	Sigmoid Approximation of Indicator Functions	125
5.3	Neural Networks	126
5.3.1	The Back-Propagation Method	126
5.3.2	The Back-Propagation Algorithm	130
5.3.3	Neural Networks for the Regression Estimation Problem	130
5.3.4	Remarks on the Back-Propagation Method	130
5.4	The Optimal Separating Hyperplane	131
5.4.1	The Optimal Hyperplane	131
5.4.2	Δ -margin hyperplanes	132
5.5	Constructing the Optimal Hyperplane	133
5.5.1	Generalization for the Nonseparable Case	136
5.6	Support Vector (SV) Machines	138
5.6.1	Generalization in High-Dimensional Space	139
5.6.2	Convolution of the Inner Product	140
5.6.3	Constructing SV Machines	141
5.6.4	Examples of SV Machines	141
5.7	Experiments with SV Machines	146
5.7.1	Example in the Plane	146
5.7.2	Handwritten Digit Recognition	147
5.7.3	Some Important Details	151
5.8	Remarks on SV Machines	154
5.9	SVM and Logistic Regression	156
5.9.1	Logistic Regression	156
5.9.2	The Risk Function for SVM	159
5.9.3	The SVM _n Approximation of the Logistic Regression	160
5.10.	Ensemble of the SVM	163
5.10.1	The AdaBoost Method	164
5.10.2	The Ensemble of SVMs	167
Informal Reasoning and Comments — 5		171
5.11	The Art of Engineering Versus Formal Inference	171
5.12	Wisdom of Statistical Models	174
5.13	What Can One Learn from Digit Recognition Experiments?	176
5.13.1	Influence of the Type of Structures and Accuracy of Capacity Control	177

5.13.2 SRM Principle and the Problem of Feature Construction	178
5.13.3 Is the Set of Support Vectors a Robust Characteristic of the Data?	179
Chapter 6 Methods of Function Estimation	181
6.1 ε -Insensitive Loss-Function	181
6.2 SVM for Estimating Regression Function	183
6.2.1 SV Machine with Convolved Inner Product	186
6.2.2 Solution for Nonlinear Loss Functions	188
6.2.3 Linear Optimization Method	190
6.3 Constructing Kernels for Estimating Real-Valued Functions	190
6.3.1 Kernels Generating Expansion on Orthogonal Polynomials	191
6.3.2 Constructing Multidimensional Kernels	193
6.4 Kernels Generating Splines	194
6.4.1 Spline of Order d With a Finite Number of Nodes	194
6.4.2 Kernels Generating Splines With an Infinite Number of Nodes	195
6.5 Kernels Generating Fourier Expansions	196
6.5.1 Kernels for Regularized Fourier Expansions	197
6.6 The Support Vector ANOVA Decomposition for Function Approximation and Regression Estimation	198
6.7 SVM for Solving Linear Operator Equations	200
6.7.1 The Support Vector Method	201
6.8 Function Approximation Using the SVM	204
6.8.1 Why Does the Value of ε Control the Number of Support Vectors?	205
6.9 SVM for Regression Estimation	208
6.9.1 Problem of Data Smoothing	209
6.9.2 Estimation of Linear Regression Functions	209
6.9.3 Estimation Nonlinear Regression Functions	216
Informal Reasoning and Comments — 6	219
6.10 Loss Functions for the Regression Estimation Problem	219
6.11 Loss Functions for Robust Estimators	221
6.12 Support Vector Regression Machine	223
Chapter 7 Direct Methods in Statistical Learning Theory	225
7.1 Problem of Estimating Densities, Conditional Probabilities, and Conditional Densities	226
7.1.1 Problem of Density Estimation: Direct Setting	226
7.1.2 Problem of Conditional Probability Estimation	227
7.1.3 Problem of Conditional Density Estimation	228

7.2	Solving an Approximately Determined Integral Equation	229
7.3	Glivenko-Cantelli Theorem	230
7.3.1	Kolmogorov-Smirnov Distribution	232
7.4	Ill-Posed Problems	233
7.5	Three Methods of Solving Ill-Posed Problems	235
7.5.1	The Residual Principle	236
7.6	Main Assertions of the Theory of Ill-Posed Problems	237
7.6.1	Deterministic Ill-Posed Problems	237
7.6.2	Stochastic Ill-Posed Problem	238
7.7	Nonparametric Methods of Density Estimation	240
7.7.1	Consistency of the Solution of the Density Estimation Problem	240
7.7.2	The Parzen's Estimators	241
7.8	SVM Solution of the Density Estimation Problem	244
7.8.1	The SVM Density Estimate: Summary	247
7.8.2	Comparison of the Parzen's and the SVM methods	248
7.9	Conditional Probability Estimation	249
7.9.1	Approximately Defined Operator	251
7.9.2	SVM Method for Conditional Probability Estimation	253
7.9.3	The SVM Conditional Probability Estimate: Summary	255
7.10	Estimation of Conditional Density and Regression	256
7.11	Remarks	258
7.11.1	One Can Use a Good Estimate of the Unknown Density	258
7.11.2	One Can Use Both Labeled (Training) and Unlabeled (Test) Data	259
7.11.3	Method for Obtaining Sparse Solutions of the Ill- Posed Problems	259
Informal Reasoning and Comments — 7		261
7.12	Three Elements of a Scientific Theory	261
7.12.1	Problem of Density Estimation	262
7.12.2	Theory of Ill-Posed Problems	262
7.13	Stochastic Ill-Posed Problems	263
Chapter 8 The Vicinal Risk Minimization Principle and the SVMs		267
8.1	The Vicinal Risk Minimization Principle	267
8.1.1	Hard Vicinity Function	269
8.1.2	Soft Vicinity Function	270
8.2	VRM Method for the Pattern Recognition Problem	271
8.3	Examples of Vicinal Kernels	275
8.3.1	Hard Vicinity Functions	276
8.3.2	Soft Vicinity Functions	279

8.4 Nonsymmetric Vicinities	279
8.5. Generalization for Estimation Real-Valued Functions	281
8.6 Estimating Density and Conditional Density	284
8.6.1 Estimating a Density Function	284
8.6.2 Estimating a Conditional Probability Function	285
8.6.3 Estimating a Conditional Density Function	286
8.6.4 Estimating a Regression Function	287
Informal Reasoning and Comments — 8	289
Chapter 9 Conclusion: What Is Important in Learning Theory?	291
9.1 What Is Important in the Setting of the Problem?	291
9.2 What Is Important in the Theory of Consistency of Learning Processes?	294
9.3 What Is Important in the Theory of Bounds?	295
9.4 What Is Important in the Theory for Controlling the Generalization Ability of Learning Machines?	296
9.5 What Is Important in the Theory for Constructing Learning Algorithms?	297
9.6 What Is the Most Important?	298
References	301
Remarks on References	301
References	302
Index	311

Introduction: Four Periods in the Research of the Learning Problem

In the history of research of the learning problem one can extract four periods that can be characterized by four bright events:

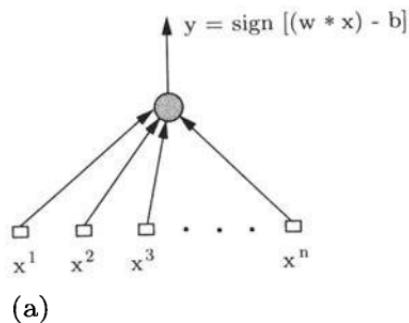
- (i) Constructing the first learning machines,
- (ii) constructing the fundamentals of the theory,
- (iii) constructing neural networks,
- (iv) constructing the alternatives to neural networks.

In different periods, different subjects of research were considered to be important. Altogether this research forms a complicated (and contradictory) picture of the exploration of the learning problem.

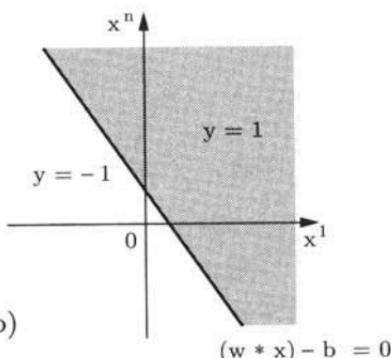
ROSENBLATT'S PERCEPTRON (THE 1960S)

More than thirty five years ago F. Rosenblatt suggested the first model of a learning machine, called the perceptron; this is when the mathematical analysis of learning processes truly began.¹ From the conceptual point of

¹Note that discriminant analysis as proposed in the 1930s by Fisher actually did not consider the problem of inductive inference (the problem of estimating the discriminant rules using the examples). This happened later, after Rosenblatt's work. In the 1930s discriminant analysis was considered a problem of constructing a decision rule separating two categories of vectors using given probability distribution functions for these categories of vectors.



(a)



(b)

FIGURE 0.1. (a) Model of a neuron. (b) Geometrically, a neuron defines two regions in input space where it takes the values -1 and 1 . These regions are separated by the hyperplane $(w \cdot x) - b = 0$.

view, the idea of the perceptron was not new. It had been discussed in the neurophysiologic literature for many years. Rosenblatt, however, did something unusual. He described the model as a program for computers and demonstrated with simple experiments that this model can be generalized. The perceptron was constructed to solve pattern recognition problems; in the simplest case this is the problem of constructing a rule for separating data of two different categories using given examples.

The Perceptron Model

To construct such a rule the perceptron uses adaptive properties of the simplest neuron model (Rosenblatt, 1962). Each neuron is described by the McCulloch–Pitts model, according to which the neuron has n inputs $x = (x^1, \dots, x^n) \in X \subset R^n$ and one output $y \in \{-1, 1\}$ (Fig. 0.1). The output is connected with the inputs by the functional dependence

$$y = \text{sign} \{(w \cdot x) - b\},$$

where $(u \cdot v)$ is the inner product of two vectors, b is a threshold value, and $\text{sign}(u) = 1$ if $u > 0$ and $\text{sign}(u) = -1$ if $u \leq 0$.

Geometrically speaking, the neurons divide the space X into two regions: a region where the output y takes the value 1 and a region where the output y takes the value -1 . These two regions are separated by the hyperplane

$$(w \cdot x) - b = 0.$$

The vector w and the scalar b determine the position of the separating hyperplane. During the learning process the perceptron chooses appropriate coefficients of the neuron.

Rosenblatt considered a model that is a composition of several neurons: He considered several levels of neurons, where outputs of neurons of the previous level are inputs for neurons of the next level (the output of one neuron can be input to several neurons). The last level contains only one neuron. Therefore, the (elementary) perceptron has n inputs and one output.

Geometrically speaking, the perceptron divides the space X into two parts separated by a piecewise linear surface (Fig. 0.2). Choosing appropriate coefficients for all neurons of the net, the perceptron specifies two regions in X space. These regions are separated by piecewise linear surfaces (not necessarily connected). Learning in this model means finding appropriate coefficients for all neurons using given training data.

In the 1960s it was not clear how to choose the coefficients simultaneously for all neurons of the perceptron (the solution came twenty five years later). Therefore, Rosenblatt suggested the following scheme: to fix the coefficients of all neurons, except for the last one, and during the training process to try to find the coefficients of the last neuron. Geometrically speaking, he suggested transforming the input space X into a new space Z (by choosing appropriate coefficients of all neurons except for the last) and to use the training data to construct a separating hyperplane in the space Z .

Following the traditional physiological concepts of learning with reward and punishment stimulus, Rosenblatt proposed a simple algorithm for iteratively finding the coefficients.

Let

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

be the training data given in input space and let

$$(z_1, y_1), \dots, (z_\ell, y_\ell)$$

be the corresponding training data in Z (the vector z_i is the transformed x_i). At each time step k , let one element of the training data be fed into the perceptron. Denote by $w(k)$ the coefficient vector of the last neuron at this time. The algorithm consists of the following:

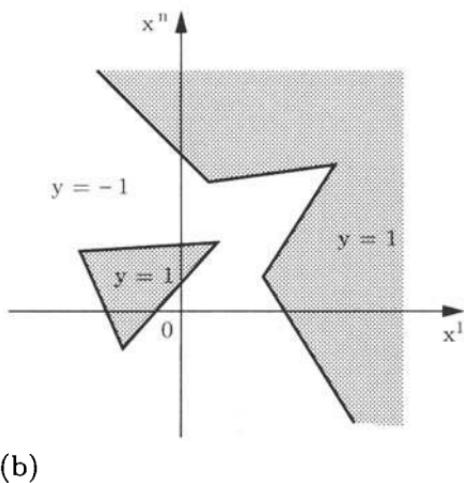
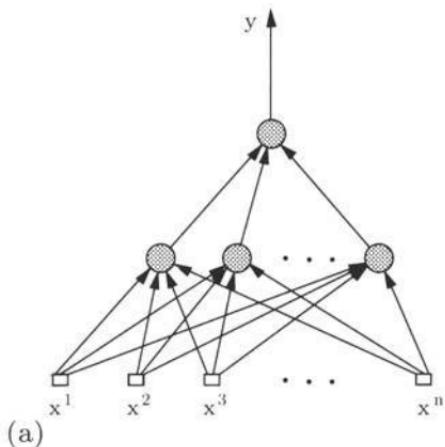


FIGURE 0.2. (a) The perceptron is a composition of several neurons. (b) Geometrically, the perceptron defines two regions in input space where it takes the values -1 and 1 . These regions are separated by a piecewise linear surface.

- (i) If the next example of the training data z_{k+1}, y_{k+1} is classified correctly, i.e.,

$$y_{k+1}(w(k) \dots z_{k+1}) > 0,$$

then the coefficient vector of the hyperplane is not changed,

$$w(k+1) = w(k).$$

- (ii) If, however, the next element is classified incorrectly, i.e.,

$$y_{k+1}(w_i(k) \cdot z_{k+1}) < 0,$$

then the vector of coefficients is changed according to the rule

$$w(k+1) = w(k) + y_{k+1}z_{k+1}.$$

- (iii) The initial vector w is zero:

$$w(1) = 0.$$

Using this rule the perceptron demonstrated generalization ability on simple examples.

Beginning the Analysis of Learning Processes

In 1962 Novikoff proved the first theorem about the perceptron (Novikoff, 1962). This theorem actually started learning theory. It asserts that if

- (i) the norm of the training vectors z is bounded by some constant R ($|z| \leq R$);
- (ii) the training data can be separated with margin ρ :

$$\sup_w \min_i y_i(z_i \cdot w) > \rho;$$

- (iii) the training sequence is presented to the perceptron a sufficient number of times,

then after at most

$$N \leq \left\lceil \frac{R^2}{\rho^2} \right\rceil$$

corrections the hyperplane that separates the training data will be constructed.

This theorem played an extremely important role in creating learning theory. It somehow connected the cause of generalization ability with the principle of minimizing the number of errors on the training set. As we will see in the last chapter, the expression $[R^2/\rho^2]$ describes an important concept that for a wide class of learning machines allows control of generalization ability.

Applied and Theoretical Analysis of Learning Processes

Novikoff proved that the perceptron can separate *training data*. Using exactly the same technique, one can prove that if the data are separable, then after a finite number of corrections, the Perceptron separates any infinite sequence of data (after the last correction the infinite tail of data will be separated without error). Moreover, if one supplies the perceptron with the following stopping rule:

perceptron stops the learning process if after the correction number k ($k = 1, 2, \dots$), the next

$$m_k = \frac{1 + 2 \ln k - \ln \eta}{-\ln(1 - \varepsilon)}$$

elements of the training data do not change the decision rule (they are recognized correctly),

then

- (i) the perceptron will stop the learning process during the first

$$\ell \leq \frac{1 + 4 \ln \frac{R}{\rho} - \ln \eta}{-\ln(1 - \varepsilon)} \left[\frac{R^2}{\rho^2} \right]$$

steps,

- (ii) by the stopping moment it will have constructed a decision rule that with probability $1 - \eta$ has a probability of error on the test set less than ε (Aizerman, Braverman, and Rozonoer, 1964).

Because of these results many researchers thought that minimizing the error on the training set is the only cause of generalization (small probability of test errors). Therefore, the analysis of learning processes was split into two branches, call them applied analysis of learning processes and theoretical analysis of learning processes.

The philosophy of applied analysis of the learning process can be described as follows:

To get a good generalization it is sufficient to choose the coefficients of the neuron that provide the minimal number of training errors. The principle of minimizing the number of training errors is a self-evident inductive principle, and from the practical point of view does not need justification. The main goal of applied analysis is to find methods for constructing the coefficients simultaneously for all neurons such that the separating surface provides the minimal number of errors on the training data.

The philosophy of theoretical analysis of learning processes is different.

The principle of minimizing the number of training errors is not self-evident and needs to be justified. It is possible that there exists another inductive principle that provides a better level of generalization ability. The main goal of theoretical analysis of learning processes is to find the inductive principle with the highest level of generalization ability and to construct algorithms that realize this inductive principle.

This book shows that indeed the principle of minimizing the number of training errors is not self-evident and that there exists another more intelligent inductive principle that provides a better level of generalization ability.

CONSTRUCTION OF THE FUNDAMENTALS OF THE LEARNING THEORY (THE 1960–1970s)

As soon as the experiments with the perceptron became widely known, other types of learning machines were suggested (such as the Madaline, constructed by B. Widrow, or the learning matrices constructed by K. Steinbuch; in fact, they started construction of special learning hardware). However, in contrast to the perceptron, these machines were considered from the very beginning as tools for solving real-life problems rather than a general model of the learning phenomenon.

For solving real-life problems, many computer programs were also developed, including programs for constructing logical functions of different types (e.g., decision trees, originally intended for expert systems), or hidden Markov models (for speech recognition problems). These programs also did not affect the study of the general learning phenomena.

The next step in constructing a general type of learning machine was done in 1986 when the so-called back-propagation technique for finding the weights simultaneously for many neurons was used. This method actually inaugurated a new era in the history of learning machines. We will discuss it in the next section. In this section we concentrate on the history of developing the fundamentals of learning theory.

In contrast to applied analysis, where during the time between constructing the perceptron (1960) and implementing back-propagation technique (1986) nothing extraordinary happened, these years were extremely fruitful for developing statistical learning theory.

Theory of the Empirical Risk Minimization Principle

As early as 1968, a philosophy of statistical learning theory had been developed. The essential concepts of the emerging theory, VC entropy and VC dimension, had been discovered and introduced for the set of indicator functions (i.e., for the pattern recognition problem). Using these concepts, the law of large numbers in functional space (necessary and sufficient conditions for uniform convergence of the frequencies to their probabilities) was found, its relation to learning processes was described, and the main nonasymptotic bounds for the rate of convergence were obtained (Vapnik and Chervonenkis, 1968); complete proofs were published by 1971 (Vapnik and Chervonenkis, 1971). The obtained bounds made the introduction of a novel inductive principle possible (structural risk minimization inductive principle, 1974), completing the development of pattern recognition learning theory. The new paradigm for pattern recognition theory was summarized in a monograph.²

Between 1976 and 1981, the results, originally obtained for the set of indicator functions, were generalized for the set of real functions: the law of large numbers (necessary and sufficient conditions for uniform convergence of means to their expectations), the bounds on the rate of uniform convergence both for the set of totally bounded functions and for the set of unbounded functions, and the structural risk minimization principle. In 1979 these results were summarized in a monograph³ describing the new paradigm for the general problem of dependencies estimation.

Finally, in 1989 necessary and sufficient conditions for consistency⁴ of the empirical risk minimization inductive principle and maximum likelihood method were found, completing the analysis of empirical risk minimization inductive inference (Vapnik and Chervonenkis, 1989).

Building on thirty years of analysis of learning processes, in the 1990s the synthesis of novel learning machines controlling generalization ability began.

These results were inspired by the study of learning processes. They are the main subject of the book.

²V. Vapnik and A. Chervonenkis, *Theory of Pattern Recognition* (in Russian), Nauka, Moscow, 1974.

German translation: W.N. Wapnik, A.Ja. Tscherwonenskis, *Theorie der Zeitdnerkennung*, Akademie-Verlag, Berlin, 1979.

³V.N. Vapnik, *Estimation of Dependencies Based on Empirical Data* (in Russian), Nauka, Moscow, 1979.

English translation: Vladimir Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer, New York, 1982.

⁴Convergence in probability to the best possible result. An exact definition of consistency is given in Section 2.1.

Theory of Solving Ill-Posed Problems

In the 1960s and 1970s, in various branches of mathematics, several groundbreaking theories were developed that became very important for creating a new philosophy. Below we list some of these theories. They also will be discussed in the Comments on the chapters.

Let us start with the regularization theory for the solution of so-called ill-posed problems.

In the early 1900s Hadamard observed that under some (very general) circumstances the problem of solving (linear) operator equations

$$Af = F, \quad f \in \mathcal{F}$$

(finding $f \in \mathcal{F}$ that satisfies the equality), is ill-posed; even if there exists a unique solution to this equation, a small deviation on the right-hand side of this equation (F_δ instead of F , where $\|F - F_\delta\| < \delta$ is arbitrarily small) can cause large deviations in the solutions (it can happen that $\|f_\delta - f\|$ is large).

In this case if the right-hand side F of the equation is not exact (e.g., it equals F_δ , where F_δ differs from F by some level δ of noise), the functions f_δ that minimize the functional

$$R(f) = \|Af - F_\delta\|^2$$

do not guarantee a good approximation to the desired solution even if δ tends to zero.

Hadamard thought that ill-posed problems are a pure mathematical phenomenon and that all real-life problems are “well-posed.” However, in the second half of the century a number of very important real-life problems were found to be ill-posed. In particular, ill-posed problems arise when one tries to reverse the cause–effect relations: to find unknown causes from known consequences. Even if the cause–effect relationship forms a one-to-one mapping, the problem of inverting it can be ill-posed.

For our discussion it is important that one of main problems of statistics, estimating the density function from the data, is ill-posed.

In the middle of the 1960s it was discovered that if instead of the functional $R(f)$ one minimizes another so-called regularized functional

$$R^*(f) = \|Af - F_\delta\|^2 + \gamma(\delta)\Omega(f),$$

where $\Omega(f)$ is some functional (that belongs to a special type of functionals) and $\gamma(\delta)$ is an appropriately chosen constant (depending on the level of noise), then one obtains a sequence of solutions that converges to the desired one as δ tends to zero (Tikhonov, 1963), (Ivanov, 1962), and (Phillips, 1962).

Regularization theory was one of the first signs of the existence of intelligent inference. It demonstrated that whereas the “self-evident” method

of minimizing the functional $R(f)$ does not work, the not “self-evident” method of minimizing the functional $R^*(f)$ does.

The influence of the philosophy created by the theory of solving ill-posed problems is very deep. Both the regularization philosophy and the regularization technique became widely disseminated in many areas of science, including statistics.

Nonparametric Methods of Density Estimation

In particular, the problem of density estimation from a rather wide set of densities is ill-posed. Estimating densities from some narrow set of densities (say from a set of densities determined by a finite number of parameters, i.e., from a so-called parametric set of densities) was the subject of the classical paradigm, where a “self-evident” type of inference (the maximum likelihood method) was used. An extension of the set of densities from which one has to estimate the desired one makes it impossible to use the “self-evident” type of inference. To estimate a density from the wide (nonparametric) set requires a new type of inference that contains regularization techniques. In the 1960s several such types of (nonparametric) algorithms were suggested (M. Rosenblatt, 1956), (Parzen, 1962), and (Chentsov, 1963); in the middle of the 1970s the general way for creating these kinds of algorithms on the basis of standard procedures for solving ill-posed problems was found (Vapnik and Stefanyuk, 1978).

Nonparametric methods of density estimation gave rise to statistical algorithms that overcame the shortcomings of the classical paradigm. Now one could estimate functions from a wide set of functions.

One has to note, however, that these methods are intended for estimating a function using large sample sizes.

The Idea of Algorithmic Complexity

Finally, in the 1960s one of the greatest ideas of statistics and information theory was suggested: the idea of algorithmic complexity (Solomonoff, 1960), (Kolmogorov, 1965), and (Chaitin, 1966). Two fundamental questions that at first glance look different inspired this idea:

- (i) *What is the nature of inductive inference (Solomonoff)?*
- (ii) *What is the nature of randomness (Kolmogorov), (Chaitin)?*

The answers to these questions proposed by Solomonoff, Kolmogorov, and Chaitin started the information theory approach to the problem of inference.

The idea of the randomness concept can be roughly described as follows: A rather large string of data forms a random string if there are no algorithms whose complexity is much less than ℓ , the length of the string, that

can generate this string. The complexity of an algorithm is described by the length of the smallest program that embodies that algorithm. It was proved that the concept of algorithmic complexity is universal (it is determined up to an additive constant reflecting the type of computer). Moreover, it was proved that if the description of the string cannot be compressed using computers, then the string possesses all properties of a random sequence.

This implies the idea that if one can significantly compress the description of the given string, then the algorithm used describes intrinsic properties of the data.

In the 1970s, on the basis of these ideas, Rissanen suggested the minimum description length (MDL) inductive inference for learning problems (Rissanen, 1978).

In Chapter 4 we consider this principle.

All these new ideas are still being developed. However, they have shifted the main understanding as to what can be done in the problem of dependency estimation on the basis of a limited amount of empirical data.

NEURAL NETWORKS (THE 1980s)

Idea of Neural Networks

In 1986 several authors independently proposed a method for simultaneously constructing the vector coefficients for all neurons of the Perceptron using the so-called back-propagation method (LeCun, 1986), (Rumelhart, Hinton, and Williams, 1986). The idea of this method is extremely simple. If instead of the McCulloch–Pitts model of the neuron one considers a slightly modified model, where the discontinuous function sign $\{(w \cdot x) - b\}$ is replaced by the continuous so-called sigmoid approximation (Fig. 0.3)

$$y = S \{(w \cdot x) - b\}$$

(here $S(u)$ is a monotonic function with the properties

$$S(-\infty) = -1, \quad S(+\infty) = 1$$

e.g., $S(u) = \tanh u$), then the composition of the new neurons is a continuous function that for any fixed x has a gradient with respect to all coefficients of all neurons. In 1986 the method for evaluating this gradient was found.⁵ Using the evaluated gradient one can apply any gradient-based technique for constructing a function that approximates the desired

⁵The back-propagation method was actually found in 1963 for solving some control problems (Brison, Denham, and Dreyfuss, 1963) and was rediscovered for perceptrons.

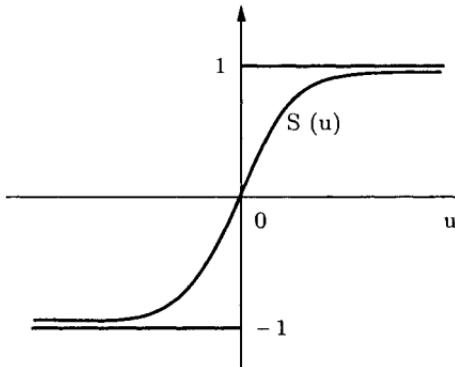


FIGURE 0.3. The discontinuous function $\text{sign}(u) = \pm 1$ is approximated by the smooth function $S(u)$.

function. Of course, gradient-based techniques only guarantee finding local minima. Nevertheless, it looked as if the main idea of applied analysis of learning processes has been found and that the problem was in its implementation.

Simplification of the Goals of Theoretical Analysis

The discovery of the back-propagation technique can be considered as the second birth of the Perceptron. This birth, however, happened in a completely different situation. Since 1960 powerful computers had appeared, moreover, new branches of science had became involved in research on the learning problem. This essentially changed the scale and the style of research.

In spite of the fact that one cannot assert for sure that the generalization properties of the Perceptron with many adjustable neurons is better than the generalization properties of the Perceptron with only one adjustable neuron and approximately the same number of free parameters, the scientific community was much more enthusiastic about this new method due to the scale of experiments.

Rosenblatt's first experiments were conducted for the problem of digit recognition. To demonstrate the generalization ability of the perceptron, Rosenblatt used training data consisting of several hundreds of vectors, containing several dozen coordinates. In the 1980s and even now in the

1990s the problem of digit recognition learning continues to be important. Today, in order to obtain good decision rules one uses tens (even hundreds) of thousands of observations over vectors with several hundreds of coordinates. This required special organization of the computational processes. Therefore, in the 1980s researchers in artificial intelligence became the main players in the computational learning game. Among artificial intelligence researchers the hardliners had considerable influence. (It is precisely they who declared that "Complex theories do not work; simple algorithms do.")

Artificial intelligence hardliners approached the learning problem with great experience in constructing "simple algorithms" for the problems where theory is very complicated. At the end of the 1960s computer natural language translators were promised within a couple of years (even now this extremely complicated problem is far from being solved); the next project was constructing a general problem solver; after this came the project of constructing an automatic controller of large systems, and so on. All of these projects had little success. The next problem to be investigated was creating a computational learning technology.

First the hardliners changed the terminology. In particular, the perceptron was renamed a neural network. Then it was declared a joint research program with physiologist, and the study of the learning problem became less general, more subject oriented. In the 1960s and 1970s the main goal of research was finding the best way for inductive inference from small sample sizes. In the 1980s the goal became constructing a model of generalization that uses the brain.⁶

The attempt to introduce theory to the artificial intelligence community was made in 1984 when the probably approximately correct (PAC) model was suggested.⁷ This model is defined by a particular case of the consistency concept commonly used in statistics in which some requirements on computational complexity were incorporated.⁸

In spite of the fact that almost all results in the PAC model were adopted from statistical learning theory and constitute particular cases of one of its four parts (namely, the theory of bounds), this model undoubtedly had the

⁶Of course it is very interesting to know how humans can learn. However, this is not necessarily the best way for creating an artificial learning machine. It has been noted that the study of birds flying was not very useful for constructing the airplane.

⁷L.G. Valiant, 1984, "A theory of learnability," *Commun. ACM* 27(11), 1134–1142.

⁸"If the computational requirement is removed from the definition then we are left with the notion of nonparametric inference in the sense of statistics, as discussed in particular by Vapnik." (L. Valiant, 1991, "A view of computational learning theory," in the book *Computation and Cognition*, Society for Industrial and Applied Mathematics, Philadelphia, p. 36.)

merit of bringing the importance of statistical analysis to the attention of the artificial intelligence community. This, however, was not sufficient to influence the development of new learning technologies.

Almost ten years have passed since the perceptron was born a second time. From the conceptual point of view, its second birth was less important than the first one. In spite of important achievements in some specific applications using neural networks, the theoretical results obtained did not contribute much to general learning theory. Also, no new interesting learning phenomena were found in experiments with neural nets. The so-called overfitting phenomenon observed in experiments is actually a phenomenon of “false structure” known in the theory for solving ill-posed problems. From the theory of solving ill-posed problems, tools were adopted that prevent overfitting — using regularization techniques in the algorithms.

Therefore, almost ten years of research in neural nets did not substantially advance the understanding of the essence of learning processes.

RETURNING TO THE ORIGIN (THE 1990s)

In the last couple of years something has changed in relation to neural networks.

More attention is now focused on the alternatives to neural nets, for example, a great deal of effort has been devoted to the study of the radial basis functions method (see the review in (Powell, 1992)). As in the 1960s, neural networks are called again multilayer perceptrons. The advanced parts of statistical learning theory now attract more researchers. In particular in the last few years both the structural risk minimization principle and the minimum description length principle have become popular subjects of analysis. The discussions on small sample size theory, in contrast to the asymptotic one, became widespread.

It looks as if everything is returning to its fundamentals.

In addition, statistical learning theory now plays a more active role: After the completion of the general analysis of learning processes, the research in the area of the synthesis of optimal algorithms (which possess the highest level of generalization ability for any number of observations) was started.

These studies, however, do not belong to history yet. They are a subject of today's research activities.⁹

⁹This remark was made in 1995. However, after the appearance of the first edition of this book important changes took place in the development of new methods of computer learning.

In the last five years new ideas have appeared in learning methodology inspired by statistical learning theory. In contrast to old ideas of constructing learning algorithms that were inspired by a biological analogy to the learning process, the new ideas were inspired by attempts to minimize theoretical bounds on the error rate obtained as a result of formal analysis of the learning processes. These ideas (which often imply methods that contradict the old paradigm) result in algorithms that have not only nice mathematical properties (such as uniqueness of the solution, simple method of treating a large number of examples, and independence of dimensionality of the input space) but also exhibit excellent performance: They outperform the state-of-the-art solutions obtained by the old methods.

Now a new methodological situation in the learning problem has developed where practical methods are the result of a deep theoretical analysis of the statistical bounds rather than the result of inventing new smart heuristics.

This fact has in many respects changed the character of the learning problem.

Chapter 1

Setting of the Learning Problem

In this book we consider the learning problem as a problem of finding a desired dependence using a *limited* number of observations.

1.1 FUNCTION ESTIMATION MODEL

We describe the general model of learning from examples through three components (Fig.1.1):

- (i) A generator (G) of random vectors $x \in R^n$, drawn independently from a fixed but unknown probability distribution function $F(x)$.
- (ii) A supervisor (S) who returns an output value y to every input vector x , according to a conditional distribution function¹ $F(y|x)$, also fixed but unknown.
- (iii) A learning machine (LM) capable of implementing a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, where Λ is a set of parameters.²

The problem of learning is that of choosing from the given set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, the one that best approximates the supervisor's response.

¹This is the general case, which includes the case where the supervisor uses a function $y = f(x)$.

²Note that the elements $\alpha \in \Lambda$ are not necessarily vectors. They can be any abstract parameters. Therefore, we in fact consider any set of functions.

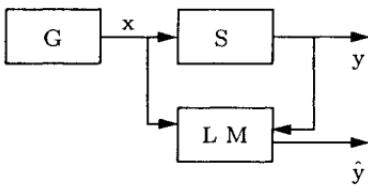


FIGURE 1.1. A model of learning from examples. During the learning process, the learning machine observes the pairs (x, y) (the training set). After training, the machine must on any given x return a value \bar{y} . The goal is to return a value \bar{y} that is close to the supervisor's response y .

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

$$(x_1, y_1), \dots, (x_\ell, y_\ell). \quad (1.1)$$

1.2 THE PROBLEM OF RISK MINIMIZATION

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

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

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

1.3 THREE MAIN LEARNING PROBLEMS

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

1.3.1 Pattern Recognition

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

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

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

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

1.3.2 Regression Estimation

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

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

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

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

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

1.3.3 Density Estimation (Fisher–Wald Setting)

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

$$L(p(x, \alpha)) = -\log p(x, \alpha). \quad (1.5)$$

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

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

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

$$x_1, \dots, x_n$$

are given.

1.4 THE GENERAL SETTING OF THE LEARNING PROBLEM

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

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

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

$$z_1, \dots, z_\ell \quad (1.7)$$

is given.

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

1.5 THE EMPIRICAL RISK MINIMIZATION (ERM) INDUCTIVE PRINCIPLE

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

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

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

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

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

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

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

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

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

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

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

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

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

1.6 THE FOUR PARTS OF LEARNING THEORY

Learning theory has to address the following four questions:

- (i) *What are (necessary and sufficient) conditions for consistency of a learning process based on the ERM principle?*
- (ii) *How fast is the rate of convergence of the learning process?*
- (iii) *How can one control the rate of convergence (the generalization ability) of the learning process?*
- (iv) *How can one construct algorithms that can control the generalization ability?*

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

- (i) Theory of consistency of learning processes.
- (ii) Nonasymptotic theory of the rate of convergence of learning processes.
- (iii) Theory of controlling the generalization ability of learning processes.
- (iv) Theory of constructing learning algorithms.

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

Informal Reasoning and Comments — 1

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

- (i) To estimate the desired function from a wide set of functions.
- (ii) To estimate the desired function on the basis of a limited number of examples.

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

1.7 THE CLASSICAL PARADIGM OF SOLVING LEARNING PROBLEMS

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

1.7.1 Density Estimation Problem (ML Method)

Let $p(x, \alpha), \alpha \in \Lambda$, be a set of density functions where (in contrast to the setting of the problem described in this chapter) the set Λ is necessarily contained in R^n (α is an n -dimensional vector). Let the unknown density $p(x, \alpha_0)$ belongs to this class. The problem is to estimate this density using i.i.d. data

$$x_1, \dots, x_\ell$$

(distributed according to this unknown density).

In the 1920s Fisher developed the ML method for estimating the unknown parameters of the density (Fisher, 1952). He suggested approximating the unknown parameters by the values that maximize the functional

$$L(\alpha) = \sum_{i=1}^{\ell} \ln p(x_i, \alpha).$$

Under some conditions the ML method is consistent. In the next chapter we use results on the law of large numbers in functional space to describe the necessary and sufficient conditions for consistency of the ML method. In the following we show how by using the ML method one can estimate a desired function.

1.7.2 Pattern Recognition (Discriminant Analysis) Problem

Using the ML technique, Fisher considered a problem of pattern recognition (he called it discriminant analysis). He proposed the following model:

There exist two categories of data distributed according to two different statistical laws $p_1(x, \alpha^*)$ and $p_2(x, \beta^*)$ (densities, belonging to parametric classes). Let the probability of occurrence of the first category of data be q_1 and the probability of the second category be $1 - q_1$. The problem is to find a decision rule that minimizes the probability of error.

Knowing these two statistical laws and the value q_1 , one can immediately construct such a rule: The smallest probability of error is achieved by the decision rule that considers vector x as belonging to the first category if the probability that this vector belongs to the first category is not less than the probability that this vector belongs to the second category. This happens if the following inequality holds:

$$q_1 p_1(x, \alpha^*) \geq (1 - q_1) p_2(x, \beta^*).$$

One considers this rule in the equivalent form

$$f(x) = \text{sign} \left\{ \ln p_1(x, \alpha^*) - \ln p_2(x, \beta^*) + \ln \frac{q_1}{(1 - q_1)} \right\}, \quad (1.9)$$

called the discriminant function (rule), which assigns the value 1 for representatives of the first category and the value -1 for representatives of the second category. To find the discriminant rule one has to estimate two densities: $p_1(x, \alpha)$ and $p_2(x, \beta)$. In the classical paradigm one uses the ML method to estimate the parameters α^* and β^* of these densities.

1.7.3 Regression Estimation Model

Regression estimation in the classical paradigm is based on another model, the so-called model of measuring a function with additive noise:

Suppose that an unknown function has the parametric form

$$f_0(x) = f(x, \alpha_0),$$

where $\alpha_0 \in \Lambda$ is an unknown vector of parameters. Suppose also that at any point x_i one can measure the value of this function with additive noise:

$$y_i = f(x_i, \alpha_0) + \xi_i,$$

where the noise ξ_i does not depend on x_i and is distributed according to a *known* density function $p(\xi)$. The problem is to estimate the function $f(x, \alpha_0)$ from the set $f(x, \alpha), \alpha \in \Lambda$, using the data obtained by measurements of the function $f(x, \alpha_0)$ corrupted with additive noise.

In this model, using the observations of pairs

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

one can estimate the parameters α_0 of the unknown function $f(x, \alpha_0)$ by the ML method, namely by maximizing the functional

$$L(\alpha) = \sum_{i=1}^{\ell} \ln p(y_i - f(x_i, \alpha)).$$

(Recall that $p(\xi)$ is a known function and that $\xi = y - f(x, \alpha_0)$.) Taking the normal law

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

with zero mean and some fixed variance as a model of noise, one obtains the least-squares method:

$$L^*(\alpha) = -\frac{1}{2\sigma^2} \sum_{i=1}^{\ell} (y_i - f(x_i, \alpha))^2 - \ell \ln(\sqrt{2\pi}\sigma).$$

Maximizing $L^*(\alpha)$ over parameters α is equivalent to minimizing the functional

$$M(\alpha) = \sum_{i=1}^{\ell} (y_i - f(x_i, \alpha))^2$$

(the so-called least-squares functional).

Choosing other laws $p(\xi)$, one can obtain other methods for parameter estimation.⁴

1.7.4 Narrowness of the ML Method

Thus, in the classical paradigm the solutions to all problems of dependency estimation described in this chapter are based on the ML method. This method, however, can fail in the simplest cases. Below we demonstrate that using the ML method it is impossible to estimate the parameters of a density that is a mixture of normal densities. To show this it is sufficient to analyze the simplest case described in the following example.

Example. Using the ML method it is impossible to estimate a density that is the simplest mixture of two normal densities

$$p(x, a, \sigma) = \frac{1}{2\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\} + \frac{1}{2\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\},$$

where the parameters (a, σ) of only one density are unknown.

Indeed for any data x_1, \dots, x_ℓ and for any given constant A , there exists such a small $\sigma = \sigma_0$ that for $a = x_1$ the likelihood will exceed A :

$$\begin{aligned} L(a = x_1, \sigma_0) &= \sum_{i=1}^{\ell} \ln p(x_i; a = x_1, \sigma_0) \\ &> \ln\left(\frac{1}{2\sigma_0\sqrt{2\pi}}\right) + \sum_{i=2}^{\ell} \ln\left(\frac{1}{2\sqrt{2\pi}} \exp\left\{-\frac{x_i^2}{2}\right\}\right) \\ &= -\ln\sigma_0 - \sum_{i=2}^{\ell} \frac{x_i^2}{2} - \ell \ln 2\sqrt{2\pi} > A. \end{aligned}$$

⁴In 1964 P. Huber extended the classical model of regression estimation by introducing the so-called robust regression estimation model. According to this model, instead of an exact model of the noise $p(\xi)$, one is given a set of density functions (satisfying quite general conditions) to which this function belongs. The problem is to construct, for the given parametric set of functions and for the given set of density functions, an estimator that possesses the minimax properties (provides the best approximation for the worst density from the set). The solution to this problem actually has the following form: Choose an appropriate density function and then estimate the parameters using the ML method (Huber, 1964).

From this inequality one concludes that the maximum of the likelihood does not exist, and therefore the ML method does not provide a solution to estimating the parameters a and σ .

Thus, the ML method can be applied only to a very restrictive set of densities.

1.8 NONPARAMETRIC METHODS OF DENSITY ESTIMATION

In the beginning of the 1960s several authors suggested various new methods, so-called nonparametric methods, for density estimation. The goal of these methods was to estimate a density from a rather wide set of functions that is not restricted to be a parametric set of functions (M. Rosenblatt, 1957), (Parzen, 1962), and (Chentsov, 1963).

1.8.1 Parzen's Windows

Among these methods the Parzen windows method probably is the most popular. According to this method, one first has to determine the so-called kernel function. For simplicity we consider a simple kernel function:

$$K(x, x_i; \gamma) = \frac{1}{\gamma^n} K\left(\frac{x - x_i}{\gamma}\right), \quad x \in R^n,$$

where $K(u)$ is a symmetric unimodal density function.

Using this function one determines the estimator

$$p(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} K(x, x_i; \gamma).$$

In the 1970s a comprehensive asymptotic theory for Parzen-type nonparametric density estimation was developed (Devroye, 1985). It includes the following two important assertions:

- (i) Parzen's estimator is consistent (in the various metrics) for estimating a density from a *very wide* class of densities.
- (ii) The *asymptotic* rate of convergence for Parzen's estimator is optimal for "smooth" densities.

The same results were obtained for other types of estimators.

Therefore, for both classical models (discriminant analysis and regression estimation) using nonparametric methods instead of parametric methods, one can obtain a good approximation to the desired dependency *if the number of observations is sufficiently large*.

Experiments with nonparametric estimators, however, did not demonstrate great advantages over old techniques. This indicates that nonparametric methods, when applied to a limited numbers of observations, do not possess their remarkable asymptotic properties.

1.8.2 The Problem of Density Estimation Is Ill-Posed

Nonparametric statistics was developed as a number of recipes for density estimation and regression estimation. To make the theory comprehensive it was necessary to find a general principle for constructing and analyzing various nonparametric algorithms. In 1978 such a principle was found (Vapnik and Stefanyuk, 1978).

By definition a density $p(x)$ (if it exists) is the solution of the integral equation

$$\int_{-\infty}^x p(t)dt = F(x), \quad (1.10)$$

where $F(x)$ is a probability distribution function. (Recall that in the theory of probability one first determines the probability distribution function, and then only if the distribution function is absolutely continuous can one define the density function.)

The general formulation of the density estimation problem can be described as follows: In the given set of functions $\{p(t)\}$, find one that is a solution to the integral equation (1.10) for the case where the probability distribution function $F(x)$ is unknown, but we are given the i.i.d. data $x_1, \dots, x_\ell, \dots$ obtained according to the unknown distribution function.

Using these data one can construct a function that is very important in statistics, the so-called empirical distribution function (Fig. 1.2)

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

where $\theta(u)$ is the step function that takes the value 1 if $u \geq 0$ and 0 otherwise.

The uniform convergence

$$\sup_x |F(x) - F_\ell(x)| \xrightarrow[\ell \rightarrow \infty]{P} 0$$

of the empirical distribution function $F_\ell(x)$ to the desired function $F(x)$ constitutes one of the most fundamental facts of theoretical statistics. We will discuss this fact several times, in the comments on Chapter 2 and in the comments on Chapter 3.

Thus, the general setting of the density estimation problem (coming from the definition of a density) is the following:

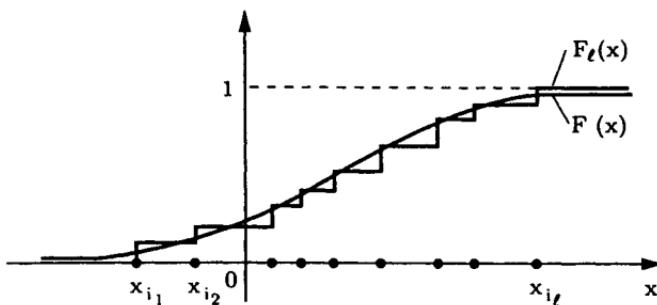


FIGURE 1.2. The empirical distribution function $F_\ell(x)$ constructed from the data x_1, \dots, x_ℓ approximates the probability distribution function $F(x)$.

Solve the integral equation (1.10) in the case where the probability distribution function is unknown, but i.i.d. $x_1, \dots, x_\ell, \dots$ data in accordance to this function are given.

Using these data one can construct the empirical distribution function $F_\ell(x)$. Therefore, one has to solve the integral equation (1.10) for the case where instead of the exact right-hand side, one knows an approximation that converges uniformly to the unknown function as the number of observations increases.

Note that the problem of solving this integral equation in a wide class of functions $\{p(t)\}$ is ill-posed. This brings us to two conclusions:

- (i) Generally speaking, the estimation of a density is a hard (ill-posed) computational problem.
- (ii) To solve this problem well one has to use regularization (i.e., not “self-evident”) techniques.

It has been shown that all proposed nonparametric algorithms can be obtained using standard regularization techniques (with different types of regularizers) and using the empirical distribution function instead of the unknown one (Vapnik, 1979, 1988).

1.9 MAIN PRINCIPLE FOR SOLVING PROBLEMS USING A RESTRICTED AMOUNT OF INFORMATION

We now formulate the main principle for solving problems using a restricted amount of information:

When solving a given problem, try to avoid solving a more general problem as an intermediate step.

Although this principle is obvious, it is not easy to follow. For our problems of dependency estimation this principle means that to solve the problem of pattern recognition or regression estimation, one must try to find the desired function “directly” (in the next section we will specify what this means) rather than first estimating the densities and then using the estimated densities to construct the desired function.

Note that estimation of densities is a universal problem of statistics (knowing the densities one can solve various problems). Estimation of densities in general is an ill-posed problem; therefore, it requires many of observations in order to be solved well. In contrast, the problems that we really need to solve (decision rule estimation or regression estimation) are quite particular ones; often they can be solved on the basis of a reasonable number of observations.

To illustrate this idea let us consider the following situation. Suppose one wants to construct a decision rule separating two sets of vectors described by two normal laws: $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$. In order to construct the discriminant rule (1.9), one has to estimate from the data two n -dimensional vectors, the means μ_1 and μ_2 , and two $n \times n$ covariance matrices Σ_1 and Σ_2 . As a result one obtains a separating polynomial of degree two:

$$f(x) = \text{sign} \left\{ \frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) - \frac{1}{2}(x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) - C \right\},$$

$$C = \ln \frac{|\Sigma_1|}{|\Sigma_2|} - \ln \frac{q_1}{1 - q_1},$$

containing $n(n + 3)/2$ coefficients. To construct a good discriminant rule from the parameters of the unknown normal densities, one needs to estimate the parameters of the covariance matrices with high accuracy, since the discriminant function uses inverse covariance matrices (in general, the estimation of a density is an ill-posed problem; for our parametric case it can give ill-conditioned covariance matrices). To estimate the high-dimensional covariance matrices well one needs an unpredictably large (depending on the properties of the actual covariance matrices) number of observations. Therefore, in high-dimensional spaces the general normal discriminant function (constructed from two different normal densities) seldom succeeds in practice. In practice, the linear discriminant function that occurs when the

two covariance matrices coincide is used, $\Sigma_1 = \Sigma_2 = \Sigma$:

$$f(x) = \text{sign} \left\{ (\mu_2 - \mu_1)^T \Sigma^{-1} x + \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1) - \frac{1}{2} (\mu_2^T \Sigma^{-1} \mu_2) + \ln \frac{q_1}{1 - q_1} \right\}$$

(in this case one has to estimate only n parameters of the discriminant function).

It is remarkable that Fisher suggested to use the *linear* discriminant function even if the two covariance matrices were different and proposed a heuristic method for constructing such functions (Fisher, 1952).⁵

In Chapter 5 we solve a specific pattern recognition problem by constructing separating polynomials (up to degree 7) in high-dimensional (256) space. This is accomplished only by avoiding the solution of unnecessarily general problems.

1.10 MODEL MINIMIZATION OF THE RISK BASED ON EMPIRICAL DATA

In what follows we argue that the setting of learning problems given in this chapter allows us not only to consider estimating problems in any given set of functions, but also to implement the main principle for using small samples: avoiding the solution of unnecessarily general problems.

1.10.1 Pattern Recognition

For the pattern recognition problem, the functional (1.2) evaluates the probability of error for any function of the admissible set of functions. The problem is to use the sample to find the function from the set of admissible functions that minimizes the probability of error. This is exactly what we want to obtain.

1.10.2 Regression Estimation

In regression estimation we minimize functional (1.2) with loss function (1.4). This functional can be rewritten in the equivalent form

$$R(\alpha) = \int (y - f(x, \alpha))^2 dF(x, y)$$

⁵In the 1960s the problem of constructing the best linear discriminant function (in the case where a quadratic function is optimal) was solved (Andersen and Bahadur, 1966). For solving real-life problems the linear discriminant functions usually are used even if it is known that the optimal solution belongs to quadratic discriminant functions.

$$= \int (f(x, \alpha) - f_0(x))^2 dF(x) + \int (y - f_0(x))^2 dF(x, y) \quad (1.11)$$

where $f_0(x)$ is the regression function. Note that the second term in (1.11) does not depend on the chosen function. Therefore, minimizing this functional is equivalent to minimizing the functional

$$R^*(\alpha) = \int (f(x, \alpha) - f_0(x))^2 dF(x).$$

The last functional equals the squared $L_2(F)$ distance between a function of the set of admissible functions and the regression. Therefore, we consider the following problem: Using the sample, find in the admissible set of functions the closest one to the regression (in metrics $L_2(F)$).

If one accepts the $L_2(F)$ metrics, then the formulation of the regression estimation problem (minimizing $R(\alpha)$) is direct. (It does not require solving a more general problem, for example, finding $F(x, y)$.)

1.10.3 Density Estimation

Finally, consider the functional

$$R(\alpha) = - \int \ln p(t, \alpha) dF(t) = - \int p_0(t) \ln p(t, \alpha) dt.$$

Let us add to this functional a constant (a functional that does not depend on the approximating functions)

$$c = \int \ln p_0(t) dF(t),$$

where $p_0(t)$ and $F(t)$ are the desired density and its probability distribution function. We obtain

$$\begin{aligned} R^*(\alpha) &= - \int \ln p(t, \alpha) dF(t) + \int \ln p_0(t) dF(t) \\ &= - \int \ln \frac{p(t, \alpha)}{p_0(t)} p_0(t) dt. \end{aligned}$$

The expression on the right-hand side is the so-called Kullback–Leibler distance that is used in statistics for measuring the distance between an approximation of a density and the actual density. Therefore, we consider the following problem: In the set of admissible densities find the closest to the desired one in the Kullback–Leibler distance using a given sample. If one accepts the Kullback–Leibler distance, then the formulation of the problem is direct.

The short form of the setting of all these problems is the general model of minimizing the risk functional on the basis of empirical data.

1.11 STOCHASTIC APPROXIMATION INFERENCE

To minimize the risk functional on the basis of empirical data, we considered in Chapter 1 the empirical risk minimization inductive principle. Here we discuss another general inductive principle, the so-called stochastic approximation method suggested in the 1950s by Robbins and Monroe (Robbins and Monroe, 1951).

According to this principle, to minimize the functional

$$R(\alpha) = \int Q(z, \alpha) dF(z)$$

with respect to the parameters α using i.i.d. data

$$z_1, \dots, z_\ell$$

one uses the following iterative procedure:

$$\alpha(k+1) = \alpha(k) - \gamma_k \operatorname{grad}_\alpha Q(z_k, \alpha(k)), \quad k = 1, 2, \dots, \ell, \quad (1.12)$$

where the number of steps is equal to the number of observations. It was proven that this method is consistent under very general conditions on the gradient $\operatorname{grad}_\alpha Q(z, \alpha)$ and the values γ_k .

Inspired by Novikoff's theorem, Ya.Z. Tsyplkin and M.A. Aizerman started discussions on consistency of learning processes in 1963 at the seminars of the Moscow Institute of Control Science. Two general inductive principles that ensure consistency of learning processes were under investigation:

- (i) principle of stochastic approximation, and
- (ii) principle of empirical risk minimization.

Both inductive principles were applied to the general problem of minimizing the risk functional (1.6) using empirical data. As a result, by 1971 two different types of general learning theories had been created:

- (i) The general *asymptotic* learning theory for *stochastic approximation* inductive inference⁶ (Aizerman, Braverman, and Rozonoer, 1965), (Tsyplkin, 1971, 1973).
- (ii) The general *nonasymptotic* theory of pattern recognition for *ERM* inductive inference (Vapnik and Chervonenkis, 1968, 1971, 1974). (By 1979 this theory had been generalized for any problem of minimization of the risk on the basis of empirical data (Vapnik, 1979).)

⁶In 1967 this theory was also suggested by S. Amari (Amari, 1967).

The stochastic approximation principle is, however, too wasteful: It uses one element of the training data per step (see (1.12)). To make it more economical, one uses the training data *many times* (using many epochs). In this case the following question arises immediately:

When does one have to stop the training process?

Two answers are possible:

- (i) When for any element of the training data the gradient is so small that the learning process cannot be continued.
- (ii) When the learning process is not saturated but satisfies some stopping criterion.

It is easy to see that in the first case the stochastic approximation method is just a special way of minimizing the empirical risk. The second case constitutes a regularization method of minimizing the risk functional.⁷ Therefore, in the “nonwasteful regimes” the stochastic approximation method can be explained as either inductive properties of the ERM method or inductive properties of the regularization method.

To complete the discussion on classical inductive inferences it is necessary to consider Bayesian inference. In order to use this inference one must possess additional *a priori* information complementary to the set of parametric functions *containing the desired one*. Namely, one must know the distribution function that describes the probability for any function from the admissible set of functions to be the desired one. Therefore, Bayesian inference is based on using strong *a priori* information (it requires that the desired function belong to the set of functions of the learning machine). In this sense it does not define a *general* method for inference. We will discuss this inference later in the comments on Chapter 4.

Thus, along with the ERM inductive principle one can use other inductive principles. However, the ERM principle (compared to other ones) looks more robust (it uses empirical data better, it does not depend on *a priori* information, and there are clear ways to implement it).

Therefore, in the analysis of learning processes, the key problem became that of exploring the ERM principle.

⁷The regularizing property of the stopping criterion in iterative procedures of solving ill-posed problems was observed in the 1950s even before the regularization theory for solving ill-posed problems was developed.

Chapter 2

Consistency of Learning Processes

The goal of this part of the theory is to describe the conceptual model for learning processes that are based on the empirical risk minimization inductive principle. This part of the theory has to explain when a learning machine that minimizes empirical risk can achieve a small value of actual risk (can generalize) and when it cannot. In other words, the goal of this part is to describe necessary and sufficient conditions for the *consistency* of learning processes that minimize the empirical risk.

The following question arises:

Why do we need an asymptotic theory (consistency is an asymptotic concept) if the goal is to construct algorithms for learning from a limited number of observations?

The answer is as follows:

To construct any theory one has to use some concepts in terms of which the theory is developed. It is extremely important to use concepts that describe necessary and sufficient conditions for consistency. This guarantees that the constructed theory is general and cannot be improved from the conceptual point of view.

The most important issue in this chapter is the concept of the VC entropy of a set of functions in terms of which the necessary and sufficient conditions for consistency of learning processes are described.

Using this concept we will obtain in the next chapter the quantitative characteristics on the rate of the learning process that we will use later for constructing learning algorithms.

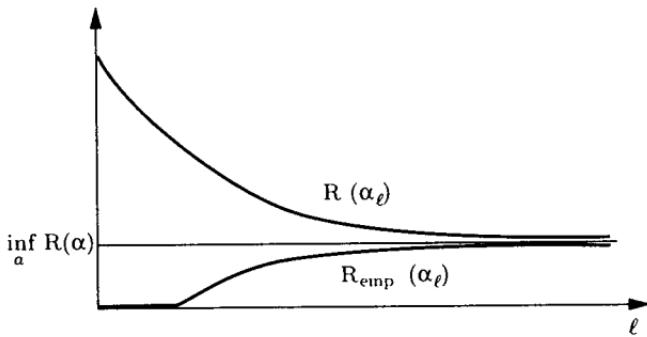


FIGURE 2.1. The learning process is consistent if both the expected risks $R(\alpha_\ell)$ and the empirical risks $R_{\text{emp}}(\alpha_\ell)$ converge to the minimal possible value of the risk, $\inf_{\alpha \in \Lambda} R(\alpha)$.

2.1 THE CLASSICAL DEFINITION OF CONSISTENCY AND THE CONCEPT OF NONTRIVIAL CONSISTENCY

Let $Q(z, \alpha_\ell)$ be a function that minimizes the empirical risk functional

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

for a given set of i.i.d. observations z_1, \dots, z_ℓ .

Definition. We say that the principle (method) of ERM is *consistent* for the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, and for the probability distribution function $F(z)$ if the following two sequences converge in probability to the same limit (see the schematic Fig.2.1):

$$R(\alpha_\ell) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha), \quad (2.1)$$

$$R_{\text{emp}}(\alpha_\ell) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha). \quad (2.2)$$

In other words, the ERM method is consistent if it provides a sequence of functions $Q(z, \alpha_\ell)$, $\ell = 1, 2, \dots$, for which both expected risk and empirical risk converge to the minimal possible value of risk. Equation (2.1) asserts that the values of achieved risks converge to the best possible, while (2.2) asserts that one can estimate on the basis of the values of empirical risk the minimal possible value of the risk.

The goal of this chapter is to describe conditions of consistency for the ERM method. We would like to obtain these conditions in terms of general characteristics of the set of functions and the probability measure.

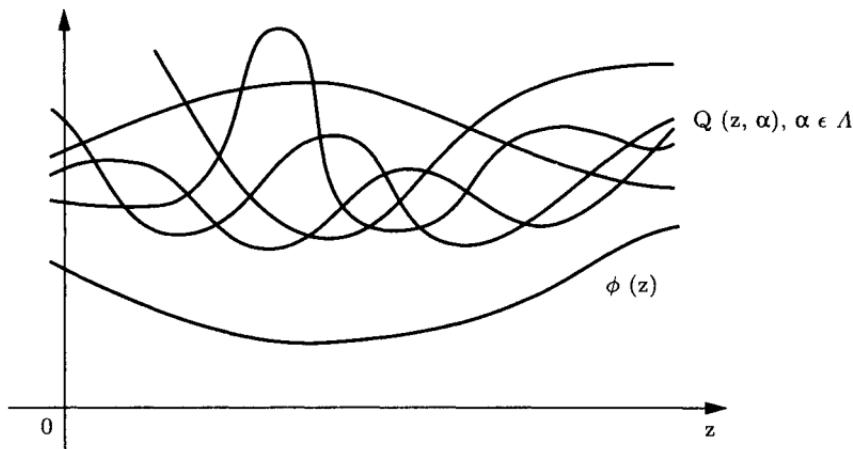


FIGURE 2.2. A case of trivial consistency. The ERM method is inconsistent on the set of functions $Q(z, \alpha), \alpha \in \Lambda$, and consistent on the set of functions $\{\phi(z)\} \cup Q(z, \alpha), \alpha \in \Lambda$.

Unfortunately, for the classical definition of consistency given above, obtaining such conditions is impossible, since this definition includes cases of *trivial* consistency.

What is a trivial case of consistency?

Suppose we have established that for some set of functions $Q(z, \alpha), \alpha \in \Lambda$, the ERM method is not consistent. Consider an extended set of functions that includes this set of functions and one additional function, $\phi(z)$. Suppose that the additional function satisfies the inequality

$$\inf_{\alpha \in \Lambda} Q(z, \alpha) > \phi(z), \quad \forall z.$$

It is clear (Fig. 2.2) that for the extended set of functions (containing $\phi(z)$) the ERM method will be consistent. Indeed, for any distribution function and for any number of observations, the minimum of the empirical risk will be attained on the function $\phi(z)$ that also gives the minimum of the expected risk.

This example shows that there exist trivial cases of consistency that depend on whether the given set of functions contains a minorizing function.

Therefore, any theory of consistency that uses the classical definition must determine whether a case of trivial consistency is possible. That means that the theory should take into account the specific functions in the given set.

In order to create a theory of consistency of the ERM method that would not depend on the properties of the elements of the set of functions,

but would depend only on the general properties (capacity) of this set of functions, we need to adjust the definition of consistency to exclude the trivial consistency cases.

Definition. We say that the ERM method is *nontrivially consistent* for the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, and the probability distribution function $F(z)$ if for any nonempty subset $\Lambda(c)$, $c \in (-\infty, \infty)$, of this set of functions defined as

$$\Lambda(c) = \{\alpha : \int Q(z, \alpha) dF(z) > c, \alpha \in \Lambda\}$$

the convergence

$$\inf_{\alpha \in \Lambda(c)} R_{\text{emp}}(\alpha) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda(c)} R(\alpha) \quad (2.3)$$

is valid.

In other words, the ERM is nontrivially consistent if it provides convergence (2.3) for the subset of functions that remain after the functions with the smallest values of the risks are excluded from this set.

Note that in the classical definition of consistency described in the previous section one uses two conditions, (2.1) and (2.2). In the definition of nontrivial consistency one uses only one condition, (2.3). It can be shown that condition (2.1) will be satisfied automatically under the condition of nontrivial consistency.

In this chapter we will study conditions for nontrivial consistency, which for simplicity we will call consistency.

2.2 THE KEY THEOREM OF LEARNING THEORY

The key theorem of learning theory is the following (Vapnik and Chervonenkis, 1989):

Theorem 2.1. *Let $Q(z, \alpha)$, $\alpha \in \Lambda$, be a set of functions that satisfy the condition*

$$A \leq \int Q(z, \alpha) dF(z) \leq B \quad (A \leq R(\alpha) \leq B).$$

Then for the ERM principle to be consistent, it is necessary and sufficient that the empirical risk $R_{\text{emp}}(\alpha)$ converge uniformly to the actual risk $R(\alpha)$ over the set $Q(z, \alpha)$, $\alpha \in \Lambda$, in the following sense:

$$\lim_{\ell \rightarrow \infty} P\{\sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \varepsilon\} = 0, \quad \forall \varepsilon > 0. \quad (2.4)$$

We call this type of uniform convergence uniform *one-sided* convergence.¹

In other words, according to the key theorem, consistency of the ERM principle is *equivalent* to existence of uniform one-sided convergence (2.4).

From the conceptual point of view this theorem is extremely important because it asserts that the conditions for consistency of the ERM principle are necessarily (and sufficiently) determined by the “worst” (in sense (2.4)) function of the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$. In other words, according to this theorem *any* analysis of the ERM principle must be a “worst case analysis.”²

2.2.1 Remark on the ML Method

As has been shown in Chapter 1, the ERM principle encompasses the ML method. However, for the ML method we define another concept of non-trivial consistency.

Definition. We say that the ML method is *nontrivially consistent* if for *any* density $p(x, \alpha_0)$, from the given set of densities $p(x, \alpha) \in \Lambda$, the convergence in probability

$$\inf_{\alpha \in \Lambda} \frac{1}{\ell} \sum_{i=1}^{\ell} (-\log p(x_i, \alpha)) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} \int (-\log p(x, \alpha)) p(x, \alpha_0) dx$$

is valid, where x_1, \dots, x_ℓ is an i.i.d. sample obtained according to the density $p_0(x)$.

In other words, we define the ML method to be nontrivially consistent if it is consistent for estimating any density from the admissible set of densities.

For the ML method the following key theorem is true (Vapnik and Chervonenkis, 1989):

Theorem 2.2. *For the ML method to be nontrivially consistent on the set of densities*

$$0 < a \leq p(x, \alpha) \leq A < \infty, \quad \alpha \in \Lambda,$$

¹In contrast to the so-called uniform two-sided convergence defined by the equation

$$\lim_{\ell \rightarrow \infty} P\{\sup_{\alpha \in \Lambda} |R(\alpha) - R_{\text{emp}}(\alpha)| > \varepsilon\} = 0, \quad \forall \varepsilon > 0.$$

²The following fact confirms the importance of this theorem. Toward the end of the 1980s and the beginning of the 1990s several alternative approaches to learning theory were attempted based on the idea that statistical learning theory is a theory of “worst-case analysis.”. In these approaches authors expressed a hope to develop a learning theory for “real-case analysis.” According to the key theorem, this type of theory for the ERM principle is impossible.

it is necessary and sufficient that uniform one-sided convergence take place for the set of risk functions

$$Q(x, \alpha) = -\ln p(x, \alpha), \quad \alpha \in \Lambda,$$

with respect to some (any) probability density $p(x, \alpha_0)$, $\alpha_0 \in \Lambda$.

2.3 NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM TWO-SIDED CONVERGENCE

The key theorem of learning theory replaced the problem of consistency of the ERM method with the problem of uniform convergence (2.4). To investigate the necessary and sufficient conditions for uniform convergence, one considers two stochastic processes that are called *empirical processes*.

Consider the sequence of random variables

$$\xi^\ell = \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right|, \quad \ell = 1, 2, \dots . \quad (2.5)$$

We call this sequence of random variables that depend both on the probability measure $F(z)$ and on the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, a *two-sided empirical process*. The problem is to describe conditions under which this empirical process converges in probability to zero. The convergence in probability of the process (2.5) means that the equality

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} = 0, \quad \forall \varepsilon > 0, \quad (2.6)$$

holds true.

Along with the empirical process ξ^ℓ , we consider the *one-sided empirical process* given by the sequence of random variables

$$\xi_+^\ell = \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right)_+, \quad \ell = 1, 2, \dots \quad (2.7)$$

where we set

$$(u)_+ = \begin{cases} u & \text{if } u > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The problem is to describe conditions under which the sequence of random variables ξ_+^ℓ converges in probability to zero. Convergence in probability of the process (2.7) means that the equality

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} \left(\int Q(\alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right)_+ > \varepsilon \right\} = 0, \quad \forall \varepsilon > 0, \quad (2.8)$$

holds true. According to the key theorem, the uniform one-sided convergence (2.8) is a necessary and sufficient condition for consistency of the ERM method.

We will see that conditions for uniform two-sided convergence play an important role in constructing conditions of uniform one-sided convergence.

2.3.1 Remark on the Law of Large Numbers and Its Generalization

Note that if the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, contains only *one* element, then the sequence of random variables ξ^ℓ defined in (2.5) always converges in probability to zero. This fact constitutes the main law of statistics, the law of large numbers:

The sequence of the means of random variables ξ^ℓ converges to zero as the (number of observations) ℓ increases.

It is easy to generalize the law of large numbers for the case where a set of functions has a finite number of elements:

The sequence of random variables ξ^ℓ converges in probability to zero if the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, contains a finite number N of elements.

This case can be interpreted as the law of large numbers in an *N -dimensional vector space* (to each function in the set corresponds one coordinate; the law of large numbers in a vector space asserts convergence in probability simultaneously for all coordinates).

The problem arises when the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, has an infinite number of elements. In contrast to the cases with a finite number of elements the sequence of random variables ξ^ℓ for a set with an infinite number of elements *does not necessarily converge to zero*. The problem is this:

To describe the properties of the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, and probability measure $F(z)$ under which the sequence of random variables ξ^ℓ converges in probability to zero.

In this case one says that the *law of large numbers in the functional space* (space of functions $Q(z, \alpha)$, $\alpha \in \Lambda$) takes place or that there exists uniform (two-sided) convergence of the means to their expectation over a given set of functions.

Thus, the problem of the existence of the law of large numbers in functional space (uniform two-sided convergence of the means to their probabilities) can be considered as a generalization of the classical law of large numbers.

Note that in classical statistics the problem of the existence of uniform one-sided convergence was not considered; it became important due to the key theorem pointing the way for analysis of the problem of consistency of the ERM inductive principle.

Necessary and sufficient conditions for both uniform one-sided convergence and uniform two-sided convergence are obtained on the basis of a concept that is called the *entropy of the set of functions* $Q(z, \alpha), \alpha \in \Lambda$, on a sample of size ℓ .

For simplicity we will introduce this concept in two steps: first for the set of indicator functions (which take only the two values 0 and 1) and then for the set of real bounded functions.

2.3.2 Entropy of the Set of Indicator Functions

Let $Q(z, \alpha), \alpha \in \Lambda$, be a set of indicator functions. Consider a sample

$$z_1, \dots, z_\ell.$$

Let us characterize the diversity of the set of functions $Q(z, \alpha), \alpha \in \Lambda$, on the given set of data by the quantity $N^\Lambda(z_1, \dots, z_\ell)$ that evaluates how many different separations of the given sample can be done using functions from the set of indicator functions.

Let us write this in a more formal way. Consider the set of ℓ -dimensional binary vectors

$$q(\alpha) = (Q(z_1, \alpha), \dots, Q(z_\ell, \alpha)), \quad \alpha \in \Lambda,$$

that one obtains when α takes various values from Λ . Then geometrically speaking, $N^\Lambda(z_1, \dots, z_\ell)$ is the number of different vertices of the ℓ -dimensional cube that can be obtained on the basis of the sample z_1, \dots, z_ℓ and the set of functions $Q(z, \alpha) \in \Lambda$ (Fig. 2.3).

Let us call the value

$$H^\Lambda(z_1, \dots, z_\ell) = \ln N^\Lambda(z_1, \dots, z_\ell)$$

the *random entropy*. The random entropy describes the diversity of the set of functions on the given data. $H^\Lambda(z_1, \dots, z_\ell)$ is a random variable, since it was constructed using the i.i.d. data. Now we consider the expectation of the random entropy over the joint distribution function $F(z_1, \dots, z_\ell)$:

$$H^\Lambda(\ell) = E \ln N^\Lambda(z_1, \dots, z_\ell).$$

We call this quantity the entropy of the set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$, on samples of size ℓ . It depends on the set of functions $Q(z, \alpha), \alpha \in \Lambda$, the probability measure, and the number of observations ℓ , and it describes the expected diversity of the given set of indicator functions on a sample of size ℓ .

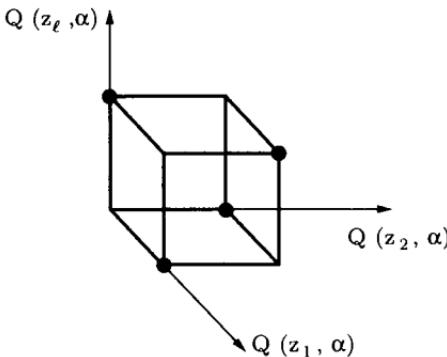


FIGURE 2.3. The set of ℓ -dimensional binary vectors $q(\alpha)$, $\alpha \in \Lambda$, is a subset of the set of vertices of the ℓ -dimensional unit cube.

2.3.3 Entropy of the Set of Real Functions

Now we generalize the definition of the entropy of the set of indicator functions on samples of size ℓ .

Definition. Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, be a set of bounded loss functions. Using this set of functions and the training set z_1, \dots, z_ℓ one can construct the following set of ℓ -dimensional vectors:

$$q(\alpha) = (Q(z_1, \alpha), \dots, Q(z_\ell, \alpha)), \quad \alpha \in \Lambda. \quad (2.9)$$

This set of vectors belongs to the ℓ -dimensional cube (Fig. 2.4) and has a finite minimal ε -net in the metric C (or in the metric L_p).³ Let $N = N^\Lambda(\varepsilon; z_1, \dots, z_\ell)$ be the number of elements of the minimal ε -net of this set

³The set of vectors $q(\alpha)$, $\alpha \in \Lambda$, has a minimal ε -net $q(\alpha_1), \dots, q(\alpha_N)$ if:

- (i) There exist $N = N^\Lambda(\varepsilon; z_1, \dots, z_\ell)$ vectors $q(\alpha_1), \dots, q(\alpha_N)$ such that for any vector $q(\alpha^*)$, $\alpha^* \in \Lambda$, one can find among these N vectors one $q(\alpha_r)$ that is ε -close to $q(\alpha^*)$ (in a given metric). For the metric C that means

$$\rho_C(q(\alpha^*), q(\alpha_r)) = \max_{1 \leq i \leq \ell} |Q(z_i, \alpha^*) - Q(z_i, \alpha_r)| \leq \varepsilon.$$

- (ii) N is the minimum number of vectors that possess this property.

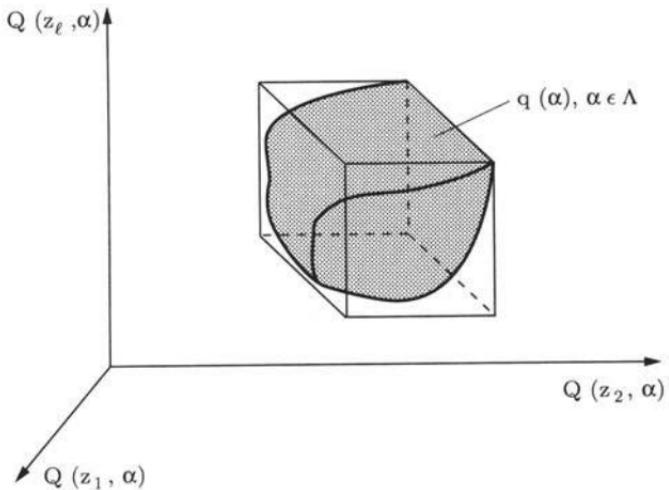


FIGURE 2.4. The set of ℓ -dimensional vectors $q(\alpha)$, $\alpha \in \Lambda$, belong to an ℓ -dimensional cube.

of vectors $q(\alpha)$, $\alpha \in \Lambda$.

Note that $N^\Lambda(\varepsilon; z_1, \dots, z_\ell)$ is a random variable, since it was constructed using random vectors z_1, \dots, z_ℓ . The logarithm of the random value $N^\Lambda(\varepsilon; z_1, \dots, z_\ell)$,

$$H^\Lambda(\varepsilon; z_1, \dots, z_\ell) = \ln N^\Lambda(\varepsilon; z_1, \dots, z_\ell),$$

is called the *random VC entropy* of the set of functions $A \leq Q(z, \alpha) \leq B$ on the sample z_1, \dots, z_ℓ . The expectation of the random VC entropy

$$H^\Lambda(\varepsilon; \ell) = EH^\Lambda(\varepsilon; z_1, \dots, z_\ell)$$

is called the *VC entropy*⁴ of the set of functions $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, on samples of size ℓ . Here the expectation is taken with respect to the product measure $F(z_1, \dots, z_\ell)$.

Note that the given definition of the entropy of a set of real functions is a generalization of the definition of the entropy given for a set of indicator

⁴The VC entropy differs from classical metrical ε -entropy

$$H(\varepsilon) = \ln N^\Lambda(\varepsilon)$$

in the following respect: $N^\Lambda(\varepsilon)$ is the cardinality of the minimal ε -net of the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, while the VC entropy is the expectation of the diversity of the set of functions on samples of size ℓ .

functions. Indeed, for a set of indicator functions the minimal ε -net for $\varepsilon < 1$ does not depend on ε and is a subset of the vertices of the unit cube. Therefore, for $\varepsilon < 1$,

$$N^\Lambda(\varepsilon; z_1, \dots, z_\ell) = N^\Lambda(z_1, \dots, z_\ell),$$

$$H^\Lambda(\varepsilon; z_1, \dots, z_\ell) = H^\Lambda(z_1, \dots, z_\ell),$$

$$H^\Lambda(\varepsilon, \ell) = H^\Lambda(\ell).$$

Below we will formulate the theory for the set of bounded real functions. The obtained general results are, of course, valid for the set of indicator functions.

2.3.4 Conditions for Uniform Two-Sided Convergence

Under some (technical) conditions of measurability on the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, the following theorem is true.

Theorem 2.3. *For uniform two-sided convergence (2.6) it is necessary and sufficient that the equality*

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon, \ell)}{\ell} = 0, \quad \forall \varepsilon > 0, \quad (2.10)$$

be valid.

In other words, the ratio of the VC entropy to the number of observations should decrease to zero with increasing numbers of observations.

Corollary. Under some conditions of measurability on the set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, necessary and sufficient condition for uniform two-sided convergence is

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell} = 0,$$

which is a particular case of equality (2.10).

This condition for uniform two-sided convergence was obtained in 1968 (Vapnik and Chervonenkis 1968, 1971). The generalization of this result for bounded sets of functions (Theorem 2.3) was found in 1981 (Vapnik and Chervonenkis 1981).

2.4 NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM ONE-SIDED CONVERGENCE

Uniform two-sided convergence can be described as follows

$$\lim_{\ell \rightarrow \infty} P \left\{ \left[\sup_{\alpha} (R(\alpha) - R_{\text{emp}}(\alpha)) > \varepsilon \right] \text{ or } \left[\sup_{\alpha} (R_{\text{emp}}(\alpha) - R(\alpha)) > \varepsilon \right] \right\} = 0. \quad (2.11)$$

The condition (2.11) includes uniform one-sided convergence and therefore forms a *sufficient condition* for consistency of the ERM method. Note, however, that when solving learning problems we face an asymmetrical situation: We require consistency in *minimizing* the empirical risk, but we do not care about consistency with respect to *maximizing* the empirical risk. So for consistency of the ERM method the second condition on the left-hand side of (2.11) can be violated.

The next theorem describes a condition under which there exists consistency in minimizing the empirical risk but not necessarily in maximizing the empirical risk (Vapnik and Chervonenkis, 1989).

Consider the set of bounded real functions $Q(z, \alpha)$, $\alpha \in \Lambda$, together with a new set of functions $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$, satisfying some conditions of measurability as well as the following conditions: For any function from $Q(z, \alpha)$, $\alpha \in \Lambda$, there exists a function in $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$, such that (Fig. 2.5)

$$\begin{aligned} Q(z, \alpha) - Q^*(z, \alpha^*) &\geq 0, \quad \forall z, \\ \int (Q(z, \alpha) - Q^*(z, \alpha^*)) dF(z) &\leq \delta. \end{aligned} \tag{2.12}$$

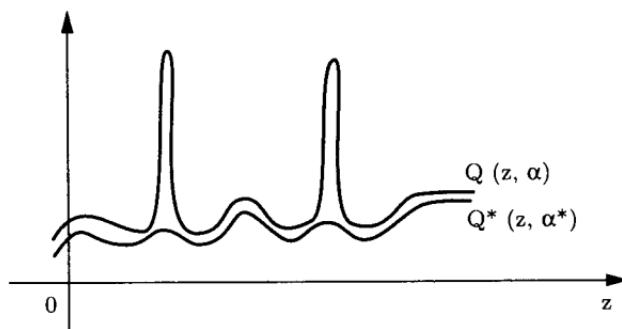


FIGURE 2.5. For any function $Q(z, \alpha)$, $\alpha \in \Lambda$, one considers a function $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$, such that $Q^*(z, \alpha^*)$ does not exceed $Q(z, \alpha)$ and is close to it.

Theorem 2.4. *In order for uniform one-sided convergence of empirical means to their expectations to hold for the set of totally bounded functions $Q(z, \alpha)$, $\alpha \in \Lambda$ (2.8), it is necessary and sufficient that for any positive δ , η , and ε there exist a set of functions $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$, satisfying (2.12) such that the following holds for the ε -entropy of the set $Q^*(z, \alpha)$, $\alpha^* \in \Lambda^*$, on samples of size ℓ :*

$$\lim_{\ell \rightarrow \infty} \frac{H^{\Lambda^*}(\varepsilon, \ell)}{\ell} < \eta. \quad (2.13)$$

In other words, for uniform one-sided convergence on the set of bounded functions $Q(z, \alpha)$, $\alpha \in \Lambda$, it is necessary and sufficient that there exist another set of functions $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$, that is close (in the sense of (2.12)) to $Q(z, \alpha)$, $\alpha \in \Lambda$, such that for this new set of functions, condition (2.13) is valid. Note that condition (2.13) is weaker than condition (2.10) in Theorem 2.3.

According to the key theorem, this is necessary and sufficient for consistency of the ERM method.

2.5 THEORY OF NONFALSIFIABILITY

From the formal point of view, Theorems 2.1, 2.3, and 2.4 give a conceptual model of learning based on the ERM inductive principle. However, both to prove Theorem 2.4 and to understand the nature of the ERM principle more deeply we have to answer the following questions:

What happens if the condition of Theorem 2.4 is not valid?

Why is the ERM method not consistent in this case?

Below, we show that if there exists an ε_0 such that

$$\lim_{\ell \rightarrow \infty} \frac{H^{\Lambda}(\varepsilon_0, \ell)}{\ell} \neq 0,$$

then the learning machine with functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is faced with a situation that in the philosophy of science corresponds to a so-called nonfalsifiable theory.

Before we describe the formal part of the theory, let us remind the reader what the idea of nonfalsifiability is.

2.5.1 Kant's Problem of Demarcation and Popper's Theory of Nonfalsifiability

Since the era of ancient philosophy, two models of reasoning have been accepted:

- (i) *deductive*, which means moving from general to particular, and

- (ii) *inductive*, which means moving from particular to general.

A model in which a system of axioms and inference rules is defined by means of which various corollaries (consequences) are obtained is ideal for the deductive approach. The deductive approach should guarantee that we obtain *true* consequences from *true* premises.

The inductive approach to reasoning consists in the formation of general judgments from particular assertions. However, the general judgments obtained from *true* particular assertions are *not always true*. Nevertheless, it is assumed that there exist such cases of inductive inference for which generalization assertions are justified.

The demarcation problem, originally proposed by Kant, is a central question of inductive theory:

What is the difference between the cases with a justified inductive step and those for which the inductive step is not justified?

The demarcation problem is usually discussed in terms of the philosophy of natural science. All theories in the natural sciences are the result of generalizations of observed real facts, and therefore theories are built using inductive inference. In the history of the natural sciences, there have been both true theories that reflect reality (say chemistry) and false ones (say alchemy) that do not reflect reality. Sometimes it takes many years of experiments to prove that a theory is false.

The question is the following:

Is there a formal way to distinguish true theories from false theories?

Let us assume that meteorology is a true theory and astrology a false one. What is the formal difference between them?

- (i) Is it in the complexity of their models?
- (ii) Is it in the predictive ability of their models?
- (iii) Is it in their use of mathematics?
- (iv) Is it in the level of formality of inference?

None of the above gives a clear advantage to either of these two theories.

- (i) The complexity of astrological models is no less than the complexity of the meteorological models.
- (ii) Both theories fail in some of their predictions.
- (iii) Astrologers solve differential equations for restoration of the positions of the planets that are no simpler than the basic equations in meteorology.

- (iv) Finally, in both theories, inference has the same level of formalization. It contains two parts: the formal description of reality and the informal interpretation of it.

In the 1930s, K. Popper suggested his famous criterion for demarcation between true and false theories (Popper, 1968). According to Popper, a necessary condition for justifiability of a theory is the feasibility of its falsification. By the falsification of a theory, Popper means the existence of a collection of particular assertions that cannot be explained by the given theory although they fall into its domain. If the given theory can be falsified it satisfies the necessary conditions of a scientific theory.

Let us come back to our example. Both meteorology and astrology make weather forecasts. Consider the following assertion:

Once, in New Jersey, in July, there was a tropical rainstorm and then snowfall.

Suppose that according to the theory of meteorology, this is impossible. Then this assertion falsifies the theory because if such a situation really should happen (note that nobody can guarantee with probability one that this is impossible⁵), the theory will not be able to explain it. In this case the theory of meteorology satisfies the necessary conditions to be viewed as a scientific theory.

Suppose that this assertion can be explained by the theory of astrology. (There are many elements in the starry sky, and they can be used to create an explanation.) In this case, this assertion does not falsify the theory. If there is no example that can falsify the theory of astrology, then astrology, according to Popper, should be considered a nonscientific theory.

In the next section we describe the theorems of nonfalsifiability. We show that if for some set of functions conditions of uniform convergence do not hold, the situation of nonfalsifiability will arise.

2.6 THEOREMS ON NONFALSIFIABILITY

In the following, we show that if uniform two-sided convergence does not take place, then the method of minimizing the empirical risk is nonfalsifiable.

⁵Recall Laplace's calculations of conditional probability that the sun has risen tomorrow given that it has risen every day up to this day. It will rise for sure according to the *models* that we use and in which we believe. However *with probability one* we can assert only that the sun has risen every day up to now during the thousands of years of recorded history.

2.6.1 Case of Complete (Popper's) Nonfalsifiability

To give a clear explanation of why this happens, let us start with the simplest case. Recall that according to the definition of VC entropy the following expressions are valid for a set of indicator functions:

$$H^\Lambda(\ell) = E \ln N^\Lambda(z_1, \dots, z_\ell) \text{ and } N^\Lambda(z_1, \dots, z_\ell) \leq 2^\ell.$$

Suppose now that for the VC entropy of the set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, the following equality is true:

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell} = \ln 2.$$

It can be shown that the ratio of the entropy to the number of observations $H^\Lambda(\ell)/\ell$ monotonically decreases as the number of observations ℓ increases.⁶ Therefore, if the limit of the ratio of the entropy to the number of observations tends to $\ln 2$, then for any finite number ℓ the equality

$$\frac{H^\Lambda(\ell)}{\ell} = \ln 2$$

holds true.

This means that for almost all samples z_1, \dots, z_ℓ (i.e., all but a set of measure zero) the equality

$$N^\Lambda(z_1, \dots, z_\ell) = 2^\ell$$

is valid.

In other words, the set of functions of the learning machine is such that almost any sample z_1, \dots, z_ℓ (of arbitrary size ℓ) can be separated in all possible ways by functions of this set. This implies that the minimum of the empirical risk for this machine equals zero. We call this learning machine nonfalsifiable because it can give a general explanation (function) for almost any data (Fig. 2.6).

Note that the minimum value of the empirical risk is equal to zero independent of the value of the expected risk.

2.6.2 Theorem on Partial Nonfalsifiability

In the case where the entropy of the set of indicator functions over the number of observations tends to a nonzero limit, the following theorem shows that there exists some subspace of the original space $Z^* \in Z$ where the learning machine is nonfalsifiable (Vapnik and Chervonenkis, 1989).

⁶This assertion is analogous to the assertion that a value of relative (with respect to the number of observations) information cannot increase with the number of observations.

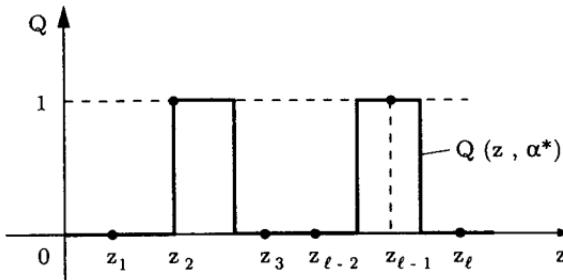


FIGURE 2.6. A learning machine with the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is *nonfalsifiable* if for almost all samples z_1, \dots, z_ℓ given by the generator of examples, and for any possible labels $\delta_1, \dots, \delta_\ell$ for these z , the machine contains a function $Q(z, \alpha^*)$ that provides equalities $\delta_i = Q(z_i, \alpha)$, $i = 1, \dots, \ell$.

Theorem 2.5. *For the set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, let the convergence*

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell} = c > 0$$

be valid.

Then there exists a subset Z^ of the set Z for which the probability measure is*

$$P(Z^*) = a(c) \neq 0$$

such that for the intersection of almost any training set

$$z_1, \dots, z_\ell$$

with the set Z^ ,*

$$z_1^*, \dots, z_k^* = (z_1, \dots, z_\ell) \cap Z^*,$$

and for any given sequence of binary values

$$\delta_1, \dots, \delta_k, \quad \delta_i \in \{0, 1\},$$

there exists a function $Q(z, \alpha^)$ for which the equalities*

$$\delta_i = Q(z_i^*, \alpha^*), \quad i = 1, 2, \dots, k,$$

hold true.

Thus, if the conditions for uniform two-sided convergence fail, then there exists some subspace of the input space where the learning machine is nonfalsifiable (Fig. 2.7).

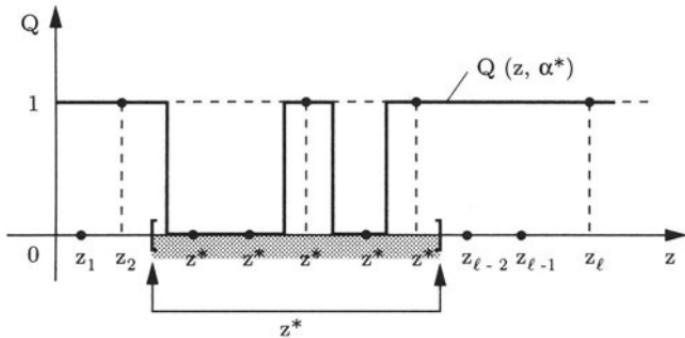


FIGURE 2.7. A learning machine with the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is *partially nonfalsifiable* if there exists a region $Z^* \subset Z$ with nonzero measure such that for almost all samples z_1, \dots, z_ℓ given by the generator of examples and for any labels $\delta_1, \dots, \delta_\ell$ for these z , the machine contains a function $Q(z, \alpha^*)$ that provides equalities $\delta_i = Q(z_i, \alpha)$ for all z_i belonging to the region Z^* .

2.6.3 Theorem on Potential Nonfalsifiability

Now let us consider the set of uniformly bounded real functions

$$|Q(z, \alpha)| \leq C, \quad \alpha \in \Lambda.$$

For this set of functions a more sophisticated model of nonfalsifiability is valid. So we give the following definition of nonfalsifiability:

Definition. We say that a learning machine that has an admissible set of real functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is *potentially nonfalsifiable* for a generator of inputs with a distribution $F(x)$ if there exist two functions⁷

$$\psi_1(z) \geq \psi_0(z)$$

such that:

- (i) There exists a positive constant c for which the equality

$$\int (\psi_1(z) - \psi_0(z)) dF(z) = c > 0$$

holds true (this equality shows that two functions $\psi_0(z)$ and $\psi_1(z)$ are essentially different).

⁷These two functions do not necessarily belong to the set $Q(z, \alpha)$, $\alpha \in \Lambda$.

(ii) For almost any sample

$$z_1, \dots, z_\ell,$$

any sequence of binary values

$$\delta(1), \dots, \delta(\ell), \quad \delta(i) \in \{0, 1\},$$

and any $\varepsilon > 0$, one can find a function $Q(z, \alpha^*)$ in the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, for which the inequalities

$$|\psi_{\delta(i)}(z_i) - Q(z_i, \alpha^*)| < \varepsilon$$

hold true.

In this definition of nonfalsifiability we use two essentially different functions $\psi_1(z)$ and $\psi_0(z)$ to generate the values y_i of the function for the given vectors z_i . To make these values arbitrary, one can switch these two functions using the arbitrary rule $\delta(i)$. The set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, forms a potentially nonfalsifiable machine for input vectors generated according to the distribution function $F(z)$ if for almost any sequence of pairs $(\psi_{\delta(i)}(z_i), z_i)$ obtained on the basis of random vectors z_i and this switching rule $\delta(i)$, one can find in this set a function $Q(z, \alpha^*)$ that describes these pairs with high accuracy (Fig. 2.8).

Note that this definition of nonfalsifiability generalizes Popper's concept:

- (i) In the simplest example considered in Section 2.6.1, for the set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, we use this concept of nonfalsifiability where $\psi_1(z) = 1$ and $\psi_0(z) = 0$,
- (ii) in Theorem 2.5 we can use the functions

$$\psi_1(z) = \begin{cases} 1 & \text{if } z \in Z^*, \\ Q(z) & \text{if } z \notin Z^*, \end{cases} \quad \psi_0(z) = \begin{cases} 0 & \text{if } z \in Z^*, \\ Q(z) & \text{if } z \notin Z^*, \end{cases}$$

where $Q(z)$ is some indicator function.

On the basis of this concept of potential nonfalsifiability, we formulate the following general theorem, which holds for an arbitrary set of uniformly bounded functions (including the sets of indicator functions) (Vapnik and Chervonenkis, 1989).

Theorem 2.6. *Suppose that for the set of uniformly bounded real functions $Q(z, \alpha)$, $\alpha \in \Lambda$, there exists an ε_0 such that the convergence*

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} = c^* > 0$$

is valid.

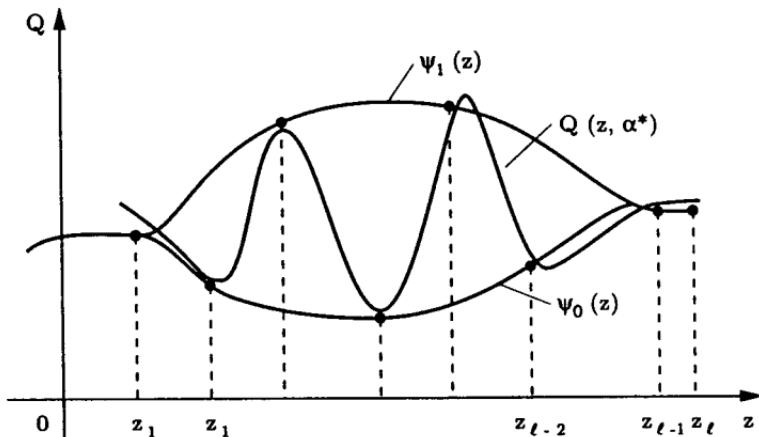


FIGURE 2.8. A learning machine with the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is *potentially nonfalsifiable* if for any $\varepsilon > 0$ there exist two essentially different functions $\psi_1(z)$ and $\psi_0(z)$ such that for almost all samples z_1, \dots, z_ℓ given by the generator of examples, and for any values u_1, \dots, u_ℓ constructed on the basis of these curves using the rule $u_i = \psi_{\delta(z_i)}(z_i)$, where $\delta(z) \subset \{0, 1\}$ is an arbitrary binary function, the machine contains a function $Q(z, \alpha^*)$ that satisfies the inequalities $|\psi_{\delta(z_i)}(z_i) - Q(z_i, \alpha^*)| \leq \varepsilon$, $i = 1, \dots, \ell$.

Then the learning machine with this set of functions is potentially non-falsifiable.

Thus, if the conditions of Theorem 2.4 fail (in this case, of course, the conditions of Theorem 2.3 will also fail), then the learning machine is non-falsifiable. This is the main reason why the ERM principle may be inconsistent.

Before continuing with the description of statistical learning theory, let me remark how amazing Popper's idea was. In the 1930s Popper suggested a general concept determining the generalization ability (in a very wide philosophical sense) that in the 1990s turned out to be one of the most crucial concepts for the analysis of consistency of the ERM inductive principle.

2.7 THREE MILESTONES IN LEARNING THEORY

Below we again consider the set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$ (i.e., we consider the problem of pattern recognition). As mentioned above, in the case of indicator functions $Q(z, \alpha), \alpha \in \Lambda$, the minimal ε -net of the vectors $q(\alpha), \alpha \in \Lambda$ (see Section 2.3.3), does not depend on ε if $\varepsilon < 1$. The number of elements in the minimal ε -net

$$N^\Lambda(z_1, \dots, z_\ell) = N^\Lambda(\varepsilon; z_1, \dots, z_\ell)$$

is equal to the number of different separations of the data z_1, \dots, z_ℓ by functions of the set $Q(z, \alpha), \alpha \in \Lambda$.

For this set of functions the VC entropy also does not depend on ε :

$$H^\Lambda(\ell) = E \ln N^\Lambda(z_1, \dots, z_\ell),$$

where expectation is taken over (z_1, \dots, z_ℓ) .

Consider two new concepts that are constructed on the basis of the values of $N^\Lambda(z_1, \dots, z_\ell)$:

(i) The *annealed VC entropy*

$$H_{\text{ann}}^\Lambda(\ell) = \ln EN^\Lambda(z_1, \dots, z_\ell);$$

(ii) The *growth function*

$$G^\Lambda(\ell) = \ln \sup_{z_1, \dots, z_\ell} N^\Lambda(z_1, \dots, z_\ell).$$

These concepts are defined in such a way that for any ℓ the inequalities

$$H^\Lambda(\ell) \leq H_{\text{ann}}^\Lambda(\ell) \leq G^\Lambda(\ell)$$

are valid.

On the basis of these functions the main milestones of learning theory are constructed.

In Section 2.3.4 we introduced the equation

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell} = 0$$

describing a *sufficient condition* for consistency of the ERM principle (the necessary and sufficient conditions are given by a slightly different construction (2.13)). This equation is the *first milestone* in learning theory: We require that any machine minimizing the empirical risk should satisfy it.

However, this equation says nothing about the rate of convergence of the obtained risks $R(\alpha_\ell)$ to the minimal one $R(\alpha_0)$. It is possible to construct examples where the ERM principle is consistent, but where the risks have an arbitrarily slow asymptotic rate of convergence.

The question is this:

Under what conditions is the asymptotic rate of convergence fast?

We say that the asymptotic rate of convergence is *fast* if for any $\ell > \ell_0$, the exponential bound

$$P\{R(\alpha_\ell) - R(\alpha_0) > \varepsilon\} < e^{-c\varepsilon^2\ell}$$

holds true, where $c > 0$ is some constant.

As it turns out, the equation

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^\Lambda(\ell)}{\ell} = 0$$

is a *sufficient condition* for a fast rate of convergence.⁸ This equation is the *second milestone* of learning theory: It guarantees a fast asymptotic rate of convergence.

Thus far, we have considered two equations: one that describes a necessary and sufficient condition for the consistency of the ERM method, and one that describes a sufficient condition for a fast rate of convergence of the ERM method. Both equations are valid for a *given* probability measure $F(z)$ on the observations (both the VC entropy $H^\Lambda(\ell)$ and the VC annealed entropy $H_{\text{ann}}^\Lambda(\ell)$ are constructed using this measure). However, our goal is to construct a learning machine capable of solving many different problems (for many different probability measures).

The question is this:

⁸The necessity of this condition for a fast rate of convergence is an open question.

Under what conditions is the ERM principle consistent and simultaneously provides a fast rate of convergence, independent of the probability measure?

The following equation describes *necessary and sufficient conditions* for consistency of ERM for *any* probability measure:

$$\lim_{\ell \rightarrow \infty} \frac{G^\Lambda(\ell)}{\ell} = 0.$$

It is also the case that if this condition holds true, then the rate of convergence is fast.

This equation is the *third milestone* in learning theory. It describes a necessary and sufficient condition under which a learning machine that implements the ERM principle has a high asymptotic rate of convergence independent of the probability measure (i.e., independent of the problem that has to be solved).

These milestones form the foundation for constructing both distribution-independent bounds for the rate of convergence of learning machines and rigorous distribution-dependent bounds, which we will consider in Chapter 3.

Informal Reasoning and Comments — 2

In the Introduction as well as in Chapter 1 we discussed the empirical risk minimization method and the methods of density estimation; however, we will not use them for constructing learning algorithms. In Chapter 4 we introduce another inductive inference, which we use in Chapter 5 for constructing learning algorithms. On the other hand, in Section 1.11 we introduced the stochastic approximation inductive principle, which we did not consider as very important in spite of the fact that some learning procedures (e.g., in neural networks) are based on this principle.

The following questions arise:

Why are the ERM principle and the methods of density estimation so important?

Why did we spend so much time describing the necessary and sufficient conditions for consistency of the ERM principle?

In these comments we will try to show that in some sense these two approaches to the problem of function estimation, one based on density estimation methods and the other based on the ERM method, reflect two quite general ideas of statistical inference.

To show this we formulate the general problem of statistics as a problem of estimating the unknown probability measure using the data. We will distinguish between two modes of estimation of probability measures, the so-called strong mode estimation and the so-called weak mode estimation. We show that methods providing strong mode estimations are based on the density estimation approach, while the methods providing weak mode estimation are based on the ERM approach.

The weak mode estimation of probability measures forms one of the most important problems in the foundations of statistics, the so-called general Glivenko–Cantelli problem. The results described in Chapter 2 provide a complete solution to this problem.

2.8 THE BASIC PROBLEMS OF PROBABILITY THEORY AND STATISTICS

In the 1930s Kolmogorov introduced an axiomatization of probability theory (Kolmogorov, 1933), and since this time probability theory has become a purely mathematical (i.e., deductive) discipline: Any analysis in this theory can be done on the basis of formal inference from the given axioms. This has allowed the development of a deep analysis of both probability theory and statistics.

2.8.1 Axioms of Probability Theory

According to Kolmogorov's axiomatization of probability theory, to every random experiment there corresponds a set Z of elementary events z that defines all possible outcomes of the experiment (the elementary events). On the set Z of elementary events, a system $\{A\}$ of subsets $A \subset Z$, which are called *events*, is defined. Considered as an event, the set Z determines a situation corresponding to a sure event (an event that always occurs). It is assumed that the set A contains the empty set \emptyset , the event that never occurs.

For the elements of $\{A\}$ the operations *union*, *complement*, and *intersection* are defined. On the set Z a σ -algebra \mathcal{F} of events $\{A\}$ is defined.⁹ The set \mathcal{F} of subsets of Z is called a σ -algebra of events $A \in \mathcal{F}$ if

- (i) $Z \in \mathcal{F}$;
- (ii) if $A \in \mathcal{F}$, then $\bar{A} \in \mathcal{F}$;
- (iii) if $A_i \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Example. Let us describe a model of the random experiments that are relevant to the following situation: Somebody throws two dice, say red and black, and observes the result of the experiment. The space of elementary events Z of this experiment can be described by pairs of integers, where the first number describes the points on the red

⁹One can read about σ -algebras in any advanced textbook on probability theory. (See, for example, A.N. Schiryaev, *Probability*, Springer, New York, p. 577.) This concept makes it possible to use the formal tools developed in measure theory for constructing the foundations of probability theory.

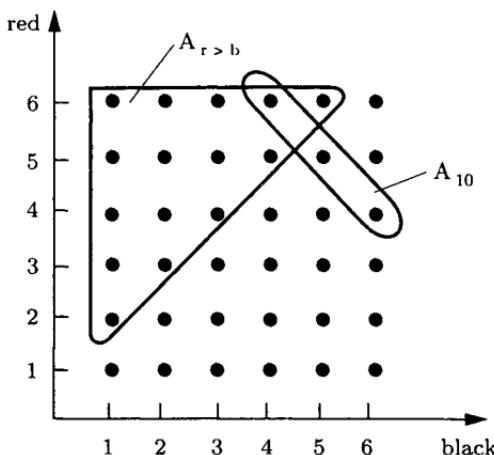


FIGURE 2.9. The space of elementary events for a two-dice throw. The events A_{10} and $A_{r>b}$ are indicated.

die and the second number describes the points on the black one. An event in this experiment can be any subset of this set of elementary events. For example, it can be the subset A_{10} of elementary events for which the sum of points on the two dice is equal to 10, or it can be the subset of elementary events $A_{r>b}$ where the red die has a larger number of points than the black one, etc. (Fig. 2.9).

The pair (Z, \mathcal{F}) consisting of the set Z and the σ -algebra \mathcal{F} of events $A \in \mathcal{F}$ is an idealization of the *qualitative* aspect of random experiments.

The *quantitative* aspect of experiments is determined by a *probability measure* $P(A)$ defined on the elements A of the set \mathcal{F} . The function $P(A)$ defined on the elements $A \in \mathcal{F}$ is called a *countably additive probability measure* on \mathcal{F} or, for simplicity, a *probability measure*, provided that

- (i) $P(A) \geq 0$;
- (ii) $P(Z) = 1$;
- (iii) $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ if $A_i, A_j \in \mathcal{F}$, and $A_i \cap A_j = \emptyset, \forall i, j$.

We say that a probabilistic model of an experiment is determined if the probability space defined by the triple (Z, \mathcal{F}, P) is determined.

Example. In our experiment let us consider a symmetrical die, where all elementary events are equally probable (have probability $1/36$).

Then the probabilities of all events are defined. (The event A_{10} has probability $3/36$, the event $A_{r>b}$ has probability $15/36$.)

In probability theory and in the theory of statistics the concept of independent trials¹⁰ plays a crucial role.

Consider an experiment containing ℓ distinct trials with probability space (Z, \mathcal{F}, P) and let

$$z_1, \dots, z_\ell \quad (2.14)$$

be the results of these trials. For an experiment with ℓ trials the model $(Z^\ell, \mathcal{F}^\ell, P^\ell)$ can be considered where Z^ℓ is a space of all possible outcomes (2.14), \mathcal{F}^ℓ is a σ -algebra on Z^ℓ that contains the sets $A_{k_1} \times \dots \times A_{k_\ell}$, and P^ℓ is a probability measure defined on the elements of the σ -algebra \mathcal{F}^ℓ .

We say that the sequence (2.14) is a sequence of ℓ *independent* trials if for any $A_{k_1}, \dots, A_{k_\ell} \in \mathcal{F}$, the equality

$$P^\ell\{z_1 \in A_{k_1}; \dots; z_\ell \in A_{k_\ell}\} = \prod_{i=1}^{\ell} P\{z_i \in A_{k_i}\}$$

is valid.

Let (2.14) be the result of ℓ independent trials with the model (Z, \mathcal{F}, P) . Consider the random variable $v(z_1, \dots, z_\ell; A)$ defined for a fixed event $A \in \mathcal{F}$ by the value

$$v_\ell(A) = v(z_1, \dots, z_\ell; A) = \frac{n_A}{\ell},$$

where n_A is the number of elements of the set z_1, \dots, z_ℓ belonging to event A . The random variable $v_\ell(A)$ is called the frequency of occurrence of an event A in a series of ℓ independent, random trials.

In terms of these concepts we can formulate the basic problems of probability theory and the theory of statistics.

The basic problem of probability theory

Given a model (Z, \mathcal{F}, P) and an event A^* , estimate the distribution (or some of its characteristics) of the frequency of occurrence of the event A^* in a series of ℓ independent random trials. Formally, this amounts to finding the distribution function

$$F(\xi; A^*, \ell) = P\{v_\ell(A^*) < \xi\} \quad (2.15)$$

(or some functionals depending on this function).

¹⁰The concept of independent trials actually is the one that makes probability theory different from measure theory. Without the concept of independent trials the axioms of probability theory define a model from measure theory.

Example. In our example with two dice it can be the following problem. What is the probability that the frequency of event A_{10} (sum of points equals 10) will be less than ξ if one throws the dice ℓ times?

In the theory of statistics one faces the *inverse* problem.

The basic problem of the theory of statistics

Given a qualitative model of random experiments (Z, \mathcal{F}) and given the i.i.d. data

$$z_1, \dots, z_\ell, \dots,$$

which occurred according to an unknown probability measure P , estimate the probability measure P defined on all subsets $A \in \mathcal{F}$ (or some functionals depending on this function).

Example. Let our two dice now be asymmetrical and somehow connected to each other (say connected by a thread). The problem is, given the results of ℓ trials (ℓ pairs), to estimate the probability measure for all events (subsets) $A \in \mathcal{F}$.

In this book we consider a set of elementary events $Z \subset R^n$ where the σ -algebra \mathcal{F} is defined to contain all Borel sets¹¹ on Z .

2.9 TWO MODES OF ESTIMATING A PROBABILITY MEASURE

One can define two modes of estimating a probability measure: A *strong mode* and A *weak mode*.

Definition:

- (i) We say that the estimator

$$\mathcal{E}_\ell(A) = \mathcal{E}_\ell(z_1, \dots, z_\ell; A), \quad A \in \mathcal{F},$$

estimates probability measure P in the strong mode if

$$\sup_{A \in \mathcal{F}} |P(A) - \mathcal{E}_\ell(A)| \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (2.16)$$

- (ii) We say that the estimator $\mathcal{E}_\ell(A)$ estimates the probability measure P in the weak mode determined by some subset $\mathcal{F}^* \subset \mathcal{F}$ if

$$\sup_{A \in \mathcal{F}^*} |P(A) - \mathcal{E}_\ell(A)| \xrightarrow[\ell \rightarrow \infty]{} 0, \quad (2.17)$$

¹¹We consider the minimal σ -algebra that contains all open parallelepipeds.

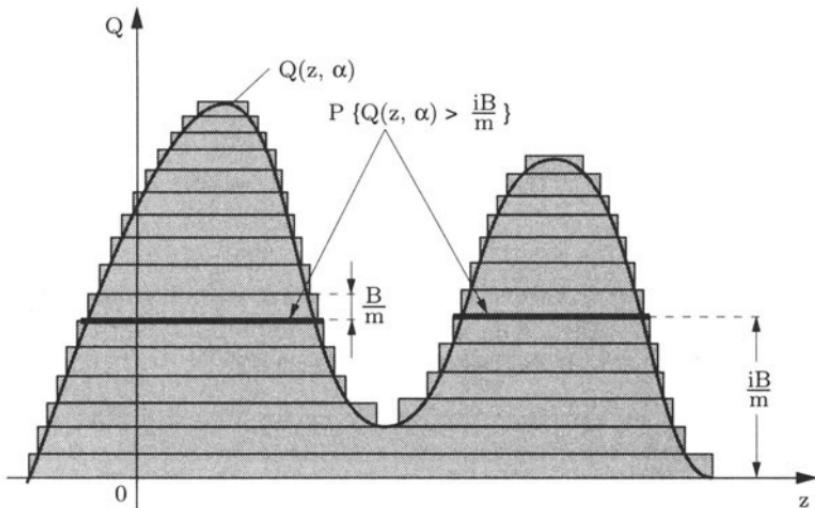


FIGURE 2.10. The Lebesgue integral defined in (2.18) is the limit of a sum of products, where the factor $P\{Q(z, \alpha) > iB/m\}$ is the (probability) measure of the set $\{z : Q(z, \alpha) > iB/m\}$, and the factor B/m is the height of a slice.

where the subset \mathcal{F}^* (of the set \mathcal{F}) does not necessarily form a σ -algebra.

For our reasoning it is important that if one can estimate the probability measure in one of these modes (with respect to a special set \mathcal{F}^* described below for the weak mode), then one can minimize the risk functional in a given set of functions.

Indeed, consider the case of bounded risk functions $0 \leq Q(z, \alpha) \leq B$. Let us rewrite the risk functional in an equivalent form, using the definition of the Lebesgue integral (Fig. 2.10):

$$R(\alpha) = \int Q(z, \alpha) dP(z) = \lim_{m \rightarrow \infty} \sum_{i=1}^m \frac{B}{m} P \left\{ Q(z, \alpha) > \frac{iB}{m} \right\}. \quad (2.18)$$

If the estimator $\mathcal{E}_\ell(A)$ approximates $P(A)$ well in the strong mode, i.e., approximates uniformly well the probability of *any* event A (including the events $A_{\alpha,i}^* = \{Q(z, \alpha) > iB/m\}$), then the functional

$$R^*(\alpha) = \lim_{m \rightarrow \infty} \sum_{i=1}^m \frac{B}{m} \mathcal{E}_\ell \left\{ Q(z, \alpha) > \frac{iB}{m} \right\} \quad (2.19)$$

constructed on the basis of the probability measure $\mathcal{E}_\ell(A)$ estimated from the data approximates uniformly well (for any α) the risk functional $R(\alpha)$. Therefore, it can be used for choosing the function that minimizes risk. The empirical risk functional $R_\ell(\alpha)$ considered in Chapters 1 and 2 corresponds to the case where estimator $\mathcal{E}_\ell(A)$ in (2.19) evaluates the frequency of event A from the given data.

Note, however, that to approximate (2.18) by (2.19) *on the given set of functions* $Q(z, \alpha)$, $\alpha \in \Lambda$, one does not need uniform approximation of P on *all events* A of the σ -algebra, one needs uniform approximation only on the events

$$A_{\alpha,i}^* = \left\{ Q(z, \alpha) > \frac{iB}{m} \right\}$$

(only these events enter in the evaluation of the risk (2.18)). Therefore, to find the function providing the minimum of the risk functional, the weak mode approximation of the probability measure with respect to the set of events

$$\left\{ Q(z, \alpha) > \frac{iB}{m} \right\}, \quad \alpha \in \Lambda,$$

is sufficient.

Thus, in order to find the function that minimizes risk (2.18) with unknown probability measure $P\{A\}$ one can minimize the functional (2.19), where instead of $P\{A\}$ an approximation $\mathcal{E}_\ell\{A\}$ that converges to $P\{A\}$ in any mode (with respect to events $A_{\alpha,i}^*$, $\alpha \in \Lambda$, $i = 1, \dots, m$, for the weak mode) is used.

2.10 STRONG MODE ESTIMATION OF PROBABILITY MEASURES AND THE DENSITY ESTIMATION PROBLEM

Unfortunately, there is no estimator that can estimate an *arbitrary* probability measure in the strong mode. One can estimate a probability measure if for this measure there exists a density function (Radon–Nikodym derivative). Let us assume that a density function $p(z)$ exists, and let $p_\ell(z)$ be an approximation to this density function. Consider an estimator

$$\mathcal{E}_\ell(A) = \int_A p_\ell(z) dz.$$

According to Scheffe's theorem, for this estimator the bound

$$\sup_A |P(A) - \mathcal{E}_\ell(A)| \leq \frac{1}{2} \int |p(z) - p_\ell(z)| dz$$

is valid, i.e., the strong mode distance between the approximation of the probability measure and the actual measure is bounded by the L_1 distance between the approximation of the density and the actual density.

Thus, to estimate the probability measure in the strong mode, it is sufficient to estimate a density function. In Section 1.8 we stressed that estimating a density function from the data forms an ill-posed problem. Therefore, generally speaking, one cannot guarantee a good approximation using a *fixed number of observations*.

Fortunately, as we saw above, to estimate the function that minimizes the risk functional one does not necessarily need to approximate the density. It is sufficient to approximate the probability measure in the weak mode, where the set of events \mathcal{F}^* depends on the admissible set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$: It must contain the events

$$\left\{ Q(z, \alpha) > \frac{iB}{m} \right\}, \quad \alpha \in \Lambda, \quad i = 1, \dots, m.$$

The “smaller” the set of admissible events considered, the “smaller” the set of events \mathcal{F}^* that must be taken into account for the weak approximation, and therefore (as we will see) minimizing the risk on a smaller set of functions requires fewer observations. In Chapter 3 we will describe bounds on the rate of uniform convergence that depend on the capacity of the set of admissible events.

2.11 THE GLIVENKO–CANTELLI THEOREM AND ITS GENERALIZATION

In the 1930s Glivenko and Cantelli proved a theorem that can be considered as the most important result in the foundation of statistics. They proved that any probability distribution function of one random variable ξ ,

$$F(z) = P\{\xi < z\},$$

can be approximated arbitrarily well by the empirical distribution function

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

where z_1, \dots, z_ℓ are i.i.d. data obtained according to an unknown density¹² (Fig. 1.2). More precisely, the Glivenko–Cantelli theorem asserts that for any $\varepsilon > 0$ the equality

$$\lim_{\ell \rightarrow \infty} P\left\{ \sup_z |F(z) - F_\ell(z)| > \varepsilon \right\} = 0$$

¹²The generalization for $n > 1$ variables was obtained later.

(convergence in probability¹³) holds true.

Let us formulate the Glivenko–Cantelli theorem in a different form. Consider the set of events

$$A_z = \{\bar{z} : \bar{z} < z\}, \quad z \in (-\infty, \infty) \quad (2.20)$$

(the set of rays on the line pointing to $-\infty$). For any event A_z of this set of events one can evaluate its probability

$$P(A_z) = \int_{-\infty}^z dF(\bar{z}) = F(z). \quad (2.21)$$

Using an i.i.d. sample of size ℓ one can also estimate the frequency of occurrence of the event A_z in independent trials:

$$v(A_z) = \frac{n_{A_z}}{\ell} = F_\ell(z). \quad (2.22)$$

In these terms, the Glivenko–Cantelli theorem asserts *weak mode convergence* of estimator (2.22) to probability measure (2.21) with respect to the set of events (2.20) (weak, because only a subset of all events is considered).

To justify the ERM inductive principle for various sets of indicator functions (for the pattern recognition problem), we constructed in this chapter a general theory of uniform convergence of frequencies to probabilities on arbitrary sets of events. This theory completed the analysis of the weak mode approximation of probability measures that was started by the Glivenko–Cantelli theory for a particular set of events (2.20).

The generalization of these results to the uniform convergence of means to their mathematical expectations over sets of functions that was obtained in 1981 actually started research on the general type of empirical processes.

2.12 MATHEMATICAL THEORY OF INDUCTION

In spite of significant results obtained in the foundation of theoretical statistics, the main conceptual problem of learning theory remained unsolved for more than twenty years (from 1968 to 1989):

Does the uniform convergence of means to their expectations form a necessary and sufficient condition for consistency of the ERM inductive principle, or is this condition only sufficient? In the latter case, might there exist another less restrictive sufficient condition?

¹³ Actually, a stronger mode of convergence holds true, the so-called convergence “almost surely.”

The answer was not obvious. Indeed, uniform convergence constitutes a global property of the set of functions, while one could have expected that consistency of the ERM principle is determined by local properties of a subset of the set of functions close to the desired one.

Using the concept of nontrivial consistency we showed in 1989 that consistency is a global property of the admissible set of functions, determined by one-sided uniform convergence (Vapnik and Chervonenkis, 1989). We found necessary and sufficient conditions for one sided convergence.

The proof of these conditions is based on a new circle of ideas — ideas on nonfalsifiability that appear in philosophical discussions on inductive inference. In these discussions, however, induction was not considered as a part of statistical inference. Induction was considered as a tool for inference in more general frameworks than the framework of statistical models.

Chapter 3

Bounds on the Rate of Convergence of Learning Processes

In this chapter we consider bounds on the rate of uniform convergence. We consider upper bounds (there exist lower bounds as well (Vapnik and Chervonenkis, 1974); however, they are not as important for controlling the learning processes as the upper bounds).

Using two different capacity concepts described in Chapter 2 (the annealed entropy function and the growth function) we describe two types of bounds on the rate of convergence:

- (i) Distribution-dependent bounds (based on the annealed entropy function), and
- (ii) distribution-independent bounds (based on the growth function).

These bounds, however, are nonconstructive, since theory does not give explicit methods to evaluate the annealed entropy function or the growth function.

Therefore, we introduce a new characteristic of the capacity of a set of functions (the VC dimension of a set of functions), which is a scalar value that can be evaluated for any set of functions accessible to a learning machine.

On the basis of the VC dimension concept we obtain

- (iii) Constructive distribution-independent bounds.

Writing these bounds in equivalent form, we find the bounds on the risk achieved by a learning machine (i.e., we estimate the generalization ability of a learning machine). In Chapter 4 we will use these bounds to control the generalization ability of learning machines.

3.1 THE BASIC INEQUALITIES

We start the description of the results of the theory of bounds with the case where $Q(z, \alpha)$, $\alpha \in \Lambda$, is a set of indicator functions and then generalize the results for sets of real functions.

Let $Q(z, \alpha)$, $\alpha \in \Lambda$, be a set of indicator functions, $H^\Lambda(\ell)$ the corresponding VC entropy, $H_{\text{ann}}^\Lambda(\ell)$ the annealed entropy and $G^\Lambda(\ell)$ the growth function (see Section 2.7).

The following two bounds on the rate of uniform convergence form the basic inequalities in the theory of bounds (Vapnik and Chervonenkis, 1968, 1971), (Vapnik, 1979, 1996).

Theorem 3.1. *The following inequality holds true:*

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{H_{\text{ann}}^\Lambda(2\ell)}{\ell} - \varepsilon^2 \right) \ell \right\}. \end{aligned} \quad (3.1)$$

Theorem 3.2. *The following inequality holds true:*

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{H_{\text{ann}}^\Lambda(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\}. \end{aligned} \quad (3.2)$$

The bounds are nontrivial (i.e., for any $\varepsilon > 0$ the right-hand side tends to zero when the number of observations ℓ goes to infinity) if

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^\Lambda(\ell)}{\ell} = 0.$$

(Recall that in Section 2.7 we called this condition the second milestone of learning theory.)

To discuss the difference between these two bounds let us recall that for any indicator function $Q(z, \alpha)$ the risk functional

$$R(\alpha) = \int Q(z, \alpha) dF(z)$$

describes the probability of event $\{z : Q(z, \alpha) = 1\}$, while the empirical functional $R_{\text{emp}}(\alpha)$ describes the frequency of this event.

Theorem 3.1 estimates the rate of uniform convergence with respect to the norm of the deviation between probability and frequency. It is clear that maximal difference more likely occurs for the events with maximal variance. For this Bernoulli case the variance is equal to

$$\sigma = \sqrt{R(\alpha)(1 - R(\alpha))},$$

and therefore the maximum of the variance is achieved for the events with probability $R(\alpha^*) \approx \frac{1}{2}$. In other words, the largest deviations are associated with functions that possess large risk.

In Section 3.3, using the bound on the rate of convergence, we will obtain a bound on the risk where the confidence interval is determined by the rate of uniform convergence, i.e., by the function with risk $R(\alpha^*) \approx \frac{1}{2}$ (the “worst” function in the set).

To obtain a smaller confidence interval one can try to construct the bound on the risk using a bound for another type of uniform convergence, namely, the uniform relative convergence

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{|R(\alpha) - R_{\text{emp}}(\alpha)|}{\sqrt{R(\alpha)(1 - R(\alpha))}} \geq \varepsilon \right\} < \Phi(\varepsilon, \ell, \cdot),$$

where the deviation is normalized by the variance. The supremum on the uniform relative convergence can be achieved on any function $Q(z, \alpha)$ including one that has a small risk.

Technically, however, it is difficult to estimate well the right-hand side for this bound. One can obtain a good bound for simpler cases, where instead of normalization by the variance one considers normalization by the function $\sqrt{R(\alpha)}$. This function is close to the variance when $R(\alpha)$ is reasonably small (this is exactly the case that we are interested in). To obtain better coefficients for the bound one considers the difference rather than the modulus of the difference in the numerator. This case of relative uniform convergence is considered in Theorem 3.2.

In Section 3.4 we will demonstrate that the upper bound on the risk obtained using Theorem 3.2 is much better than the upper bound on the risk obtained on the basis of Theorem 3.1.

The bounds obtained in Theorems 3.1 and 3.2 are distribution-dependent: They are valid for a given distribution function $F(z)$ on the observations (the distribution was used in constructing the annealed entropy function $H_{\text{ann}}^\Lambda(\ell)$).

To construct distribution independent bounds it is sufficient to note that for any distribution function $F(z)$ the growth function is not less than the annealed entropy

$$H_{\text{ann}}^\Lambda(\ell) \leq G^\Lambda(\ell).$$

Therefore, for any distribution function $F(z)$, the following inequalities hold

true:

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{G^\Lambda(2\ell)}{\ell} - \varepsilon^2 \right) \ell \right\}, \end{aligned} \quad (3.3)$$

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{G^\Lambda(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\}. \end{aligned} \quad (3.4)$$

These inequalities are nontrivial if

$$\lim_{\ell \rightarrow \infty} \frac{G^\Lambda(\ell)}{\ell} = 0. \quad (3.5)$$

(Recall that in Section 2.7 we called this equation the third milestone in learning theory).

It is important to note that conditions (3.5) are necessary and sufficient for distribution-free uniform convergence (3.3). In particular,

if condition (3.5) is violated, then there exist probability measures $F(z)$ on Z for which uniform convergence

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} = 0$$

does not take place.

3.2 GENERALIZATION FOR THE SET OF REAL FUNCTIONS

There are several ways to generalize the results obtained for the set of indicator functions to the set of real functions. Below we consider the simplest and most effective (it gives better bounds and is valid for the set of *unbounded* real functions) (Vapnik 1979, 1996).

Let $Q(z, \alpha)$, $\alpha \in \Lambda$, now be a set of real functions, with

$$A = \inf_{\alpha, z} Q(z, \alpha) \leq Q(z, \alpha) \leq \sup_{\alpha, z} Q(z, \alpha) = B$$

(here A can be $-\infty$ and/or B can be $+\infty$). We denote the open interval

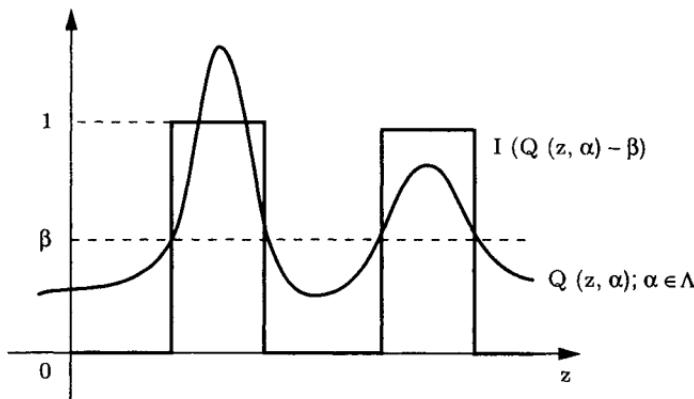


FIGURE 3.1. The indicator of level β for the function $Q(z, \alpha)$ shows for which z the function $Q(z, \alpha)$ exceeds β and for which it does not. The function $Q(z, \alpha)$ can be described by the set of all its indicators.

(A, B) by \mathcal{B} . Let us construct a *set of indicators* (Fig. 3.1) of the set of real functions $Q(z, \alpha)$, $\alpha \in \Lambda$:

$$I(z, \alpha, \beta) = \theta\{Q(z, \alpha) - \beta\}, \quad \alpha \in \Lambda, \quad \beta \in \mathcal{B}.$$

For a given function $Q(z, \alpha^*)$ and for a given β^* the indicator $I(z, \alpha^*, \beta^*)$ indicates by 1 the region $z \in Z$ where $Q(z, \alpha^*) \geq \beta^*$ and indicates by 0 the region $z \in Z$ where $Q(z, \alpha^*) < \beta^*$.

In the case where $Q(z, \alpha)$, $\alpha \in \Lambda$, are indicator functions, the set of indicators $I(z, \alpha, \beta)$, $\alpha \in \Lambda$, $\beta \in (0, 1)$, coincides with this set $Q(z, \alpha)$, $\alpha \in \Lambda$.

For any given set of real functions $Q(z, \alpha)$, $\alpha \in \Lambda$, we will extend the results of the previous section by considering the corresponding set of indicators $I(z, \alpha, \beta)$, $\alpha \in \Lambda$, $\beta \in \mathcal{B}$.

Let $H^{\Lambda, \mathcal{B}}(\ell)$ the VC entropy for the set of indicators, $H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$ the annealed entropy for the set, and $G^{\Lambda, \mathcal{B}}(\ell)$ the growth function.

Using these concepts we obtain the basic inequalities for the set of real functions as generalizations of inequalities (3.1) and (3.2). In our generalization we distinguish three cases:

- (i) Totally bounded functions $Q(z, \alpha)$, $\alpha \in \Lambda$.
- (ii) Totally bounded nonnegative functions $Q(z, \alpha)$, $\alpha \in \Lambda$.
- (iii) Nonnegative (not necessarily bounded) functions $Q(z, \alpha)$, $\alpha \in \Lambda$.

Below we consider the bounds for all three cases.

(i) Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, be a set of totally bounded functions. Then the following inequality holds true:

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{(B-A)^2} \right) \ell \right\}. \quad (3.6) \end{aligned}$$

(ii) Let $0 \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, be a set of totally bounded nonnegative functions. Then the following inequality holds true:

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{4B} \right) \ell \right\}. \quad (3.7) \end{aligned}$$

These inequalities are direct generalizations of the inequalities obtained in Theorems 3.1 and 3.2 for the set of indicator functions. They coincide with inequalities (3.1) and (3.2) when $Q(z, \alpha) \in \{0, 1\}$.

(iii) Let $0 \leq Q(z, \alpha)$, $\alpha \in \Lambda$ be a set of functions such that for some $p > 2$ the p th normalized moments¹ of the random variables $\xi_{\alpha} = Q(z, \alpha)$ exist:

$$m_p(\alpha) = \sqrt[p]{\int Q^p(z, \alpha) dF(z)}.$$

Then the following bound holds true:

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt[p]{\int Q^p(z, \alpha) dF(z)}} > a(p)\varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\}, \quad (3.8) \end{aligned}$$

where

$$a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}. \quad (3.9)$$

The bounds (3.6), (3.7), and (3.8) are nontrivial if

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)}{\ell} = 0.$$

¹We consider $p > 2$ only to simplify the formulas. Analogous results hold true for $p > 1$ (Vapnik, 1979, 1996).

3.3 THE MAIN DISTRIBUTION-INDEPENDENT BOUNDS

The bounds (3.6), (3.7), and (3.8) were distribution-dependent: The right-hand sides of the bounds use the annealed entropy $H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$ that is constructed on the basis of the distribution function $F(z)$. To obtain distribution-independent bounds one replaces the annealed entropy $H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$ on the right-hand sides of bounds (3.6), (3.7), (3.8) with the growth function $G^{\Lambda, \mathcal{B}}(\ell)$. Since for any distribution function the growth function $G^{\Lambda, \mathcal{B}}(\ell)$ is not smaller than the annealed entropy $H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$, the new bound will be truly independent of the distribution function $F(x)$.

Therefore, one can obtain the following distribution-independent bounds on the rate of various types of uniform convergence:

- (i) For the set of totally bounded functions $-\infty < A \leq Q(z, \alpha) \leq B < \infty$,

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{G^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{(B-A)^2} \right) \ell \right\}. \end{aligned} \quad (3.10)$$

- (ii) For the set of nonnegative totally bounded functions $0 \leq Q(z, \alpha) \leq B < \infty$,

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{G^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{4B} \right) \ell \right\}. \end{aligned} \quad (3.11)$$

- (iii) For the set of nonnegative real functions $0 \leq Q(z, \alpha)$ whose p th normalized moment exists for some $p > 2$,

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt[p]{\int Q^p(z, \alpha) dF(z)}} > a(p)\varepsilon \right\} \\ \leq 4 \exp \left\{ \left(\frac{G^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\}. \end{aligned} \quad (3.12)$$

These inequalities are nontrivial if

$$\lim_{\ell \rightarrow \infty} \frac{G^{\Lambda, \mathcal{B}}(\ell)}{\ell} = 0. \quad (3.13)$$

Using these inequalities one can establish bounds on the generalization ability of different learning machines.

3.4 BOUNDS ON THE GENERALIZATION ABILITY OF LEARNING MACHINES

To describe the generalization ability of learning machines that implement the ERM principle one has to answer two questions:

- (A) *What actual risk $R(\alpha_\ell)$ is provided by the function $Q(z, \alpha_\ell)$ that achieves minimal empirical risk $R_{\text{emp}}(\alpha_\ell)$?*
- (B) *How close is this risk to the minimal possible $\inf_\alpha R(\alpha)$, $\alpha \in \Lambda$, for the given set of functions?*

Answers to both questions can be obtained using the bounds described above. Below we describe distribution-independent bounds on the generalization ability of learning machines that implement sets of totally bounded functions, totally bounded nonnegative functions, and arbitrary sets of non-negative functions. These bounds are another form of writing the bounds given in the previous section.

To describe these bounds we use the notation

$$\mathcal{E} = 4 \frac{G^{\Lambda, \mathcal{B}}(2\ell) - \ln(\eta/4)}{\ell}. \quad (3.14)$$

Note that the bounds are nontrivial when $\mathcal{E} < 1$.

Case 1. The set of totally bounded functions

Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, be a set of totally bounded functions. Then:

- (A) The following inequalities hold with probability at least $1 - \eta$ simultaneously for all functions of $Q(z, \alpha)$, $\alpha \in \Lambda$ (including the function that minimizes the empirical risk):

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{(B - A)}{2} \sqrt{\mathcal{E}}, \quad (3.15)$$

$$R_{\text{emp}}(\alpha) - \frac{(B - A)}{2} \sqrt{\mathcal{E}} \leq R(\alpha).$$

(These bounds are equivalent to the bound on the rate of uniform convergence (3.10).)

- (B) The following inequality holds with probability at least $1 - 2\eta$ for the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk:

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \leq (B - A) \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{(B - A)}{2} \sqrt{\mathcal{E}}. \quad (3.16)$$

Case 2. The set of totally bounded nonnegative functions

Let $0 \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, be a set of nonnegative bounded functions. Then:

- (A) The following inequality holds with probability at least $1 - \eta$ simultaneously for all functions $Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ (including the function that minimizes the empirical risk):

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B\mathcal{E}}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{B\mathcal{E}}} \right). \quad (3.17)$$

(This bound is equivalent to the bound on the rate of uniform convergence (3.11).)

- (B) The following inequality holds with probability of at least $1 - 2\eta$ for the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \leq B \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{B\mathcal{E}}{2} \left(1 + \sqrt{1 + \frac{4}{\mathcal{E}}} \right). \quad (3.18)$$

Case 3. The set of unbounded nonnegative functions

Finally, consider the set of unbounded nonnegative functions $0 \leq Q(z, \alpha)$, $\alpha \in \Lambda$.

It is easy to show (by constructing examples) that without additional information about the set of unbounded functions and/or probability measures it is impossible to obtain any inequalities describing the generalization ability of learning machines. Below we assume the following information: We are given a pair (p, τ) such that the inequality

$$\sup_{\alpha \in \Lambda} \frac{\left(\int Q^p(z, \alpha) dF(z) \right)^{1/p}}{\int Q(z, \alpha) dF(z)} \leq \tau < \infty \quad (3.19)$$

holds true,² where $p > 1$.

The main result of the theory of learning machines with unbounded sets of functions is the following assertion, which for simplicity we will describe for the case $p > 2$ (the results for the case $p > 1$ can be found in (Vapnik, 1979, 1996)):

²This inequality describes some general properties of the distribution functions of the random variables $\xi_\alpha = Q(z, \alpha)$ generated by $F(z)$. It describes the “tails of the distributions” (the probability of large values for the random variables ξ_α). If the inequality (3.19) with $p \geq 2$ holds, then the distributions have so-called “light tails” (large values of ξ_α do not occur very often). In this case a fast rate of convergence is possible. If, however, the inequality (3.19) holds only for $p < 2$ (large values ξ_α occur rather often), then the rate of convergence will be slow (it will be arbitrarily slow if p is sufficiently close to one).

(A) With probability at least $1 - \eta$ the inequality

$$R(\alpha) \leq \frac{R_{\text{emp}}(\alpha)}{\left(1 - a(p)\tau\sqrt{\mathcal{E}}\right)_+}, \quad (3.20)$$

where

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

holds true simultaneously for all functions satisfying (3.19), where $(u)_+ = \max(u, 0)$. (This bound is a corollary of the bound on the rate of uniform convergence (3.12) and constraint (3.19).)

(B) With probability at least $1 - 2\eta$ the inequality

$$\frac{R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha)}{\inf_{\alpha \in \Lambda} R(\alpha)} \leq \frac{\tau a(p)\sqrt{\mathcal{E}}}{\left(1 - \tau a(p)\sqrt{\mathcal{E}}\right)_+} + O\left(\frac{1}{\ell}\right) \quad (3.21)$$

holds for the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk.

The inequalities (3.15), (3.17), and (3.20) bound the risks for all functions in the set $Q(z, \alpha)$, $\alpha \in \Lambda$, including the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk. The inequalities (3.16), (3.18), and (3.21) evaluate how close the risk obtained using the ERM principle is to the smallest possible risk.

Note that if $\mathcal{E} < 1$, then bound (3.17) obtained from the rate of uniform relative deviation is much better than bound (3.15) obtained from the rate of uniform convergence: For a small value of empirical risk the bound (3.17) has a confidence interval whose order of magnitude is \mathcal{E} , but not $\sqrt{\mathcal{E}}$, as in bound (3.15).

3.5 THE STRUCTURE OF THE GROWTH FUNCTION

The bounds on the generalization ability of learning machines presented above are to be thought of as conceptual rather than constructive. To make them constructive one has to find a way to evaluate the annealed entropy $H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$ and/or the growth function $G^\Lambda(\ell)$ for the given set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$.

We will find constructive bounds by using the concept of *VC dimension* of the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$ (abbreviation for Vapnik–Chervonenkis dimension).

The remarkable connection between the concept of VC dimension and the growth function was discovered in 1968 (Vapnik and Chervonenkis, 1968, 1971).

Theorem 3.3. Any growth function either satisfies the equality

$$G^\Lambda(\ell) = \ell \ln 2$$

or is bounded by the inequality

$$G^\Lambda(\ell) \leq h \left(\ln \frac{\ell}{h} + 1 \right),$$

where h is an integer such that when $\ell = h$,

$$G^\Lambda(h) = h \ln 2,$$

$$G^\Lambda(h+1) < (h+1) \ln 2.$$

In other words, the growth function is either linear or is bounded by a logarithmic function. (The growth function cannot, for example, be of the form $G^\Lambda(\ell) = c\sqrt{\ell}$ (Fig. 3.2).)

Definition. We will say that the VC dimension of the set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$ is infinite if the growth function for this set of functions is linear.

We will say that the VC dimension of the set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$, is finite and equals h if the corresponding growth function is bounded by a logarithmic function with coefficient h .

Since the inequalities

$$\frac{H^\Lambda(\ell)}{\ell} \leq \frac{H_{\text{ann}}^\Lambda(\ell)}{\ell} \leq \frac{G^\Lambda(\ell)}{\ell} \leq \frac{h(\ln \frac{\ell}{h} + 1)}{\ell} \quad (\ell > h)$$

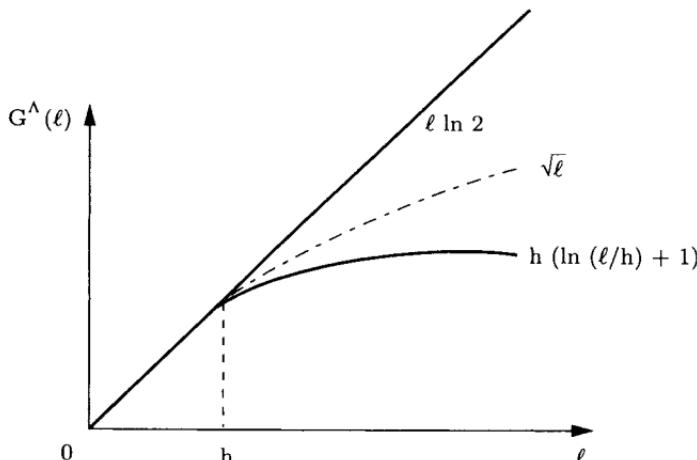


FIGURE 3.2. The growth function is either linear or bounded by a logarithmic function. It cannot, for example, behave like the dashed line.

are valid, the finiteness of the VC dimension of the set of indicator functions implemented by a learning machine is a sufficient condition for consistency of the ERM method independent of the probability measure. Moreover, a finite VC dimension implies a fast rate of convergence.

Finiteness of the VC dimension is also a necessary and sufficient condition for distribution-independent consistency of ERM learning machines. The following assertion holds true (Vapnik and Chervonenkis, 1974):

If uniform convergence of the frequencies to their probabilities over some set of events (set of indicator functions) is valid for any distribution function $F(x)$, then the VC dimension of the set of functions is finite.

3.6 THE VC DIMENSION OF A SET OF FUNCTIONS

Below we give an equivalent definition of the VC dimension for sets of indicator functions and then generalize this definition for sets of real functions. These definitions stress the method of evaluating the VC dimension.

The VC dimension of a set of indicator functions (Vapnik and Chervonenkis, 1968, 1971)

The *VC dimension of a set of indicator functions* $Q(z, \alpha)$, $\alpha \in \Lambda$, is the maximum number h of vectors z_1, \dots, z_h that can be separated into two classes in all 2^h possible ways using functions of the set³ (i.e., the maximum number of vectors that can be shattered by the set of functions). If for any n there exists a set of n vectors that can be shattered by the set $Q(z, \alpha)$, $\alpha \in \Lambda$, then the VC dimension is equal to infinity.

The VC dimension of a set of real functions (Vapnik, 1979)

Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, be a set of real functions bounded by constants A and B (A can be $-\infty$ and B can be ∞).

Let us consider along with the set of real functions $Q(z, \alpha)$, $\alpha \in \Lambda$, the set of indicators (Fig. 3.1)

$$I(z, \alpha, \beta) = \theta\{Q(z, \alpha) - \beta\}, \quad \alpha \in \Lambda, \quad \beta \in (A, B), \quad (3.22)$$

where $\theta(z)$ is the step function

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

The *VC dimension of a set of real functions* $Q(z, \alpha)$, $\alpha \in \Lambda$, is defined to be the VC dimension of the set of corresponding indicators (3.22) with parameters $\alpha \in \Lambda$ and $\beta \in (A, B)$.

³Any indicator function separates a given set of vectors into two subsets: the subset of vectors for which this indicator function takes the value 0 and the subset of vectors for which this indicator function takes the value 1.

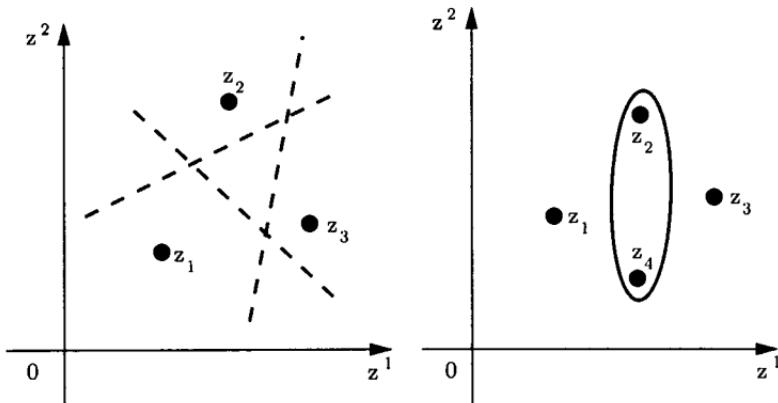


FIGURE 3.3. The VC dimension of the lines in the plane is equal to 3, since they can shatter three vectors, but not four: The vectors z_2, z_4 cannot be separated by a line from the vectors z_1, z_3 .

Example 1.

- (i) The VC dimension of the set of *linear indicator functions*

$$Q(z, \alpha) = \theta \left\{ \sum_{p=1}^n \alpha_p z_p + \alpha_0 \right\}$$

in n -dimensional coordinate space $Z = (z_1, \dots, z_n)$ is equal to $h = n + 1$, since by using functions of this set one can shatter at most $n + 1$ vectors (Fig. 3.3).

- (ii) The VC dimension of the set of *linear functions*

$$Q(z, \alpha) = \sum_{p=1}^n \alpha_p z_p + \alpha_0, \quad \alpha_0, \dots, \alpha_n \in (-\infty, \infty),$$

in n -dimensional coordinate space $Z = (z_1, \dots, z_n)$ is equal to $h = n + 1$, because the VC dimension of the corresponding linear indicator functions is equal to $n + 1$. (Note: Using $\alpha_0 - \beta$ instead of α_0 does not change the set of indicator functions.)

Note that for the set of linear functions the VC dimension equals the number of free parameters $\alpha_0, \alpha_1, \dots, \alpha_n$. In the general case this is not true.

Example 2.

- (i) The VC dimension of the set of functions

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

is infinite: The points on the line

$$z_1 = 10^{-1}, \dots, z_\ell = 10^{-\ell}$$

can be shattered by functions from this set.

Indeed, to separate these data into two classes determined by the sequence

$$\delta_1, \dots, \delta_\ell, \quad \delta_i \in \{0, 1\},$$

it is sufficient to choose the value of the parameter α to be

$$\alpha = \pi \left(\sum_{i=1}^{\ell} (1 - \delta_i) 10^i + 1 \right).$$

This example reflects the fact that by choosing an appropriate coefficient α one can for any number of appropriately chosen points approximate values of any function bounded by $(-1, +1)$ (Fig. 3.4) using $\sin \alpha x$.

In Chapter 5 we will consider a set of functions for which the VC dimension is much less than the number of parameters.

Thus, generally speaking, the VC dimension of a set of functions does not coincide with the number of parameters. It can be either larger than

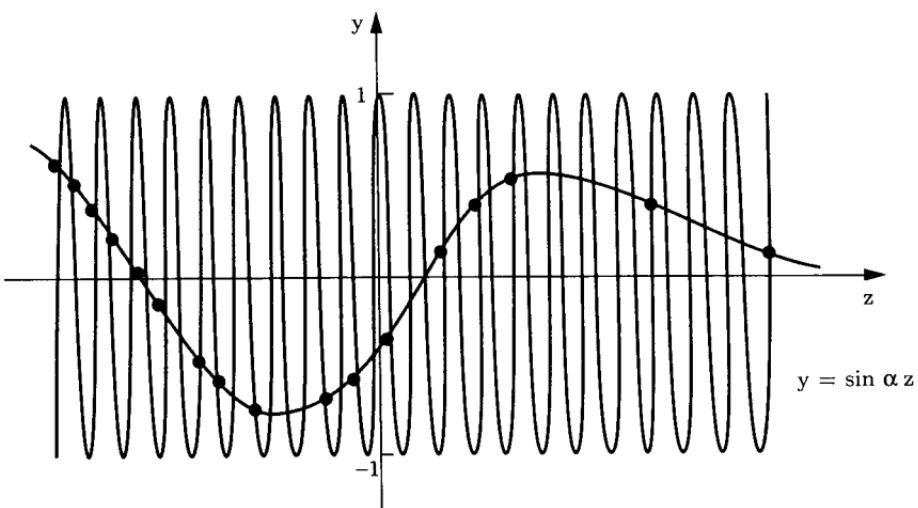


FIGURE 3.4. Using a high-frequency function $\sin(\alpha z)$, one can approximate well the value of any function $-1 \leq f(z) \leq 1$ at ℓ appropriately chosen points.

the number of parameters (as in Example 2) or smaller than the number of parameters (we will use sets of functions of this type in Chapter 5 for constructing a new type of learning machine).

In the next section we will see that the VC dimension of the set of functions (rather than number of parameters) is responsible for the generalization ability of learning machines. This opens remarkable opportunities to overcome the “curse of dimensionality”: to generalize well on the basis of a set of functions containing a huge number of parameters but possessing a small VC dimension.

3.7 CONSTRUCTIVE DISTRIBUTION-INDEPENDENT BOUNDS

In this section we will present the bounds on the risk functional that in Chapter 4 we use for constructing the methods for controlling the generalization ability of learning machines.

Consider sets of functions that possess a finite VC dimension h . In this case Theorem 3.3 states that the bound

$$G^\Lambda(\ell) \leq h \left(\ln \frac{\ell}{h} + 1 \right), \quad \ell > h, \quad (3.23)$$

holds. Therefore, in all inequalities of Section 3.3 the following constructive expression can be used:

$$\mathcal{E} = 4 \frac{h \left(\ln \frac{2\ell}{h} + 1 \right) - \ln(\eta/4)}{\ell}. \quad (3.24)$$

We also will consider the case where the set of loss functions $Q(z, \alpha)$, $\alpha \in \Lambda$, contains a finite number N of elements. For this case one can use the expression

$$\mathcal{E} = 2 \frac{\ln N - \ln \eta}{\ell}. \quad (3.25)$$

Thus, the following constructive bounds hold true, where in the case of the finite VC dimension one uses the expression for \mathcal{E} given in (3.24), and in the case of a finite number of functions in the set one uses the expression for \mathcal{E} given in (3.25).

Case 1. The set of totally bounded functions

Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, be a set of totally bounded functions. Then:

- (A) The following inequalities hold with probability at least $1 - \eta$ simultaneously for all functions $Q(z, \alpha)$, $\alpha \in \Lambda$ (including the function that

minimizes the empirical risk):

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{(B - A)}{2} \sqrt{\mathcal{E}}, \quad (3.26)$$

$$R(\alpha) \geq R_{\text{emp}}(\alpha) - \frac{(B - A)}{2} \sqrt{\mathcal{E}}.$$

- (B) The following inequality holds with probability at least $1 - 2\eta$ for the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk:

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \leq (B - A) \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{(B - A)}{2} \sqrt{\mathcal{E}}. \quad (3.27)$$

Case 2. The set of totally bounded nonnegative functions

Let $0 \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$, be a set of nonnegative bounded functions. Then

- (A) The following inequality holds with probability at least $1 - \eta$ simultaneously for all functions $Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ (including the function that minimizes the empirical risk):

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B\mathcal{E}}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{B\mathcal{E}}} \right). \quad (3.28)$$

- (B) The following inequality holds with probability at least $1 - 2\eta$ for the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk:

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \leq B \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{B\mathcal{E}}{2} \left(1 + \sqrt{1 + \frac{4}{\mathcal{E}}} \right). \quad (3.29)$$

Case 3. The set of unbounded nonnegative functions

Finally, consider the set of unbounded nonnegative functions $0 \leq Q(z, \alpha)$, $\alpha \in \Lambda$.

- (A) With probability at least $1 - \eta$ the inequality

$$R(\alpha) \leq \frac{R_{\text{emp}}(\alpha)}{\left(1 - a(p)\tau\sqrt{\mathcal{E}}\right)_+}, \quad (3.30)$$

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

holds true simultaneously for all functions satisfying (3.19), where $(u)_+ = \max(u, 0)$.

(B) With probability at least $1 - 2\eta$ the inequality

$$\frac{R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha)}{\inf_{\alpha \in \Lambda} R(\alpha)} \leq \frac{\tau a(p)\sqrt{\mathcal{E}}}{(1 - \tau a(p)\sqrt{\mathcal{E}})_+} + O\left(\frac{1}{\ell}\right) \quad (3.31)$$

holds for the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk.

These bounds cannot be significantly improved.⁴

3.8 THE PROBLEM OF CONSTRUCTING RIGOROUS (DISTRIBUTION-DEPENDENT) BOUNDS

To construct rigorous bounds on the risk one has to take into account information about the probability measure. Let \mathcal{P}_0 be the set of all probability measures on Z^ℓ and let $\mathcal{P} \subset \mathcal{P}_0$ be a subset of the set \mathcal{P}_0 . We say that one has *a priori* information about the unknown probability measure $F(z)$ if one knows a set of measures \mathcal{P} that contains $F(z)$.

Consider the following generalization of the growth function:

$$\mathcal{G}_{\mathcal{P}}^{\Lambda}(\ell) = \ln \sup_{F \in \mathcal{P}} E_F N^{\Lambda}(z_1, \dots, z_\ell).$$

For the extreme case where $\mathcal{P} = \mathcal{P}_0$, the generalized growth function $\mathcal{G}_{\mathcal{P}}^{\Lambda}(\ell)$ coincides with the growth function $G^{\Lambda}(\ell)$ because the measure that assigns probability one on z_1, \dots, z_ℓ is contained in \mathcal{P} . For another extreme case where \mathcal{P} contains only one function $F(z)$, the generalized growth function coincides with the annealed VC entropy.

Rigorous bounds for the risk can be derived in terms of the generalized growth function. They have the same functional form as the distribution-independent bounds (3.15), (3.17), and (3.21) but a different expression for \mathcal{E} . The new expression for \mathcal{E} is

$$\mathcal{E} = 4 \frac{\mathcal{G}_{\mathcal{P}}^{\Lambda}(2\ell) - \ln \eta/4}{\ell}.$$

However, these bounds are nonconstructive because no general methods have yet been found to evaluate the generalized growth function (in contrast to the original growth function, where constructive bounds were obtained on the basis of the VC dimension of the set of functions).

⁴There exist lower bounds on the rate of uniform convergence where the order of magnitude is close to the order of magnitude obtained for the upper bounds ($\sqrt{h/\ell}$ in the lower bounds instead of $\sqrt{(h/\ell) \ln(\ell/h)}$ in the upper bounds; see (Vapnik and Chervonenkis, 1974) for lower bounds).

To find rigorous constructive bounds one has to find a way of evaluating the Generalized Growth function for different sets \mathcal{P} of probability measures. The main problem here is to find a subset \mathcal{P} different from \mathcal{P}_0 for which the generalized growth function can be evaluated on the basis of some constructive concepts (much as the growth function was evaluated using the VC dimension of the set of functions).

Informal Reasoning and Comments — 3

A particular case of the bounds obtained in this chapter was already under investigation in classical statistics. These bounds are known as Kolmogorov–Smirnov distributions, widely used in both applied and theoretical statistics.

The bounds obtained in learning theory are different from the classical ones in two respects:

- (i) They are more general (they are valid for any set of indicator functions with finite VC dimension).
- (ii) They are valid for a finite number of observations (the classical bounds are asymptotic.)

3.9 KOLMOGOROV–SMIRNOV DISTRIBUTIONS

As soon as the Glivenko–Cantelli theorem became known, Kolmogorov obtained asymptotically exact estimates on the rate of uniform convergence of the empirical distribution function to the actual one (Kolmogorov, 1933). He proved that if the distribution function for a scalar random variable $F(z)$ is continuous and if ℓ is sufficiently large, then for any $\varepsilon > 0$ the following equality holds:

$$P \left\{ \sup_z |F(z) - F_\ell(z)| > \varepsilon \right\} = 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp\{-2\varepsilon^2 k^2 \ell\}. \quad (3.32)$$

This equality describes one of the main statistical laws, according to which the distribution of the random variable

$$\xi_\ell = \sup_z |F(z) - F_\ell(z)|$$

does not depend on the distribution function $F(z)$ and has the form of (3.32).

Simultaneously, Smirnov found the distribution function for one-sided deviations of the empirical distribution function from the actual one (Smirnov, 1933). He proved that for continuous $F(z)$ and sufficiently large ℓ the following equalities hold asymptotically:

$$P \left\{ \sup_z (F(z) - F_\ell(z)) > \varepsilon \right\} = \exp\{-2\varepsilon^2\ell\},$$

$$P \left\{ \sup_z (F_\ell(z) - F(z)) > \varepsilon \right\} = \exp\{-2\varepsilon^2\ell\}.$$

The random variables

$$\xi_1 = \sqrt{\ell}|F(x) - F_\ell(x)|,$$

$$\xi_2 = \sqrt{\ell}(F(x) - F_\ell(x))$$

are called the Kolmogorov–Smirnov statistics.

When the Glivenko–Cantelli theorem was generalized for multidimensional distribution functions,⁵ it was proved that for any $\varepsilon > 0$ there exists a sufficiently large ℓ_0 such that for $\ell > \ell_0$ the inequality

$$P \left\{ \sup_{\bar{z}} |F(\bar{z}) - F_\ell(\bar{z})| > \varepsilon \right\} < 2 \exp\{-a\varepsilon^2\ell\}$$

holds true, where a is any constant smaller than 2.

The results obtained in learning theory generalize the results of Kolmogorov and Smirnov in two directions:

- (i) The obtained bounds are valid for any set of events (not only for sets of rays, as in the Glivenko–Cantelli case).
- (ii) The obtained bounds are valid for any ℓ (not only asymptotically for sufficiently large ℓ).

⁵For an n -dimensional vector space Z the distribution function of the random vectors $z = (z^1, \dots, z^n)$ is determined as follows:

$$F(\bar{z}) = P \{ z^1 < \bar{z}^1, \dots, z^n < \bar{z}_n \}.$$

The empirical distribution function $F_\ell(z)$ estimates the frequency of (occurrence of) the event $A_z = \{ z^1 < \bar{z}^1, \dots, z^n < \bar{z}_n \}$.

3.10 RACING FOR THE CONSTANT

Note that the results obtained in learning theory have the form of inequalities, rather than equalities as obtained for a particular case by Kolmogorov and Smirnov. For this particular case it is possible to evaluate how close to the exact values the obtained general bounds are.

Let $Q(z, \alpha)$, $\alpha \in \Lambda$, be the set of indicator functions with VC dimension h . Let us rewrite the bound (3.3) in the form

$$\begin{aligned} P \left\{ \sup_{\alpha} \left| \int Q(z, \alpha) dP(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \\ < 4 \exp \left\{ - \left(a\varepsilon^2 - \frac{h(\ln 2\ell/h + 1)}{\ell} \right) \ell \right\}, \quad (3.33) \end{aligned}$$

where the coefficient a equals one. In the Glivenko–Cantelli case (for which the Kolmogorov–Smirnov bounds are valid) we actually consider a set of indicator functions $Q(z, \alpha) = \theta(z - \alpha)$. (For these indicator functions

$$F(\alpha) = \int \theta(z - \alpha) dF(z),$$

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

where z_1, \dots, z_ℓ are i.i.d. data.) Note that for this set of indicator functions the VC dimension is equal to one: Using indicators of rays (with one direction) one can shatter only one point. Therefore, for a sufficiently large ℓ , the second term in parentheses of the exponent on the right-hand side of (3.33) is arbitrarily small, and the bound is determined by the first term in the exponent. This term in the general formula coincides with the (main) term in the Kolmogorov–Smirnov formulas up to a constant: Instead of $a = 1$ Kolmogorov–Smirnov bounds have constant⁶ $a = 2$.

In 1988 Devroye found a way to obtain a nonasymptotic bound with the constant $a = 2$ (Devroye, 1988). However, in the exponent of the right-hand side of this bound the second term is

$$\frac{h(\ln \ell^2/h + 1)}{\ell}$$

⁶In the first result obtained in 1968 the constant was $a = 1/8$ (Vapnik and Chervonenkis, 1968, 1971); then in 1979 it was improved to $a = 1/4$ (Vapnik, 1979). In 1991 L. Bottou showed me a proof with $a = 1$. This bound also was obtained by J.M. Parrondo and C. Van den Broeck (Parrondo and Van den Broeck, 1993).

instead of

$$\frac{h(\ln 2\ell/h + 1)}{\ell}. \quad (3.34)$$

For the case that is important in practice, namely, where

$$-\ln \eta < h(\ln h - 1),$$

the bound with coefficient $a = 1$ and term (3.34) described in this chapter is better.

3.11 BOUNDS ON EMPIRICAL PROCESSES

The bounds obtained for the set of real functions are generalizations of the bounds obtained for the set of indicator functions. These generalizations were obtained on the basis of a generalized concept of VC dimension that was constructed for the set of real functions.

There exist, however, several ways to construct a generalization of the VC dimension concept for sets of real functions that allow us to derive the corresponding bounds.

One of these generalizations is based on the concept of a VC subgraph introduced by Dudley (Dudley, 1978) (in the AI literature, this concept was renamed pseudo-dimension). Using the VC subgraph concept Dudley obtained a bound on the metric ε -entropy for the set of bounded real functions. On the basis of this bound, Pollard derived a bound for the rate of uniform convergence of the means to their expectation (Pollard, 1984). This bound was used by Haussler for learning machines.⁷

Note that the VC dimension concept for the set of real functions described in this chapter forms a slightly stronger requirement on the capacity of the set of functions than Dudley's VC subgraph. On the other hand, using the VC dimension concept one obtains more attractive bounds:

- (i) They have a form that has a clear physical sense (they depend on the ratio ℓ/h).
- (ii) More importantly, using this concept one can obtain bounds on uniform relative convergence for sets of bounded functions as well as for sets of *unbounded* functions. The rate of uniform convergence (or uniform relative convergence) of the empirical risks to actual risks for the unbounded set of loss functions is the basis for an analysis of the regression problem.

⁷D. Haussler (1992), "Decision theoretic generalization of the PAC model for neural net and other applications," *Inform. Comp.* **100** (1) pp. 78–150.

The bounds for uniform relative convergence have no analogy in classical statistics. They were derived for the first time in learning theory to obtain rigorous bounds on the risk.

Chapter 4

Controlling the Generalization Ability of Learning Processes

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

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

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

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

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

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

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

where

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

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

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

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

4.1 STRUCTURAL RISK MINIMIZATION (SRM) INDUCTIVE PRINCIPLE

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

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

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

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

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

$$S_1 \subset S_2 \subset \dots \subset S_n \dots, \quad (4.3)$$

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

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

$$h_1 \leq h_2 \dots \leq h_n \dots .$$

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

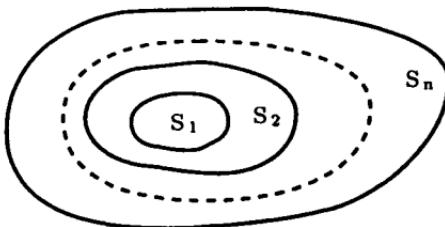


FIGURE 4.1. A structure on the set of functions is determined by the nested subsets of functions.

(ii) Any element S_k of the structure contains either

a set of totally bounded functions,

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

or a set of functions satisfying the inequality

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

for some pair (p, τ_k) .

We call this structure an *admissible structure*.

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

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

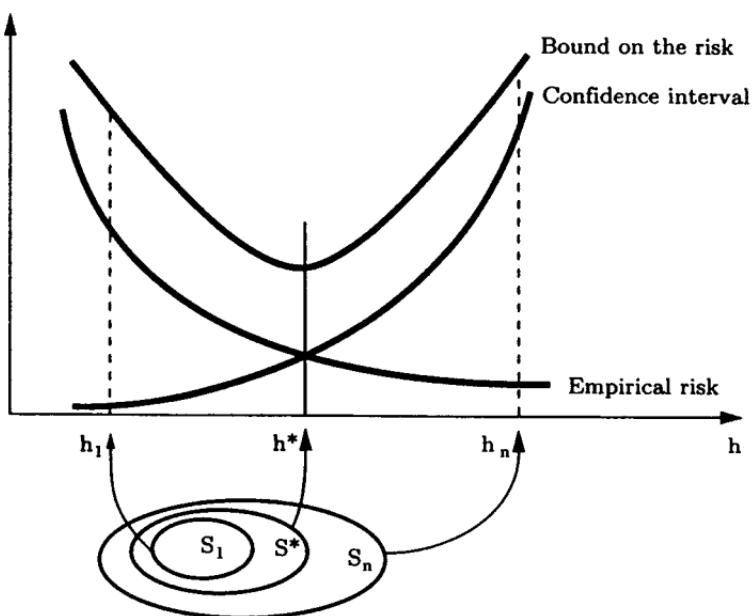


FIGURE 4.2. The bound on the risk is the sum of the empirical risk and the confidence interval. The empirical risk decreases with the index of the element of the structure, while the confidence interval increases. The smallest bound of the risk is achieved on some appropriate element of the structure.

4.2 ASYMPTOTIC ANALYSIS OF THE RATE OF CONVERGENCE

Denote by S^* the set of functions

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

Suppose that the set of functions S^* is everywhere dense² in S (recall $S = \{Q(z, \alpha), \alpha \in \Lambda\}$) with respect to the metric

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

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

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

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

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

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

with asymptotic rate of convergence³

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

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

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

holds true.

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

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

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

$$\lim_{\ell \rightarrow \infty} \frac{T_{n(\ell)}^2 h_{n(\ell)} \ln \ell}{\ell} = 0, \quad (4.7)$$

where

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

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

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

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

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

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

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

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

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

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

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

4.3 THE PROBLEM OF FUNCTION APPROXIMATION IN LEARNING THEORY

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

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

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

$$r_n = n^{-\frac{s}{N}}, \quad (4.10)$$

where N is the dimensionality of the input space (Lorentz, 1966). Note that this implies that a high asymptotic rate of convergence⁵ in high-dimensional spaces can be guaranteed only for very smooth functions.

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

- (i) $Q(z, \alpha)$, $\alpha \in \Lambda$, is a set of high-dimensional functions.
- (ii) The elements S_k of the structure are not necessarily linear manifolds. (They can be any set of functions with finite VC dimension.)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.4 EXAMPLES OF STRUCTURES FOR NEURAL NETS

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

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

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

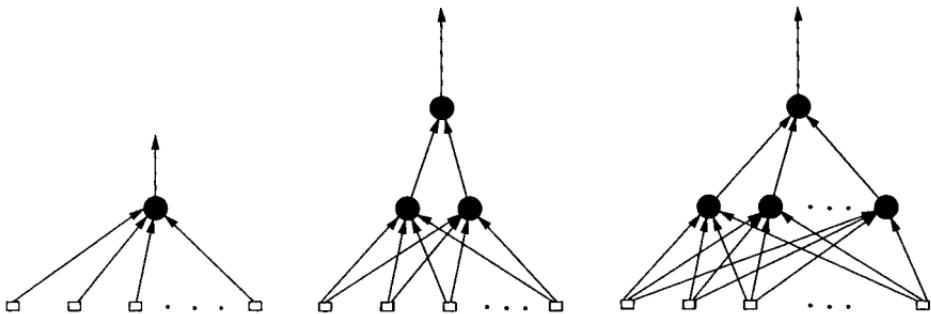


FIGURE 4.3. A structure determined by the number of hidden units.

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

2. A structure given by the learning procedure

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

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

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

3. A structure given by preprocessing

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

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

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

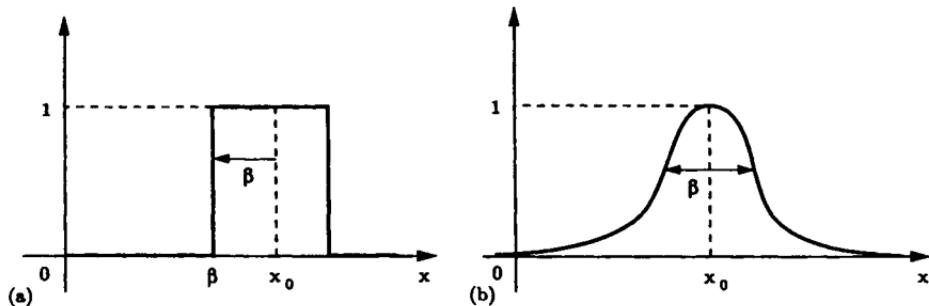


FIGURE 4.4. Examples of vicinity functions: (a) shows a hard-threshold vicinity function and (b) shows a soft-threshold vicinity function.

4.5 THE PROBLEM OF LOCAL FUNCTION ESTIMATION

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

$$\begin{aligned} 0 \leq K(x, x_0; \beta) \leq 1, \\ K(x_0, x_0; \beta) = 1. \end{aligned} \quad (4.15)$$

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

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

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

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

meet these conditions.

Let us define a value

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

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

risk functional

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

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

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

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

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

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

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

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

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

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

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

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

4.6 THE MINIMUM DESCRIPTION LENGTH (MDL) AND SRM PRINCIPLES

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

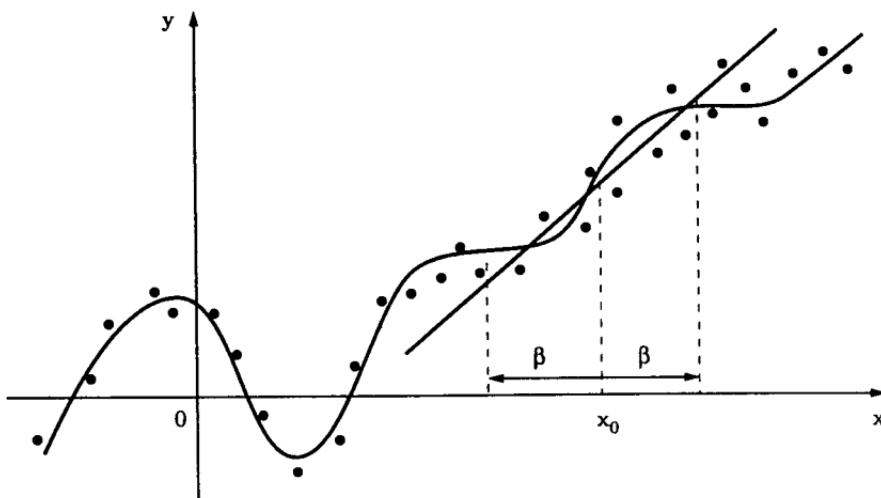


FIGURE 4.5. Using linear functions one can estimate an unknown smooth function in the vicinity of any point of interest.

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

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

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

The main idea of Kolmogorov is this:

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

Ten years after the concept of algorithmic complexity was introduced, Rissanen suggested using Kolmogorov's concept as the main tool of inductive inference of learning machines; he suggested the so-called MDL principle⁶ (Rissanen, 1978).

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

4.6.1 The MDL Principle

Suppose that we are given a training set of pairs

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

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

$$\omega_1, \dots, \omega_\ell \quad (4.20)$$

and the string of vectors

$$x_1, \dots, x_\ell. \quad (4.21)$$

The question is,

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

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

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

Consider the following model: Suppose that we are given some fixed codebook C_b with $N \ll 2^\ell$ different tables T_i , $i = 1, \dots, N$. Any table T_i describes some function⁷ from x to ω .

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

$$\omega_1^*, \dots, \omega_\ell^* \quad (4.22)$$

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

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

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

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

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

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

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

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

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

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

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

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

4.6.2 Bounds for the MDL Principle

The important question is the following:

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

The answer is yes.

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

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

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

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

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

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

$$\begin{aligned} & \frac{d}{\ell} + \frac{\ln N - \ln \eta}{\ell} \left(1 + \sqrt{1 + \frac{2d}{\ln N - \ln \eta}} \right) \\ & < 2 \left(\frac{\lceil \ln N \rceil + \lceil \ln C_\ell^d \rceil + \lceil \lg_2 d \rceil + \Delta_d}{\ell} - \frac{\ln \eta}{\ell} \right) \end{aligned} \quad (4.26)$$

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

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

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

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

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

4.6.3 The SRM and MDL Principles

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

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

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

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

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

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

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

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

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

- (i) How many examples we used,
- (ii) how the structure of the codebooks was organized,
- (iii) which codebook was used,
- (iv) how many tables were in the codebook,
- (v) how many training errors were made using this table.

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

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

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

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

4.6.4 A Weak Point of the MDL Principle

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

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

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

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

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

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

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

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

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

Informal Reasoning and Comments — 4

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

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

- (i) Methods of quasi-solutions (Ivanov, 1962),
- (ii) methods of regularization (Tikhonov, 1963)).

It then appeared in the method for nonparametric density estimation:

- (i) Parzen windows (Parzen, 1962),
- (ii) projection methods (Chentsov, 1963),
- (iii) conditional maximum likelihood method (the method of sieves (Grenander, 1981)),
- (iv) maximum penalized likelihood method (Tapia and Thompson, 1978)), etc.

The idea then appeared in methods for regression estimation:

- (i) Ridge regression (Hoerl and Kennard, 1970),
- (ii) model selection (see review in (Miller, 1990)).

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

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

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

4.7 METHODS FOR SOLVING ILL-POSED PROBLEMS

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

$$Af = F, \quad f \in M, \quad (4.29)$$

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

$$M_1 \subset M_2 \subset \cdots \subset M_k, \dots, \quad (4.30)$$

$$\overline{\bigcup_{i=1}^{\infty} M_i} = M, \quad (4.31)$$

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

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

Ivanov proved that under some general conditions the sequence of solutions

$$f_1^*, \dots, f_k^*, \dots$$

converges to the desired one.

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

- (i) The domain of the functional coincides with M (the domain to which the solution of (4.29) belongs).
- (ii) The region for which the inequality

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

holds forms a compactum in the metric of space E_1 .

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

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

Tikhonov suggested finding a sequence of functions f_γ minimizing the functionals

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

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

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

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

In this case, in minimizing the functionals

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

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

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

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

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

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

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

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

- (i) Each of these approximations F_δ is a random function.⁹
 - (ii) The sequence of approximations converges in probability (in the metric of the space E_2) to the unknown function F as δ converges to zero.
- In other words, the sequence of random functions F_δ has the property

$$P\{\rho_{E_2}(F, F_\delta) > \varepsilon\} \xrightarrow[\delta \rightarrow 0]{} 0, \quad \forall \varepsilon > 0.$$

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

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

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

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

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

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

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

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

Let us construct the empirical distribution function

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

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

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

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

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

Therefore, if one minimizes the regularized functional

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

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

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

Therefore, the conditions for consistency of the obtained estimators are

$$\begin{aligned} \gamma_\ell &\xrightarrow[\ell \rightarrow \infty]{} 0, \\ \ell\gamma_\ell &\xrightarrow[\ell \rightarrow \infty]{} \infty. \end{aligned} \quad (4.35)$$

Thus, minimizing functionals of type (4.34) under the constraint (4.35) gives consistent estimators. Using various norms E_2 and various functionals $\Omega(p)$ one can obtain various types of density estimators (including all classical estimators¹⁰). For our reasoning it is important that all nonparametric density estimators implement the SRM principle. By choosing the functional $\Omega(p)$, one defines a structure on the set of admissible solutions (the nested set of functions $M_c = \{p : \Omega(p) \leq c\}$ determined by constant c); using the law γ_ℓ one determines the appropriate element of the structure.

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

4.9 THE PROBLEM OF POLYNOMIAL APPROXIMATION OF THE REGRESSION

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

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

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

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

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

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

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

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

4.10 THE PROBLEM OF CAPACITY CONTROL

4.10.1 Choosing the Degree of the Polynomial

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

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

the second element contains polynomials of degree two:

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

and so on.

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

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

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

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

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

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

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

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

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

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

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

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

To find the best approximating polynomial, one has to choose both the degree m of the polynomial and the coefficients α minimizing functional¹¹ (4.36).

4.10.2 Choosing the Best Sparse Algebraic Polynomial

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

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

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

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

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

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

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

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

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

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

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

4.10.3 Structures on the Set of Trigonometric Polynomials

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

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

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

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

is infinite (see Example 2, Section 3.6).

¹²Trigonometric polynomials of degree m have the form

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

4.10.4 The Problem of Feature Selection

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

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

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

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

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

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

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

4.11 THE PROBLEM OF CAPACITY CONTROL AND BAYESIAN INFERENCE

4.11.1 The Bayesian Approach in Learning Theory

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

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

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

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

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

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

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

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

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

equals

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

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

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

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

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

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

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

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

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

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

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

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

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

Then from (4.37) one obtains the functional

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

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

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

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

4.11.2 Discussion of the Bayesian Approach and Capacity Control Methods

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

- (i) The given set of functions of the learning machine coincides with the set of problems to be solved.
- (ii) The *a priori* distribution on the set of problems is described by the given expression $P(\alpha)$.¹⁴

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

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

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

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

The Bayesian approach also works in general situations.

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

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

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

Chapter 5

Methods of Pattern Recognition

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

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

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

5.1 WHY CAN LEARNING MACHINES GENERALIZE?

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

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

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

$$R(\alpha_\ell^k) \leq R_{\text{emp}}(\alpha_\ell^k) + \Phi\left(\frac{\ell}{h_k}\right), \quad (5.1)$$

where the first term is the empirical risk and the second term is the confidence interval.

There are two *constructive* approaches to minimizing the right-hand side of inequality (5.1).

In the first approach, during the design of the learning machine one determines a set of admissible functions with some VC dimension h^* . For a given amount ℓ of training data, the value h^* determines the confidence interval $\Phi(\frac{\ell}{h^*})$ for the machine. Choosing an appropriate element of the structure is therefore a problem of designing the machine for a specific amount of data.

During the learning process this machine minimizes the first term of the bound (5.1) (the number of errors on the training set).

If for a given amount of training data one designs too complex a machine, the confidence interval $\Phi(\frac{\ell}{h^*})$ will be large. In this case even if one could minimize the empirical risk down to zero, the number of errors on the test set could still be large. This phenomenon is called *overfitting*.

To avoid overfitting (to get a small confidence interval) one has to construct machines with small VC dimension. On the other hand, if the set of functions has a small VC dimension, then it is difficult to approximate the training data (to get a small value for the first term in inequality (5.1)). To obtain a small approximation error and simultaneously keep a small confidence interval one has to choose the architecture of the machine to reflect *a priori* knowledge about the problem at hand.

Thus, to solve the problem at hand by these types of machines, one first has to find the appropriate architecture of the learning machine (which is a result of the trade off between overfitting and poor approximation) and second, find in this machine the function that minimizes the number of errors on the training data. This approach to minimizing the right-hand side of inequality (5.1) can be described as follows:

Keep the confidence interval fixed (by choosing an appropriate construction of machine) and minimize the empirical risk.

The second approach to the problem of minimizing the right-hand side of inequality (5.1) can be described as follows:

Keep the value of the empirical risk fixed (say equal to zero) and minimize the confidence interval.

Below we consider two different types of learning machines that implement these two approaches:

- (i) neural networks (which implement the first approach), and
- (ii) support vector machines (which implement the second approach).

Both types of learning machines are generalizations of the learning machines with a set of linear indicator functions constructed in the 1960s.

5.2 SIGMOID APPROXIMATION OF INDICATOR FUNCTIONS

Consider the problem of minimizing the empirical risk on the set of *linear indicator functions*

$$f(x, w) = \text{sign} \{(w \cdot x)\}, \quad w \in R^n, \quad (5.2)$$

where $(w \cdot x)$ denotes an inner product between vectors w and x . Let

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

be a training set, where x_j is a vector, and $y_j \in \{1, -1\}$, $j = 1, \dots, \ell$.

The goal is to find the vector of parameters w_0 (weights) that minimize the empirical risk functional

$$R_{\text{emp}}(w) = \frac{1}{\ell} \sum_{j=1}^{\ell} (y_j - f(x_j, w))^2. \quad (5.3)$$

If the training set is separable without error (i.e., the empirical risk can become zero), then there exists a finite-step procedure that allows us to find such a vector w_0 , for example the procedure that Rosenblatt proposed for the perceptron (see the Introduction).

The problem arises when the training set cannot be separated without errors. In this case the problem of separating the training data with the smallest number of errors is NP-complete. Moreover, one cannot apply regular gradient-based procedures to find a local minimum of functional (5.3), since for this functional the gradient is either equal to zero or undefined.

Therefore, the idea was proposed to approximate the indicator functions (5.2) by the so-called *sigmoid functions* (see Fig. 0.3)

$$\bar{f}(x, w) = S \{(w \cdot x)\}, \quad (5.4)$$

where $S(u)$ is a smooth monotonic function such that

$$S(-\infty) = -1, \quad S(+\infty) = 1,$$

for example,

$$S(u) = \tanh u = \frac{\exp(u) - \exp(-u)}{\exp(u) + \exp(-u)}.$$

For the set of sigmoid functions, the empirical risk functional

$$R_{\text{emp}}(w) = \frac{1}{\ell} \sum_{j=1}^{\ell} (y_j - S\{(w \cdot x_i)\})^2$$

is smooth in w . It has gradient

$$\text{grad}_w R_{\text{emp}}(w) = -\frac{2}{\ell} \sum_{j=1}^{\ell} [y_j - S((w \cdot x_j))] S' \{(w \cdot x_j)\} x_j^T,$$

and therefore it can be minimized using standard gradient-based methods, for example, the *gradient descent method*:

$$w_{\text{new}} = w_{\text{old}} - \gamma(\cdot) \text{grad} R_{\text{emp}}(w_{\text{old}}),$$

where $\gamma(\cdot) = \gamma(n) \geq 0$ is a value that depends on the iteration number n . For convergence of the gradient descent method to local minima it is sufficient that the values of the gradient be bounded and that the coefficients $\gamma(n)$ satisfy the following conditions:

$$\sum_{n=1}^{\infty} \gamma(n) = \infty, \quad \sum_{n=1}^{\infty} \gamma^2(n) < \infty.$$

Thus, the idea is to use the sigmoid approximation at the stage of estimating the coefficients, and use the threshold functions (with the obtained coefficients) for the last neuron at the stage of recognition.

5.3 NEURAL NETWORKS

In this section we consider classical neural networks, which implement the first strategy: Keep the confidence interval fixed and minimize the empirical risk.

This idea is used to estimate the weights of all neurons of a multilayer perceptron (neural network). Instead of linear indicator functions (single neurons) in the networks one considers a set of sigmoid functions.

The method for calculating the gradient of the empirical risk for the sigmoid approximation of neural networks, called the *back-propagation method*, was proposed in 1986 (Rumelhart, Hinton, and Williams, 1986), (LeCun, 1986). Using this gradient, one can iteratively modify the coefficients (weights) of a neural net on the basis of standard gradient-based procedures.

5.3.1 The Back-Propagation Method

To describe the back-propagation method we use the following notation (Fig. 5.1):

- (i) The neural net contains $m + 1$ layers: the first layer $x(0)$ describes the input vector $x = (x^1, \dots, x^n)$. We denote the input vector by

$$x_i = (x_i^1(0), \dots, x_i^n(0)), \quad i = 1, \dots, \ell,$$

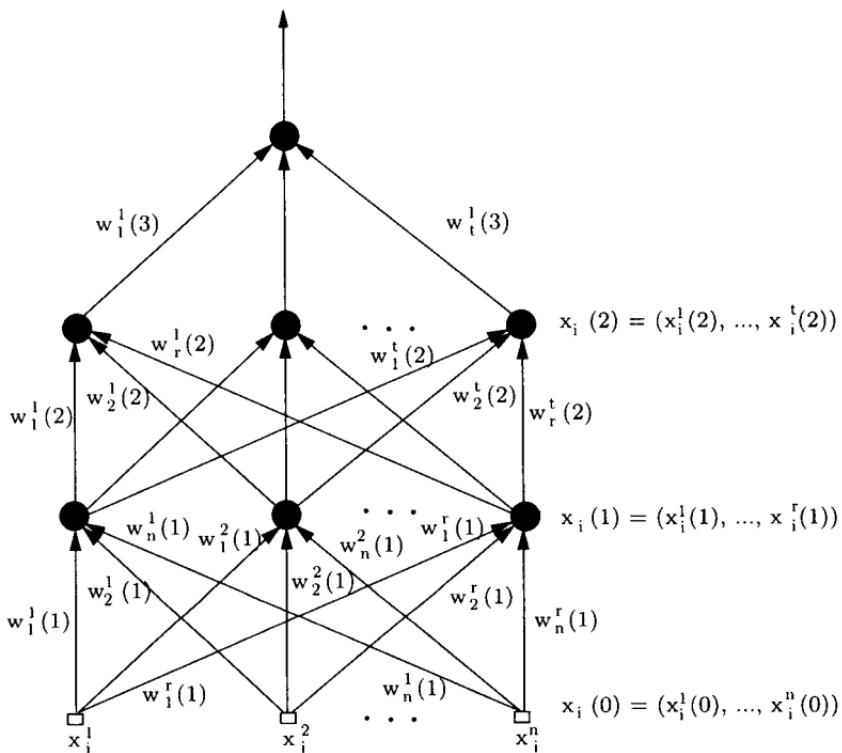


FIGURE 5.1. A neural network is a combination of several levels of sigmoid elements. The outputs of one layer form the inputs for the next layer.

and the image of the input vector $x_i(0)$ on the k th layer by

$$x_i(k) = (x_i^1(k), \dots, x_i^{n_k}(k)), \quad i = 1, \dots, \ell,$$

where we denote by n_k the dimensionality of the vectors $x_i(k)$, $i = 1, \dots, \ell$ (n_k , $k = 1, \dots, m - 1$ can be any number, but $n_m = 1$).

- (ii) Layer $k - 1$ is connected with layer k through the $(n_k \times n_{k-1})$ matrix $w(k)$

$$x_i(k) = S\{w(k)x_i(k - 1)\}, \quad k = 1, 2, \dots, m, \quad i = 1, \dots, \ell, \quad (5.5)$$

where $S\{w(k)x_i(k - 1)\}$ defines the sigmoid function of the vector

$$u_i(k) = w(k)x_i(k - 1) = (u_i^1(k), \dots, u_i^{n_k}(k))$$

as the vector coordinates transformed by the sigmoid:

$$S(u_i(k)) = (S(u_i^1(k)), \dots, S(u_i^{n_k}(k))).$$

The goal is to minimize the functional

$$I(w(1), \dots, w(m)) = \sum_{i=1}^{\ell} (y_i - x_i(m))^2 \quad (5.6)$$

under conditions (5.5).

This optimization problem is solved by using the standard technique of Lagrange multipliers for equality type constraints. We will minimize the Lagrange function

$$L(W, X, B)$$

$$= \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - x_i(m))^2 - \sum_{i=1}^{\ell} \sum_{k=1}^m (b_i(k) \cdot [x_i(k) - S\{w(k)x_i(k - 1)\}]),$$

where $b_i(k) \geq 0$ are Lagrange multipliers corresponding to the constraints (5.5) that describe the connections between vectors $x_i(k - 1)$ and vectors $x_i(k)$.

It is known that

$$\nabla L(W, X, B) = 0$$

is a necessary condition for a local minimum of the performance function (5.6) under the constraints (5.5) (the gradient with respect to all parameters from $b_i(k)$, $x_i(k)$, $w(k)$, $i = 1, \dots, \ell$, $k = 1, \dots, m$, is equal to zero).

This condition can be split into three subconditions:

$$(i) \quad \frac{\partial L(W, X, B)}{\partial b_i(k)} = 0 \quad \forall i, k,$$

$$(ii) \quad \frac{\partial L(W, X, B)}{\partial x_i(k)} = 0 \quad \forall i, k,$$

$$(iii) \quad \frac{\partial L(W, X, B)}{\partial w(k)} = 0 \quad \forall w(k).$$

The solution of these equations determines a stationary point (W_0, X_0, B_0) that includes the desired matrices of weights $W_0 = (w^0(1), \dots, w^0(m))$. Let us rewrite these three subconditions in explicit form

(i) The first subcondition

The first subcondition gives a set of equations:

$$x_i(k) = S \{w(k)x_i(k-1)\}, \quad i = 1, \dots, \ell, \quad k = 1, \dots, m,$$

with initial conditions

$$x_i(0) = x_i,$$

the equation of the so-called *forward dynamics*.

(ii) The second subcondition

We consider the second subconditions for two cases: The case $k = m$ (for the last layer) and the case $k \neq m$ (for hidden layers).

For the last layer we obtain

$$b_i(m) = 2(y_i - x_i(m)), \quad i = 1, \dots, \ell.$$

For the general case (hidden layers) we obtain

$$b_i(k) = w^T(k+1) \nabla S \{w(k+1)x_i(k)\} b_i(k+1),$$

$$i = 1, \dots, \ell, \quad k = 1, \dots, m-1,$$

where $\nabla S \{w(k+1)x_i(k)\}$ is a diagonal $n_{k+1} \times n_{k+1}$ matrix with diagonal elements $S'(u_r)$, where u_r is the r th coordinate of the $(n_{k+1}$ -dimensional) vector $w(k+1)x_i(k)$. This equation describes the *backward dynamics*.

(iii) The third subcondition

Unfortunately, the third subcondition does not give a direct method for computing the matrices of weights $w(k)$, $k = 1, \dots, m$. Therefore, to estimate the weights, one uses steepest gradient descent:

$$w(k) \leftarrow w(k) - \gamma(\cdot) \frac{\partial L(W, X, B)}{\partial w(k)}, \quad k = 1, \dots, m.$$

In explicit form this equation is

$$w(k) \leftarrow w(k) - \gamma(\cdot) \sum_{i=1}^{\ell} b_i(k) \nabla S \{w(k)x_i(k-1)\} w(k)x_i^T(k-1),$$

$$k = 1, 2, \dots, m.$$

This equation describes the rule for weight update.

5.3.2 The Back-Propagation Algorithm

Therefore, the back-propagation algorithm contains three elements:

(i) *Forward pass:*

$$x_i(k) = S\{w(k)x_i(k-1)\}, \quad i = 1, \dots, \ell, \quad k = 1, \dots, m,$$

with the boundary conditions

$$x_i(0) = x_i, \quad i = 1, \dots, \ell.$$

(II) *Backward pass:*

$$b_i(k) = w^T(k+1) \nabla S\{w(k+1)x_i(k)\} b_i(k+1),$$

$$i = 1, \dots, \ell, \quad k = 1, \dots, m-1,$$

with the boundary conditions

$$b_i(m) = 2(y_i - x_i(m)), \quad i = 1, \dots, \ell.$$

(iii) *Weight update for weight matrices $w(k)$, $k = 1, 2, \dots, m$:*

$$w(k) \leftarrow w(k) - \gamma(\cdot) \sum_{i=1}^{\ell} b_i(k) \nabla S\{w(k)x_i(k-1)\} w(k)x_i^T(k-1).$$

Using the back-propagation technique one can achieve a local minimum for the empirical risk functional.

5.3.3 Neural Networks for the Regression Estimation Problem

To adapt neural networks for solving the regression estimation problem, it is sufficient to use in the last layer a linear function instead of a sigmoid one. This implies only the following changes in the equations described above:

$$x_i(m) = w(m)x_i(m-1),$$

$$\nabla S\{w(m), x_i(m-1)\} = 1, \quad i = 1, \dots, \ell.$$

5.3.4 Remarks on the Back-Propagation Method

The main problems with the neural net approach are:

(i) The empirical risk functional has many local minima. Standard optimization procedures guarantee convergence to one of them. The quality of the obtained solution depends on many factors, in particular on the initialization of weight matrices $w(k)$, $k = 1, \dots, m$.

The choice of initialization parameters to achieve a “small” local minimum is based on heuristics.

- (ii) The convergence of the gradient-based method is rather slow. There are several heuristics to speedup the rate of convergence.
- (iii) The sigmoid function has a scaling factor that affects the quality of the approximation. The choice of the scaling factor is a trade-off between the quality of approximation and the rate of convergence. There are empirical recommendations for choosing the scaling factor.

Therefore, neural networks are not well-controlled learning machines. Nevertheless, in many practical applications, neural networks demonstrate good results.

5.4 THE OPTIMAL SEPARATING HYPERPLANE

Below we consider a new type of universal learning machine that implements the second strategy: Keep the value of the empirical risk fixed and minimize the confidence interval.

As in the case of neural networks, we start by considering linear decision rules (the separating hyperplanes). However, in contrast to previous considerations, we use a special type of hyperplane, the so-called optimal separating hyperplanes (Vapnik and Chervonenkis, 1974), (Vapnik, 1979). First we consider the optimal separating hyperplane for the case where the training data are linearly separable. Then, in Section 5.5.1 we generalize the idea of optimal separating hyperplanes to the case of nonseparable data. Using a technique for constructing optimal hyperplanes, we describe a new type of universal learning machine, the support vector machine. Finally, we construct the support vector machine for solving regression estimation problems.

5.4.1 The Optimal Hyperplane

Suppose the training data

$$(x_1, y_1), \dots, (x_\ell, y_\ell), \quad x \in R^n, \quad y \in \{+1, -1\},$$

can be separated by a hyperplane

$$(w \cdot x) - b = 0. \quad (5.7)$$

We say that this set of vectors is separated by the *optimal hyperplane (or the maximal margin hyperplane)* if it is separated without error and the distance between the closest vector to the hyperplane is maximal (Fig. 5.2).

To describe the separating hyperplane let us use the following form:

$$(w \cdot x_i) - b \geq 1 \quad \text{if } y_i = 1,$$

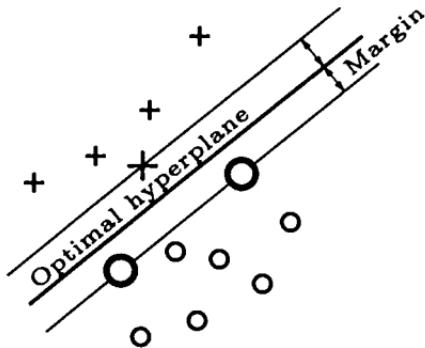


FIGURE 5.2. The optimal separating hyperplane is the one that separates the data with maximal margin.

$$(w \cdot x_i) - b \leq -1 \quad \text{if } y_i = -1.$$

In the following we use a compact notation for these inequalities:

$$y_i[(w \cdot x_i) - b] \geq 1, \quad i = 1, \dots, \ell. \quad (5.8)$$

It is easy to check that the optimal hyperplane is the one that satisfies the conditions (5.8) and minimizes

$$\Phi(w) = \|w\|^2. \quad (5.9)$$

(The minimization is taken with respect to both the vector w and the scalar b .)

5.4.2 Δ -Margin Separating Hyperplanes

We call a hyperplane

$$(w^* \cdot x) - b = 0, \quad |w^*| = 1$$

a Δ -margin separating hyperplane if it classifies vectors x as follows:

$$y = \begin{cases} 1 & \text{if } (w^* \cdot x) - b \geq \Delta, \\ -1 & \text{if } (w^* \cdot x) - b \leq -\Delta. \end{cases}$$

It is easy to check that the optimal hyperplane defined in canonical form (5.8) is the Δ -margin separating hyperplane with $\Delta = 1/\|w^*\|$. The following theorem is true.

Theorem 5.1. *Let vectors $x \in X$ belong to a sphere of radius R . Then the set of Δ -margin separating hyperplanes has VC dimension h bounded*

by the inequality

$$h \leq \min \left(\left[\frac{R^2}{\Delta^2} \right], n \right) + 1.$$

In Section 3.5 we stated that the VC dimension of the set of separating hyperplanes is equal to $n + 1$, where n is the dimension of the space. However, the VC dimension of the Δ -margin separating hyperplanes can be less.¹

Corollary. *With probability $1 - \eta$ one can assert that the probability that a test example will not be separated correctly by the Δ -margin hyperplane has the bound*

$$P_{\text{error}} \leq \frac{m}{\ell} + \frac{\mathcal{E}}{2} \left(1 + \sqrt{1 + \frac{4m}{\ell \mathcal{E}}} \right),$$

where

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

m is the number of training examples that are not separated correctly by this Δ -margin hyperplane, and h is the bound of the VC dimension given in Theorem 5.1.

On the basis of this theorem one can construct the SRM method where in order to obtain a good generalization one chooses the appropriate value of Δ .

5.5 CONSTRUCTING THE OPTIMAL HYPERPLANE

To construct the optimal hyperplane one has to separate the vectors x_i of the training set

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

belonging to two different classes $y \in \{-1, 1\}$ using the hyperplane with the smallest norm of coefficients.

To find this hyperplane one has to solve the following quadratic programming problem: Minimize the functional

$$\Phi(w) = \frac{1}{2}(w \cdot w) \tag{5.10}$$

under the constraints of inequality type

$$y_i[(x_i \cdot w) - b] \geq 1, \quad i = 1, 2, \dots, \ell. \tag{5.11}$$

¹In Section 5.7 we describe a separating hyperplane in 10^{13} -dimensional space with relatively small estimate of the VC dimension ($\approx 10^3$).

The solution to this optimization problem is given by the saddle point of the Lagrange functional (Lagrangian):

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{\ell} \alpha_i \{[(x_i \cdot w) - b]y_i - 1\}, \quad (5.12)$$

where the α_i are Lagrange multipliers. The Lagrangian has to be minimized with respect to w and b and maximized with respect to $\alpha_i > 0$.

At the saddle point, the solutions w_0 , b_0 , and α^0 should satisfy the conditions

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial b} = 0,$$

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial w} = 0.$$

Rewriting these equations in explicit form, one obtains the following properties of the optimal hyperplane:

- (i) The coefficients α_i^0 for the optimal hyperplane should satisfy the constraints

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i = 0, \quad \alpha_i^0 \geq 0, \quad i = 1, \dots, \ell \quad (5.13)$$

(first equation).

- (ii) The Optimal hyperplane (vector w_0) is a linear combination of the vectors of the training set.

$$w_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i, \quad \alpha_i^0 \geq 0, \quad i = 1, \dots, \ell \quad (5.14)$$

(second equation).

- (iii) Moreover, only the so-called *support vectors* can have nonzero coefficients α_i^0 in the expansion of w_0 . The support vectors are the vectors for which in inequality (5.11) equality is achieved. Therefore, we obtain

$$w_0 = \sum_{\text{support vectors}} y_i \alpha_i^0 x_i, \quad \alpha_i^0 \geq 0. \quad (5.15)$$

This fact follows from the classical Kühn–Tucker theorem, according to which necessary and sufficient conditions for the optimal hyperplane are that the separating hyperplane satisfy the conditions

$$\alpha_i^0 \{[(x_i \cdot w_0) - b_0]y_i - 1\} = 0, \quad i = 1, \dots, \ell. \quad (5.16)$$

Putting the expression for w_0 into the Lagrangian and taking into account the Kühn–Tucker conditions, one obtains the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j}^{\ell} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j). \quad (5.17)$$

It remains to maximize this functional in the nonnegative quadrant

$$\alpha_i \geq 0, \quad i = 1, \dots, \ell, \quad (5.18)$$

under the constraint

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0. \quad (5.19)$$

According to (5.15), the Lagrange multipliers and support vectors determine the optimal hyperplane. Thus, to construct the optimal hyperplane one has to solve a simple quadratic programming problem: Maximize the quadratic form (5.17) under constraints² (5.18) and (5.19).

Let $\alpha_0 = (\alpha_1^0, \dots, \alpha_\ell^0)$ be a solution to this quadratic optimization problem. Then the norm of the vector w_0 corresponding to the Optimal hyperplane equals

$$|w_0|^2 = 2W(\alpha_0) = \sum_{\text{support vectors}} \alpha_i^0 \alpha_j^0 (x_i \cdot x_j) y_i y_j.$$

The separating rule, based on the optimal hyperplane, is the following indicator function

$$f(x) = \text{sign} \left(\sum_{\text{support vectors}} y_i \alpha_i^0 (x_i \cdot x) - b_0 \right), \quad (5.20)$$

where x_i are the support vectors, α_i^0 are the corresponding Lagrange coefficients, and b_0 is the constant (threshold)

$$b_0 = \frac{1}{2} [(w_0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))],$$

where we denote by $x^*(1)$ some (any) support vector belonging to the first class and we denote by $x^*(-1)$ a support vector belonging to the second class (Vapnik and Chervonenkis, 1974), (Vapnik, 1979).

²This quadratic programming problem is simple because it has simple constraints. For the solution of this problem, one can use special methods that are fast and applicable for the case with a large number of support vectors ($\approx 10^4$ support vectors) (More and Toraldo, 1991). Note that in the training data the support vectors constitute only a small part of the training vectors (in our experiments 3% to 5%).

5.5.1 Generalization for the Nonseparable Case

To construct the optimal-type hyperplane in the case when the data are linearly nonseparable, we introduce nonnegative variables $\xi_i \geq 0$ and a function

$$F_\sigma(\xi) = \sum_{i=1}^{\ell} \xi_i^\sigma$$

with parameter $\sigma > 0$.

Let us minimize the functional $F_\sigma(\xi)$ subject to constraints

$$y_i((w \cdot x_i) - b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, \ell, \quad (5.21)$$

and one more constraint,

$$(w \cdot w) \leq \Delta^{-2}. \quad (5.22)$$

For sufficiently small $\sigma > 0$ the solution to this optimization problem defines a hyperplane that minimizes the number of training errors under the condition that the parameters of this hyperplane belong to the subset (5.22) (to the element of the structure

$$S_n = \{(w \cdot x) - b : (w \cdot w) \leq \Delta^{-2}\}$$

determined by the constant $c_n = 1/\Delta^{-2}$.

For computational reasons, however, we consider the case $\sigma = 1$. This case corresponds to the smallest $\sigma > 0$ that is still computationally simple. We call this hyperplane the Δ -margin separating hyperplane.

1. Constructing Δ -margin separating hyperplanes. One can show (using the technique described above) that the Δ -margin hyperplane is determined by the vector

$$w = \frac{1}{C^*} \sum_{i=1}^{\ell} \alpha_i y_i x_i,$$

where the parameters α_i , $i = 1, \dots, \ell$, and C^* are the solutions to the following convex optimization problem:

Maximize the functional

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

subject to constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \quad C^* \geq 0,$$

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

2. Constructing soft-margin separating hyperplanes. To simplify computations one can introduce the following (slightly modified) concept of the soft-margin optimal hyperplane (Cortes and Vapnik, 1995). The soft-margin hyperplane (also called the generalized optimal hyperplane) is determined by the vector w that minimizes the functional

$$\Phi(w, \xi) = \frac{1}{2}(w \cdot w) + C \left(\sum_{i=1}^{\ell} \xi_i \right)$$

(here C is a *given* value) subject to constraint (5.21).

The technique of solution of this quadratic optimization problem is almost equivalent to the technique used in the separable case: To find the coefficients of the generalized optimal hyperplane

$$w = \sum_{i=1}^{\ell} \alpha_i y_i x_i,$$

one has to find the parameters α_i , $i = 1, \dots, \ell$, that maximize the same quadratic form as in the separable case

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

under slightly different constraints:

$$0 \leq \alpha_i \leq C, \quad i = 1, \dots, \ell,$$

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

As in the separable case, only some of the coefficients α_i , $i = 1, \dots, \ell$, differ from zero. They determine the support vectors.

Note that if the coefficient C in the functional $\Phi(w, \xi)$ is equal to the optimal value of the parameter C^* for minimization of the functional $F_1(\xi)$,

$$C = C^*,$$

then the solutions to both optimization problems (defined by the functional $F_1(\xi)$ and by the functional $\Phi(w, \xi)$) coincide.

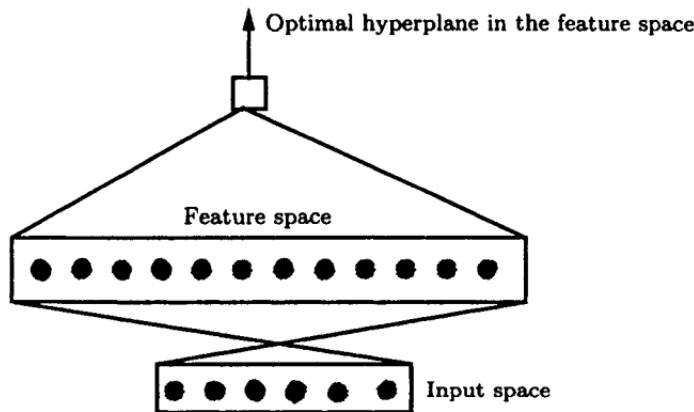


FIGURE 5.3. The SV machine maps the input space into a high-dimensional feature space and then constructs an Optimal hyperplane in the feature space.

5.6 SUPPORT VECTOR (SV) MACHINES

The support vector (SV) machine implements the following idea: It maps the input vectors x into a high-dimensional feature space Z through some nonlinear mapping, chosen *a priori*. In this space, an optimal separating hyperplane is constructed (Fig. 5.3).

Example. To construct a decision surface corresponding to a polynomial of degree two, one can create a feature space Z that has $N = \frac{n(n+3)}{2}$ coordinates of the form

$$z^1 = x^1, \dots, z^n = x^n, \quad n \text{ coordinates},$$

$$z^{n+1} = (x^1)^2, \dots, z^{2n} = (x^n)^2, \quad n \text{ coordinates},$$

$$z^{2n+1} = x^1 x^2, \dots, z^N = x^n x^{n-1}, \quad \frac{n(n-1)}{2} \text{ coordinates},$$

where $x = (x^1, \dots, x^n)$. The separating hyperplane constructed in this space is a second degree polynomial in the input space. To construct polynomials of degree $d \ll n$ in n -dimensional space one needs more than $\approx (n/d)^d$ features.

Two problems arise in the above approach: one conceptual and one technical.

- (i) *How does one find a separating hyperplane that will generalize well?*
(The conceptual problem)

The dimensionality of the feature space will be huge, and a hyperplane that separates the training data will not necessarily generalize well.³

- (ii) *How does one treat computationally such high-dimensional spaces?*
 (The technical problem)

To construct a polynomial of degree 4 or 5 in a 200-dimensional space it is necessary to construct hyperplanes in a billion-dimensional feature space. How can this “curse of dimensionality” be overcome?

5.6.1 Generalization in High-Dimensional Space

The conceptual part of this problem can be solved by constructing both the Δ -margin separating hyperplane and soft margin separating hyperplane.

According to Theorem 5.1 the VC dimension of the set of Δ -margin separating hyperplanes with large Δ is small. Therefore, according to the corollary to Theorem 5.1 the generalization ability of the constructed hyperplane is high.

For the maximal margin hyperplane the following theorem holds true.

Theorem 5.2. *If training sets containing ℓ examples are separated by the maximal margin hyperplanes, then the expectation (over training sets) of the probability of test error is bounded by the expectation of the minimum of three values: the ratio m/ℓ , where m is the number of support vectors, the ratio $[R^2|w|^2]/\ell$, where R is the radius of the sphere containing the data and $|w|^{-2}$ is the value of the margin, and the ratio n/ℓ , where n is the dimensionality of the input space:*

$$EP_{\text{error}} \leq E \min \left(\frac{m}{\ell}, \frac{[R^2|w|^2]}{\ell}, \frac{n}{\ell} \right). \quad (5.23)$$

Equation (5.23) gives three reasons why optimal hyperplanes can generalize:

1. Because the expectation of the data compression is large⁴.

³Recall Fisher’s concern about the small amount of data for constructing a quadratic discriminant function in classical discriminant analysis (Section 1.9).

⁴One can compare the result of this theorem to the result of analysis of the following compression scheme. To construct the optimal separating hyperplane one needs only to specify among the training data the support vectors and their classification. This requires $\approx \lceil \lg_2 m \rceil$ bits to specify the number m of support vectors, $\lceil \lg_2 C_\ell^m \rceil$ bits to specify the support vectors, and $\lceil \lg_2 C_{m_1}^{m_1} \rceil$ bits to specify m_1 representatives of the first class among the support vectors. Therefore, for $m \ll \ell$ and $m_1 \approx m/2$ the compression coefficient is

$$K \approx \frac{m(\lg_2 \ell/m + 1)}{\ell}.$$

According to Theorem 4.3 the probability of error for the general compression

2. Because the expectation of the margin is large.
3. Because the input space is small.

Classical approaches ignore the first two reasons for generalization and rely on the third one. In support vector machines we ignore the dimensionality factor and rely on the first two factors.

5.6.2 Convolution of the Inner Product

However, even if the optimal hyperplane generalizes well and can theoretically be found, the technical problem of how to treat the high-dimensional feature space remains.

In 1992 it was observed (Boser, Guyon, and Vapnik, 1992) that for constructing the optimal separating hyperplane in the feature space Z one does not need to consider the feature space in *explicit form*. One has only to be able to calculate the inner products between support vectors and the vectors of the feature space ((5.17) and (5.20)).

Consider a general expression for the inner product in Hilbert space⁵

$$(z_i \cdot z) = K(x, x_i),$$

where z is the image in feature space of the vector x in input space.

According to Hilbert–Schmidt theory, $K(x, x_i)$ can be any symmetric function satisfying the following general conditions (Courant and Hilbert, 1953):

Theorem 5.3. (Mercer) *To guarantee that the symmetric function $K(u, v)$ from L_2 has an expansion*

$$K(u, v) = \sum_{k=1}^{\infty} a_k \psi_k(u) \psi_k(v) \quad (5.24)$$

with positive coefficients $a_k > 0$ (i.e., $K(u, v)$ describes an inner product in some feature space), it is necessary and sufficient that the condition

$$\int \int K(u, v) g(u) g(v) du dv > 0$$

scheme is proportional to K . From Theorem 5.2 it follows that $EP_{\text{error}} \leq Em/\ell$.

Therefore, the bound obtained for the SV machine is much better than the bound obtained for the general compression scheme even if the random value m in (5.23) is always the smallest one.

⁵This idea was used in 1964 by Aizerman, Braverman, and Rozonoer in their analysis of the convergence properties of the method of potential functions (Aizerman, Braverman, and Rozonoer, 1964, 1970). It happened at the same time (1965) as the method of the optimal hyperplane was developed (Vapnik and Chervonenkis 1965). However, combining these two ideas, which lead to the SV machines, was done only in 1992.

be valid for all $g \neq 0$ for which

$$\int g^2(u)du < \infty.$$

5.6.3 Constructing SV Machines

The convolution of the inner product allows the construction of decision functions that are nonlinear in the input space,

$$f(x) = \text{sign} \left(\sum_{\text{support vectors}} y_i \alpha_i K(x_i, x) - b \right), \quad (5.25)$$

and that are equivalent to linear decision functions in the high-dimensional feature space $\psi_1(x), \dots, \psi_N(x)$ ($K(x_i, x)$ is a convolution of the inner product for this feature space).

To find the coefficients α_i in the separable case (analogously in the non-separable case) it is sufficient to find the maximum of the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (5.26)$$

subject to the constraints

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \dots, \ell. \quad (5.27)$$

This functional coincides with the functional for finding the optimal hyperplane, except for the form of the inner products: Instead of inner products $(x_i \cdot x_j)$ in (5.17), we now use the convolution of the inner products $K(x_i, x_j)$.

The learning machines that construct decision functions of the type (5.25) are called *support vector (SV) Machines*. (With this name we stress the idea of expanding the solution on support vectors. In SV machines the complexity of the construction depends on the number of support vectors rather than on the dimensionality of the feature space.) The scheme of SV machines is shown in Figure 5.4.

5.6.4 Examples of SV Machines

Using different functions for convolution of the inner products $K(x, x_i)$, one can construct learning machines with different types of nonlinear decision surfaces in input space. Below, we consider three types of learning machines:

- (i) polynomial learning machines,

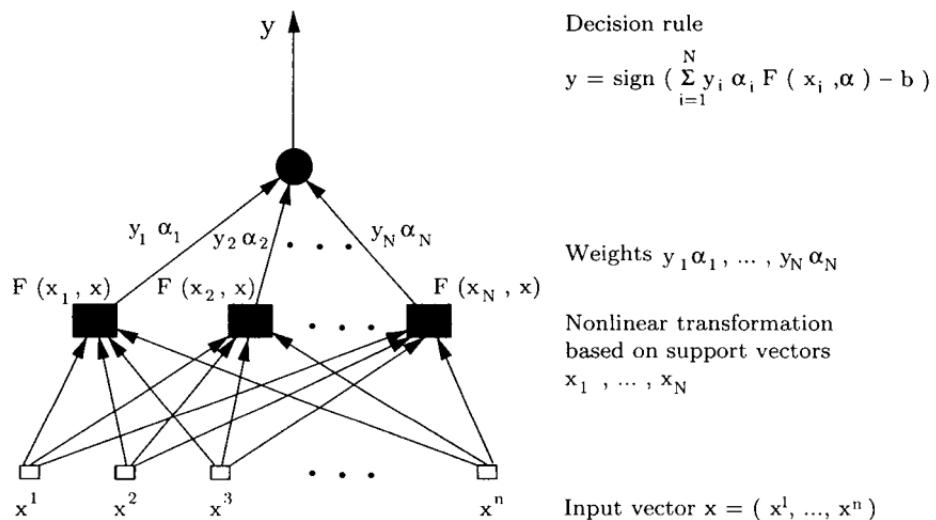


FIGURE 5.4. The two-layer SV machine is a compact realization of an optimal hyperplane in the high-dimensional feature space Z .

- (ii) radial basis functions machines, and
- (iii) two layer neural networks.

For simplicity we consider here the regime where the training vectors are separated without error.

Note that the support vector machines implement the SRM principle. Indeed, let

$$\Psi(x) = (\psi_1(x), \dots, \psi_N(x))$$

be a feature space and $w = (w_1, \dots, w_N)$ be a vector of weights determining a hyperplane in this space. Consider a structure on the set of hyperplanes with elements S_k containing the functions satisfying the conditions

$$R^2|w|^2 \leq k,$$

where R is the radius of the smallest sphere that contains the vectors $\Psi(x)$, and $|w|$ is the norm of the weights (we use canonical hyperplanes in feature space with respect to the vectors $z = \Psi(x_i)$, where x_i are the elements of the training data).

According to Theorem 5.1 (now applied in the feature space), k gives an estimate of the VC dimension of the set of functions S_k .

The SV machine separates without error the training data

$$y_i [(\Psi(x_i) \cdot w) - b] \geq 1, \quad y_i = \{+1, -1\}, \quad i = 1, 2, \dots, \ell,$$

and has minimal norm $|w|$.

In other words, the SV machine separates the training data using functions from the element S_k with the smallest estimate of the VC dimension.

Recall that in the feature space the equality

$$|w_0|^2 = \sum_i^{\ell} \alpha_i^0 \alpha_j^0 K(x_i, x_j) y_i y_j = \sum_i^{\ell} \alpha_i^0 \quad (5.28)$$

holds true. To control the generalization ability of the machine (to minimize the probability of test errors) one has to construct the separating hyperplane that minimizes the functional

$$\Phi(R, w_0, \ell) = \frac{R^2|w_0|^2}{\ell}. \quad (5.29)$$

With probability $1 - \eta$ the hyperplane that separates data without error has the following bound on the test error

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

where h is the VC dimension of the set of hyperplanes. We approximate the VC dimension h of the maximal margin hyperplanes by $h_{\text{est}} = R^2|w_0|^2$. To estimate this functional it is sufficient to estimate $|w_0|^2$ (say by expression (5.28)) and estimate R^2 by finding

$$R^2 = R^2(K) = \min_a \max_{x_i} [K(x_i, x_i) + K(a, a) - 2K(x_i, a)]. \quad (5.30)$$

Polynomial learning machine

To construct polynomial decision rules of degree d , one can use the following function for convolution of the inner product:

$$K(x, x_i) = [(x \cdot x_i) + 1]^d. \quad (5.31)$$

This symmetric function satisfies the conditions of Theorem 5.3, and therefore it describes a convolution of the inner product in the feature space that contains all products $x_i \cdot x_j \cdot x_k$ up to degree d . Using the technique described, one constructs a decision function of the form

$$f(x, \alpha) = \text{sign} \left(\sum_{\text{support vectors}} y_i \alpha_i [(x_i \cdot x) + 1]^d - b \right),$$

which is a factorization of d -dimensional polynomials in n -dimensional input space.

In spite of the very high dimensionality of the feature space (polynomials of degree d in n -dimensional input space have $O(n^d)$ free parameters) the estimate of the VC dimension of the subset of polynomials that solve real-life problems can be low.

As described above, to estimate the VC dimension of the element of the structure from which the decision function is chosen, one has only to estimate the radius R of the smallest sphere that contains the training data, and the norm of weights in feature space (Theorem 5.1).

Note that both the radius $R = R(d)$ and the norm of weights in the feature space depend on the degree of the polynomial.

This gives the opportunity to choose the best degree of the polynomial for the given data.

To make a *local polynomial* approximation in the neighborhood of a point of interest x_0 , let us consider the hard-threshold neighborhood function (4.16). According to the theory of local algorithms, one chooses a ball with radius R_β around point x_0 in which ℓ_β elements of the training set fall, and then using only these training data, one constructs the decision function that minimizes the probability of errors in the chosen neighborhood. The solution to this problem is a radius R_β that minimizes the functional

$$\Phi(R_\beta, w_0, \ell_\beta) = \frac{R_\beta^2 |w_0|^2}{\ell_\beta} \quad (5.32)$$

(the parameter $|w_0|$ depends on the chosen radius as well). This functional describes a trade-off between the chosen radius R_β , the value of the minimum of the norm $|w_0|$, and the number of training vectors ℓ_β that fall into radius R_β .

Radial basis function machines

Classical radial basis function (RBF) machines use the following set of decision rules:

$$f(x) = \text{sign} \left(\sum_{i=1}^N a_i K_\gamma(|x - x_i|) - b \right), \quad (5.33)$$

where $K_\gamma(|x - x_i|)$ depends on the distance $|x - x_i|$ between two vectors. For the theory of RBF machines see (Micchelli, 1986), (Powell, 1992).

The function $K_\gamma(|x - x_i|)$ is for any fixed γ a nonnegative monotonic function; it tends to zero as z goes to infinity. The most popular function of this type is

$$K_\gamma(|x - x_i|) = \exp\{-\gamma|x - x_i|^2\}. \quad (5.34)$$

To construct the decision rule (5.33) one has to estimate

- (i) The value of the parameter γ ,
- (ii) the number N of the centers x_i ,
- (iii) the vectors x_i , describing the centers,
- (iv) the value of the parameters a_i .

In the classical RBF method the first three steps (determining the parameters γ , N , and vectors (centers) x_i , $i = 1, \dots, N$) are based on heuristics, and only the fourth step (after finding these parameters) is determined by minimizing the empirical risk functional.

The radial function can be chosen as a function for the convolution of the inner product for an SV machine. In this case, the SV machine will construct a function from the set (5.33). One can show (Aizerman, Braverman, and Rozonoer, 1964, 1970) that radial functions (5.34) satisfy the condition of Theorem 5.3.

In contrast to classical RBF methods, in the SV technique all four types of parameters are chosen to minimize the bound on the probability of test error by controlling the parameters R, w_0 in the functional (5.29). By minimizing the functional (5.29) one determines

- (i) N , the number of support vectors,
- (ii) x_i , (the pre-images of) support vectors;
- (iii) $a_i = \alpha_i y_i$, the coefficients of expansion, and

- (iv) γ , the width parameter of the kernel function.

Two-layer neural networks

Finally, one can define two-layer neural networks by choosing kernels:

$$K(x, x_i) = S[v(x \cdot x_i) + c],$$

where $S(u)$ is a sigmoid function. In contrast to kernels for polynomial machines or for radial basis function machines that always satisfy Mercer conditions, the sigmoid kernel $\tanh(vu + c)$, $|u| \leq 1$, satisfies Mercer conditions only for some values of the parameters v , c . For these values of the parameters one can construct SV machines implementing the rules

$$f(x, \alpha) = \text{sign} \left\{ \sum_{i=1}^N \alpha_i S(v(x \cdot x_i) + c) + b \right\}.$$

Using the technique described above, the following are found automatically:

- (i) the architecture of the two layer machine, determining the number N of hidden units (the number of support vectors),
- (ii) the vectors of the weights $w_i = x_i$ in the neurons of the first (hidden) layer (the support vectors), and
- (iii) the vector of weights for the second layer (values of α).

5.7 EXPERIMENTS WITH SV MACHINES

In the following we will present two types of experiments constructing the decision rules in the pattern recognition problem:⁶

- (i) Experiments in the plane with artificial data that can be visualized, and
- (ii) experiments with real-life data.

5.7.1 Example in the Plane

To demonstrate the SV technique we first give an artificial example (Fig.

⁶The experiments were conducted in the Adaptive System Research Department, AT&T Bell Laboratories.

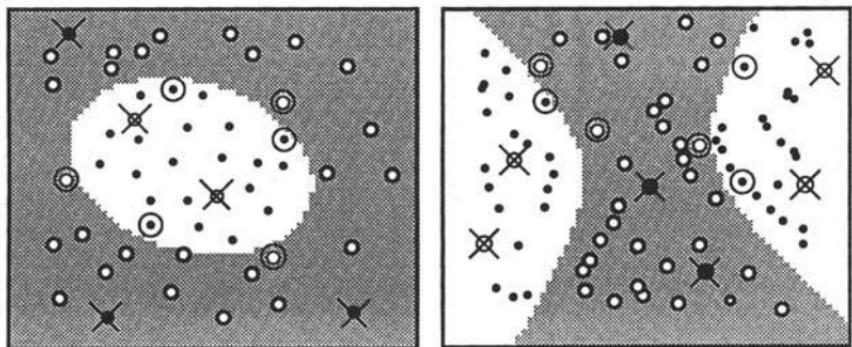


FIGURE 5.5. Two classes of vectors are represented in the picture by black and white balls. The decision boundaries were constructed using an inner product of polynomial type with $d = 2$. In the pictures the examples cannot be separated without errors; the errors are indicated by crosses and the support vectors by double circles.

5.5).

The two classes of vectors are represented in the picture by black and white balls. The decision boundaries were constructed using an inner product of polynomial type with $d = 2$. In the pictures the examples cannot be separated without errors; the errors are indicated by crosses and the support vectors by double circles.

Notice that in both examples the number of support vectors is small relative to the number of training data and that the number of training errors is minimal for polynomials of degree two.

5.7.2 Handwritten Digit Recognition

Since the first experiments of Rosenblatt, the interest in the problem of learning to recognize handwritten digits has remained strong. In the following we describe results of experiments on learning the recognition of handwritten digits using different SV machines. We also compare these results to results obtained by other classifiers. In these experiments, the U.S. Postal Service database (LeCun et al., 1990) was used. It contains 7,300 training patterns and 2,000 test patterns collected from real-life zip codes. The resolution of the database is 16×16 pixels; therefore, the dimensionality of the input space is 256. Figure 5.6 gives examples from this data base.

2 6 0 1 4 9 6 3 5 7 1 4 6 3 7 1 0 3 7 1 1 4 4 9 7
 8888 8881 8882 8883 8884 8885 8886 8887 8888 8889 8818 8811 8812 8813 8814 8815 8816 8817 8818 8819 8820 8821 8822 8823 8824
 1 1 0 5 7 1 1 1 2 9 9 8 1 1 0 2 8 6 0 0 2 8 7 0
 8825 8826 8827 8828 8829 8830 8831 8832 8833 8834 8835 8836 8837 8838 8839 8840 8841 8842 8843 8844 8845 8846 8847 8848 8849
 3 3 0 1 0 3 3 0 1 0 2 9 0 6 0 2 8 4 0 0 2 9 0 1 2
 8850 8851 8852 8853 8854 8855 8856 8857 8858 8859 8860 8861 8862 8863 8864 8865 8866 8867 8868 8869 8870 8871 8872 8873 8874
 9 4 0 5 2 9 0 6 7 2 9 8 0 1 2 9 6 5 0 2 9 9 0 5 5
 8875 8876 8877 8878 8879 8880 8881 8882 8883 8884 8885 8886 8887 8888 8889 8890 8891 8892 8893 8894 8895 8896 8897 8898 8899
 5 1 0 1 2 9 2 0 1 8 0 3 2 2 7 0 1 2 4 4 3 1 0 6 4
 8900 8901 8902 8903 8904 8905 8906 8907 8908 8909 8910 8911 8912 8913 8914 8915 8916 8917 8918 8919 8920 8921 8922 8923 8924
 1 1 6 1 1 7 6 0 5 7 1 8 8 6 0 0 1 5 8 7 0 1 8 9 9
 8925 8926 8927 8928 8929 8930 8931 8932 8933 8934 8935 8936 8937 8938 8939 8940 8941 8942 8943 8944 8945 8946 8947 8948 8949
 1 1 5 7 5 8 7 2 1 2 5 7 0 6 8 3 2 7 4 9 9 8 1 6
 8950 8951 8952 8953 8954 8955 8956 8957 8958 8959 8960 8961 8962 8963 8964 8965 8966 8967 8968 8969 8970 8971 8972 8973 8974
 9 9 5 0 5 1 2 0 0 1 5 3 6 2 7 2 2 0 3 2 4 2 3 7 2
 8975 8976 8977 8978 8979 8980 8981 8982 8983 8984 8985 8986 8987 8988 8989 8990 8991 8992 8993 8994 8995 8996 8997 8998 8999
 3 5 0 7 2 7 1 2 7 2 3 1 5 3 9 3 0 5 3 8 8 0 3 1 1
 9000 9001 9002 9003 9004 9005 9006 9007 9008 9009 9010 9011 9012 9013 9014 9015 9016 9017 9018 9019 9020 9021 9022 9023 9024
 1 3 7 1 9 1 4 1 1 9 1 2 9 1 9 2 5 1 1 9 1 7 0 1 4
 9025 9026 9027 9028 9029 9030 9031 9032 9033 9034 9035 9036 9037 9038 9039 9040 9041 9042 9043 9044 9045 9046 9047 9048 9049
 1 0 1 1 9 1 3 4 8 5 7 2 6 8 0 3 2 2 6 4 1 4 1 8 6
 9050 9051 9052 9053 9054 9055 9056 9057 9058 9059 9060 9061 9062 9063 9064 9065 9066 9067 9068 9069 9070 9071 9072 9073 9074
 6 3 5 9 7 2 0 2 9 9 2 9 9 7 2 2 5 1 0 0 4 6 7 0 1
 9075 9076 9077 9078 9079 9080 9081 9082 9083 9084 9085 9086 9087 9088 9089 9090 9091 9092 9093 9094 9095 9096 9097 9098 9099
 3 0 8 4 1 1 1 5 9 1 0 1 0 6 1 5 4 0 6 1 0 3 6 3 1
 9100 9101 9102 9103 9104 9105 9106 9107 9108 9109 9110 9111 9112 9113 9114 9115 9116 9117 9118 9119 9120 9121 9122 9123 9124
 1 0 6 4 1 1 1 0 3 0 4 7 5 2 6 2 0 0 9 9 7 9 9 6 6
 9125 9126 9127 9128 9129 9130 9131 9132 9133 9134 9135 9136 9137 9138 9139 9140 9141 9142 9143 9144 9145 9146 9147 9148 9149
 8 9 1 2 0 5 4 7 0 8 5 5 7 1 2 1 4 2 7 9 5 5 4 6 0
 9150 9151 9152 9153 9154 9155 9156 9157 9158 9159 9160 9161 9162 9163 9164 9165 9166 9167 9168 9169 9170 9171 9172 9173 9174
 1 0 1 8 2 3 0 1 8 7 1 1 2 9 9 1 0 8 9 9 7 0 9 8 4
 9175 9176 9177 9178 9179 9180 9181 9182 9183 9184 9185 9186 9187 9188 9189 9190 9191 9192 9193 9194 9195 9196 9197 9198 9199
 0 1 0 9 7 0 7 5 9 7 3 3 1 9 7 2 0 1 5 5 1 9 0 5 5
 9200 9201 9202 9203 9204 9205 9206 9207 9208 9209 9210 9211 9212 9213 9214 9215 9216 9217 9218 9219 9220 9221 9222 9223 9224
 1 0 7 5 5 1 8 2 5 5 1 8 2 8 1 4 3 5 8 0 9 0 9 6 3
 9225 9226 9227 9228 9229 9230 9231 9232 9233 9234 9235 9236 9237 9238 9239 9240 9241 9242 9243 9244 9245 9246 9247 9248 9249
 1 7 8 7 5 4 1 6 5 5 4 6 0 3 5 4 6 0 3 5 4 6 0 5 5
 9250 9251 9252 9253 9254 9255 9256 9257 9258 9259 9260 9261 9262 9263 9264 9265 9266 9267 9268 9269 9270 9271 9272 9273 9274
 1 8 2 5 5 1 0 8 5 0 3 0 4 7 5 2 0 4 3 9 4 0 1
 9275 9276 9277 9278 9279 9280 9281 9282 9283 9284 9285 9286 9287 9288 9289 9290 9291 9292 9293 9294 9295 9296 9297

FIGURE 5.6. Examples of patterns (with labels) from the U.S. Postal Service database.

Classifier	Raw error%
Human performance	2.5
Decision tree, C4.5	16.2
Best two-layer neural network	5.9
Five-layer network (LeNet 1)	5.1

TABLE 5.1. Human performance and performance of the various learning machines in solving the problem of digit recognition on U.S. Postal Service data.

Table 5.1 describes the performance of various classifiers, solving this problem⁷

For constructing the decision rules three types of SV machines were used:⁸

- (i) A polynomial machine with convolution function

$$K(x, x_i) = \left(\frac{(x \cdot x_i)}{256} \right)^d, \quad d = 1, \dots, 7.$$

- (ii) A radial basis function machine with convolution function

$$K(x, x_i) = \exp \left\{ -\frac{(x - x_i)^2}{256\sigma^2} \right\}.$$

- (iii) A two-layer neural network machine with convolution function

$$K(x, x_i) = \tanh \left(\frac{b(x \cdot x_i)}{256} - c \right).$$

All machines constructed ten classifiers, each one separating one class from the rest. The ten-class classification was done by choosing the class with the largest classifier output value.

The results of these experiments are given in Table 5.2. For different types of SV machines, Table 5.2 shows the best parameters for the machines (column 2), the average (over one classifier) of the number of support vectors, and the performance of the machine.

⁷The result of human performance was reported by J. Bromley and E. Säckinger; the result of C4.5 was obtained by C. Cortes; the result for the two-layer neural net was obtained by B. Schölkopf; the results for the special purpose neural network architecture with five layers (LeNet 1), was obtained by Y. LeCun *et al.*

⁸The results were obtained by C. Burges, C. Cortes, and B. Schölkopf.

Type of SV classifier	Parameters of classifier	Number of support vectors	Raw error
Polynomials	$d=3$	274	4.0
RBF classifiers	$\sigma^2 = 0.3$	291	4.1
Neural network	$b = 2, c = 1$	254	4.2

TABLE 5.2. Results of digit recognition experiments with various SV machines using the U.S. Postal Service database. The number of support vectors means the average per classifier.

	Poly	RBF	NN	Common
total # of sup.vect.	1677	1727	1611	1377
% of common sup. vect.	82	80	85	100

TABLE 5.3. Total number (in ten classifiers) of support vectors for various SV machines and percentage of common support vectors.

Note that for this problem, all types of SV machines demonstrate approximately the same performance. This performance is better than the performance of any other type of learning machine solving the digit recognition problem by constructing the entire decision rule on the basis of the U.S. Postal Service database.⁹

In these experiments one important singularity was observed: Different types of SV machines use approximately the same set of support vectors. The percentage of common support vectors for three different classifiers exceeded 80%.

Table 5.3 describes the total number of different support vectors for ten classifiers of different machines: polynomial machine (Poly), radial basis function machine (RBF), and Neural Network machine (NN). It shows also the number of common support vectors for all machines.

⁹Note that using the local approximation approach described in Section 4.5 (which does not construct the entire decision rule but approximates the decision rule of any point of interest) one can obtain a better result: 3.3% error rate (L. Bottou and V. Vapnik, 1992).

The best result for this database, 2.7, was obtained by P. Simard, Y. LeCun, and J. Denker without using any learning methods. They suggested a special method of elastic matching with 7200 templates using a smart concept of distance (so-called tangent distance) that takes into account invariance with respect to small translations, rotations, distortions, and so on (P. Simard, Y. LeCun, and J. Denker, 1993).

	Poly	RBF	NN
Poly	100	84	94
RBF	87	100	88
NN	91	82	100

TABLE 5.4. Percentage of common (total) support vectors for two SV machines.

Table 5.4 describes the percentage of support vectors of the classifier given in the columns contained in the support vectors of the classifier given in the rows.

This fact, if it holds true for a wide class of real-life problems, is very important.

5.7.3 Some Important Details

In this subsection we give some important details on solving the digit recognition problem using a polynomial SV machine.

The training data are not linearly separable. The total number of misclassifications on the training set for linear rules is equal to 340 ($\approx 5\%$ errors). For second degree polynomial classifiers the total number of misclassifications on the training set is down to four. These four mis-classified examples (with desired labels) are shown in Fig. 5.7. Starting with polynomials of degree three, the training data are separable.

Table 5.5 describes the results of experiments using decision polynomials (ten polynomials, one per classifier in one experiment) of various degrees. The number of support vectors shown in the table is a mean value per classifier.

Note that the number of support vectors increases slowly with the degree of the polynomials. The seventh degree polynomial has only 50% more support vectors than the third degree polynomial.¹⁰

¹⁰The relatively high number of support vectors for the linear separator is due to nonseparability: The number 282 includes both support vectors and misclas-

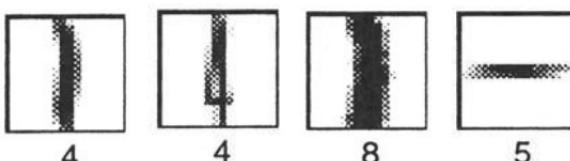


FIGURE 5.7. Labeled examples of training errors for the second degree polynomials.

degree of polynomial	dimensionality of feature space	support vectors	raw error
1	256	282	8.9
2	≈ 33000	227	4.7
3	$\approx 1 \times 10^6$	274	4.0
4	$\approx 1 \times 10^9$	321	4.2
5	$\approx 1 \times 10^{12}$	374	4.3
6	$\approx 1 \times 10^{14}$	377	4.5
7	$\approx 1 \times 10^{16}$	422	4.5

TABLE 5.5. Results of experiments with polynomials of different degrees.

The dimensionality of the feature space for a seventh degree polynomial is, however, 10^{10} times larger than the dimensionality of the feature space for a third degree polynomial classifier. Note that the performance does not change significantly with increasing dimensionality of the space — indicating no overfitting problems.

To choose the degree of the best polynomials for one specific classifier we estimate the VC dimension (using the estimate $[R^2 A^2]$) for all constructed polynomials (from degree two up to degree seven) and choose the one with the smallest estimate of the VC dimension. In this way we found the ten best classifiers (with different degrees of polynomials) for the ten two-class problems. These estimates are shown in Figure 5.8, where for all ten two-class decision rules the estimated VC dimension is plotted versus the degree of the polynomials. The question is this:

Do the polynomials with the smallest estimate of the VC dimension provide the best classifier?

To answer this question we constructed Table 5.6, which describes the performance of the classifiers for each degree of polynomial.

Each row describes one two-class classifier separating one *digit* (stated in the first column) from all the other digits.

The remaining columns contain:

deg.: the degree of the polynomial as chosen (from two up to seven) by the described procedure,

dim.: the dimensionality of the corresponding feature space, which is also the maximum possible VC dimension for linear classifiers in that space,

h_{est.}: the VC dimension estimate for the chosen polynomial (which is much smaller than the number of free parameters),

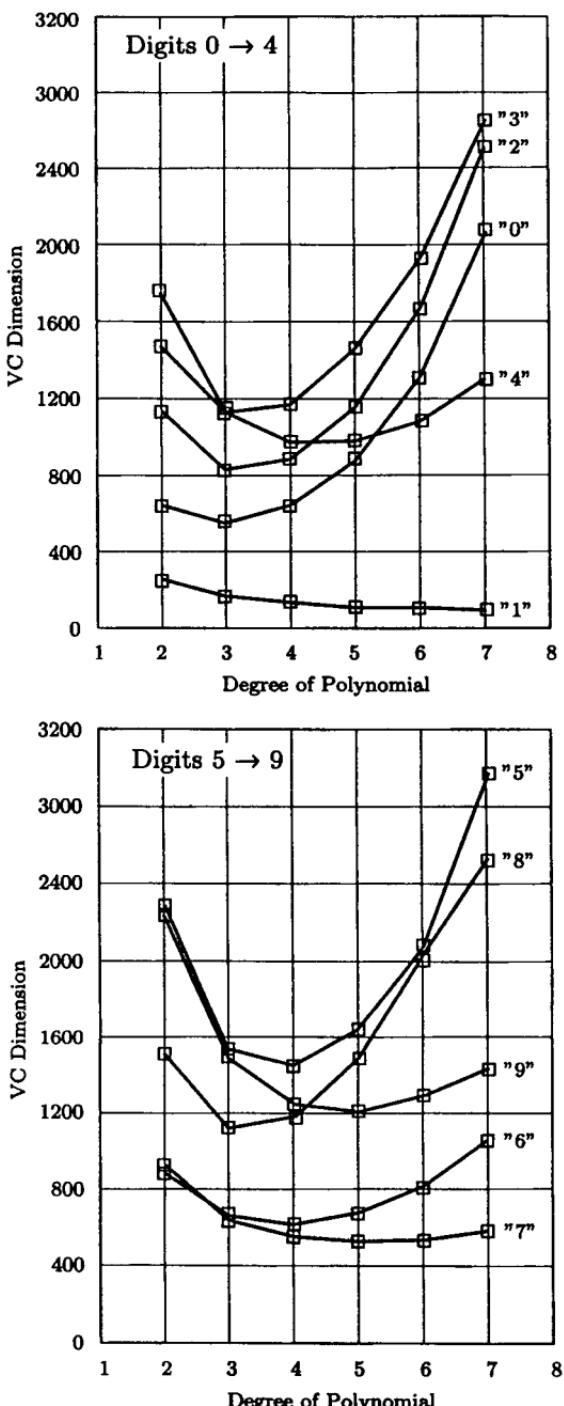


FIGURE 5.8. The estimate of the VC dimension of the best element of the structure (defined on the set of canonical hyperplanes in the corresponding feature space) versus the degree of the polynomial for various two-class digit recognition problems (denoted digit versus the rest).

Digit	Chosen classifier			Number of test errors						
	deg.	dim.	$h_{est.}$	1	2	3	4	5	6	7
0	3	$\sim 10^6$	530	36	14	11	11	11	12	17
1	7	$\sim 10^{16}$	101	17	15	14	11	10	10	10
2	3	$\sim 10^6$	842	53	32	28	26	28	27	32
3	3	$\sim 10^6$	1157	57	25	22	22	22	22	23
4	4	$\sim 10^9$	962	50	32	32	30	30	29	33
5	3	$\sim 10^6$	1090	37	20	22	24	24	26	28
6	4	$\sim 10^9$	626	23	12	12	15	17	17	19
7	5	$\sim 10^{12}$	530	25	15	12	10	11	13	14
8	4	$\sim 10^9$	1445	71	33	28	24	28	32	34
9	5	$\sim 10^{12}$	1226	51	18	15	11	11	12	15

TABLE 5.6. Experiments on choosing the best degree of polynomial.

Number of test errors: the number of test errors, using the constructed polynomial of corresponding degree; the boxes show the number of errors for the chosen polynomial.

Thus, Table 5.5 shows that for the SV polynomial machine there are no overfitting problems with increasing degree of polynomials, while Table 5.6 shows that even in situations where the difference between the best and the worst solutions is small (for polynomials starting from degree two up to degree seven), the theory gives a method for approximating the best solutions (finding the best degree of the polynomial).

Note also that Table 5.6 demonstrates that the problem is essentially nonlinear. The difference in the number of errors between the best polynomial classifier and the linear classifier can be as much as a factor of four (for digit 9).

5.8 REMARKS ON SV MACHINES

The quality of any learning machine is characterized by three main components:

- (i) *How universal is the learning machine?*
How rich is the set of functions that it can approximate?
- (ii) *How well can the machine generalize?*
How close is the upper bound on the error rate that this machine achieves (implementing a given set of functions and a given structure on this set of functions) to the smallest possible?

(iii) *How fast does the learning process for this machine converge?*

How many operations does it take to find the decision rule, using a given number of observations?

We address these in turn below.

(i) SV machines implement the sets of functions

$$f(x, \alpha, w) = \operatorname{sign} \left(\sum_{i=1}^N \alpha_i K(x, w_i) - b \right), \quad (5.35)$$

where N is any integer ($N < \ell$), α_i , $i = 1, \dots, N$, are any scalars, and w_i , $i = 1, \dots, N$, are any vectors. The kernel $K(x, w)$ can be any symmetric function satisfying the conditions of Theorem 5.3.

As was demonstrated, the best guaranteed risk for these sets of functions is achieved when the vectors of weights w_1, \dots, w_N are equal to some of the vectors x from the training data (support vectors).

Using the set of functions

$$\bar{f}(x, \alpha, w) = \sum_{\text{support vectors}} y_i \alpha_i K(x, w_i) - b$$

with convolutions of polynomial, radial basis function, or neural network type, one can approximate a continuous function to any degree of accuracy.

Note that for the SV machine one does not need to construct the architecture of the machine by choosing *a priori* the number N (as is necessary in classical neural networks or in classical radial basis function machines).

Furthermore, by changing only the function $K(x, w)$ in the SV machine one can change the type of learning machine (the type of approximating functions).

(ii) SV machines minimize the upper bound on the error rate for the structure given on a set of functions in a feature space. For the best solution it is *necessary* that the vectors w_i in (5.35) coincide with some vectors of the training data (support vectors).¹¹ SV machines find the functions from the set (5.35) that separate the training data and belong to the subset with the smallest bound of the VC dimension. (In the more general case they minimize the bound of the risk (5.1).)

(iii). Finally, to find the desired function, the SV machine has to maximize a nonpositive quadratic form in the nonnegative quadrant. This problem is a particular case of a special quadratic programming problem: to maximize a nonpositive quadratic form $Q(x)$ with bounded constraints

$$a_i \leq x^i \leq b_i, \quad i = 1, \dots, n,$$

¹¹This assertion is a direct corollary of the necessity of the Kühn–Tucker conditions for solving the quadratic optimization problem described in Section 5.4. The Kühn–Tucker conditions are necessary and sufficient for the solution of this problem.

where x^i , $i = 1, \dots, n$, are the coordinates of the vector x , and a_i , b_i are given constants. For this specific quadratic programming problem fast algorithms exist.

5.9 SVM AND LOGISTIC REGRESSION

5.9.1 Logistic Regression

Often it is important not only to construct a decision rule but also to find a function that for any given input vector x defines the probability $P\{y = 1|x\}$ that the vector x belongs to the first class. This problem is more general than the problem of constructing a decision rule with good performance. Knowing the conditional probability function one can construct the Bayesian (optimal) decision rule

$$r(x) = \text{sign} \left\{ \ln \left(\frac{P\{y = 1|x\}}{1 - P\{y = 1|x\}} \right) \right\}.$$

Below we consider the following (parametric) problem of estimating the conditional probability.¹² Suppose that the logarithm of the ratio of the following two probabilities is a function $f(x, w_0)$ from a given parametric set $f(x, w)$, $w \in W$

$$\ln \left(\frac{P\{y = 1|x\}}{1 - P\{y = 1|x\}} \right) = f(x, w_0).$$

From this equation it follows that the conditional probability function $P\{y = 1|x\}$ has the following form:

$$P\{y = 1|x\} = \frac{e^{f(x, w_0)}}{1 + e^{f(x, w_0)}}. \quad (5.36)$$

The function (5.36) is called logistic regression.

Our goal is given data

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

to estimate the parameters w_0 of the logistic regression.¹³ First we show that the minimum of the functional

$$R_x(w) = E_y \ln \left(1 + e^{-yf(x, w)} \right) \quad (5.37)$$

¹²The more general nonparametric setting of this problem we discuss in Chapter 7.

¹³Note that (5.36) is a form of sigmoid function considered in Section 5.2. Therefore a one-layer neural network with sigmoid function (5.36) is often considered as an estimate of the logistic regression.

$(E_y$ is expectation over y for a fixed value of x) defines the desired parameters.

Indeed, the necessary condition for a minimum is

$$\frac{\partial R_x(w)}{\partial w} = \left[\frac{\partial}{\partial w} E_y \ln \left(1 + e^{-yf(x,w)} \right) \right]_{w_0} = 0.$$

Taking the derivative over w and using expression (5.36) we obtain

$$\begin{aligned} \frac{\partial R(w)}{\partial w} &= \frac{\partial}{\partial w} E_y \ln \left(1 + e^{-yf(x,w)} \right) \\ &= \left(\frac{-f'_w(x,w)e^{-f(x,w)}}{1 + e^{-f(x,w)}} \right) P\{y = 1|x\} + \left(\frac{f'_w(x,w)}{1 + e^{f(x,w)}} \right) P\{y = -1|x\} \\ &= \left(\frac{-f'_w(x,w)e^{-f(x,w)}}{1 + e^{-f(x,w)}} \right) \left(\frac{e^{f(x,w_0)}}{1 + e^{f(x,w_0)}} \right) + \left(\frac{f'_w(x,w)e^{f(x,w)}}{1 + e^{f(x,w)}} \right) \left(\frac{1}{1 + e^{f(x,w_0)}} \right) \end{aligned}$$

This expression is equal to zero when $w = w_0$. That is, the minimum of the functional (5.37) defines the parameters of the logistic regression.

Below we assume that the desired logistic regression is a linear function

$$f(x, w) = (x \cdot w_0) + b$$

whose parameters w_0 and b we will estimate by minimizing the functional

$$R(w) = E_{y,x} \ln \left(1 + e^{-y[(x \cdot w) + b]} \right) \quad (5.38)$$

using observations

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

To minimize the functional (5.38) we use the structural risk minimization method with the structure defined as follows:

$$(w \cdot w) \leq r.$$

We consider this minimization problem in the following form: Minimize the functional

$$R_{\text{emp}}(w, b) = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} \ln \left(1 + e^{-y_i[(w \cdot x_i) + b]} \right). \quad (5.39)$$

One can show that the minimum of (5.39) defines the following approximation to the logistic regression:

$$P\{y = 1|x\} = \frac{\exp \left\{ \sum_{i=1}^{\ell} y_i [C\alpha_i^0(x_i, x) + b_0] \right\}}{1 + \exp \left\{ \sum_{i=1}^{\ell} y_i [C\alpha_i^0(x_i, x) + b^0] \right\}}, \quad (5.40)$$

where the coefficients α_i^0 and b_0 are the solution of the equations

$$\alpha_i = \frac{\exp\{-y_i[\sum_{j=1}^{\ell} C\alpha_j y_j(x_j, x_i) + b]\}}{1 + \exp\{-y_i[\sum_{j=1}^{\ell} C\alpha_j y_j(x_j, x_i) + b]\}},$$

$$\sum_{i=1}^{\ell} y_i \frac{\exp\{-y_i[\sum_{j=1}^{\ell} C\alpha_j y_j(x_j, x_i) + b]\}}{1 + \exp\{-y_i[\sum_{j=1}^{\ell} C\alpha_j y_j(x_j, x_i) + b]\}} = 0.$$

Indeed, a necessary condition for the point (w_0, b_0) to minimize the functional (5.39) is

$$\frac{\partial R(w, b)}{\partial w} \Big|_{w_0, b_0} = w - C \sum_{i=1}^{\ell} y_i x_i \frac{\exp\{-y_i[(w, x_i) + b]\}}{1 + \exp\{-y_i[(w, x_i) + b]\}} \Big|_{w_0, b_0} = 0,$$

$$\frac{\partial R(w, b)}{\partial b} \Big|_{w_0, b_0} = -C \sum_{i=1}^{\ell} y_i \frac{\exp\{-y_i[(w, x_i) + b]\}}{1 + \exp\{-y_i[(w, x_i) + b]\}} \Big|_{w_0, b_0} = 0. \quad (5.41)$$

Using the notation

$$\frac{\exp\{-y_i[(w_0, x_i) + b_0]\}}{1 + \exp\{-y_i[(w_0, x_i) + b_0]\}} = \alpha_i^0, \quad (5.42)$$

we can rewrite expressions (5.41) as follows:

$$w_0 = C \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i,$$

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 = 0. \quad (5.43)$$

Putting expressions (5.43) and back into (5.37) we obtain the approximation (5.40).

Note that from (5.42) and (5.43) we have

$$0 < \alpha^0 < 1.$$

That is, this solution is not sparse.

To find the logistic regression one can rewrite the functional (5.39) (using expression (5.43)) in the equivalent form

$$R_{\text{emp}}(\alpha, b)$$

$$= \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (x_i, x_j) + C \sum_{i=1}^{\ell} \ln \left(1 + \exp\{-y_i [\sum_{j=1}^{\ell} y_j y_j \alpha_j (x_i, x_j) + b]\} \right).$$

Since this functional is convex with respect to the parameters α and b , one can use the gradient descent method to find its minimum.

5.9.2 The Risk Function for SVM

Let us introduce the following notation

$$z = (w \cdot x) + b.$$

Using this notation we can rewrite the risk functional for the logistic regression as follows

$$Q(z) = \ln(1 + e^{-yz}).$$

Consider the loss function

$$Q^*(z) = c_1 (1 - z)_+, \quad (5.44)$$

where c_1 is some constant (in constructing the SVM we used $c_1 = 1$) and $(a)_+ = \max(0, a)$ (the linear spline function with one node, for more about spline approximations see Section 6.3).

Figure 5.9 shows this loss function with $c_1 = 0.8$ (the bold lines) and the logistic loss (dashed curve).

It is easy to see that the SVM minimize the following functional:

$$R_{\text{emp}}(w, b) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{\ell} (1 - y_i[(w \cdot x_i) + b])_+. \quad (5.45)$$

Indeed, denote by the ξ_i the expression

$$\xi_i = (1 - y_i[(w \cdot x_i) + b])_+,$$

which is equivalent to the inequality

$$y_i[(w \cdot x_i) + b] \geq 1 - \xi_i. \quad (5.46)$$

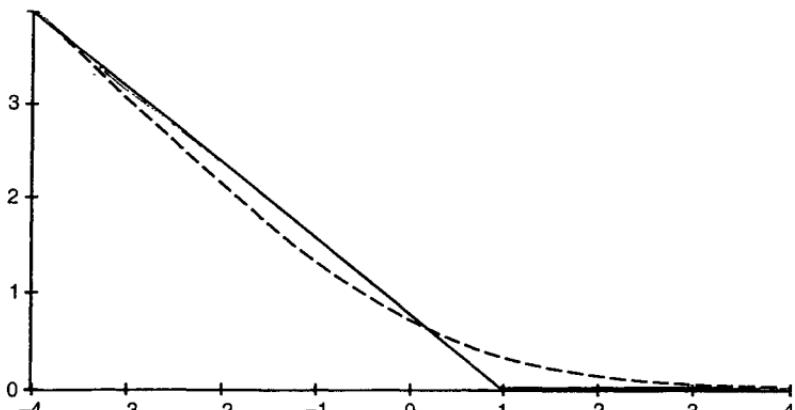


FIGURE 5.9. The logistic loss function (dashed line) and its approximation by a linear spline with one node (bold line).

Now we can rewrite our optimization problem (5.45) as follows: Minimize the functional

$$R(w, b) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{\ell} \xi_i \quad (5.47)$$

subject to constraints (5.46) and constraints

$$\xi_i \geq 0$$

This problem coincides with one that was suggested in Section 5.5.1 for constructing the optimal separating hyperplane for the nonseparable case.

5.9.3 The SVM_n Approximation of the Logistic Regression

One can construct better SVM approximations to the logistic loss function using linear spline functions with $n > 1$ nodes.

Suppose we are given the following spline approximation to the logistic loss:

$$F(z) = \sum_{k=1}^n c_k (a_k - z)_+,$$

where

$$z = y[(w \cdot x) + b],$$

a_k , $k = 1, \dots, n$ are nodes of the spline and $c_k \geq 0$, $k = 1, \dots, n$, are coefficients of the spline. (Since the logistic loss function is convex monotonic function, one can approximate it with any degree of accuracy using a linear spline with nonnegative coefficients c_k .)

Figure 5.10 shows an approximation of the logistic loss (dashed curve) by (a) spline function with two nodes and (b) by spline function with three nodes (bold lines).

Let us minimize the functional

$$R(w, b) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{\ell} \sum_{k=1}^n c_k (a_k - z_i)_+$$

which is our approximation to the functional (5.38).

Set

$$(a_k - z_i)_+ = (a_k - y_i[(w \cdot x_i) + b])_+ = \xi_i^k, \quad \xi_i^k \geq 0, \quad k = 1, \dots, n, \quad i = 1, \dots, \ell$$

Using this notation we can rewrite our problem as follows:

Minimize the functional

$$R(w, b) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{\ell} \sum_{k=1}^n \xi_i^k$$

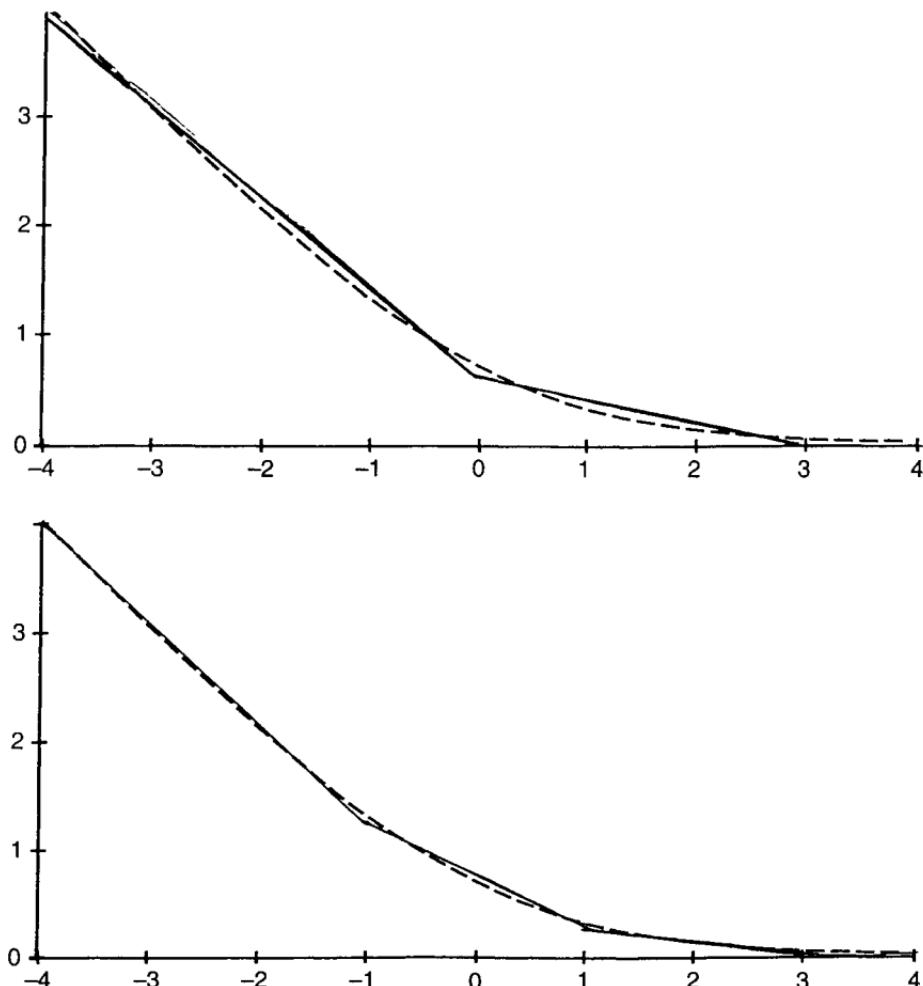


FIGURE 5.10. The logistic loss function (dashed line) and its approximations: (a) by a linear spline with two nodes and (b) by a linear spline with three nodes (bold lines).

subject to the constraints

$$y_i[(w \cdot x_i) + b] \geq a_k - \xi_i^k, \quad i = 1, \dots, \ell, \quad k = 1, \dots, n,$$

and constraints

$$\xi_i^k \geq 0, \quad i = 1, \dots, \ell, \quad k = 1, \dots, n.$$

As before, to solve this quadratic optimization problem in the dual space we construct the Lagrangian

$$L = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^{\ell} \sum_{k=1}^n \xi_i^k - \sum_{i=1}^{\ell} \sum_{k=1}^n \beta_i^k (y_i[(w \cdot x_i) + b] - a_k + \xi_i^k) - \sum_{i=1}^{\ell} \sum_{k=1}^n \xi_i^k r_i^k.$$

Taking the minimum over w , b , and ξ_i^k we obtain

$$w = \sum_{i=1}^{\ell} \left(\sum_{k=1}^n \beta_i^k \right) y_i x_i, \quad (5.48)$$

$$\sum_{i=1}^{\ell} \left(\sum_{k=1}^n \beta_i^k \right) y_i = 0, \quad (5.49)$$

$$0 \leq \beta_i^k \leq C c_k, \quad k = 1, \dots, n. \quad (5.50)$$

Substituting the expression for w back into the Lagrangian and taking into account (5.49) we obtain the functional

$$W(\beta) = \sum_{i=1}^{\ell} \left(\sum_{k=1}^n \beta_i^k a_k \right) - \frac{1}{2} \sum_{i,j=1}^{\ell} \left(\sum_{k=1}^n \beta_i^k \right) \left(\sum_{k=1}^n \beta_j^k \right) y_i y_j (x_i \cdot x_j), \quad (5.51)$$

where a_1, \dots, a_n are nodes in our spline approximation to the logistic loss function.

To find the parameters $\beta_i^1, \dots, \beta_i^n$, $i = 1, \dots, \ell$ that specify the expansion (5.48) of the optimal vector w we have to maximize the functional (5.51) subject to constraints (5.49) and (5.50).

We also can find the parameter b from the Kuhn-Tucker conditions

$$\beta_i^k \{y_i[(w \cdot x_i) + b] - a_k + \xi_i^k\} = 0, \quad i = 1, \dots, \ell, \quad k = 1, \dots, n.$$

Using these parameters one can construct the linear function

$$l(x) = \sum_{j=1}^{\ell} y_j \left(\sum_{k=1}^n \beta_j^k \right) (x_j \cdot x) + b \quad (5.52)$$

that defines the approximation

$$P\{y = 1|x\} = \frac{\exp \left\{ \sum_{j=1}^{\ell} y_j \left(\sum_{k=1}^n \beta_j^k \right) (x_j \cdot x) + b \right\}}{\left(1 + \exp \left\{ \sum_{j=1}^{\ell} y_j \left(\sum_{k=1}^n \beta_j^k \right) (x_j \cdot x) + b \right\} \right)} \quad (5.53)$$

to the logistic regression (5.36). As before, to define the vector w in the exponent of the logistic regression we need only calculate the inner products between two vectors x . Therefore, using kernels $K(x, x_i)$ satisfying the Mercer condition one can construct an approximation to the logistic regression of the form

$$P\{y = 1|x\} = \frac{\exp \left\{ \sum_{j=1}^{\ell} y_j (\sum_{k=1}^n \beta_j^k) K(x_j, x) + b \right\}}{\left(1 + \exp \left\{ \sum_{j=1}^{\ell} y_j (\sum_{k=1}^n \beta_j^k) K(x_j, x) + b \right\} \right)},$$

where the coefficients β_j^k are the solution of the following quadratic optimization problem: Maximize the functional

$$W(\beta) = - \sum_{i=1}^{\ell} \left(\sum_{k=1}^n \beta_i^k a_k \right) - \frac{1}{2} \sum_{i,j=1}^{\ell} \left(\sum_{k=1}^n \beta_i^k \right) \left(\sum_{k=1}^n \beta_j^k \right) y_i y_j K(x_i, x_j), \quad (5.54)$$

subject to constraints

$$\sum_{i=1}^{\ell} \left(\sum_{k=1}^n \beta_i^k \right) y_i = 0,$$

$$0 \leq \beta_i^k \leq C c_k, \quad k = 1, \dots, n.$$

Note that a larger number of nodes is used in the approximation of the logistic loss, a larger number of support vectors will be used for the constructing corresponding hyperplane. With increasing accuracy of approximation (number of nodes) the SVM_n loses sparsity.

However, with increasing n in the SVM_n one cannot guarantee a better performance for the solution obtained using a given sample size. The problem of estimating well the logistic regression is more general than the problem of estimating a good decision rule, and therefore, in order to be solved well it requires more data for its solution.

Our experiments did not show an advantage of logistic regression or SVM_n compared to SVM₁.

5.10 ENSEMBLE OF THE SVM

In 1996 Y. Freund and R. Schapire proposed the AdaBoost algorithm for combining several weak rules¹⁴ (features) in one linear decision rule that can perform much better than any weak rule.

Later it was shown that in fact, AdaBoost minimizes (using a greedy optimization procedure) some functional whose minimum defines the logistic

¹⁴That is, indicator functions that classify test data at least slightly better than a random guess.

regression (Friedman, Hestie, and Tibshirany (1998)). Also, it was shown that the optimal hyperplane constructed on top of the weak (indicator) rules chosen by the AdaBoost often outperforms the AdaBoost solution.

Therefore, in the AdaBoost algorithm we distinguish two parts:

1. The choice of N appropriate features from a given set of indicator features.
2. The construction of a separating hyperplane using the chosen features.

In this section we introduce a two-stage method for constructing an ensemble of SVMs. In the first stage, using given training data, we find N indicator functions (features), which on the one hand are SVM solutions of the given pattern recognition problem, and on the other hand are the result of greedy minimization of the same functional that minimizes AdaBoost algorithm.

In the second stage using training data we construct on top of the features obtained the SVM decision rules. Therefore, we will construct N different SVM solutions of the same pattern recognition problem and then combine them into one decision rule.

5.10.1 The AdaBoost Method

In Section 5.9.1 we introduced the risk functional (5.37) whose minimum defined parameters of the logistic regression. Below we consider another risk functional

$$R(\alpha) = Ee^{-yf(x, \alpha)} \quad (5.55)$$

defined on a set of functions $f(x, \alpha)$ that contain the function

$$f(x, \alpha_0) = \frac{1}{2} \ln \frac{P(y = 1|x)}{P(y = -1|x)}. \quad (5.56)$$

It is easy to see that the function $f(x, \alpha_0)$ provides the minimum to functional (5.55).

Indeed, equation (5.56) is equivalent to the equations

$$\begin{aligned} P(y = 1|x) &= \frac{e^{2f(x, \alpha_0)}}{1 + e^{2f(x, \alpha_0)}} = \frac{e^{f(x, \alpha_0)}}{e^{-f(x, \alpha_0)} + e^{f(x, \alpha_0)}}, \\ P(y = -1|x) &= \frac{1}{1 + e^{2f(x, \alpha_0)}} = \frac{e^{-f(x, \alpha_0)}}{e^{-f(x, \alpha_0)} + e^{f(x, \alpha_0)}}. \end{aligned} \quad (5.57)$$

Since

$$E(e^{-yf(x, \alpha)} | x) = P(y = 1|x)e^{-f(x, \alpha)} + P(y = -1|x)e^{f(x, \alpha)},$$

we have

$$\frac{\partial E(e^{-yf(x,\alpha)}|x)}{\partial f(x,\alpha)} = -P(y=1|x)e^{-f(x,\alpha)} + P(y=-1)e^{f(x,\alpha)}. \quad (5.58)$$

At the point α_0 the derivative (5.58) is equal to zero as soon as (5.57) takes place.

Let us instead of (5.55) use the empirical risk functional

$$R_{\text{emp}}(\alpha) = \sum_{i=1}^{\ell} e^{-y_i f(x_i, \alpha)}, \quad (5.59)$$

which we minimize iteratively, using the following greedy optimization procedure.

Greedy optimization procedure:

1. We minimize functional (5.59) iteratively constructing on the k th iteration a function of the form

$$f(x, \beta_k) = \sum_{r=1}^k d_r \phi_r(x), \quad d_1 = 1,$$

where $\phi_r(x)$, $r = 1, \dots, N$, belong to a given (maybe infinite) set of indicator functions, k is the number of iteration, and $\beta_k = (d_1, \dots, d_k)$ is a k -dimensional vector.

On the first iteration we choose the feature $\phi_1(x)$ that minimizes the number of training errors.

2. Suppose that at the k th iteration we achieved the following value of the empirical risk:

$$R_{\text{emp}}(\beta_k) = \sum_{i=1}^{\ell} e^{-y_i f_k(x_i, \beta_k)}.$$

At the next $(k+1)$ iteration we continue to minimize the empirical risk functional in the set of one-parameter functions

$$f(x, \beta_{(k+1)}) = f(x, \beta_k) + d_{(k+1)} \phi_{(k+1)}(x). \quad (5.60)$$

For function (5.60) we obtain the following value of the empirical risk

$$R_{\text{emp}}(\beta_{(k+1)}) = \sum_{i=1}^{\ell} e^{-y_i f(x_i, \beta_{k+1})} = \sum_{i=1}^{\ell} c_i^{k+1} e^{-d_{(k+1)} y_i \phi_{(k+1)}(x_i)}, \quad (5.61)$$

where we have set

$$c_i^{k+1} = e^{-y_i f_k(x_i, \beta_k)}.$$

Suppose that for the $(k+1)$ st iteration we have chosen the indicator function $\phi_{(k+1)}(x)$ (later we will define how to choose this function). Then in order to minimize the empirical risk (5.61) we have to choose the following value of the parameter:

$$d_{(k+1)} = \frac{1}{2} \ln \frac{\mathcal{C}_+^{k+1}}{\mathcal{C}_-^{k+1}}, \quad (5.62)$$

where we set

$$\mathcal{C}_+^{k+1} = \sum_{\{i: y_i \phi_{(k+1)}(x_i) = 1\}} c_i^{k+1},$$

$$\mathcal{C}_-^{k+1} = \sum_{\{i: y_i \phi_{(k+1)}(x_i) = -1\}} c_i^{k+1}.$$

This follows from the facts that $y_i \phi_{(k+1)}(x_i) \in \{1, -1\}$ and that at the optimal point $d_{(k+1)}$ the derivative over d of the empirical functional (5.61) must be equal to zero

$$\begin{aligned} & \frac{\partial}{\partial d} \sum_{i=1}^{\ell} e^{-y_i [f(x_i, \beta_k) + d\phi_{(k+1)}(x_i)]} \\ &= - \sum_{i=1}^{\ell} c_i^{k+1} y_i \phi_{(k+1)}(x_i) e^{-d y_i \phi_{(k+1)}(x_i)} = 0. \end{aligned} \quad (5.63)$$

3. To choose the appropriate function $\phi_{(k+1)}(x)$ for the $(k+1)$ st iteration, note that after the k th iteration, according to (5.63), the equality

$$-\sum_{i=1}^{\ell} c_i^k y_i \phi_k(x_i) e^{-d_k y_i \phi_k(x_i)} = -\sum_{i=1}^{\ell} c_i^{k+1} y_i \phi_k(x_i) = 0.$$

holds true.

Suppose that coefficients c_i^{k+1} are normalized to 1:

$$c_i^{k+1} \leftarrow \frac{c_i^{k+1}}{\sum_{i=1}^{\ell} c_i^{k+1}}.$$

This does not change the result. However, normalization allows us to propose a nice statistical interpretation of equation (5.63): Normalized coefficients c_i^{k+1} , $i = 1, \dots, \ell$ can be considered as a probability measure assigned on the given training data for the $(k+1)$ th iteration and indicator function function $\phi_k(x)$ as the worst solution for

our training data assign with this probability measure (for this probability measure the rule $\phi_k(x)$ has a 50% error rate). That is, after every iteration, the algorithm assigns to a given training set a new probability measure that is the most difficult for the last weak rule.

Therefore, for the next, $(k+1)$ st, iteration we choose the function $\phi_{(k+1)}(x)$ that minimizes the error rate for the assigned probability measure. That is, we choose the function $\phi_{(k+1)}(x)$ that minimizes the functional

$$R(\phi) = - \sum_{i=1}^{\ell} c_i^{k+1} y_i \phi(x_i). \quad (5.64)$$

4. The indicator function

$$\Phi(x) = \text{sign} \left(\sum_{k=1}^N d_k \phi_k(x) \right), \quad (5.65)$$

obtained as result of the greedy minimization procedure described, is the AdaBoost decision rule.

5.10.2 The Ensemble of SVMs

Let us use the greedy optimization idea described above for constructing the ensemble of SVMs. We start with the case where weak features are linear decision rules

$$\phi_k(x) = \text{sign}\{(x \cdot w_k) + b_k\}.$$

Our goal is to find N optimal hyperplanes that in greedy fashion minimize the functional

$$R(w, b) = \sum_{i=1}^{\ell} \exp\{-y_i \sum_{k=1}^N d_k \text{sign}[(x_i \cdot w_k) + b_k]\} \quad (5.66)$$

and then using these linear decision rules as the features construct the desired ensemble.

Constructing the features. To construct N features we need to specify in the general scheme described in the previous section only the method for minimizing the functional (5.64) in the set of linear decision functions:

$$\phi_k(x) = \text{sign}\{(w_k \cdot x) + b_k\}$$

(defined by the optimal hyperplane).

As before, we replace this problem with the following problem: Minimize the functional

$$R(w_k) = \frac{1}{2}(w_k \cdot w_k) + C \sum_{i=1}^{\ell} c_i^k \xi_i^k, \quad c_i^1 = 1, \quad (5.67)$$

subject to constraints

$$y_i((w_k \cdot x_i) + b_k) \geq 1 - \xi_i^k, \quad \xi_i^k \geq 0. \quad (5.68)$$

The only difference in the problem of constructing this hyperplane compared to the problem of constructing the soft-margin hyperplane described in Section 5.5.1 is that in the case of the soft-margin hyperplane all coefficients c_i^k were equal to 1. Now the second term in (5.67) is a weighted sum.

We solve this optimization problem using the same technique with Lagrange multipliers. We obtain the following solution:

$$w_k = \sum_{i=1}^{\ell} y_i \alpha_i^k x_i,$$

where the coefficients α_i^k maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (5.69)$$

subject to the constraints

$$0 \leq \alpha_i \leq C c_i^k \quad (5.70)$$

and the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i c_i^k = 0. \quad (5.71)$$

The coefficient b_k can be defined from Kuhn-Tucker conditions

$$\alpha_i (y_i (w_k \cdot x_i) + b_k - 1 - \xi_i^k) = 0.$$

Therefore, the difference in decision rules is defined by the coefficients c_i^k . These coefficients are calculated iteratively as it was described in the greedy optimization procedure (Section 5.10.1):

$$c_i^1 = 1, \quad i = 1, \dots, \ell,$$

$$c_i^{(k+1)} = \exp \left\{ -y_i \sum_{r=1}^k d_r \phi_r(x_i) \right\} = c_i^k \exp \left\{ -y_i d_k \phi_k(x_i) \right\}, \quad (5.72)$$

where

$$d_k = \frac{1}{2} \ln \frac{\sum_{\{i: y_i \phi_k(x_i) = 1\}} c_i^k}{\sum_{\{i: y_i \phi_k(x_i) = -1\}} c_i^k}. \quad (5.73)$$

Remark. Note that if the training data are separable, then the denominator of equation (5.73) is equal to zero, and therefore, according to (5.72),

$c_i^k = 0$, $i = 1, \dots, \ell$ for all $k > 1$. That is, the set of features has only one decision rule. To prevent this situation one can choose a sufficiently small value of C (large regularization parameter). If, however, for sufficiently small C the training data are still separable, then the obtained hyperplane has a good generalization ability.

The choice of the constant C plays an important role in constructing an ensemble of SVMs.

Constructing the decision rule. To obtain the decision rule one constructs the optimal hyperplane in N -dimensional binary space

$$z = (\phi_1(x), \dots, \phi_N(x)).$$

Using the given set of training data one obtains the new set of training data

$$(y_1, z_1,) \dots (y_\ell, z_\ell) \quad (5.74)$$

($z_i = (\phi_1(x_i), \dots, \phi_N(x_i))$), based on which one constructs the optimal hyperplane.

Ensemble of SVMs As before we can use kernels to obtain features using general type of SVMs. We can use features of the form

$$\phi_k(x) = \text{sign} \left(\sum_{i=1}^{\ell} y_i \alpha_i K(x, x_i) \right)$$

where the coefficients α_i are solution of the following optimization problem:
Maximize the functional

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

subject to the constraints

$$0 \leq \alpha_i \leq C c_i^k$$

and the constraint

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

Using obtained N features $\phi_k(x)$, $k = 1, \dots, N$ that define a binary space Z one constructs the training set (5.74). On the basis of this training set using a kernel $K^*(z, z_i)$ defined in Z space one constructs the SVM solution

$$r(x) = \text{sign} \left(\sum_{i=1}^{\ell} y_i \beta_i K^*(z(x), z(x_i)) \right).$$

Informal Reasoning and Comments — 5

5.11 THE ART OF ENGINEERING VERSUS FORMAL INFERENCE

The existence of neural networks can be considered a challenge for theoreticians.

From the formal point of view one cannot guarantee that neural networks generalize well, since according to theory, in order to control generalization ability one should control two factors: the value of the empirical risk and the value of the confidence interval. Neural networks, however, cannot control either of the two.

Indeed, to minimize the empirical risk, a neural network must minimize a functional that has many local minima. Theory offers no constructive way to prevent ending up with unacceptable local minima. In order to control the confidence interval one has first to construct a structure on the set of functions that the neural network implements and then to control capacity using this structure. There are no accurate methods to do this for neural networks.

Therefore, from the formal point of view it seems that there should be no question as to what type of machine should be used for solving real-life problems.

The reality, however, is not so straightforward. The designers of neural networks compensate the mathematical shortcomings with the high art of engineering. Namely, they incorporate various heuristic algorithms that

make it possible to attain reasonably local minima using a reasonable small number of calculations.

Moreover, for given problems they create special network architectures that both have an appropriate capacity and contain “useful” functions for solving the problem. Using these heuristics, neural networks demonstrate surprisingly good results.

In Chapter 5, describing the best results for solving the digit recognition problem using the U.S. Postal Service database by constructing an entire (not local) decision rule, we gave two figures:

5.1% error rate for the neural network LeNet 1 (designed by Y. LeCun),

4.0% error rate for a polynomial SV machine.

We also mentioned the two best results:

3.3% error rate for the local learning approach, and the record

2.7% error rate for tangent distance matching to templates given by the training set.

In 1993, responding to the community’s need for benchmarking, the U.S. National Institute of Standards and Technology (NIST) provided a database of handwritten characters containing 60,000 training images and 10,000 test data, where characters are described as vectors in $20 \times 20 = 400$ pixel space.

For this database a special neural network (LeNet 4) was designed. The following is how the article reporting the benchmark studies (Léon Bottou *et al*, 1994) describes the construction of LeNet 4:

“For quite a long time, LeNet 1 was considered the state of the art. The local learning classifier, the SV classifier, and tangent distance classifier were developed to improve upon LeNet 1 — and they succeeded in that. However, they in turn motivated a search for an improved neural network architecture. This search was guided in part by estimates of the capacity of various learning machines, derived from measurements of the training and test error (on the large NIST database) as a function of the number of training examples.¹⁵ We discovered that more capacity was needed. Through a series of experiments in architecture, combined with an analysis of the characteristics of recognition errors, LeNet 4 was crafted.”

¹⁵V. Vapnik, E. Levin, and Y. LeCun (1994) “Measuring the VC dimension of a learning machine,” *Neural Computation*, 6(5), pp. 851-876.

In these benchmarks, two learning machines that construct entire decision rules,

- (i) LeNet 4,
- (ii) Polynomial SV machine (polynomial of degree four),

provided the same performance: 1.1% test error.¹⁶

The local learning approach and tangent distance matching to 60,000 templates also gave the same performance: 1.1% test error.

Recall that for a small (U.S. Postal Service) database the best result (by far) was obtained by the tangent distance matching method which uses *a priori* information about the problem (incorporated in the concept of tangent distance). As the number of examples increases to 60,000 the advantage of *a priori* knowledge decreased. The advantage of the local learning approach also decreased with the increasing number of observations.

LeNet 4, crafted for the NIST database demonstrated remarkable improvement in performance comparing to LeNet 1 (which has 1.7% test errors for the NIST database¹⁷).

The standard polynomial SV machine also did a good job. We continue the quotation (Léon Bottou, *et al*, 1994):

“The SV machine has excellent accuracy, which is most remarkable, because unlike the other high performance classifiers it *does not include knowledge about the geometry of the problem*.

In fact this classifier would do just as well if the image pixel were encrypted, e.g., by a fixed random permutation.”

However, the performance achieved by these learning machines is not the record for the NIST database. Using models of characters (the same that was used for constructing the tangent distance) and 60,000 examples of training data, H. Drucker, R. Schapire, and P. Simard generated more than 1,000,000 examples, which they used to train three LeNet 4 neural networks, combined in the special “boosting scheme” (Drucker, Schapire, and Simard, 1993) which achieved a 0.7% error rate.

Now the SV machines have a challenge — to cover this gap (between 1.1% to 0.7%). Probably the use of only brute force SV machines and 60,000 training examples will not be sufficient to cover the gap. Probably one has to incorporate some *a priori* information about the problem at hand.

¹⁶Unfortunately, one cannot compare these results to the results described in Chapter 5. The digits from the NIST database are “easier” for recognition than the ones from the U.S. Postal Service database.

¹⁷Note that LeNet 4 has an advantage for a large 60,000 training examples (NIST) database. For a small (U.S. Postal Service) database containing 7,000 training examples, the network with smaller capacity, LeNet 1, is better.

There are several ways to do this. The simplest one is use the same 1,000,000 examples (constructed from the 60,000 NIST prototypes). However, it is more interesting to find a way for directly incorporating the invariants that were used for generating the new examples. For example, for polynomial machines one can incorporate *a priori* information about invariance by using the convolution of an inner product in the form $(x^T Ax^*)^d$, where x and x^* are input vectors and A is a symmetric positive definite matrix reflecting the invariants of the models.¹⁸

One can also incorporate another (geometrical) type of *a priori* information using only features (monomial) $x_i x_j x_k$ formed by pixels that are close each to other (this reflects our understanding of the geometry of the problem — important features are formed by pixels that are connected to each other, rather than pixels far from each other). This essentially reduces (by a factor of millions) the dimensionality of feature space.

Thus, although the theoretical foundations of support vector machines look more solid than those of neural networks, the practical advantages of the new type of learning machines still needs to be proved.¹⁹

5.12 WISDOM OF STATISTICAL MODELS

In this chapter we introduced the support vector machines, which realize the structural risk minimization inductive principle by:

- (i) Mapping the input vector into a high-dimensional feature space using a nonlinear transformation.
- (ii) Constructing in this space a structure on the set of linear decision rules according to the increasing norm of weights of canonical hyperplanes.
- (iii) Choosing the best element of the structure and the best function within this element in order to minimize the bound on error probability.

¹⁸B. Schölkopf considered an intermediate way: He constructed an SV machine, generated new examples by transforming the SV images (translating them in the four principal directions), and retrained on the support vectors and the new examples. This improves the performance from 4.0% to 3.2% for the U.S. Postal Service database and from 1.1% to 0.8% for the NIST database.

¹⁹In connection with heuristics incorporated in neural networks let me recall the following remark by R. Feynman: "We must make it clear from the beginning that if a thing is not a science, it is not necessarily bad. For example, love is not science. So, 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." *The Feynman Lectures on Physics*, Addison-Wesley, 3-1, 1975.

The implementation of this scheme in the algorithms described in this chapter, however, contained one violation of the SRM principle. To define the structure on the set of linear functions we use the set of canonical hyperplanes constructed with respect to vectors x from the training data. According to the SRM principle, the structure has to be defined *a priori* before the training data appear.

The attempt to implement the SRM principle *in toto* brings us to a new statement of the learning problem that forms a new type of inference. For simplicity we consider this model for the pattern recognition problem.

Let the learning machine that implements a set of functions linear in feature space be given $\ell + k$ vectors

$$x_1, \dots, x_{\ell+k} \quad (5.75)$$

drawn randomly and independently according to some distribution function.

Suppose now that these $\ell + k$ vectors are randomly divided into two subsets: the subset

$$x_1, \dots, x_\ell$$

for which the string

$$y_1, \dots, y_\ell, \quad y \in \{-1, +1\},$$

describing classification of these vectors is given (the training set), and the subset

$$x_{\ell+1}, \dots, x_{\ell+k}$$

for which the classification string should be found by the machine (test set). The goal of the machine is to find the rule that gives the string with the minimal number of errors on the given test set.

In contrast to the model of function estimation considered in this book, this model looks for the rule that minimizes the number of errors on the *given test set* rather than for the rule minimizing the probability of error on the admissible test set. We call this problem the *estimation of the values of the function at given points*. For the problem of estimating the values of a function at given points the SV machines will realize the SRM principle *in toto* if one defines the canonical hyperplanes with respect to all $\ell + k$ vectors (5.78). (One can consider the data (5.78) as *a priori* information. *A posteriori* information is any information about separating this set into two subsets.)

Estimating the values of a function at given points has both a solution and a method of solution that differ from those based on estimating an unknown function.

Consider, for example, the five-digit zipcode recognition problem.²⁰ The existing technology based on estimating functions suggests recognizing the five digits x_1, \dots, x_5 of the zipcode independently: First one uses the rules constructed during the learning procedures to recognize digit x_1 , then one uses the same rules to recognize digit x_2 , and so on.

The technology of estimating the values of a function suggests recognizing all five digits jointly: The recognition of one digit, say x_1 , depends not only on the training data and vector x_1 , but also on vectors x_2, \dots, x_5 . In this technology one uses the rules that are in a special way adapted to solving a given specific task. One can prove that this technology gives more accurate solutions.²¹

It should be noted that for the first time this new view of the learning problem was found due to attempts to justify a structure defined on the set of canonical hyperplanes for the SRM principle.

5.13 WHAT CAN ONE LEARN FROM DIGIT RECOGNITION EXPERIMENTS?

Three observations should be discussed in connection with the experiments described in this chapter:

- (i) The structure constructed in the feature space reflects *real-life problems* well.
- (ii) The quality of decision rules obtained does not strongly depend on the type of SV machine (polynomial machine, RBF machine, two-layer NN). It does, however, strongly depend on the accuracy of the VC dimension (capacity) control.
- (iii) Different types of machines use the same elements of training data as support vectors.

²⁰For simplicity we do not consider the segmentation problem. We suppose that all five digits of a zipcode are segmented.

²¹Note that the local learning approach described in Section 4.5 can be considered as an intermediate model between function estimation and estimation of the values of a function at points of interest. Recall that for a small (Postal Service) database the local learning approach gave significantly better results (3.3% error rate) than the best result based on the entire function estimation approach (5.1% obtained by LeNet 1, and 4.0% obtained by the polynomial SV machine).

5.13.1 Influence of the Type of Structures and Accuracy of Capacity Control

The classical approach to estimating multidimensional functional dependencies is based on the following belief:

Real-life problems are such that there exists a small number of “strong features,” simple functions of which (say linear combinations) approximate well the unknown function. Therefore, it is necessary to carefully choose a low-dimensional feature space and then to use regular statistical techniques to construct an approximation.

This approach stresses, that one should be careful at the stage of feature selection (this is an informal operation) and then use routine statistical techniques.

The new technique is based on a different belief:

Real-life problems are such that there exist a large number of “weak features” whose “smart” linear combination approximates the unknown dependency well. Therefore, it is not very important what kind of “weak feature” one uses, it is more important to form “smart” linear combinations.

This approach stresses, that one should choose any reasonable “weak feature space” (this is an informal operation), but be careful at the point of making “smart” linear combinations. From the perspective of SV machines, “smart” linear combinations correspond to the capacity control method.

This belief in the structure of real-life problems has been expressed many times both by theoreticians and by experimenters.

In 1940, Church made a claim that is known as the Turing–Church Thesis:²²

All (sufficiently complex) computers compute the same family of functions.

In our specific case we discuss the even stronger belief that linear functions in various feature spaces associated with different convolutions of the inner product approximate the same set of functions if they possess the same capacity.

Church made his claim on the basis of pure theoretical analysis. However, as soon as computer experiments became widespread, researchers unexpectedly faced a situation that could be described in the spirit of Church’s claim.

In the 1970s and 1980s a considerable amount of experimental research was conducted in solving various operator equations that formed ill-posed

²²Note that the thesis does not reflect some proved fact. It reflects the belief in the existence of some law that is hard to prove (or formulate in exact terms).

problems, in particular, in density estimation. A common observation was that the choice of the type of regularizers $\Omega(f)$ in (4.32) (determining a type of structure) is not as important as choosing the correct regularization constant $\gamma(\delta)$ (determining capacity control).

In particular, in density estimation using the Parzen window

$$p(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{1}{\gamma^n} K\left(\frac{x - x_i}{\gamma}\right),$$

a common observation was the following: If the number of observations is not “very small,” the type of kernel function $K(u)$ in the estimator is not as important as the value of the constant γ . (Recall that the kernel $K(u)$ in Parzen’s estimator is determined by the functional $\Omega(f)$, and γ is determined by the regularization constant.)

The same was observed in the regression estimation problem, where one tries to use expansions in different series to estimate the regression function: If the number of observations is not “very small,” the type of series used is not as important as the number of terms in the approximation. All these observations were done solving low-dimensional (mostly one-dimensional) problems.

In the experiments described we observed the same phenomena in very high-dimensional space.

5.13.2 SRM Principle and the Problem of Feature Construction

The “smart” linear combination of the large number of features used in the SV machine has an important structure: The set of support vectors. We can describe this structure as follows: Along with the set of weak features (weak feature space) there exists a set of complex features associated with support vectors. Let us denote this space by

$$u = (K(x, x_1), \dots, K(x, x_N)) \in U,$$

where

$$x_1, \dots, x_N$$

are the support vectors. In the space of complex features U , we constructed a linear decision rule. Note that in the bound obtained in Theorem 5.2 the expectation of the number of complex features plays the role of the dimensionality of the problem. Therefore, one can describe the difference between the support vector approach and the classical approach in the following way:

To perform the classical approach well requires the human selection (construction) of a relatively small number of “smart features,” while the support vector approach selects (constructs) a small number of “smart features” automatically.

Note that the SV machines construct the optimal hyperplane in the space Z (space of weak features) but not in the space of complex features. It is easy, however, to find the coefficients that provide optimality for the hyperplane in the space U (after the complex features are chosen). Moreover, one can construct in the U space a new SV machine (using the same training data). Therefore, one can construct a two- (or several-) layer SV machine. In other words, one can suggest a multistage selection of “smart features.” As we remarked in Section 4.10, the problem of feature selection is, however, quite delicate (recall the difference between constructing sparse algebraic polynomials and sparse trigonometric polynomials).

5.13.3 Is the Set of Support Vectors a Robust Characteristic of the Data?

In our experiments we observed an important phenomenon: Different types of SV machines optimal in parameters use almost the same support vectors. There exists a small subset of the training data (in our experiments less than 3% to 5% of the data) that for the problem of constructing the best decision rule is equivalent to the complete set of training data, and this subset of the training data is almost the same for different types of optimal SV machines (polynomial machine with the best degree of polynomials, RBF machine with the best parameter γ , and NN machine with the best parameter b).

The important question is whether this is true for a wide set of real-life problems. There exists indirect theoretical evidence that this is quite possible. One can show that if a majority vote scheme, based on various support vector machines, does not improve performance, then the percentage of common support vectors of these machines must be high.

It is too early to discuss the properties of SV machines: The analysis of these properties has now just started.²³ Therefore, I would like to finish

²³After this book had been completed, C. Burges demonstrated that one can approximate the obtained decision rule

$$f(x) = \text{sign} \left\{ \sum_{i=1}^N \alpha_i K(x, x_i) + \alpha_0 \right\}$$

by the much simpler decision rules

$$f^*(x) = \text{sign} \left\{ \sum_{i=1}^M \beta_i K(x, T_i) + \beta_0 \right\}, \quad M \ll N,$$

using the so-called generalized support vectors T_1, \dots, T_M (a specially constructed set of vectors).

these comments with the following remark.

The SV machine is a very suitable object for theoretical analysis. It unifies various conceptual models:

- (i) The SRM model. (That is how the SV machine initially was obtained. Theorem 5.1.)
- (ii) The data compression model. (The bound in Theorem 5.2 can be described in terms of the compression coefficient.)
- (iii) A universal model for constructing complex features. (The convolution of the inner product in Hilbert space can be considered as a standard method for feature construction.)
- (iv) A model of real-life data. (A small set of support vectors might be sufficient to characterize the whole training set for different machines.)

In a few years it will be clear whether such unification of models reflects some intrinsic properties of learning mechanisms or whether it is the next cul-de-sac.²⁴

To obtain approximately the same performance for the digit recognition problem, described in Section 5.7, it was sufficient to use an approximation based on $M = 11$ generalized support vectors per classifier instead of $N = 270$ (initially obtained) support vectors per classifier.

This means that for support vector machines there exists a regular way to synthesize the decision rules possessing optimal complexity.

²⁴Four years have passed since this remark was made in 1995. Since then we have had a lot of evidence, including experimental evidence (see, for example, Section 5.7) that the SV method is a general approach to various problems of function estimation in high-dimensional spaces.

Chapter 6

Methods of Function Estimation

In this chapter we generalize results obtained for estimating indicator function (for the pattern recognition problem) to the problem of estimating real-valued functions (regressions). We introduce a new type of loss function (the so-called ε -insensitive loss function) that makes our estimates not only robust but also sparse. As we will see, in this and in the next chapter, the sparsity of the solution is very important for estimating dependencies in high-dimensional spaces using a large number of data.

6.1 ε -INSENSITIVE LOSS FUNCTIONS

In Chapter 1, Section 1.7, to describe the problem of estimation of the supervisor rule $F(y|x)$ in the class of real-valued functions $\{f(x, \alpha), \alpha \in \Lambda\}$ we considered a quadratic loss function

$$L(y, f(x, \alpha)) = (y - f(x, \alpha))^2. \quad (6.1)$$

Under conditions where y is the result of measuring a regression function with normal additive noise ξ the ERM principle provides (for this loss function) an efficient (best unbiased) estimator of the regression $f(x, \alpha_0)$.

It is known, however, that if additive noise is generated by other distributions, better approximations to the regression (for the ERM principle) give estimators based on other loss functions (associated with these distributions)

$$L(y, f(x, \alpha)) = L(|y - f(x, \alpha)|) \quad (6.2)$$

$(L(\xi) = -\ln p(\xi)$ for the symmetric density function $p(\xi))$.

In 1964, Huber developed a theory that allows finding the best strategy for choosing the loss function using only general information about the model of the noise. In particular, he showed that if one knows only that the density describing the noise is a symmetric function, then the best minimax strategy for regression approximation (the best L_2 approximation for the worst possible model of noise $p(x)$) provides the loss function

$$L(y, f(x, \alpha)) = |y - f(x, \alpha)|. \quad (6.3)$$

Minimizing the empirical risk with respect to this loss function is called the *least modulus* method. It belongs to the so-called *robust regression* family. This, however, is an extreme case where one has minimal information about the unknown density. Huber also considered the model based on mixture of some fixed noise (below we consider the normal noise) with an arbitrary noise that is described by a symmetric continuous density function. He showed that the optimal (minimax strategy) for this type of noise is achieved when one uses the following loss function:

$$L(|y - f(x, \alpha)|) = \begin{cases} c|y - f(x, \alpha)| - \frac{c^2}{2} & \text{for } |y - f(x, \alpha)| > c, \\ \frac{1}{2}|y - f(x, \alpha)|^2 & \text{for } |y - f(x, \alpha)| \leq c. \end{cases} \quad (6.4)$$

The constant c is defined by the proportion of the mixture.

To construct an SVM for real-valued functions we use a new type of loss functions, the so-called ε -insensitive loss functions

$$L(y, f(x, \alpha)) = L(|y - f(x, \alpha)|_\varepsilon), \quad (6.5)$$

where we set

$$|y - f(x, \alpha)|_\varepsilon = \begin{cases} 0, & \text{if } |y - f(x, \alpha)| \leq \varepsilon, \\ |y - f(x, \alpha)| - \varepsilon, & \text{otherwise.} \end{cases} \quad (6.6)$$

These loss functions describe the ε -insensitive model: The loss is equal to 0 if the discrepancy between the predicted and the observed values is less than ε . It coincides with Huber's loss functions when $\varepsilon = 0$ and is close to loss function (6.4) when c is small.

Below we consider three loss functions:

1. The linear ε -insensitive loss function

$$L(y - f(x, \alpha)) = |y - f(x, \alpha)|_\varepsilon \quad (6.7)$$

(it coincides with the robust loss function (6.3) if $\varepsilon = 0$).

2. The quadratic ε -insensitive loss function

$$L(y - f(x, \alpha)) = |y - f(x, \alpha)|_\varepsilon^2 \quad (6.8)$$

(it coincides with the quadratic loss function (6.1) if $\varepsilon = 0$).

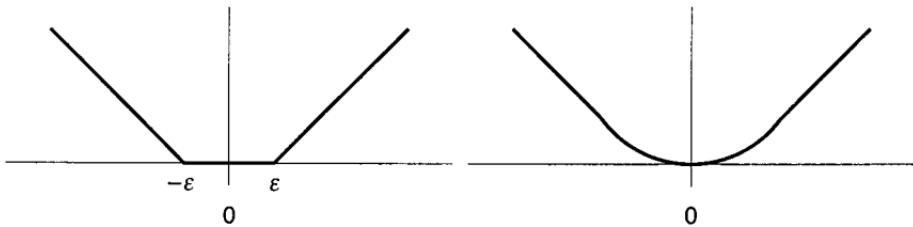


FIGURE 6.1. ε -insensitive linear loss function and Huber's loss function.

3. The Huber loss function

$$L(|y - f(x, \alpha)|) = \begin{cases} c|y - f(x, \alpha)| - \frac{c^2}{2} & \text{for } |y - f(x, \alpha)| > c, \\ \frac{1}{2}|y - f(x, \alpha)|^2 & \text{for } |y - f(x, \alpha)| \leq c. \end{cases} \quad (6.9)$$

Using the same technique one can consider any convex loss function $L(u)$. However, the above three are special: They lead to the same simple optimization task that we used for the pattern recognition problem.

6.2 SVM FOR ESTIMATING REGRESSION FUNCTION

The support vector approximation to regression takes place if:

- (i) One estimates the regression in the set of linear functions

$$f(x, \alpha) = (w \cdot x) + b.$$

- (ii) One defines the problem of regression estimation as the problem of risk minimization with respect to an ε -insensitive ($\varepsilon \geq 0$) loss function (6.8).
- (iii) One minimizes the risk using the *SRM principle*, where elements of the structure S_n are defined by the inequality

$$(w \cdot w) \leq c_n. \quad (6.10)$$

1. Solution for a given element of the structure. Suppose we are given training data

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

Then the problem of finding the w_ℓ and b_ℓ that minimize the empirical risk

$$R_{\text{emp}}(w, b) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - (w \cdot x_i) - b|_{\varepsilon}$$

under constraint (6.10) is equivalent to the problem of finding the pair w , b that minimizes the quantity defined by slack variables ξ_i , ξ_i^* , $i = 1, \dots, \ell$,

$$F(\xi, \xi^*) = \sum_{i=1}^{\ell} \xi_i^* + \sum_{i=1}^{\ell} \xi_i, \quad (6.11)$$

under the constraints

$$\begin{aligned} y_i - (w \cdot x_i) - b &\leq \varepsilon + \xi_i^*, \quad i = 1, \dots, \ell, \\ (w \cdot x_i) + b - y_i &\leq \varepsilon + \xi_i, \quad i = 1, \dots, \ell, \\ \xi_i^* &\geq 0, \quad i = 1, \dots, \ell, \\ \xi_i &\geq 0, \quad i = 1, \dots, \ell, \end{aligned} \quad (6.12)$$

and constraint (6.10).

As before, to solve the optimization problem with constraints of inequality type one has to find a saddle point of the Lagrange functional

$$\begin{aligned} L(w, \xi^*, \xi; \alpha^*, \alpha, C^*, \gamma, \gamma^*) &= \sum_{i=1}^{\ell} (\xi_i^* + \xi_i) - \sum_{i=1}^{\ell} \alpha_i [y_i - (w \cdot x_i) - b + \varepsilon + \xi_i] \\ &- \sum_{i=1}^{\ell} \alpha_i^* [(w \cdot x_i) + b - y_i + \varepsilon + \xi_i^*] - \frac{C^*}{2} (c_n - (w \cdot w)) \\ &- \sum_{i=1}^{\ell} (\gamma_i^* \xi_i^* + \gamma_i \xi_i) \end{aligned} \quad (6.13)$$

(minimum with respect to elements w , b , ξ_i^* , and ξ_i and maximum with respect to Lagrange multipliers $C^* \geq 0$, $\alpha_i^* \geq 0$, $\alpha_i \geq 0$, $\gamma_i^* \geq 0$, and $\gamma_i \geq 0$, $i = 1, \dots, \ell$).

Minimization with respect to w , b , and ξ_i^* , ξ_i implies the following three conditions:

$$w = \sum_{i=1}^{\ell} \frac{\alpha_i^* - \alpha_i}{C^*} x_i, \quad (6.14)$$

$$\sum_{i=1}^{\ell} \alpha_i^* = \sum_{i=1}^{\ell} \alpha_i, \quad (6.15)$$

$$0 \leq \alpha_i^* \leq 1, \quad i = 1, \dots, \ell,$$

$$0 \leq \alpha_i \leq 1, \quad i = 1, \dots, \ell, \quad (6.16)$$

Putting (6.14) and (6.15) into (6.13) one obtains that for the solution of this optimization problem, one has to find the maximum of the convex functional

$$\begin{aligned} W(\alpha, \alpha^*, C^*) = & -\varepsilon \sum_{i=1}^{\ell} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ & - \frac{1}{2C^*} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) - \frac{c_n C^*}{2} \end{aligned} \quad (6.17)$$

subject to constraints (6.15), (6.16), and the constraint

$$C^* \geq 0.$$

As in pattern recognition, here only some of the parameters in expansion (6.14),

$$\beta_i = \frac{\alpha_i^* - \alpha_i}{C^*}, \quad i = 1, \dots, \ell,$$

differ from zero. They define the support vectors of the problem.

2. The basic solution. One can reduce the convex optimization problem of finding the vector w to a quadratic optimization problem if instead of minimizing the functional (6.11), subject to constraints (6.12) and (6.10), one minimizes

$$\Phi(w, \xi^*, \xi) = \frac{1}{2}(w \cdot w) + C \left(\sum_{i=1}^{\ell} \xi_i^* + \sum_{i=1}^{\ell} \xi_i \right)$$

(with given value C) subject to constraints (6.12). In this case to find the desired vector

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

one has to find coefficients α_i^* , α_i , $i = 1, \dots, \ell$, that maximize the quadratic form

$$W(\alpha, \alpha^*) = -\varepsilon \sum_{i=1}^{\ell} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) \quad (6.18)$$

subject to the constraints

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

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

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

As in the pattern recognition case, the solutions to these two optimization problems coincide if $C = C^*$.

One can show that for any $i = 1, \dots, \ell$ the equality

$$\alpha_i^* \times \alpha_i = 0$$

holds true. Therefore, for the particular case where $\varepsilon = 1 - \delta$ (δ is small) and $y_i \in \{-1, 1\}$ these optimization problems coincide with those described for pattern recognition.

To derive the bound on the generalization of the SVM, suppose that the distribution $F(x, y) = F(y|x)F(x)$ is such that for any fixed w, b the corresponding distribution of the random variable $|y - (w \cdot x) - b|_\varepsilon$ has a "light tail" (see Section 3.4):

$$\sup_{w,d} \frac{(E|y - (w \cdot x) - b|)_\varepsilon^{1/p}}{E|y - (w \cdot x) - b|_\varepsilon} \leq \tau, \quad p > 2.$$

Then according to equation (3.30) one can assert that the solution w_ℓ, b_ℓ of the optimization problem provides a risk (with respect to the chosen loss function) such that with probability at least $1 - \eta$ the bound

$$R(w_\ell, b_\ell) \leq \varepsilon + \frac{R_{\text{emp}}(w_\ell, b_\ell) - \varepsilon}{\left(1 - a(p)\tau\sqrt{\mathcal{E}}\right)_+}$$

holds true, where

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

and

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

Here h_n is the VC dimension of the set of functions

$$S_n = \{|y - (w \cdot x) - b|_\varepsilon : (w \cdot w) \leq c_n\}.$$

6.2.1 SV Machine with Convolved Inner Product

Using the same argument with mapping input vectors into high-dimensional space that was considered for the pattern recognition case in Chapter 5 one can construct the best approximation of the form

$$f(x; v, \beta) = \sum_{i=1}^N \beta_i K(x, v_i) + b, \quad (6.19)$$

where β_i , $i = 1, \dots, N$, are scalars, v_i , $i = 1, \dots, N$, are vectors, and $K(\cdot, \cdot)$ is a given function satisfying Mercer's conditions.

1. Solution for a given element of the structure. Using the convex optimization approach one evaluates coefficients β_i , $i = 1, \dots, \ell$, in (6.19) as

$$\beta_i = \frac{\alpha_i^* - \alpha_i}{C^*}, \quad i = 1, \dots, \ell,$$

where α_i^* , α_i , C are parameters that maximize the function

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

subject to the constraint

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

and to the constraints

$$0 \leq \alpha_i^* \leq 1, \quad 0 \leq \alpha_i \leq 1, \quad i = 1, \dots, \ell,$$

$$0 \leq \alpha_i \leq 1, \quad i = 1, \dots, \ell,$$

and

$$C^* \geq 0.$$

2. The basic solution. Using the quadratic optimization approach one evaluates the vector w (5.48) with coordinates

$$\beta_i = \alpha_i^* - \alpha_i, \quad i = 1, \dots, \ell,$$

where α_i^* , α_i are parameters that maximize the function

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

subject to the constraint

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

and to the constraints

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

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

By controlling the two parameters C and ε in the quadratic optimization approach one can control the generalization ability, even of the SVM in a high-dimensional space.

6.2.2 Solution for Nonlinear Loss Functions

Along with linear loss functions one can obtain the solution for convex loss functions $L(\xi_i^*)$, $L(\xi_i)$.

In general, when $L(\xi)$ is a concave function, one can find the solution using the corresponding optimization technique. However, for a quadratic loss function $L(\xi) = \xi^2$ or Huber's loss function one can obtain a solution using a simple quadratic optimization technique.

1. Quadratic loss function. To find the solution (coefficients of expansion α_i^* , α_i of the hyperplane on support vectors) one has to maximize the quadratic form

$$W(\alpha, \alpha^*) = - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i)$$

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

subject to the constraints

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

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

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

When $\varepsilon = 0$ and

$$K(x_i, x_j) = \text{Cov}\{f(x_i), f(x_j)\}$$

is the covariance function of stochastic processes with

$$Ef(x) = 0,$$

the obtained solution coincides with the so-called *kriging* method developed in geostatistics (see Matheron, 1987).

2. Solution for the Huber loss function. Lastly, consider the SVM for the Huber loss function

$$F(\xi) = \begin{cases} c|\xi| - \frac{c^2}{2} & \text{for } |\xi| \leq c, \\ \frac{1}{2}\xi^2 & \text{for } |\xi| > c. \end{cases}$$

For this loss function, to find the desired function

$$\phi(x) = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) K(x_i * x) + b$$

one has to find the coefficients α_i^*, α_i that maximize the quadratic form

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

subject to the constraints

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

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

When $c = \varepsilon < 1$, the solution obtained for the Huber loss function is close to the solution obtained for the ε -insensitive loss function. However, the expansion of the solution for the ε -insensitive loss function uses fewer support vectors.

3. Spline approximation of the loss functions. If $F(\xi)$ is a concave function that is symmetric with respect to zero then one can approximate it to any degree of accuracy using linear splines

$$F(\xi) = \sum_{k=1}^n c_k (\xi - a_k)_+, \quad 0 < a_1 = \varepsilon < a_2 < \dots < a_n.$$

In this case using the same technique that was used in pattern recognition for SVM logistic regression approximation one can obtain the solution on the basis of the quadratic optimization technique.

6.2.3 Linear Optimization Method

As in the pattern recognition case one can simplify the optimization problem even more by reducing it to a linear optimization task. Suppose we are given data

$$(y_i, x_i), \dots, (x_\ell, x_\ell).$$

Let us approximate functions using functions from the set

$$y(x) = \sum_{j=1}^{\ell} \beta_j K(x_j, x) + b,$$

where β_i is some real value, x_i is a vector from a training set, and $K(x_i, x)$ is a kernel function. We call the vectors from the training set that correspond to nonzero β_i the support vectors. Let us rewrite β_i in the form

$$\beta_i = \alpha_i^* - \alpha_i,$$

where $\alpha_i^* > 0$, $\alpha_i > 0$.

One can use as an approximation the function that minimizes the functional

$$W(\alpha, \xi_i) = \sum_{i=1}^{\ell} \alpha_i + \sum_{i=1}^{\ell} \alpha_i^* + C \sum_{i=1}^{\ell} \xi_i + C \sum_{i=1}^{\ell} \xi_i^*$$

subject to the constraints

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

$$\xi_i \geq 0, \quad \xi_i^* \geq 0,$$

$$y_i - \sum_{j=1}^{\ell} (\alpha_j^* - \alpha_j) K(x_i, x_j) - b \leq \varepsilon - \xi_i^*,$$

$$\sum_{j=1}^{\ell} (\alpha_j^* - \alpha_j) K(x_i, x_j) + b - y_i \leq \varepsilon - \xi_i.$$

The solution to this problem requires only linear optimization techniques.

6.3 CONSTRUCTING KERNELS FOR ESTIMATING REAL-VALUED FUNCTIONS

To construct different types of SVM one has to choose different kernels $K(x, x_i)$ satisfying Mercer's condition.

In particular, one can use the same kernels that were used for approximation of indicator functions:

- (i) kernels generating polynomials

$$K(x, x_i) = [(x * x^*) + 1]^d,$$

- (ii) kernels generating radial basis functions

$$K(x, x_i) = K(|x - x_i|),$$

for example

$$K(|x - x_i|) = \exp \{-\gamma |x - x_i|^2\},$$

- (iii) kernels generating two-layer neural networks

$$K(x, x_i) = S(v(x * x_i) + c).$$

On the basis of these kernels one can obtain the approximation

$$f(x, \alpha_0) = \sum_{i=1}^{\ell} \beta_i K(x, x_i) + b \quad (6.20)$$

using the optimization techniques described above.

These kernels imply approximating functions $f(x, \alpha)$ that were used in the pattern recognition problem under discrimination sign; namely, we considered functions $\text{sign}[f(x, \alpha)]$.

However, the problem of approximation of real-valued functions is more delicate than the approximation of indicator functions (the absence of $\text{sign}\{\cdot\}$ in front of function $f(x, \alpha)$ significantly changes the problem of approximation).

Various real-valued function estimation problems need various sets of approximating functions. Therefore, it is important to construct special kernels that reflect special properties of approximating functions.

To construct such kernels we will use two main techniques:

- (i) constructing kernels for approximating one-dimensional functions, and
- (ii) composition of multidimensional kernels using one-dimensional kernels.

6.3.1 Kernels Generating Expansion on Orthogonal Polynomials

To construct kernels that generate expansion of one-dimensional functions in the first N terms of the orthonormal polynomials $P_i(x), i = 1, \dots, N$

(Chebyshev, Legendre, Hermite polynomials, etc.), one can use the Christoffel-Darboux formula

$$\begin{aligned} K_n(x, y) &= \sum_{k=1}^n P_k(x)P_k(y) = a_n \frac{P_{n+1}(x)P_n(y) - P_n(x)P_{n+1}(y)}{x - y}, \\ K_n(x, x) &= \sum_{k=1}^n P_k^2(x) = a_n [P'_{n+1}(x)P_n(x) - P'_n(x)P_{n+1}(x)], \end{aligned} \quad (6.21)$$

where a_n is a constant that depends on the type of polynomial and the number n of elements in the orthonormal basis.

It is clear, however, that with increasing n the kernels $K(x, y)$ approach the δ -function. However, we can modify the generating kernels to reproduce a regularized function. Consider the kernel

$$K(x, y) = \sum_{i=1}^{\infty} r_i \psi_i(x) \psi_i(y), \quad (6.22)$$

where r_i converges to zero as i increases. This kernel defines a regularized expansion on polynomials.

We can choose values r_i such that they improve the convergence properties of the series (6.22). For example, we can choose $r_i = q^i$, $0 \leq q \leq 1$.

Example. Consider the (one-dimensional) Hermite polynomials

$$H_k(x) = \mu_k P_k(x) e^{-x^2}, \quad (6.23)$$

where

$$P_k(x) = (-1)^k e^{x^2} \left(\frac{d}{dx} \right)^k e^{-x^2}$$

and μ_k are normalization constants.

For these polynomials one can obtain the kernels

$$\begin{aligned} K(x, y) &= \sum_{i=0}^{\infty} q^i H_i(x) H_i(y) \\ &= \frac{1}{\sqrt{\pi(1-q^2)}} \exp \left\{ \frac{2xyq}{1+q} - \frac{(x-y)^2 q^2}{1-q^2} \right\} \end{aligned} \quad (6.24)$$

(Mikhlin (1964)). From (6.24) one can see that the closer q is to one, the closer the kernel $K(x, y)$ is to the δ -function.

To construct our kernels we do not even need to use orthonormal bases. In the next section, to construct kernels for spline approximations we will use linearly independent bases that are not orthogonal.

Such generality (any linearly independent system with any smoothing parameters) opens wide opportunities to construct kernels for SVMs.

6.3.2 Constructing Multidimensional Kernels

Our goal, however, is to construct kernels for approximating multidimensional functions defined on the vector space $X \subset R^n$ where all coordinates of the vector $x = (x^1, \dots, x^n)$ are defined on the same finite or infinite interval I .

Suppose now that for any coordinate x^k the complete orthonormal basis $b_{i_k}(x^k)$, $i = 1, 2, \dots$, is given. Consider the set of basis functions

$$b_{i_1, i_2, \dots, i_n}(x^1, \dots, x^n) = b_{i_1}(x^1)b_{i_2}(x^2) \cdots b_{i_n}(x^n) \quad (6.25)$$

in n -dimensional space. These functions are constructed from the coordinatewise basis functions by direct multiplication (tensor products) of the basis functions, where all indices i_k take all possible integer values from 0 to ∞ . It is known that the set of functions (6.25) is a complete orthonormal basis in $X \subset R^n$.

Now let us consider the more general situation where a (finite or infinite) set of coordinatewise basis functions is not necessarily orthonormal. Consider as a basis of n dimensional space the tensor products of the coordinatewise basis.

For this structure of multidimensional spaces the following theorem is true.

Theorem 6.1. *Let a multidimensional set of functions be defined by the basis functions that are tensor products of the coordinatewise basis functions. Then the kernel that defines the inner product in the n -dimensional basis is the product of one-dimensional kernels.*

Continuation of example. Now let us construct a kernel for the regularized expansion on n -dimensional Hermite polynomials. In the example discussed above we constructed a kernel for one-dimensional Hermite polynomials. According to Theorem 6.1 if we consider as a basis of n -dimensional space the tensor product of one-dimensional basis functions, then the kernel for generating the n -dimensional expansion is the product of n one-dimensional kernels

$$\begin{aligned} K(x, y) &= \prod_{i=1}^n \frac{1}{\sqrt{\pi(1-q^2)}} \exp \left\{ \frac{2x^i y^i q}{1+q} - \frac{(x^i - y^i)^2 q^2}{1-q^2} \right\} \\ &= \frac{1}{(1-q^2)^{n/2}} \exp \left\{ \frac{2(x * y)q}{1+q} - \frac{|x-y|^2 q^2}{1-q^2} \right\} \end{aligned} \quad (6.26)$$

Thus, we have obtained a kernel for constructing semilocal approximations

$$K(x, y) = C \exp \{2(x * y)\delta\} \exp \{-|x-y|^2 \sigma^2\}, \quad \delta, \sigma > 0, \quad (6.27)$$

where the factor containing the inner product of two vectors defines a “global” approximation, since the Gaussian defines the vicinity of approximation.

6.4 KERNELS GENERATING SPLINES

Below we introduce the kernels that can be used to construct a spline approximation of high-dimensional functions. We will construct splines with both a fixed number of nodes and with an infinite number of nodes. In all cases the computational complexity of the solution depends on the number of support vectors that one needs to approximate the desired function with ϵ -accuracy, rather than on the dimensionality of the space or on the number of nodes.

6.4.1 Spline of Order d With a Finite Number of Nodes

Let us start with describing the kernel for the approximation of one-dimensional functions on the interval $[0, a]$ by splines of order $d \geq 0$ with m nodes,

$$(t_1, \dots, t_m), \quad t_i = \frac{ia}{m}, \quad i = 1, \dots, m.$$

By definition, spline approximations have the form

$$f(x) = \sum_{r=0}^d a_r^* x^r + \sum_{i=1}^m a_i (x - t_i)_+^d. \quad (6.28)$$

Consider the following mapping of the one-dimensional variable x into an $(m + d + 1)$ -dimensional vector u :

$$x \longrightarrow u = (1, x, \dots, x^d, (x - t_1)_+^d, \dots, (x - t_m)_+^d),$$

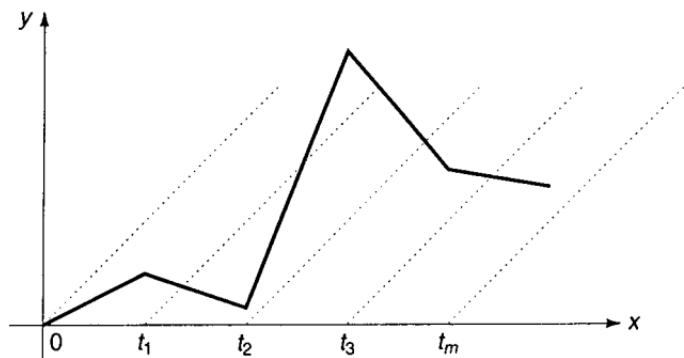


FIGURE 6.2. Using an expansion on the functions $1, x, (x - t_1)_+, \dots, (x - t_m)_+$ one can construct a piecewise linear approximation of a function. Analogously an expansion on the functions $1, x, \dots, x^d, (x - t_1)_+^d, \dots, (x - t_m)_+^d$ provides piecewise polynomial approximation.

where we set

$$(x - t_k)_+^d = \begin{cases} 0 & \text{if } x \leq t_k, \\ (x - t_k)^d & \text{if } x > t_k. \end{cases}$$

Since spline approximation (6.28) can be considered as the inner product of two vectors,

$$f(x) = (a * u)$$

(where $a = (a_0, \dots, a_{m+d})$), one can define the kernel that generates the inner product in feature space as follows:

$$K(x, x_t) = (u * u_t) = \sum_{r=0}^d x^r x_i^r + \sum_{i=1}^m (x - t_i)_+^d (x_t - t_i)_+^d. \quad (6.29)$$

Using the generating kernel (6.29) the SVM constructs the function

$$f(x, \beta) = \sum_{i=1}^{\ell} \beta_i K(x, x_i) + b,$$

that is, a spline of order d defined on m nodes.

To construct kernels generating splines in n -dimensional spaces note that n -dimensional splines are defined as an expansion on the basis functions that are tensor products of one-dimensional basis functions. Therefore, according to Theorem 6.1, kernels generating n -dimensional splines are the product of n one-dimensional kernels:

$$K(x, x_i) = \prod_{k=1}^n K(x^k, x_i^k),$$

where we have set $x = (x^1, \dots, x^k)$.

6.4.2 Kernels Generating Splines With an Infinite Number of Nodes

In applications of SVMs the number of nodes does not play an important role (more important are the values of ε_i). Therefore, to simplify the calculation, we use splines with an infinite number of nodes defined on the interval $(0, a)$, $0 < a < \infty$, as the expansion

$$f(x) = \sum_{i=0}^d a_i x^i + \int_0^a a(t)(x - t)_+ dt,$$

where a_i , $i = 0, \dots, d$, are unknown values and $a(t)$ is an unknown function that defines the expansion. One can consider this expansion as an inner product. Therefore, one can construct the following kernel for

generating splines of order d with an infinite number of nodes and then use the following inner product in this space:

$$\begin{aligned}
 K(x_j, x_i) &= \int_0^a (x_j - t)_+^d (x_i - t)_+^d dt + \sum_{r=0}^d x_j^r x_i^r \\
 &= \int_0^{(x_j \wedge x_i)} (x_j - t)^d (x_i - t)^d dt + \sum_{r=0}^d x_j^r x_i^r \\
 &= \int_0^{(x_j \wedge x_i)} u^d (u + |x_j - x_i|)^d du + \sum_{r=1}^d x_j^r x_i^r \\
 &= \sum_{r=0}^d \frac{C_d^r}{2d-r+1} (x_j \wedge x_i)^{2d-r+1} |x_j - x_i|^r + \sum_{r=0}^d x_j^r x_i^r,
 \end{aligned} \tag{6.30}$$

where we set $\min(x, x_i) = (x \wedge x_i)$. In particular, for a linear spline ($d = 1$) we have

$$K_1(x_j, x_i) = 1 + x_j x_i + \frac{1}{2} |x_j - x_i| (x_j \wedge x_i)^2 + \frac{(x_j \wedge x_i)^3}{3}.$$

Again the kernel for n -dimensional splines with an infinite number of nodes is the product of n kernels for one-dimensional splines.

On the basis of this kernel one can construct a spline approximation (using the techniques described in the previous section) that has the form

$$f(x, \beta) = \sum_{i=1}^{\ell} \beta_i K(x, x_i).$$

6.5 KERNELS GENERATING FOURIER EXPANSIONS

An important role in signal processing belongs to Fourier expansions. In this section we construct kernels for Fourier expansions in multidimensional spaces. As before, we start with the one-dimensional case.

Suppose we would like to analyze a one-dimensional signal in terms of Fourier series expansions.

Let us map the input variable x into the $(2N + 1)$ -dimensional vector

$$u = (1/\sqrt{2}, \sin x, \dots, \sin Nx, \cos x, \dots, \cos Nx).$$

Then for any fixed x the Fourier expansion can be considered as the inner product in this $(2N + 1)$ -dimensional feature space

$$f(x) = (a * u) = \frac{a}{\sqrt{2}} + \sum_{k=1}^N (a_k \sin kx + b_k^* \cos kx). \tag{6.31}$$

Therefore, the inner product of two vectors in this space has the form

$$K_N(x, x_i) = \frac{1}{2} + \sum_{k=1}^N (\sin kx \sin kx_i + \cos kx \cos kx_i).$$

After obvious transformations and taking into account the Dirichlet function we obtain

$$K_N(x, x_i) = \frac{1}{2} + \sum_{k=1}^N \cos k(x - x_i) = \frac{\sin \frac{(2N+1)}{2}(x - x_i)}{\sin \frac{(x - x_i)}{2}}.$$

To define the signal in terms of the Fourier expansion, the SVM uses the representation

$$f(x, \beta) = \sum_{i=1}^{\ell} \beta_i K_N(x, x_i).$$

Again, to construct the SVM for the d -dimensional vector space $x = (x^1, \dots, x^n)$, it is sufficient to use the generating kernel that is the product of one-dimensional kernels

$$K(x, x_i) = \prod_{k=1}^n K(x^k, x_i^k).$$

6.5.1 Kernels for Regularized Fourier Expansions

It is known, however, that Fourier expansions do not possess good approximation properties. Therefore, below we introduce two regularizing kernels, which we use for approximation of multidimensional functions with SVMs.

Consider the following regularized Fourier expansion:

$$f(x) = \frac{a_0}{\sqrt{2}} + \sum_{k=1}^{\infty} q^k (a_k \cos kx + b_k \sin kx), \quad 0 < q < 1,$$

where a_k, b_k are coefficients of the Fourier expansion. This expansion differs from expansion (6.31) by factors q^k that provide regularization. The corresponding kernel for this regularizing expansion is

$$K(x_i, x_j) = \frac{1}{2} + \sum_{k=1}^{\infty} q^k (\cos kx_i \cos kx_j + \sin kx_i \sin kx_j)$$

$$= \frac{1}{2} + \sum_{k=1}^{\infty} q^k \cos k(x_i - x_j) = \frac{1 - q^2}{2(1 - 2q \cos(x_i - x_j) + q^2)}. \quad (6.32)$$

(For the last equality see Gradshteyn and Ryzhik (1980).) Another type of

regularization was obtained using the following regularization of the Fourier expansion:

$$f(x) = \frac{a_0}{\sqrt{2}} + \sum_{k=1}^{\infty} \frac{a_k \cos kx + b_k \sin kx}{1 + \gamma^2 k^2},$$

where a_k, b_k are coefficients of the Fourier expansion. For this type of regularized Fourier expansion we have the following kernel:

$$\begin{aligned} K(x_i, x_j) &= \frac{1}{2} + \sum_{k=1}^{\infty} \frac{\cos kx_i \cos kx_j + \sin kx_i \sin kx_j}{1 + \gamma^2 k^2} \\ &= \frac{\pi}{2\gamma} \frac{\operatorname{ch} \frac{\pi - |x_i - x_j|}{\gamma}}{\operatorname{sh} \frac{\pi}{\gamma}}, \quad 0 \leq |x_i - x_j| \leq 2\pi. \end{aligned} \quad (6.33)$$

(For last equality see Gradshteyn and Ryzhik (1980).)

Again the kernel for a multidimensional Fourier expansion is the product of the kernels for one-dimensional Fourier expansions.

6.6 THE SUPPORT VECTOR ANOVA DECOMPOSITION (SVAD) FOR FUNCTION APPROXIMATION AND REGRESSION ESTIMATION

The kernels defined in the previous sections can be used both for approximating multidimensional functions and for estimating multidimensional regression. However, they can define too rich a set of functions. Therefore, to control generalization one needs to make a structure on this set of functions, in order to choose the function from an appropriate element of the structure. Note also that when the dimensionality of the input space is large (say 100), the values of an n -dimensional kernel (which is the product of n

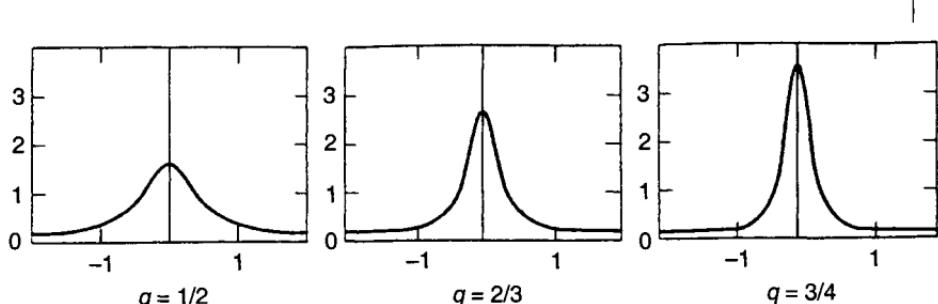


FIGURE 6.3. Kernels for a strong mode of regularization with various q .

one-dimensional kernels) can have order of magnitude q^n . These values are inappropriate for both cases $q > 1$ and $q < 1$.

Classical statistics considered the following structure on the set of multi-dimensional functions from L_2 , the so-called ANOVA decomposition (acronym for “analysis of variances”).

Suppose that an n -dimensional function $f(x) = f(x^1, \dots, x^n)$ is defined on the set $I \times I \times \dots \times I$, where I is a finite or infinite interval.

The ANOVA decomposition of the function $f(x)$ is an expansion

$$f(x^1, \dots, x^n) = F_0 + F_1(x^1, \dots, x^n) + F_2(x^1, \dots, x^n) + \dots + F_n(x^1, \dots, x^n),$$

where

$$F_0 = C,$$

$$F_1(x^1, \dots, x^n) = \sum_{1 \leq k \leq n} \phi_k(x^k),$$

$$F_2(x^1, \dots, x^n) = \sum_{1 \leq k_1 < k_2 \leq n} \phi_{k_1, k_2}(x^{k_1}, x^{k_2}),$$

...

$$F_r(x^1, \dots, x^n) = \sum_{1 \leq k_1 < k_2 < \dots < k_r \leq n} \phi_{k_1, \dots, k_r}(x^{k_1}, x^{k_2}, \dots, x^{k_r}),$$

$$F_n(x^1, \dots, x^n) = \phi_{k_1, \dots, k_n}(x^1, \dots, x^n).$$

The classical approach to the ANOVA decompositions has a problem with exponential explosion of the number of summands with increasing order of approximation. In support vector techniques we do not have this problem. To construct the kernel for the ANOVA decomposition of order p using a sum of products of one-dimensional kernels $K(x^i, x_r^i)$, $i = 1, \dots, n$,

$$K_p(x, x_r) = \sum_{1 \leq i_1 < i_2 < \dots < i_p \leq n} K(x^{i_1}, x_r^{i_1}) \times \dots \times K(x^{i_p}, x_r^{i_p}),$$

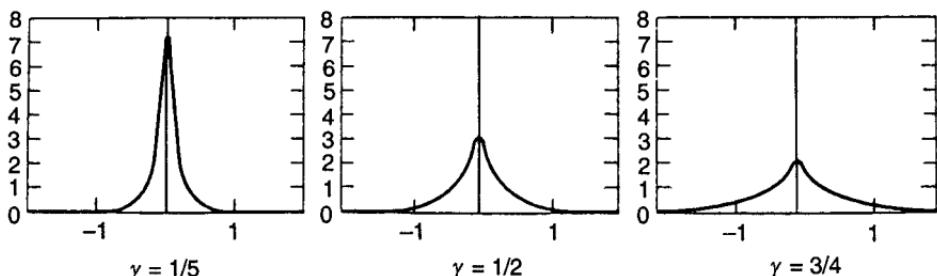


FIGURE 6.4. Kernels for a weak mode of regularization with various γ .

one can introduce a recurrent procedure for computing $K_p(x, x_r)$, $p = 1, \dots, n$.

Let us define

$$K^s(x, x_r) = \sum_{i=1}^n K^s(x^i, x_r^i).$$

One can easily check that the following recurrent procedure defines the kernels $K_p(x, x_r)$, $p = 1, \dots, n$:

$$\begin{aligned} K_0(x, x_r) &= 1, \\ K_1(x, x_r) &= \sum_{1 \leq i \leq n} K(x^i, x_r^i) = K^1(x, x_r), \\ K_2(x, x_r) &= \sum_{1 \leq i_1 < i_2 \leq n} K(x^{i_1}, x_r^{i_1}) K(x^{i_2}, x_r^{i_2}) \\ &= \frac{1}{2} [K_1(x, x_r) K^1(x, x_r) - K^2(x, x_r)], \\ K_3(x, x_r) &= \sum_{1 \leq i_1 < i_2 < i_3 \leq n} K_1(x^{i_1}, x_r^{i_1}) K_2(x^{i_2}, x_r^{i_2}) K(x^{i_3}, x_r^{i_3}) \\ &= \frac{1}{3} [K_2(x, x_r) K^1(x, x_r) - K_1(x, x_r) K^2(x, x_r) + K^3(x, x_r)]. \end{aligned}$$

In the general case we have¹

$$K_p(x, x_r) = \frac{1}{p} \sum_{s=1}^p (-1)^{s+1} K_{p-s}(x, x_r) K^s(x, x_r).$$

Using such kernels and the SVM with L_2 loss functions one can obtain an approximation of any order.

6.7 SVM FOR SOLVING LINEAR OPERATOR EQUATIONS

In this section we use the SVM for solving linear operator equations

$$Af(t) = F(x), \quad (6.34)$$

where the operator A realizes a one-to-one mapping from a Hilbert space E_1 into a Hilbert space E_2 .

¹“A new method for constructing artificial neural networks” Interim Technical Report ONR Contract N00014-94-C-0186 Data Item A002. May 1, 1995. Prepared by C. Burges and V. Vapnik.

We will solve equations in the situation where instead of a function $F(x)$ on the right-hand side of (6.34) we are given measurements of this function (generally with errors)

$$(x_1, F_1), \dots, (x_\ell, F_\ell). \quad (6.35)$$

It is necessary to estimate the solution of equation (6.34) from the data (6.35).

Below we will show that the support vector technique realizes the classical ideas of solving ill-posed problems where the choice of the kernel is equivalent to the choice of the regularization functional. Using this technique one can solve operator equations in high-dimensional spaces.

6.7.1 The Support Vector Method

In the next chapter we discuss the regularization method of solving operator equations, where in order to solve operator equation (6.34) one minimizes the functional

$$R_\gamma(f, F) = \rho^2(Af, F) + \gamma W(f),$$

where the solution belongs to some compact $W(f) \leq C$ (C is an unknown constant). When one solves operator equation (6.34) using data (6.35) one considers the functional

$$R_\gamma(f, F) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(Af(t)|_{x_i} - F_i) + \gamma(Pf * Pf)$$

with some loss function $L(Af - F)$ and regularizer of the form

$$W(f) = (Pf * Pf)$$

defined by some nongenerating operator P . Let

$$\varphi_1(t), \dots, \varphi_n(t), \dots,$$

$$\lambda_1, \dots, \lambda_n, \dots,$$

be eigenfunctions and eigenvalues of the self-conjugate operator $P^* P$:

$$P^* P \varphi_i = \lambda_i \varphi_i.$$

Consider the solution of equation (6.34) as the expansion

$$f(t) = \sum_{k=1}^{\infty} \frac{w_k}{\sqrt{\lambda_k}} \varphi_k(t).$$

Putting this expansion into the functional $R_\gamma(f, F)$, we obtain

$$R_\gamma(f, F) = \frac{1}{\ell} \sum_{i=1}^{\ell} L\left(A\left\{\sum_{k=1}^{\infty} \frac{w_k}{\sqrt{\lambda_k}} \varphi_k(t)\right\}|_{x_i} - F_i\right) + \gamma \sum_{k=1}^{\infty} w_k^2.$$

Writting

$$\phi_k(t) = \frac{\varphi_k(t)}{\sqrt{\lambda_k}},$$

we can rewrite our problem in a familiar form: Minimize the functional

$$R_\gamma(w, F) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(|A(w * \Phi(t))|_{x=x_i} - F_i) + \gamma(w * w)$$

in the set of functions

$$f(t, w) = \sum_{r=1}^{\infty} w_r \phi_r(t) = (w * \Phi(t)), \quad (6.36)$$

where we have set

$$\begin{aligned} w &= (w_1, \dots, w_N, \dots), \\ \Phi(t) &= (\phi_1(t), \dots, \phi_N(t), \dots). \end{aligned} \quad (6.37)$$

The operator A maps the set of functions (6.36) into the set of functions

$$F(x, w) = Af(t, w) = \sum_{r=1}^{\infty} w_r A\phi_r(t) = \sum_{r=1}^{\infty} w_r \psi_r(x) = (w * \Psi(x)), \quad (6.38)$$

linear in another feature space

$$\Psi(x) = (\psi_1(x), \dots, \psi_N(x), \dots),$$

where

$$\psi_r(x) = A\phi_r(t).$$

To find the solution of equation (6.34) in a set of functions $f(t, w)$ (to find the vector coefficients w) one can minimize the functional

$$D(F) = C \sum_{i=1}^{\ell} (|F(x_i, w) - F_i|_{\varepsilon})^k + (w * w), \quad k = 1, 2,$$

in the space of functions $F(x, w)$ (that is, in the image space) and then use the parameters w to define the solution (6.36) (in preimage space). To realize this idea we use along with the kernel function the so-called cross-kernel function. Let us define the generating kernel in the image space

$$K(x_i, x_j) = \sum_{r=1}^{\infty} \psi_r(x_i) \psi_r(x_j) \quad (6.39)$$

(here we suppose that the right-hand side converges for any fixed x_i and x_j) and the cross-kernel function

$$\mathcal{K}(x_i, t) = \sum_{r=1}^{\infty} \psi_r(x_i) \phi_r(t) \quad (6.40)$$

(here we also suppose that the operator A is such that the right-hand side converges for any fixed x and t).

Note that in the case considered the problem of finding the solution to the operator equation (finding the corresponding vector of coefficients w) is equivalent to the problem of finding the vector w for the linear regression function (6.38) in the image space using measurements (6.35).

Let us solve this regression problem using the quadratic optimization support vector technique. That is, using the kernel (6.39) one finds both the support vectors x_i , $i = 1, \dots, N$, and the corresponding coefficients $\alpha_i^* - \alpha_i$ that define the vector w for the support vector regression approximation

$$w = \sum_{i=1}^N (\alpha_i^* - \alpha_i) \Psi(x_i)$$

(to do this it is sufficient to use the standard quadratic optimization support vector technique). Since the same coefficients w define the approximation to the solution of the operator equation, one can put these coefficients in expression (6.36), obtaining

$$f(t, \alpha, \alpha^*) = \sum_{i=1}^N (\alpha_i^* - \alpha_i) (\Psi(x_i) * \Phi(t)) = \sum_{i=1}^N (\alpha_i^* - \alpha_i) K(x_i, t).$$

That is, we find the solution to our problem of solving the operator equation using the cross-kernel function as an expansion on support vectors.

Thus, in order to solve a linear operator equation using the support vector method one must:

1. Define the corresponding regression problem in image space.
2. Construct the kernel function $K(x_i, x_j)$ for solving the regression problem using the support vector method.
3. Construct the corresponding cross-kernel function $K(x_i, t)$.
4. Using the kernel function $K(x_i, x_j)$ solve the regression problem by the support vector method (i.e., find the support vectors x_i^* , $i = 1, \dots, N$, and the corresponding coefficients $\beta_i = \alpha_i^* - \alpha_i$, $i = 1, \dots, N$).
5. Using these support vectors and the corresponding coefficients define the solution

$$f(t) = \sum_{r=1}^N \beta_r K(x_r, t). \quad (6.41)$$

In these five steps the first three steps (constructing the regression, the constructing the kernel in image space, and constructing the corresponding cross-kernel function) reflect the singularity of the problem at hand (they

depend on the operator A). The last two steps (solving the regression problem by an SVM and constructing the solution to the desired problem) are routine.

The main problem with solving an operator equation using the support vector technique is for a given operator equation to obtain both the explicit expression for the kernel function in image space and an explicit expression for the corresponding cross-kernel function. For many problems such as the density estimation problem or the problem of solving Radon equation such functions are easy to find.

6.8 FUNCTION APPROXIMATION USING THE SVM

Consider examples of solving the function approximation problem using the SVM. With the required level of accuracy ε we approximate one- and two-dimensional functions defined on a uniform lattice $x_i = ia/\ell$ by its values

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

Our goal is to demonstrate that the number of support vectors that are used to construct the SV approximation depends on the required accuracy ε : The less accurate the approximation, the fewer support vectors are needed.

In this section, to approximate real-valued functions we use linear splines with the infinite number of nodes.

First we describe experiments for approximating the one-dimensional *sinc* function

$$f(x) = \frac{\sin(x - 10)}{x - 10} \quad (6.42)$$

defined on 100 uniform lattice points on the interval $0 \leq x \leq 200$.

Then we approximate the two-dimensional sinc function

$$f(x, y) = \frac{\sin \sqrt{(x - 10)^2 + (y - 10)^2}}{\sqrt{(x - 10)^2 + (y - 10)^2}} \quad (6.43)$$

defined on the uniform lattice points on $0 \leq x \leq 20$, $0 \leq y \leq 20$.

To construct the one-dimensional linear spline approximation we use the kernel defined in Section 6.3:

$$K_1(x, x_i) = 1 + x_i x + \frac{1}{2}|x - x_i|(x \wedge x_i)^2 + \frac{(x \wedge x_i)^3}{3}.$$

We obtain an approximation of the form

$$y = \sum_{i=1}^N (\alpha_i^* - \alpha_i) K_1(x, x_i) + b,$$

where the coefficients α_i^* , α_i are the result of solving a quadratic optimization problem.

Figure 6.5 shows the approximation of the function (6.42) with different levels of accuracy. The black dots on the figures indicate the support vectors; the circles are nonsupport vectors. One can see that with a decrease in the required accuracy of the approximation, the number of support vectors decreases.

To approximate the two-dimensional *sinc* function (6.43) we used the kernel

$$\begin{aligned} K(x, y; x_i, y_i) &= K(x, x_i)K(y, y_i) \\ &= \left(1 + xx_i + \frac{1}{2}|x - x_i|(\wedge x_i)^2 + \frac{(x \wedge x_i)^3}{3}\right) \times \\ &\quad \times \left(1 + yy_i + \frac{1}{2}|y - y_i|(y \wedge y_i)^2 + \frac{(y \wedge y_i)^3}{3}\right), \end{aligned}$$

which is defined by multiplication of the two one-dimensional kernels.

We obtain an approximation in the form

$$y = \sum_{i=1}^N (\alpha_i^* - \alpha_i) K(x, x_i) K(y, y_i) + b,$$

where the coefficients α^* , α are defined by solving the same quadratic optimization problem as in the one-dimensional case.

Figure 6.6 shows the approximations to the two-dimensional *sinc* function with the required accuracy $\varepsilon = 0.03$ conducted using lattices with different numbers of grid points: 400 in figure *a*, 2025 in figure *b*, and 7921 in figure *c*. One can see that changing the number of grid points by a factor of 20 increases the number of support vectors by less than a factor of 2: 153 SV in approximation *a*, 234 SV in approximation *b*, and 285 SV in approximation *c*.

6.8.1 Why Does the Value of ε Control the Number of Support Vectors?

The following model describes a mechanism for choosing the support vectors for function approximation using the SV machine with an ε -insensitive loss function. This mechanism explains why the choice of ε controls the number of support vectors.

Suppose one would like to approximate a function $f(x)$ with accuracy ε , that is, to describe the function $f(x)$ by another function $f^*(x)$ such that the function $f(x)$ is situated in the ε -tube of $f^*(x)$. To construct such a function let us take an elastic ε -tube (a tube that tends to be flat) and put the function $f(x)$ into the ε -tube. Since the elastic tube tends to become

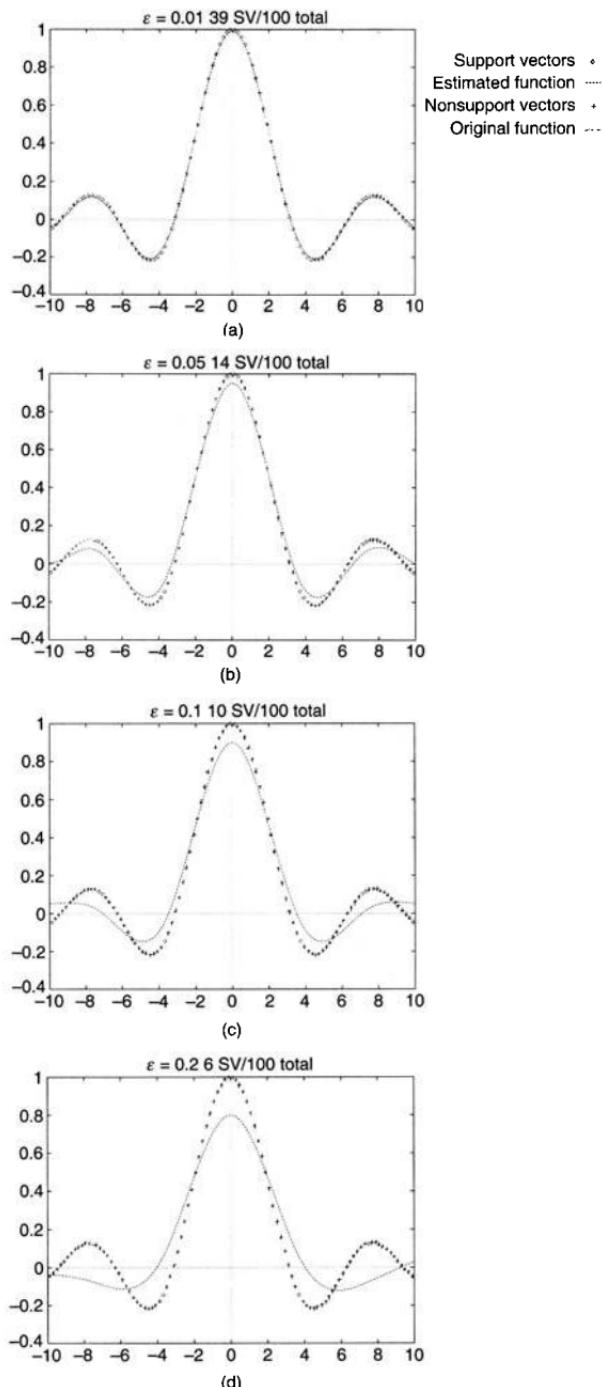


FIGURE 6.5. Approximations with different levels of accuracy require different numbers of support vectors: 39 SV for $\epsilon = 0.01$ (figure a), 14 SV for $\epsilon = 0.05$ (figure b), 10 SV for $\epsilon = 0.1$ (figure c) and 6 SV for $\epsilon = 0.2$ (figure d).

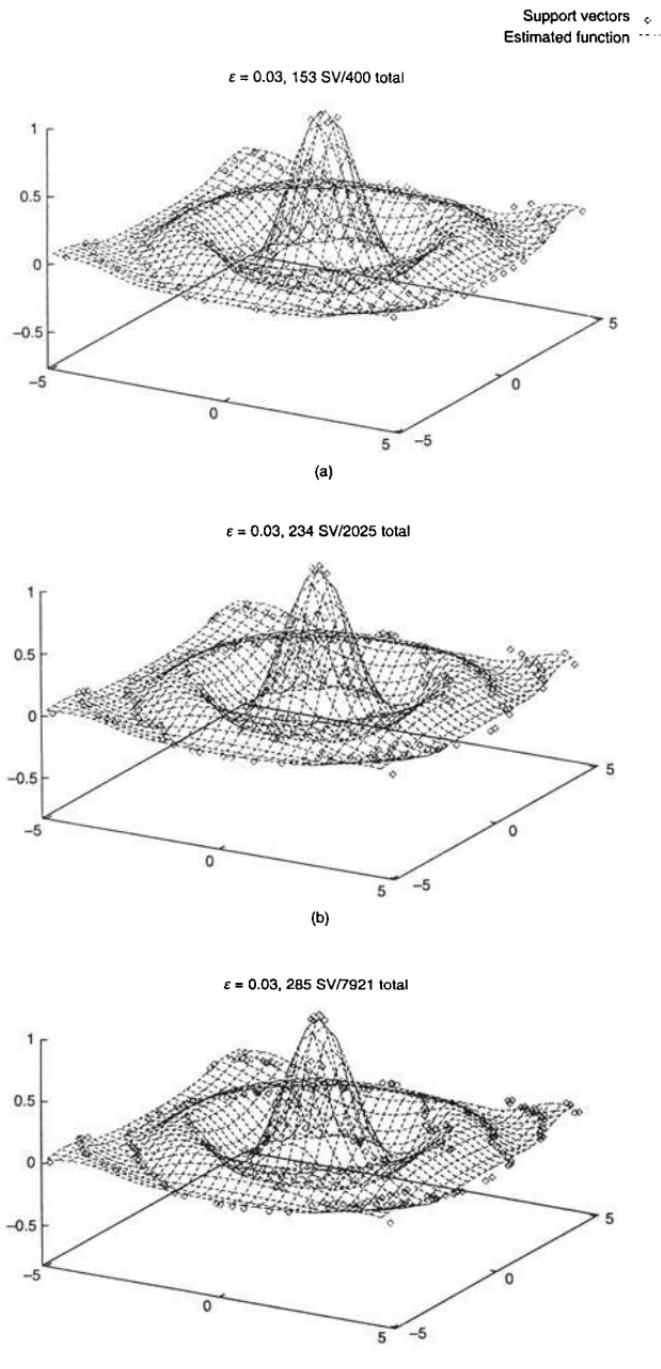


FIGURE 6.6. Approximations to the two-dimensional *sinc* function defined on lattices containing different numbers of grid points with the same accuracy $\varepsilon = 0.03$ do not require large differences in the number of support vectors: 153 SV (grey squares) for the approximation constructed using 400 grid points (figure a), 234 SV for the approximation constructed using 2025 grid points, and 285 SV

flat, it will touch some points of the function $f(x)$. Let us fasten the tube at these points. Then the axis of the tube defines an ε -approximation $f^*(x)$ of the function $f(x)$, and the coordinates of the points where the ε -tube touches the function $f(x)$ define the support vectors. The kernel $K(x_i, x_j)$ describes the law of elasticity.

Indeed, since the function $f(x)$ is in the ε -tube, there are no points of the function with distance of more than ε to axis. Therefore, the axis describes the required approximation.

To prove that touching points define the support vectors it is sufficient to note that we obtained our approximation by solving an optimization problem defined in Section 6.2 for which the Kuhn-Tucker conditions hold. By definition, the support vectors are those for which in the Kuhne-Tucker condition the Lagrange multipliers are different from zero, and hence the second multiplier must be zero. This multiplier defines the border points in an optimization problem of inequality type, i.e., coordinates where the function $f(x)$ touches the ε -tube. The wider the ε -tube, the fewer touching points there are.

This model is valid for the function approximation problem in a space of arbitrary dimension. It explains why with increasing ε -insensitivity the number of support vectors decreases.

Figure 6.7 shows the ε -tube approximation that corresponds to the case of approximating the one-dimensional *sinc* function with accuracy $\varepsilon = 0.2$. Compare this figure to Figure 6.5d.

6.9 SVM FOR REGRESSION ESTIMATION

We start this section with simple examples of regression estimation tasks where regressions are defined by one- and two-dimensional *sinc* functions. Then we consider estimating multidimensional linear regression functions

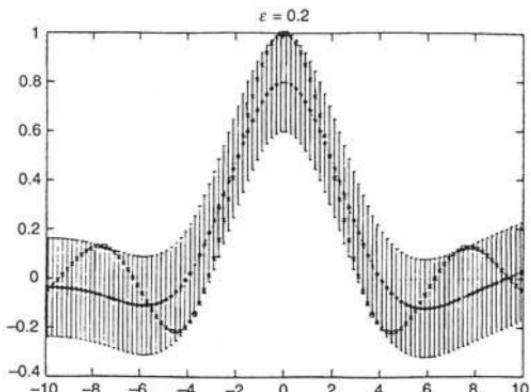


FIGURE 6.7. The ε -tube model of function approximation

using the SVM. We construct a linear regression task that is extremely favorable for a feature selection method and compare results obtained for the forward feature selection method with results obtained by the SVM. Then we compare the support vector regression method with new nonlinear techniques for three multidimensional artificial problems suggested by J. Friedman and one multidimensional real-life (Boston housing) problem (these problems are usually used in benchmark studies of different regression estimation methods).

6.9.1 Problem of Data Smoothing

Let the set of data

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

be defined by the one-dimensional *sinc* function on the interval $[-10, 10]$; the values y_i are corrupted by noise with normal distribution

$$y_i = \frac{\sin x}{x} + \xi_i, \quad E\xi_i = 0, \quad E\xi_i^2 = \sigma^2.$$

The problem is to estimate the regression function

$$y = \frac{\sin x}{x}$$

from 100 such observations on a uniform lattice on the interval $[-10, 10]$.

Figures 6.8 and 6.9 show the results of SV regression estimation experiments from data corrupted by different levels of noise. The rectangles in the figure indicate the support vectors. The approximations were obtained using linear splines with an infinite number of nodes.

Figures 6.10, 6.11, and 6.12 show approximations of the two-dimensional regression function

$$y = \frac{\sin \sqrt{x^2 + y^2}}{\sqrt{x^2 + y^2}}$$

defined on a uniform lattice on the square $[-5, 5] \times [-5, 5]$. The approximations were obtained using two dimensional linear splines with an infinite number of notes.

6.9.2 Estimation of Linear Regression Functions

Below we describe experiments with SVMs in estimating linear regression functions (Drucker et al. (1997)).

We compare the SVM to two different methods for estimating the linear regression function, namely the ordinary least-squares method (OLS) and the forward stepwise feature selection (FSFS) method.

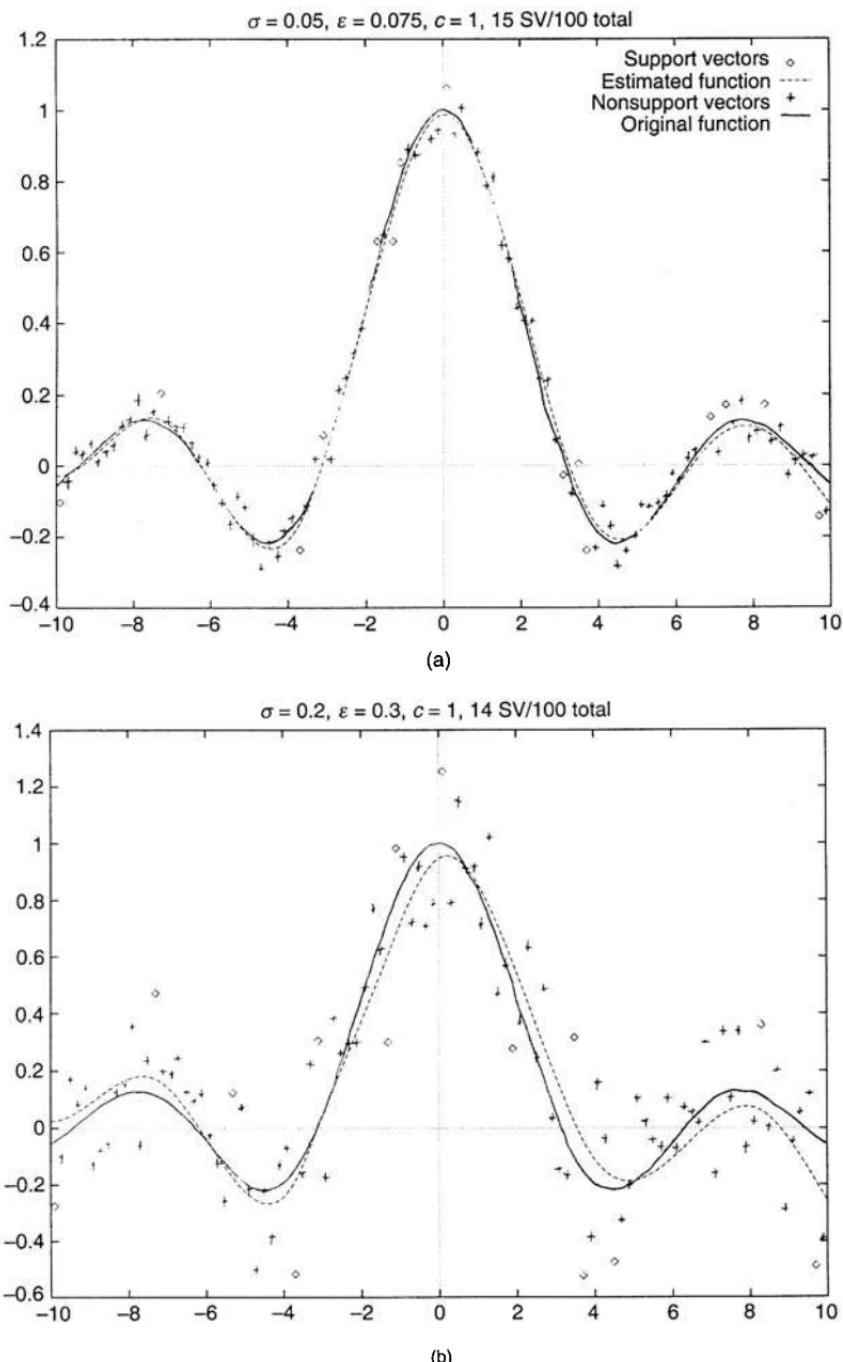


FIGURE 6.8. The regression function and its approximations obtained from the data with different levels of noise and different values ε ($\sigma = 0.05$ and $\varepsilon = 0.075$ in part (a); $\sigma = 0.2$ and $\varepsilon = 0.3$ in part (b)). Note that the approximations were constructed using approximately the same number of support vectors (15 in part (a) and 14 in part (b)).

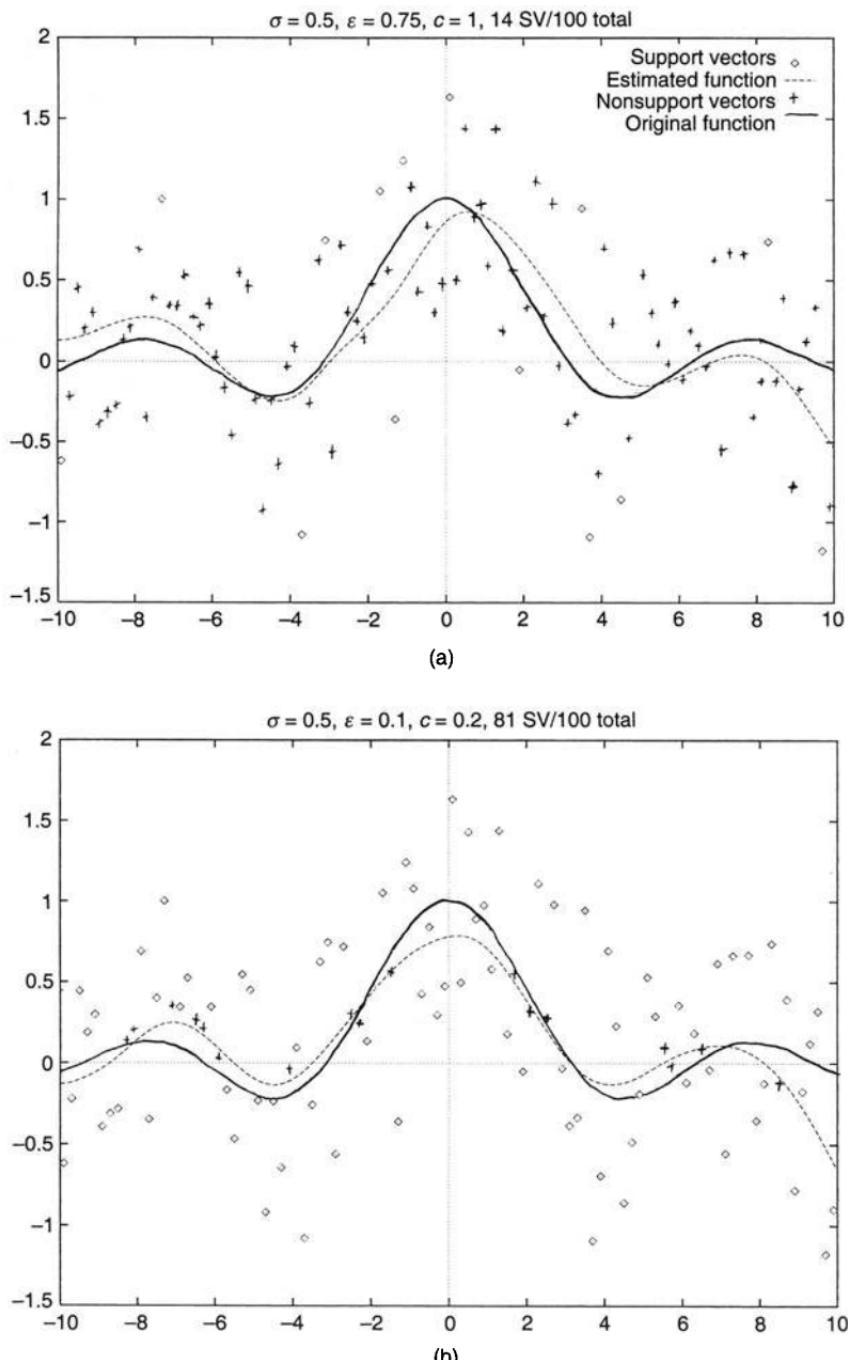


FIGURE 6.9. The regression function and its approximations obtained from the data with the same level of noise $\sigma = 0.5$ and different values of ε ($\varepsilon = 0.25$ in part (a) and $\varepsilon = 0.15$ in part (b)). Note that different values of ε imply a different number of support vectors in the approximating function (14 in part (a) and 81 in part (b)).

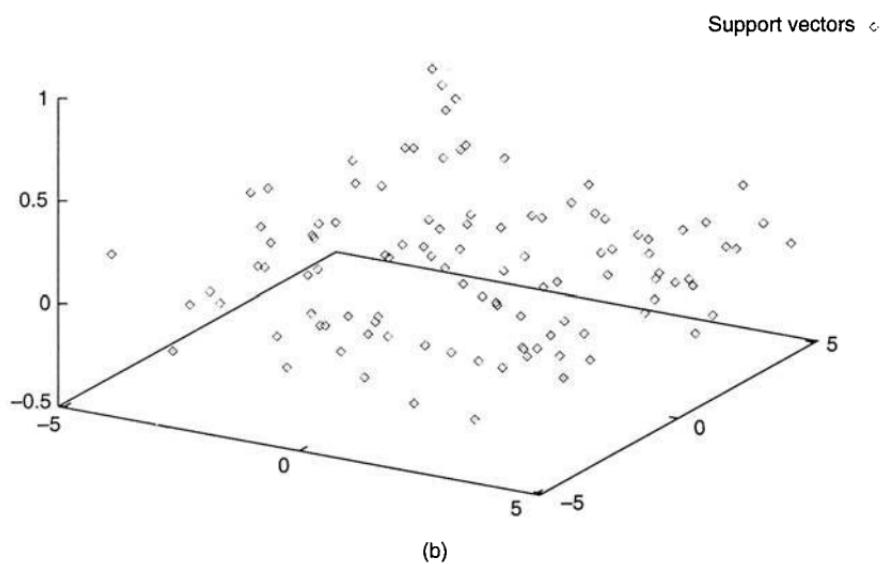
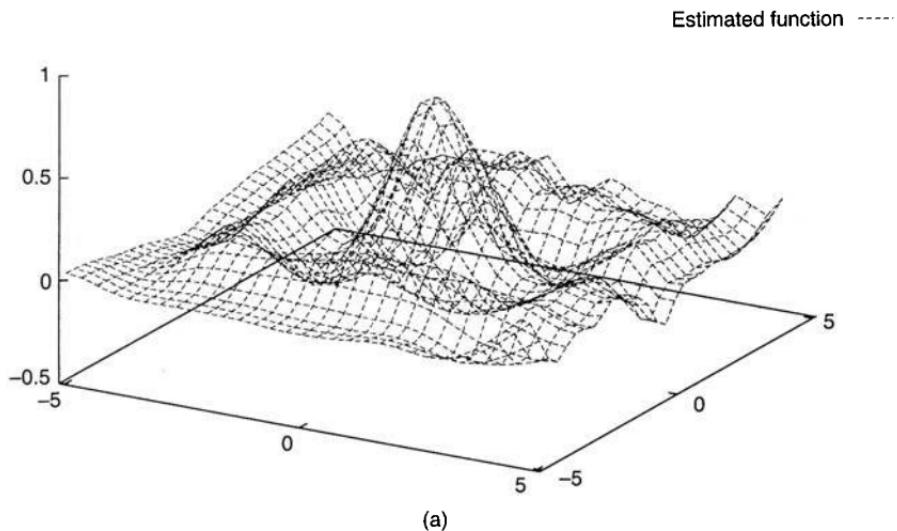
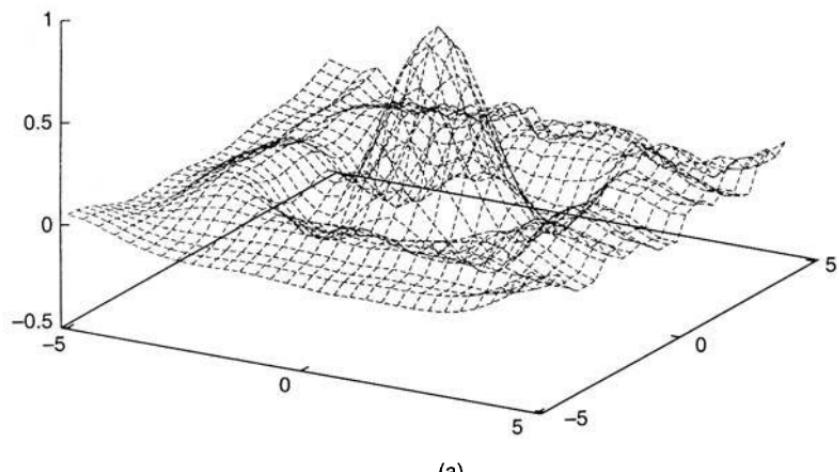
$\sigma = 0.1, \varepsilon = 0.15, 107 \text{ SV/400 total}$


FIGURE 6.10. The approximation to the regression (part (a)) and 107 support vectors (part (b)) obtained from a data set of size 400 with noise $\sigma = 0.1$ and $\varepsilon = 0.15$.

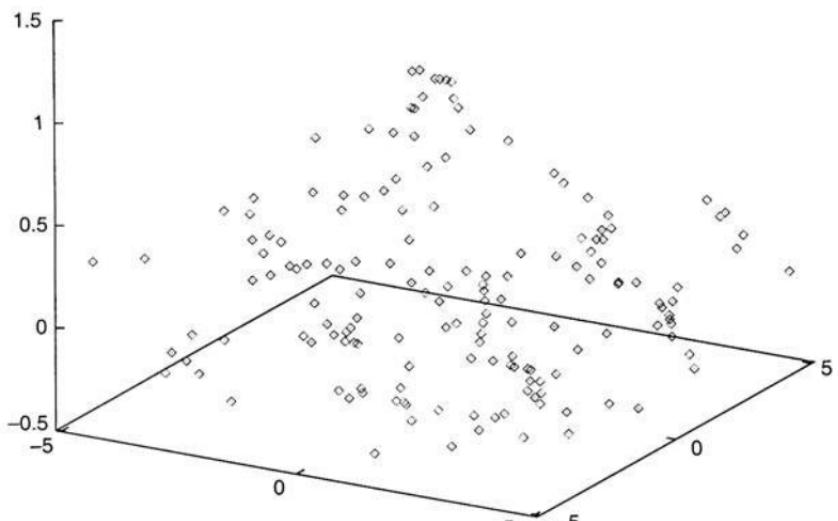
$\sigma = 0.1, \varepsilon = 0.25, 159 \text{ SV}/3969 \text{ total}$

Estimated function -----



(a)

Support vectors ◇



(b)

FIGURE 6.11. The approximation to the regression (part (a)) and 159 support vectors (part (b)) obtained from a data set of size 3969 with the same noise $\sigma = 0.1$ and $\varepsilon = 0.25$.

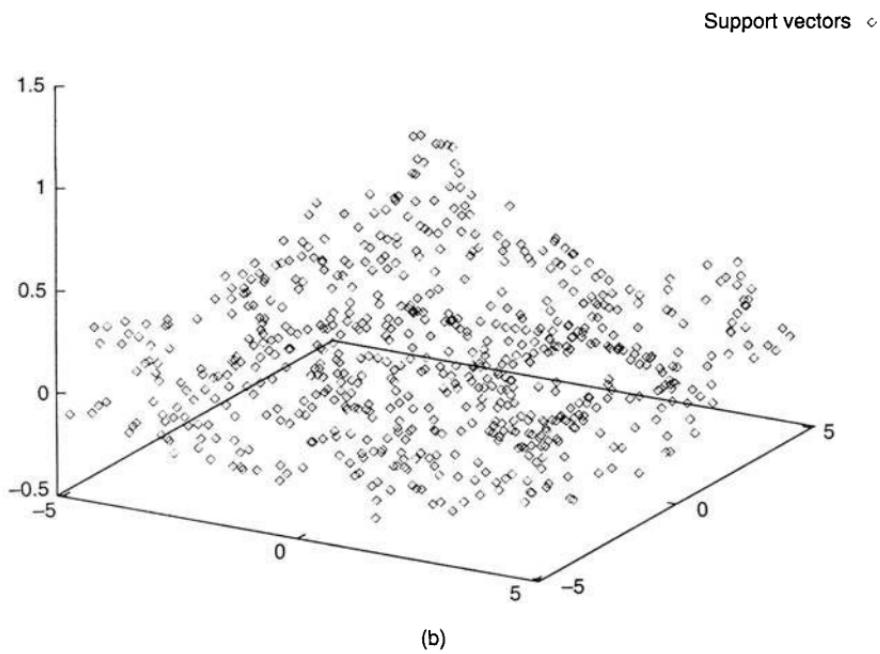
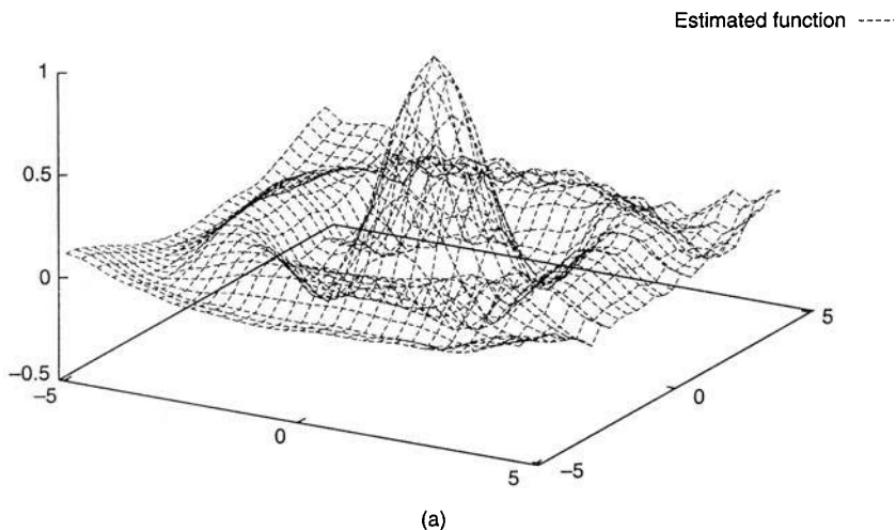
$\sigma = 0.1, \varepsilon = 0.15, 649 \text{ SV}/3969 \text{ total}$


FIGURE 6.12. The approximation to the regression (part (a)) and 649 support vectors (part (b)) obtained from a data set of size 3969 with the same noise $\sigma = 0.1$ and $\varepsilon = 0.15$.

Recall that the OLS method is a method that estimates the coefficients of a linear regression function by minimizing the functional

$$R(a) = \sum_{i=1}^{\ell} (y_i - (a * x_i))^2.$$

The FSFS method is a method that first chooses one coordinate of the vector that gives the best approximation to the data. Then it fixes this coordinate and adds a second coordinate such that these two define the best approximation to the data, and so on. One uses some technique to choose the appropriate number of coordinates.

We consider the problem of linear regression estimation from the data

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

in the 30-dimensional vector space $x = (x^{(1)}, \dots, x^{(30)})$, where the regression function depends only on three coordinates,

$$y(x) = 2x_i^{(1)} + x^{(2)} + x_i^{(3)} + 0 \sum_{i=4}^{30} x^{(k)},$$

and the data are obtained as measurements of this function at randomly chosen points x . The measurements are taken with additive noise

$$y = y(x_i) + \xi_i$$

that is independent of x_i .

Table 6.1 describes the results of experiments of estimating this regression function by the above three methods for different signal-to-noise ratios, different models of noise, and 60 observations. The data in the table are an average of 100 experiments. The table shows that for large noise (small SNR) the support vector regression gives results that are close to (favorable for this model) the FSFS method that are significantly better than the OLS method.

SNR	Normal			Laplacian			Uniform		
	OLS	FSFS	SV	OLS	FSFS	SV	OLS	FSFS	SV
0.8	45.8	28.0	29.3	40.8	24.5	25.4	39.7	24.1	28.1
1.2	20.0	12.8	14.9	18.1	11.0	12.5	17.6	11.7	12.8
2.5	4.6	3.1	3.9	4.2	2.5	3.2	4.1	2.8	3.6
5.0	1.2	0.77	1.3	1.0	0.60	0.52	1.0	0.62	1.0

TABLE 6.1. Comparison results for ordinary least-squares (OLS), forward step feature selection (FSFS), and support vector (SV) methods.

The experiments with the model

$$y_i = \sum_{i=1}^{30} x_i^{(k)} + \xi_i$$

demonstrated the advantage of the SV technique for all levels of signal-to-noise ratio defined in Table 6.1.

6.9.3 Estimation Nonlinear Regression Functions

For these regression estimation experiments we chose regression functions suggested by J. Friedman that were used in many benchmark studies:

1. Friedman's target function #1 is a function of 10 nominal variables

$$y = 10 \sin(\pi x^{(1)} x^{(2)}) + 20(x^{(3)} - 0.5)^2 + 10x^{(4)} + 5x^{(5)} + \xi. \quad (6.44)$$

However, it depends on only 5 variables. In this model the 10 variables are uniformly distributed in $[0, 1]$, and the noise is normal with parameters $N(0, 1)$.

2. Friedman's target function #2,

$$y = \sqrt{(x^{(1)})^2 + [x^{(2)}x^{(3)} - 1/(x^{(2)}x^{(3)})]^2},$$

has four independent variables uniformly distributed in the following region

$$\begin{aligned} 0 &\leq x^{(1)} \leq 100, \\ 40\pi &\leq x^{(2)} \leq 560\pi, \\ 0 &\leq x^{(3)} \leq 1, \\ 1 &\leq x^{(4)} \leq 11. \end{aligned} \quad (6.45)$$

The noise is adjusted for a 3:1 signal-to-noise ratio.

3. Friedman's target function # 3 also has four independent variables

$$y = \tan^{-1} \left[\frac{x^{(2)}x^{(3)} - 1/x^{(2)}x^{(4)}}{x^{(1)}} \right] + \xi, \quad (6.46)$$

that are uniformly distributed in the same region (6.45). The noise was adjusted for a 3:1 signal-to-noise ratio.

Below we compare the advanced regression techniques called bagging (L. Brieman, 1996) and AdaBoost² that construct different types of committee

²The AdaBoost algorithm was proposed for the pattern recognition problem see Section 5.10). It was adapted for regression estimation by H. Drucker (1997).

	Bagging	Boosting	SV
Friedman #1	2.2	1.65	0.67
Friedman #2	11,463	11,684	5,402
Friedman #3	0.0312	0.0218	0.026

TABLE 6.2. Comparison of Bagging and Boosted regression trees with SVM regression for three artificial data sets.

machine by combining given in the comments to Chapter 13) with the support vector regression machine.

The experiments were conducted using the same format as in (Drucker, 1997, Drucker et al. 1997).

Table 6.2 shows results of experiments for estimating Friedman's functions using bagging, boosting, and polynomial ($d = 2$) SVMs. The experiments were conducted using 240 training examples. Table 6.2 shows an average (over 10 runs) of the model error (mean squared deviation between the true target function and obtained approximation).

Table 6.3 shows performance obtained for the Boston housing data set where 506 examples of 13-dimensional real-life data were used as follows: 401 random chosen examples as the training set, 80 as the validation set, and 25 as test set. Table 6.3 shows results of averaging over 100 runs. The SV machine constructed polynomials (mostly of degree 4 and 5) chosen on the basis of the validation set. For the Boston housing data the performance index is the mean squared error between the predicted and actual values y on the test set.

Bagging	Boosting	SV
12.4	10.7	7.2

TABLE 6.3. Performance of different methods for the Boston housing data.

Informal Reasoning and Comments — 6

6.10 LOSS FUNCTIONS FOR THE REGRESSION ESTIMATION PROBLEM

The methods for estimating functional dependencies based on empirical data have a long history. They were begun by great mathematicians: Gauss (1777–1855) and Laplace (1749–1827), who suggested two different methods for estimating dependencies from results of measurements in astronomy and physics.

Gauss proposed the least-squares method (LSM), while Laplace proposed the least modulo method (LMM). Since that time the question has arisen as to which method is better. In the nineteenth century and beginning of the twentieth century preference was given to the least-squares method: The solution with this method for linear functions has a closed form. Also, it was proven that among *linear and unbiased* estimates the LSM is the best.

Later, in the second part of the twentieth century, it was noted that in many situations the set of linear and unbiased estimates is too narrow to be sure that the best estimate in this set is really good (it is quite possible that the whole set contains only “bad” estimators).

In the 1920s R. Fisher discovered the maximum likelihood (ML) method and introduced the model of measurements with additive noise. According to this model the measurement of a function $f(x, \alpha_0)$ at any point x^* is corrupted by the additive noise (described by the known symmetric density

$p_0(\xi)$; ξ is uncorrelated with x^*)

$$y^* = f(x, \alpha_0) + \xi.$$

Since

$$\xi = y - f(x, \alpha_0),$$

to estimate the parameter α_0 of density $p_0(\xi)$ (the unknown function $f(x, \alpha_0)$) from the data

$$x_1, \dots, x_\ell$$

using maximum likelihood one has to maximize the functional

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

In 1953 L. Le Cam defined conditions under which the ML method is consistent. He found some sufficient conditions on uniform convergence (over the set of $\alpha \in \Lambda$) under which the empirical functional $R_\ell(\alpha)$ converges to the functional

$$R(\alpha) = \int \ln p(y - f(x, \alpha)) dP(y, x)$$

(they are a particular case of the necessary and sufficient conditions considered in Chapter 2); this immediately implies that the following assertion holds true:

$$-\int \ln \left(\frac{p(y - f(x, \alpha_\ell))}{p(y - f(x, \alpha_0))} \right) \xrightarrow[\ell \rightarrow \infty]{P} 0.$$

That is, the ML solutions are consistent in the Kulbac-Leibler distance. It is also in the set of unbiased estimators (not necessary linear) that the LM method has the smallest variance (the unbiased estimate with the smallest variance is called *effective*).

This implies that if the noise is described by Gaussian (normal) law, then the LSM gives the best solution. If, however, the noise is defined by the Laplacian law

$$p(x, \Delta) = \frac{1}{2\Delta} \exp \left\{ -\frac{|\xi|}{\Delta} \right\},$$

then the best solution defines the least modulo estimate. From these results it also follows that the loss function for the best (effective) estimate is defined by the distribution of noise.

In practice (even if the additive model of measurements is valid), the form of noise is usually unknown. In the 1960s Tukey demonstrated that in real-life situations the form of noise is far from both the Gaussian and the Laplacian laws.

Therefore, it became important to create the best strategy for estimating functions in real-life situations (when the form of noise is unknown). Such a strategy was suggested by P. Huber, who created the concept of *robust* estimators.

6.11 LOSS FUNCTIONS FOR ROBUST ESTIMATORS

Consider the following situation. Suppose our goal is to estimate the expectation m of the random variable ξ using i.i.d. data

$$\xi_1, \dots, \xi_\ell.$$

Suppose also that the corresponding unknown density $p_0(\xi - m_0)$ is a smooth function, is symmetric with respect to the position m_0 , and possesses a second moment.

It is known that in this situation the maximum likelihood estimator

$$m = \mathcal{M}(\xi_1, \dots, \xi_\ell | p_0)$$

that maximizes

$$L(m) = \sum_{i=1}^{\ell} \ln p_0(\xi_i - m)$$

is an effective estimator. This means that among all possible *unbiased* estimators³ this estimator achieves the smallest variance, or in other words, estimator $\mathcal{M}(\xi_1, \dots, \xi_\ell | p_0)$ minimizes the functional

$$V(\mathcal{M}) = \int (\mathcal{M}(\xi_1, \dots, \xi_\ell) - m)^2 dp_0(\xi_1 - m) \cdots dp_0(\xi_\ell - m). \quad (6.47)$$

Suppose now that although the density $p_0(\xi - m)$ is unknown, it is known that it belongs to some admissible set of densities $p_0(\xi - m) \in \mathcal{P}$. How do we choose an estimator in this situation? Let the unknown density be $p_0(\xi - m)$. However, we construct an estimator that is optimal for density $p_1(\xi - m) \in \mathcal{P}$, i.e., we define the estimator $\mathcal{M}(\xi_1, \dots, \xi_\ell | p_1)$ that maximizes the functional

$$L_1(m) = \sum_{i=1}^{\ell} \ln p_1(\xi_i - m). \quad (6.48)$$

The quality of this estimator now depends on two densities, the actual one $p_0(\xi - m)$ and the one used for constructing estimator (11.8):

$$V(p_0, p_1) = \int (\mathcal{M}(\xi_1, \dots, \xi_\ell | p_1) - m)^2 dp_0(\xi_1 - m) \cdots dp_0(\xi_\ell - m).$$

Huber proved that for a wide set of admissible densities \mathcal{P} there exists a saddle point of the functional $V(p_0, p_1)$. That is, for any admissible set of

³The estimator $\mathcal{M}(\xi_1, \dots, \xi_\ell)$ is called unbiased if

$$E\mathcal{M}(\xi_1, \dots, \xi_\ell) = m.$$

densities there exists such a density $p_r(\xi - m)$ that the inequalities

$$V(p, p_r) \leq V(p_r, p_r) \leq V(p_r, p) \quad (6.49)$$

hold true for any function $p(\xi - m) \in \mathcal{P}$.

Inequalities (11.9) assert that for any admissible set of densities there exists the minimax density, the so-called *robust density*, which in the worst scenario guarantees the smallest loss.

Using the robust density one constructs the so-called *robust regression estimator*. Namely, the robust regression estimator is the one that minimizes the functional

$$R_h(w) = - \sum_{i=1}^{\ell} \ln p_r(y_i - f(x_i, \alpha)).$$

Below we formulate the Huber theorem that is a foundation of the theory of robust estimation.

Consider the class H of densities formed by mixtures

$$p(\xi) = (1 - \epsilon)g(\xi) + \epsilon h(\xi)$$

of a certain fixed density $g(\xi)$ and an arbitrary density $h(\xi)$, where both densities are symmetric with respect to the origin. The weights in the mixture are $1 - \epsilon$ and ϵ respectively. For the class of these densities the following theorem is valid.

Theorem. (Huber) *Let $-\ln g(\xi)$ be a twice continuously differentiable function. Then the class H possesses the following robust density:*

$$p_r(\xi) = \begin{cases} (1 - \epsilon)g(\xi_0) \exp\{-c(\xi_0 - \xi)\}, & \text{for } \xi < \xi_0, \\ (1 - \epsilon)g(\xi), & \text{for } \xi_0 \leq \xi < \xi_1, \\ (1 - \epsilon)g(\xi_1) \exp\{-c(\xi - \xi_1)\}, & \text{for } \xi \geq \xi_1, \end{cases} \quad (6.50)$$

where ξ_0 and ξ_1 are endpoints of the interval $[\xi_0, \xi_1]$ on which the monotone (due to convexity of $-\ln g(\xi)$) function

$$-\frac{d \ln g(\xi)}{d\xi} = -\frac{g'(\xi)}{g(\xi)}$$

is bounded in absolute value by a constant c determined by the normalization condition

$$1 = (1 - \epsilon) \left(\int_{\xi_0}^{\xi_1} g(\xi) d\xi + \frac{g(\xi_0) + g(\xi_1)}{c} \right).$$

This theorem allows us to construct various robust densities. In particular, if we choose for $g(\xi)$ the normal density

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

and consider the class H of densities

$$p(\xi) = \frac{1-\epsilon}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} + \epsilon h(\xi),$$

then according to the theorem, the density

$$p_r(\xi) = \begin{cases} \frac{1-\epsilon}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2} - \frac{c}{\sigma}|\xi|\right\} & \text{for } |\xi| > c\sigma, \\ \frac{1-\epsilon}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} & \text{for } |\xi| \leq c\sigma, \end{cases} \quad (6.51)$$

will be robust in the class, where c is determined from the normalization condition

$$1 = \frac{1-\epsilon}{\sqrt{2\pi}\sigma} \left(\int_{-c\sigma}^{c\sigma} \exp\left\{-\frac{\xi^2}{2}\right\} d\xi + \frac{2 \exp\left\{-\frac{c^2}{2}\right\}}{c} \right).$$

The loss function derived from this robust density is

$$L(\xi) = -\ln p(\xi) = \begin{cases} c|\xi| - \frac{c^2}{2} & \text{for } |\xi| > c, \\ \frac{\xi^2}{2} & \text{for } |\xi| \leq c. \end{cases} \quad (6.52)$$

It smoothly combines two functions: quadratic and linear. In one extreme case (when c tends to infinity) it defines the least-squares method; in the other extreme case (when c tends to zero), it defines the least modulo method. In the general case, the loss functions for robust regression are combinations of two functions one of which is $f(u) = |u|$ and the other is much less sensitive to deviations of u (the derivative of the nonlinear part of the function $f(u)$ is less than the derivative of the linear part).

6.12 SUPPORT VECTOR REGRESSION MACHINE

Our construction of SVMs for the regression problem is based on the ϵ -insensitive loss function. This loss function has the same structure as robust loss functions: It combines two functions one of which is $f(u) = |u|$ and the constant function⁴: $f(u) = \text{const}$ (we considered case $\text{const} = 0$).

The ϵ -insensitivity implies some new properties of the SVM solutions, namely the sparsity of solutions. By changing (increasing) the value of ϵ one controls (increases) the sparsity of the SVM solutions.

However, the difference between the robust approach and SVM approach reflects also the fact that the loss function for the SVM regression is more

⁴Formally it does not belong to the family of Huber's robust estimators, since the uniform distribution function does not possess a smooth derivative.

complicated than the loss function for robust regression. For linear functions it has the form⁵

$$L(a) = \frac{1}{C} (w, w) + \sum_{i=1}^{\ell} |y_i - (w, x)|_{\varepsilon},$$

where (w, w) is the regularization functional and $1/C$ is the regularization parameter (we will discuss the regularization techniques in the next chapter).

The addition of the regularization term into the functional dramatically changes the situation: On one hand it connected SVM regression to regularization techniques introduced for solving ill-posed problems, and on the other hand it increases the number of free parameters.

Now, in order to estimate the regression function we have to specify three free parameters: the value of ε -insensitivity, the regularization parameter C , and the kernel parameter (the order of the polynomial for polynomial kernels, the width parameter for radial basis kernels, the order of the spline for spline generating kernels, and so on).

In the next chapter we show that using some general ideas developed in classical statistics and general principles for solving ill-posed problems developed in the theory of ill-posed problems we will be able not only to specify how these parameters should be connected, in order to provide optimal estimates, but also to describe effective algorithms for evaluating the best possible parameters for solving the main problem of statistical learning theory: estimating density functions, conditional probability (this is more general solution to the pattern recognition problem than was described before), and regression functions. The ε -insensitive estimators will play a crucial part in these algorithms.

⁵In the main part of this chapter we used an equivalent form of this functional.

Chapter 7

Direct Methods in Statistical Learning Theory

In this chapter we introduce a new approach to the main problems of statistical learning theory: pattern recognition, regression estimation, and density estimation.

We introduce the so-called direct approach, which requires solving operator equations that define the desired functions. The solutions of these equations are based on solving stochastic ill-posed problems. To solve them effectively we combine ideas that were originated within three different branches of mathematics: the theory of ill-posed problems, classical non-parametric statistics, and statistical learning theory. The results obtained in the first two branches were not considered in the main part of the book (they were only briefly discussed in the informal reasoning and comments to the chapters).

In this chapter we introduce the necessary results from these branches and combine corresponding techniques to obtain a new type of algorithms.

7.1 PROBLEM OF ESTIMATING DENSITIES, CONDITIONAL PROBABILITIES, AND CONDITIONAL DENSITIES

7.1.1 Problem of Density Estimation: Direct Setting

We start this chapter with the problem of density estimation. Let ξ be a random variable. The probability of a random event

$$F(x) = P\{\xi < x\}$$

we call a *probability distribution function* of the random variable ξ . A random vector $\bar{\xi}$ is a generalization of the notion of a random variable. The function

$$F(x) = P\{\bar{\xi} < \bar{x}\},$$

where the inequality is interpreted coordinatewise, is called a *probability distribution function of the random vector $\bar{\xi}$* . We say that the random variable ξ (random vector $\bar{\xi}$) has a density if there exists a nonnegative function $p(x)$ such that for all x the equality

$$F(x) = \int_{-\infty}^x p(x') dx'$$

is valid.

The function $p(x)$ is called a *probability density* of the random variable (random vector). So, by definition, to estimate a probability density from the data we need to obtain a solution of the integral equation¹

$$\int_{-\infty}^x p(x', \alpha) dx' = F(x) \quad (7.1)$$

on a given set of densities $p(x, \alpha)$, $\alpha \in \Lambda$, under the condition that the distribution function $F(x)$ is unknown and a random independent sample

$$x_1, \dots, x_\ell \quad (7.2)$$

obtained in accordance with $F(x)$ is given.

¹When $x = (x^1, \dots, x^n)$ is a vector, this notation defines coordinatewise integration

$$\int_{-\infty}^x p(x, \alpha) dx \equiv \int_{-\infty}^{x^1} \cdots \int_{-\infty}^{x^n} p(x^1, \dots, x^n; \alpha) dx^1 \cdots dx^n.$$

One can construct approximations to the distribution function $F(x)$ using data (7.2), for example, the so-called *empirical distribution function* (7.2):

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i), \quad (7.3)$$

where we define for the vector ² u the step function

$$\theta(u) = \begin{cases} 1 & \text{all coordinates of the vector } u \text{ are positive,} \\ 0 & \text{otherwise.} \end{cases}$$

In the next section we will show that the empirical distribution function $F_\ell(x)$ is a good approximation to the actual distribution function $F(x)$.

Thus, the problem of density estimation is to find an approximation to the solution of the integral equation (7.1) if the probability distribution function is unknown; however, an approximation to this function can be defined.

We call this setting of the density estimation problem the *direct setting* because it is based on the definition of a density. In the following sections we shall discuss the problem of solving integral equations with an approximate right-hand side and approximate operator, but now we turn to the direct setting of the problem of estimating the conditional probability $P(\omega|x)$ that defines the probability of class ω given the vector x .

7.1.2 Problem of Conditional Probability Estimation

Consider pairs (ω, x) , where x is a vector and ω is a scalar that takes on only k values $\{0, 1, \dots, k-1\}$. According to the definition, the conditional probability $P(\omega|x)$ is the solution of the integral equation

$$\int_{-\infty}^x P(\omega|x') dF(x') = F(\omega, x), \quad (7.4)$$

where $F(x)$ is a distribution function of random vectors x , and $F(\omega, x)$ is the joint distribution function of pairs (ω, x) . Indeed, since $dF(x) = p(x)dx$ (we suppose that the density does exist) and

$$P(\omega|x') = \frac{p(\omega, x')}{p(x')},$$

the solution of (7.4) defines the conditional probability.

The problem of estimating the conditional probability in the set of functions $P_\alpha(\omega|x)$, $\alpha \in \Lambda$, is to obtain an approximation to the solution of the

²Including scalars as one-dimensional vectors.

integral equation (7.4) when both distribution functions $F(x)$ and $F(\omega, x)$ are unknown but the data

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

are given. As in the case of density estimation, we can approximate the unknown distribution functions $F(x)$ and $F(\omega, x)$ by the empirical distribution function (7.3) and the function

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

where

$$\delta(\omega, x) = \begin{cases} 1 & \text{if the vector } x \text{ belongs to the class } \omega, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the problem is to obtain an approximation to the solution of the integral equation (7.4) in the set of functions $P_\alpha(\omega|x)$, $\alpha \in \Lambda$, when the probability distribution functions $F(x)$ and $F(\omega, x)$ are unknown, but approximations $F_\ell(x)$ and $F_\ell(\omega, x)$ are given.

Note that estimation of the conditional probability function $P(\omega|x)$ is a stronger solution to the pattern recognition problem than the one considered in Chapter 1. In Chapter 1, the goal was to find the best decision rule from the *given set of decision rules*; it did not matter whether this set did or did not contain a good approximation to the supervisor's decision rule. In this statement the goal is to find the best approximation to the supervisor's decision rule (which is the conditional probability function according to the statement of the problem. See Chapter 1). Of course, if the approximation of the supervisor's operator $P(\omega|x)$ is known, then one can easily construct the optimal decision rule. For the case where $\omega \in \{0, 1\}$ and the *a priori* probabilities of the classes are equal it has the form

$$f(x) = \theta\left(P(\omega = 1|x) - \frac{1}{2}\right).$$

This is the so-called Bayes rule; it assigns the vector x to the class 1 if the probability that this vector belongs to the first class is larger than $\frac{1}{2}$ and assigns 0 otherwise. However, the knowledge of the conditional probability not only gives the best solution to the pattern recognition problem but also provides an estimate of the error probability for any specific vector x .

7.1.3 Problem of Conditional Density Estimation

Finally, consider the problem of conditional density estimation. In the pair (y, x) , let the variable y be scalar and let x be a vector. Consider the equality

$$\int_{-\infty}^y \int_{-\infty}^x p(y'|x') dF(x') dy' = F(y, x), \quad (7.5)$$

where $F(x)$ is a probability distribution function that has a density and $F(y, x)$ is the joint probability distribution function³ defined on the pairs (y, x) .

As before, we are looking for an approximation to the conditional density $p(y|x)$ by solving the integral equation (7.5) on the given set of functions when both distribution functions $F(x)$ and $F(y, x)$ are unknown and the random i.i.d. pairs

$$(y_1, x_1), \dots, (y_\ell, x_\ell) \quad (7.6)$$

are given. As before, we can approximate $F(x)$ by the empirical distribution function (7.3) and the distribution function $F(y, x)$ by the empirical distribution function

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

Thus, our problem is to obtain an approximation to the solution of the integral equation (7.5) in the set of functions $p_\alpha(y|x)$, $\alpha \in \Lambda$, when the probability distribution functions are unknown but we can construct the approximations $F_\ell(x)$ and $F_\ell(y, x)$ using data (7.6).

Note that the conditional density $p(y|x)$ contains much more information about the behavior of the random value y for a given x than the regression function. The regression function can be easily obtained from the conditional density. According to its definition the regression function is

$$r(x) = \int y p(y|x) dy.$$

7.2 THE PROBLEM OF SOLVING AN APPROXIMATELY DETERMINED INTEGRAL EQUATION

All three problems of estimating stochastic dependencies can be described in the following general way. It is necessary to solve a linear operator equation

$$Af = F, \quad f \in \mathcal{F}, \quad (7.7)$$

where some functions that form the equation are unknown, but data are given. Using these data the approximations to the unknown functions can be obtained.

³Actually, the solution of this equation is the definition of conditional density. Suppose that $p(x)$ and $p(y, x)$ are the densities corresponding to probability distribution functions $F(x)$ and $F(y, x)$. Then equality (7.5) is equivalent to the equality $p(y|x)p(x) = p(y, x)$.

A difference exists between the problem of density estimation and the problems of conditional probability and conditional density estimation. In the problem of density estimation, instead of the right-hand side of the equation we are given its approximation. We would like to obtain an approximation to the solution of equation (7.7) from the relationship

$$Af \approx F_\ell, \quad f \in \mathcal{F}.$$

In the problems of conditional probability and conditional density estimation, not only is the right-hand side of the equation (7.7) known approximately, but also the operator A known approximately (on the left-hand side of integral equations (7.4) and (7.5), instead of the distribution functions we use their approximations). So our problem is to obtain an approximation to the solution of equation (7.7) from the relationship

$$A_\ell f \approx F_\ell, \quad f \in \mathcal{F},$$

where A_ℓ is an approximation of the operator A .

There is good news and bad news about solving these problems. The good news is that the empirical distribution function forms a good approximation to the unknown distribution function. In the next section we show that as the number of observations tends to infinity, the empirical distribution function converges to the desired one at the fast rate $1/\sqrt{\ell}$. In the one-dimensional case, there is known an asymptotically exact description of the rate of convergence for different metrics determining different definitions of a distance between empirical and true distribution functions.

In particular, for the one-dimensional case the Kolmogorov-Smirnov distribution of distances (in the uniform metric C) between approximations and the desired function is known. In the multidimensional case one can calculate any quantile of this distribution [Paramasamy, 1992].

The bad news is that the problem of solving operator equation (7.7) belongs to the so-called *ill-posed* problems. In Section 7.4 we shall define the concept of “ill-posed” problems and describe the difficulties that arise when one needs to solve ill-posed problems. We will describe the main results of the classical theory for solving ill-posed problems and its generalizations to the case of stochastic ill-posed problems. The theory of solving stochastic ill-posed problems will be used for solving our integral equations.

7.3 GLIVENKO-CANTELLI THEOREM

As we mention in the 1930s Glivenko and Cantelli proved one of the most important theorems in statistics. They proved that when the number of observations tends to infinity, the empirical distribution function $F_\ell(x)$

converges to the actual distribution function $F(x)$. This theorem plays an important part in the foundations of theoretical statistics.

Theorem. (Glivenko-Cantelli). *The convergence*

$$\sup_x |F(x) - F_\ell(x)| \xrightarrow[\ell \rightarrow \infty]{P} 0$$

takes place.

In this formulation, the Glivenko-Cantelli theorem asserts the convergence in probability⁴ (in the uniform metric) of the empirical distribution function $F_\ell(x)$ to the actual distribution function $F(x)$.

One can formulate this theorem in terms of uniform convergence described in Chapter 2. Indeed, consider the following set of events:

$$e(\alpha) = \theta(\alpha - x), \quad \alpha \in \Lambda. \quad (7.8)$$

For any fixed α it defines the set of x that are less than α . Now, let a probability measure be defined on the set of x . Then the expectation

$$R(\alpha) = E\theta(\alpha - x)$$

as a function of α defines a probability distribution function, while the empirical functional

$$R_\ell(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(\alpha - x_i), \quad \alpha \in R^1,$$

calculated from i.i.d. data x_1, \dots, x_ℓ defines an empirical distribution function. Therefore, in fact, the Glivenko-Cantelli theory is the theory of uniform convergence for a specific set of events (7.8) defined in R^1 .

In the n -dimensional case where $\alpha = (\alpha^1, \dots, \alpha^n)$ and $x = (x^1, \dots, x^n)$ the Glivenko-Cantelli theorem describes the uniform convergence of the frequencies to their probabilities over the following sets of events:

$$e(\alpha) = \prod_{k=1}^n \theta(x^k - \alpha^k), \quad \alpha \in R^n \quad (7.9)$$

In Chapter 3 we analyzed the conditions for uniform convergence over any given set of events (not necessarily defined by (7.9)). Therefore, the theory of uniform convergence developed in statistical learning theory includes the Glivenko-Cantelli theory as a particular case.

⁴The convergence almost surely takes place as well.

7.3.1 Kolmogorov-Smirnov Distribution

As soon as the Glivenko-Cantelli theorem had been proved, the problem of the rate of convergence of $F_\ell(x)$ to $F(x)$ emerged.

Investigations of the rate of convergence of $F_\ell(x)$ to $F(x)$ for the one-dimensional continuous functions $F(x)$ resulted in the establishment of the following important statistical law:

Kolmogorov-Smirnov distribution. The random variable

$$\xi_\ell = \sqrt{\ell} \sup_x |F(x) - F_\ell(x)|$$

has the following limiting probability distribution (Kolmogorov):

$$\lim_{\ell \rightarrow \infty} P\{\sqrt{\ell} \sup_x |F(x) - F_\ell(x)| \geq \varepsilon\} = 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2\varepsilon^2 k^2}. \quad (7.10)$$

The random variables

$$\xi_\ell^+ = \sqrt{\ell} \sup_x (F(x) - F_\ell(x)),$$

$$\xi_\ell^- = \sqrt{\ell} \sup_x (F_\ell(x) - F(x)),$$

have the following limiting probability distributions (Smirnov):

$$\lim_{\ell \rightarrow \infty} P\{\sqrt{\ell} \sup_x (F(x) - F_\ell(x)) \geq \varepsilon\} = e^{-2\varepsilon^2},$$

$$\lim_{\ell \rightarrow \infty} P\{\sqrt{\ell} \sup_x (F_\ell(x) - F(x)) \geq \varepsilon\} = e^{-2\varepsilon^2}. \quad (7.11)$$

As we mentioned in the previous section, the Glivenko-Cantelli theory (originally developed for the one-dimensional case) is a particular case of the statistical learning theory. In Chapter 3 we described bounds on uniform convergence that are valid for any specific ℓ and set of events in a space of arbitrary dimension.

In particular, this theory can be applied to the set of events (7.9). Since the VC dimension of this set defined in R^n is equal to n (the dimensionality of the space), we can obtain a bound for uniform convergence over the set of events (7.9) as well. Therefore, using results from statistical learning theory one can obtain nonasymptotic bounds of inequality type.

There exists, however, something in the analysis of uniform convergence of events (7.9) that was not obtained in statistical learning theory for general types of events. For the set of events (7.9) there exists an exact description of the rate of uniform convergence that does not depend on the

probability measure (universal distribution). This exact distribution was obtained by Kolmogorov and Smirnov (for sufficiently large ℓ) for the one-dimensional case. For the multidimensional case this type of distribution is unknown. However, it is known that such a distribution does exist.⁵.

In Section 7.5 we will see how important it is for our estimation problem to have universal equality-type characteristics of this distribution. In spite of the fact that for the multidimensional case and/or for a finite number of observations the analytical expression for this distribution is unknown, one can easily create a table that for any number of observations ℓ and for any reasonable dimension n (say $n < 100$) defines any quantile of this distribution. In sections 7.8, 7.9, and 7.10 we will estimate optimal parameters of our algorithms using such a table.

7.4 ILL-POSED PROBLEMS

Let the operator equation

$$Af(t) = F(x) \quad (7.12)$$

be defined by the continuous operator A that maps in a one-to-one manner the elements f of the metric space E_1 into elements F of the metric space E_2 .

We say that the solution of the operator equation (7.12) is *stable* if a small variation in the right-hand side $F(x) \in F(x, \alpha)$ results in a small change in the solution; i.e., if for any $\varepsilon > 0$ there exists $\delta(\varepsilon)$ such that the inequality

$$\rho_{E_1}(f(t, \alpha_1), f(t, \alpha_2)) \leq \varepsilon$$

is valid as long as the inequality

$$\rho_{E_2}(F(x, \alpha_1), F(x, \alpha_2)) \leq \delta(\varepsilon)$$

holds.

We say that the problem of solving the operator equation (7.12) is *well-posed in the Hadamard sense* if the solution of the equation

- *exists*,
- *is unique*, and
- *is stable*.

⁵It is interesting to describe sets of events that possess a *universally* (independent of probability measure) exact distribution of the rate of uniform convergence.

The problem of solving an operator equation is considered *ill-posed* if the solution of this equation violates at least one of the above-mentioned requirements. Below we consider ill-posed problems for which the solution of the operator equation exists, is unique, but is not stable. We consider ill-posed problems defined by the Fredholm integral equation of type 1:

$$\int_a^b K(t, x)f(t)dt = F(x).$$

However, all the results obtained will also be valid for equations defined by any other linear continuous operator.

Thus, consider Fredholm's integral equation of type 1,

$$\int_a^b K(t, x)f(t)dt = F(x), \quad (7.13)$$

defined by the kernel $K(t, x)$, which is continuous almost everywhere on $a \leq t \leq b$, $a \leq x \leq b$. This kernel maps the set of functions $\{f(t)\}$ continuous on $[a, b]$ onto the set of functions $\{F(x)\}$ also continuous on $[a, b]$.

It is easy to show that the problem of solving equation (7.13) is an ill-posed one. For this purpose we note that the continuous function $G_\nu(x)$ that is formed by means of the kernel $K(t, x)$,

$$G_\nu(x) = \int_a^b K(t, x) \sin \nu t dt$$

possesses the property

$$\lim_{\nu \rightarrow \infty} G_\nu(x) = 0.$$

Consider the integral equation

$$\int_a^b K(t, x)f^*(t)dt = F(x) + G_\nu(x).$$

Since the Fredholm equation is linear, the solution of this equation has the form

$$f^*(t) = f(t) + \sin \nu t,$$

where $f(t)$ is the solution of equation (7.13). For sufficiently large ν , the right hand side of this equation differs from the right hand side of (7.13) only by the small amount $G_\nu(x)$, while its solution differs by the amount $\sin \nu t$.

Note that our equations (7.1), (7.4), and (7.5) also belong to the Fredholm equation of type 1. One can rewrite them as follows:

$$\int_I \theta(x - x')p(x')dx' = F(x')$$

$$\int_I \theta(x - x') P(\omega|x') dF(x') = F(\omega, x),$$

$$\int \int_I \theta(y - y') \theta(x - x') p(y'|x') dF(x') dy' = F(y, x)$$

Recall that for simplicity we suppose that x (pairs (x, y)) belongs to the unit cube I .

7.5 THREE METHODS OF SOLVING ILL-POSED PROBLEMS

In the 1960s three methods of solving ill-posed problems were proposed. All of them are based on introducing the so-called regularization functional $\Omega(f)$.

The regularization functional $\Omega(f)$ is a semicontinuous, positive functional for which $\Omega(f) \leq c$, $c > 0$, is a compactum (in the space of functions f). It is defined on the set of functions $f \in \mathcal{F}$, the domain of solution of the equations.

Below, to impose uniqueness of the solution we consider regularization functionals possessing the following properties:

1. $\Omega(f)$ is a nonnegative convex functional. That is, for any $0 \leq \lambda \leq 1$ the inequality

$$\Omega(\lambda f_1 + (1 - \lambda) f_2) \leq \lambda \Omega(f_1) + (1 - \lambda) \Omega(f_2), \quad f_1, f_2 \in \mathcal{F},$$

is valid.

2. The following equality holds:

$$\Omega(0) = 0.$$

3. For each fixed f the function

$$r(\gamma) = \Omega(\gamma f)$$

is a strictly increasing function of γ .

On the basis of the regularization functional the following three methods were proposed:

1. *Tikhonov's Variation Method* (Method T) [Tikhonov, 1963].

Minimize the functional

$$W_T(f) = \|Af - F\|_{E_2}^2 + \gamma \Omega(f),$$

where $\gamma > 0$ is some predefined constant.

2. *Phillips Residual Method* (Method P) [Phillips, 1962].

Minimize the functional

$$W_P(f) = \Omega(f),$$

subject to the constraint

$$\|Af - F\|_{E_2} \leq \mu,$$

where $\mu > 0$ is some predefined constant.

3. *Ivanov's Quasi-Solution Method* (Method I) [Ivanov, 1962].

Minimize the functional

$$W_I(f) = \|Af - F\|_{E_2}$$

subject to the constraint

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

where $C > 0$ is some predefined constant.

It was shown (Vasin, (1970)) that these methods are equivalent in the sense that if one of the methods (say Method T) for a given value of the parameter (say γ^*) produces a solution f^* , then there exist corresponding values of parameters of the other two methods that produce the same solution.

7.5.1 The Residual Principle

All three methods for solving ill-posed problems contain one free parameter (parameter γ for Method T, parameter σ for Method P, and parameter C for Method I). The choice of the appropriate value of the parameter is crucial for obtaining a good solution of an ill-posed problem.

In the theory of solving ill-posed problems there exists a general principle for choosing such a parameter, the so-called *residual principle* [Morozov, 1983].

Suppose that we know the accuracy of approximation of the right-hand side F of equation (7.12) by a function F_δ , that is we know the value σ for which the following equality holds:

$$\|F - F_\delta\|_{E_2} = \sigma.$$

Then the residual principle suggests that we choose a parameter (γ_ℓ for Method T or C_ℓ for Method I) that produces the solution f_δ satisfying the equality

$$\|Af_\delta - F_\delta\|_{E_2} = \sigma \tag{7.14}$$

(for Method P one chooses the solution that exactly satisfies the constraint (7.14) with σ).

Usually, it is not easy to obtain an accurate estimate of the discrepancy between the exact right-hand side and a given approximation.

Fortunately, it is not the case for our problems of estimating the density, conditional probability, and conditional density. For these problems there exist accurate estimates of the value $\sigma = \sigma_\ell$, which depends on the number of examples ℓ and the dimensionality of the space n .

Note that common to all our three problems is the fact that the right-hand sides of the equations are probability distribution functions. In our solution, instead of actual distribution functions we use empirical distribution functions. As we discuss in Section 7.3, for any fixed number of observations ℓ and any fixed dimensionality n of the space there exists a universal distribution of discrepancy

$$\xi = \sqrt{\ell} \sup_x |F(x) - F_\ell(x)|.$$

Let us take an appropriate quantile q^* of this distribution (say 50% quantile) and choose

$$\sigma = \sigma_\ell = \frac{q^*}{\sqrt{\ell}}. \quad (7.15)$$

In the following we will choose solutions that satisfy the residual principle with constant (7.15).

7.6 MAIN ASSERTIONS OF THE THEORY OF ILL-POSED PROBLEMS

In this section we will describe the main theorem for the Tikhonov method. Since all methods are equivalent, analogous assertions are valid for the two other methods.

7.6.1 Deterministic Ill-Posed Problems

Suppose that instead of the exact right hand side of the operator equation

$$Af = F$$

we are given approximations F_δ such that

$$\|F_\delta - F\|_{E_2} \leq \delta. \quad (7.16)$$

Our goal is to specify the relationship between the value $\delta > 0$ and the regularization parameter $\gamma_\delta > 0$ in such a way that the solution of our

regularization method converges to the desired one as soon as δ converges to zero.

The following theorem establishes these relations [Tikhonov and Arsenin 1977].

Theorem 7.1 *Let E_1 and E_2 be metric spaces, and suppose for $F \in E_2$ there exists a solution $f \in E_1$ of equation (7.12). Let instead of an exact right-hand side F of equation (7.12), approximations $F_\delta \in E_2$ be given such that $\rho_{E_2}(F, F_\delta) \leq \delta$. Suppose the values of the parameter $\gamma(\delta)$ are chosen in such a manner that*

$$\gamma(\delta) \rightarrow 0 \text{ for } \delta \rightarrow 0,$$

$$\lim_{\delta \rightarrow 0} \frac{\delta^2}{\gamma(\delta)} \leq r < \infty. \quad (7.17)$$

Then the elements $f_\delta^{\gamma(\delta)}$ minimizing the functionals $W_T(f)$ on E_1 converge to the exact solution f as $\delta \rightarrow 0$.

In a Hilbert space the following theorem is valid.

Theorem 7.2. *Let E_1 be a Hilbert space and $\Omega(f) = \|f\|^2$. Then for $\gamma(\delta)$ satisfying the relations*

$$\gamma(\delta) \rightarrow 0 \text{ for } \delta \rightarrow 0,$$

$$\lim_{\delta \rightarrow 0} \frac{\delta^2}{\gamma(\delta)} = 0, \quad (7.18)$$

the functions $f_\delta^{\gamma(\delta)}$ minimizing the functional

$$W_T^*(f) = \rho_{E_2}^2(Af, F_\delta) + \gamma\Omega(f) \quad (7.19)$$

converge as $\delta \rightarrow 0$ to the exact solution f in the metric of the space E_1 .

7.6.2 Stochastic Ill-Posed Problem

Consider now the situation where instead of the right-hand side of the equation

$$Af = F \quad (7.20)$$

we are given a sequence of random functions F_ℓ that converge in probability to F . That is, we are given a sequence $F_1, \dots, F_\ell, \dots$ for which the following equation holds true:

$$\lim_{\ell \rightarrow \infty} P\{\rho_{E_2}(F_\ell, F) > \varepsilon\} = 0, \quad \forall \varepsilon > 0.$$

Our goal is to use the sequence $F_1, \dots, F_\ell, \dots$ to find a sequence of solutions of equation (7.20) that converge in probability to the true solution. We call this problem the *stochastic ill-posed problem*, since we are solving our equation using random functions $F_\ell(x)$.

To solve these stochastic ill-posed problems we use Method T. For any F_ℓ we minimize the functional

$$W_T(f) = \rho_{E_2}^2(Af, F_\ell) + \gamma_\ell \Omega(f),$$

finding the sequence $f_1, \dots, f_\ell, \dots$. Below we consider the case where

$$\gamma_\ell \rightarrow 0 \quad \text{as } \ell \rightarrow \infty.$$

Under these conditions the following theorems describing the relationship between the distributions of two random variables, the random variable $\rho_{E_2}(F, F_\ell)$ and the random variable $\rho_{E_1}(f, f_\ell)$ hold true [Vapnik and Stefnyuk, 1978].

Theorem 7.3. *For any positive numbers ε and μ there exists a positive number $n(\varepsilon, \mu)$ such that for all $\ell > n(\varepsilon, \mu)$ the inequality*

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \leq P\{\rho_{E_2}(F_\ell, F) > \sqrt{\gamma_\ell \mu}\} \quad (7.21)$$

is satisfied.

For the case where E_1 is a Hilbert space the following theorem holds true.

Theorem 7.4. *Let E_1 be a Hilbert space, A in (7.20) be a linear operator, and*

$$W(f) = \|f\|^2 = (f, f).$$

Then for any positive ε there exists a number $n(\varepsilon)$ such that for all $\ell > n(\varepsilon)$ the inequality

$$P\{\|f_\ell - f\|^2 > \varepsilon\} < 2P\{\rho_{E_2}^2(F_\ell, F) > \frac{\varepsilon}{2} \gamma_\ell\}$$

is satisfied.

These theorems are generalizations of Theorem 7.1 and Theorem 7.2 for the stochastic case.

Corollary. From Theorems 7.3 and 7.4 it follows that if approximations F_ℓ of the right-hand side of the operator equation (7.20) converge in probability to the true function $F(x)$ in the metric of space E_2 with the rate

$$\rho_{E_2}(F(x), F_\ell(x)) \leq r(\ell),$$

then the sequence of the solutions to equation (7.20) converges in probability to the desired one if

$$\lim_{\ell \rightarrow \infty} \frac{r(\ell)}{\sqrt{\gamma_\ell}} = 0$$

and γ_ℓ converges to zero with $\ell \rightarrow \infty$.

7.7 NONPARAMETRIC METHODS OF DENSITY ESTIMATION

7.7.1 Consistency of the Solution of the Density Estimation Problem

Consider now our integral equation

$$\int_{-\infty}^x f(x') dx' = F(x).$$

Let us solve this equation using empirical distribution functions $F_1, \dots, F_\ell, \dots$ instead of the actual distribution function. For different ℓ we minimized the functional

$$W_T(f) = \rho_{E_2}^2(Af, F_\ell) + \gamma_\ell \Omega(f),$$

where we chose the metric $\rho_{E_2}(Af, F_\ell)$ such that

$$\rho_{E_2}(Af, F_\ell) \leq \sup_x |(Af)x - F_\ell(x)|. \quad (7.22)$$

Suppose that

$$f_1, \dots, f_\ell, \dots$$

is a sequence of the solutions obtained.

Then according to Theorem 7.3, for any ε and any μ the inequality

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \leq P\{\sup_x |F_\ell(x) - F(x)| > \sqrt{\gamma_\ell \mu}\}$$

holds true for sufficiently large ℓ .

Since the VC dimension of the set of events (7.9) is bounded (equal to the dimensionality of the space) for sufficiently large ℓ , the inequality

$$P\{\sup_x |F_\ell(x) - F(x)| > \epsilon\} \leq C \exp\{-\epsilon^2 \ell\}$$

holds true (see bounds (3.3) and (3.23)). Therefore, there exists an $\ell(\varepsilon, \mu)$ such that for $\ell > \ell(\varepsilon, \mu)$ the inequality

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \leq C \exp\{-\gamma_\ell \mu \ell\} \quad (7.23)$$

is satisfied.

If $f(x) \in L_2$, it then follows from Theorem 7.4 and from the VC bound that for sufficiently large ℓ , the inequality

$$P\left\{\int (f_\ell(x) - f(x))^2 dx > \varepsilon\right\} \leq C \exp\{-\gamma_\ell \mu \ell\} \quad (7.24)$$

holds.

Inequalities (7.23) and (7.24) imply that the solution f_ℓ converges in probability to the desired one (in the metric $\rho_{E_1}(f_\ell, f)$) if

$$\begin{aligned} \gamma_\ell &\xrightarrow{\ell \rightarrow \infty} 0, \\ \frac{\ell}{\ln \ell} \gamma_\ell &\xrightarrow{\ell \rightarrow \infty} \infty. \end{aligned} \quad (7.25)$$

(In this case the right-hand sides of equations (7.23) and (7.24) converge to zero.)

One can also show (using the Borel-Cantelli lemma) that solutions converge with probability one if

$$\begin{aligned} \gamma_\ell &\xrightarrow{\ell \rightarrow \infty} 0, \\ \ell \gamma_\ell &\xrightarrow{\ell \rightarrow \infty} \infty. \end{aligned}$$

Note that this assertion is true for any regularization functional $\Omega(f)$ and for any metric $\rho_{E_1}(f, f_\ell)$ satisfying (7.22). Choosing specific functionals $\Omega(f)$ and a specific metric $\rho_{E_2}(F, F_\ell)$ satisfying the condition

$$\rho_{E_2}(F, F_\ell) \leq \sup_x |F_1(x) - F_2(x)|,$$

one constructs a specific estimator of the density.

7.7.2 The Parzen's Estimators

Let us specify the metric $\rho_{E_2}(F, F_\ell)$ and such functionals $\Omega(f)$ for which Method T minimizing the functional

$$W(f) = \rho_2^2(Af, F_\ell) + \gamma_\ell \Omega(f) \quad (7.26)$$

produces Parzen's estimators.

Consider L_2 metrics in the set of functions F ,

$$\rho_{E_2}(F, F_\ell) = \sqrt{\int_{-\infty}^{\infty} (F(x) - F_\ell(x))^2 dx},$$

and the regularization functional.

$$\Omega(f) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} R(z-x)f(x)dx \right)^2 dz.$$

Here $R(z-x)$ is the kernel that defined the linear operator

$$Bf = \int_{-\infty}^{\infty} R(z-x)f(x)dx.$$

In particular, if $R(z-x) = \delta^p(z-x)$, the operator

$$Bf = \int_{-\infty}^{\infty} \delta^p(z-x)f(x)dx = f^{(p)}(x)$$

defines the p th derivative of the function $f(x)$.

For these elements we have the functional

$$W_T(f)$$

$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^x f(t)dt - F_\ell(x) \right)^2 dx + \gamma_\ell \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} R(z-x)f(x)dx \right)^2 dz. \quad (7.27)$$

Below we show that the estimator f_γ that minimizes this functional is the Parzen's estimator

$$f_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} G_\gamma(x - x_i),$$

where the kernel function $G_\gamma(u)$ is defined by the kernel function $R(u)$.

Indeed, let us denote by $\bar{f}(\omega)$ the Fourier transform of the function $f(t)$ and by $\bar{R}(\omega)$ the Fourier transform of the function $R(x)$. Then one can evaluate the Fourier transform for the function $F(x)$,

$$\begin{aligned} \bar{F}(\omega) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(x)e^{-i\omega x} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\omega x} \int_{-\infty}^x f(t)dt = \frac{\bar{f}(\omega)}{i\omega}, \end{aligned}$$

and for the function $F_\ell(x)$,

$$\bar{F}_\ell(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_\ell(x)e^{-i\omega x} dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\ell} \sum_{j=1}^{\ell} \theta(x - x_j) e^{-i\omega x} dx$$

$$= \frac{1}{\ell} \sum_{j=1}^{\ell} \frac{e^{-i\omega x_j}}{i\omega}.$$

Note that the Fourier transform for the convolution of two functions is equal to the product of the Fourier transforms of these two functions. In our case this means that

$$\begin{aligned} & \frac{1}{2\pi} \int_{-\infty}^{\infty} (R(x) * f(x)) e^{-i\omega x} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} R(z-x) f(x) dx \right) e^{-i\omega z} dz = \bar{R}(\omega) \bar{f}(\omega). \end{aligned}$$

Lastly, recall that according to Parseval's equality the L_2 norm of any function $f(x)$ is equal (within the constant $1/2\pi$) to the L_2 norm of its Fourier transform $\bar{f}(\omega)$ (here $\bar{f}(\omega)$ is the Fourier transform of the function $f(x)$). Therefore, one can rewrite (7.27) in the form

$$\bar{W}_T(f) = \left\| \frac{\bar{f}(\omega) - \frac{1}{\ell} \sum_{j=1}^{\ell} e^{-i\omega x_j}}{i\omega} \right\|_{L_2}^2 + \gamma_\ell \|\bar{R}(\omega) \bar{f}_\ell(\omega)\|_{L_2}^2.$$

This functional is quadratic with respect to $\bar{f}(\omega)$.

Therefore, the condition for its minimum is

$$\frac{\bar{f}_\ell(\omega)}{\omega^2} - \frac{1}{\ell\omega^2} \sum_{j=1}^{\ell} e^{i\omega x_j} + \gamma_\ell \bar{R}(\omega) \bar{R}(-\omega) \bar{f}(\omega) = 0. \quad (7.28)$$

Solving this equation with respect to $\bar{f}_\ell(\omega)$, one obtains

$$\bar{f}_\ell(\omega) = \left(\frac{1}{1 + \gamma_\ell \omega^2 \bar{R}(\omega) \bar{R}(-\omega)} \right) \frac{1}{\ell} \sum_{j=1}^{\ell} e^{-i\omega x_j}.$$

Let us introduce the notation

$$g_{\gamma_\ell}(\omega) = \frac{1}{1 + \gamma_\ell \omega^2 \bar{R}(\omega) \bar{R}(-\omega)}$$

and

$$G_{\gamma_\ell}(x) = \int_{-\infty}^{\infty} g_{\gamma_\ell}(\omega) e^{i\omega x} d\omega.$$

To obtain an approximation to the density one has to evaluate the inverse Fourier transform

$$\begin{aligned} f_\ell(x) &= \int_{-\infty}^{\infty} \bar{f}_\ell(\omega) e^{i\omega x} d\omega = \int_{-\infty}^{\infty} g_{\gamma_\ell}(\omega) \left(\frac{1}{\ell} \sum_{j=1}^{\ell} e^{-i\omega x_j} \right) e^{i\omega x} d\omega \\ &= \frac{1}{\ell} \sum_{j=1}^{\ell} \int_{-\infty}^{\infty} g_{\gamma_\ell}(\omega) e^{i\omega(x-x_j)} d\omega = \frac{1}{\ell} \sum_{j=1}^{\ell} G_{\gamma_\ell}(x - x_j). \end{aligned}$$

The last expression is the Parzen's estimator with kernel function $G_{\gamma_\ell}(u)$.

7.8 SVM SOLUTION OF THE DENSITY ESTIMATION PROBLEM

Now we consider another solution of the operator equation (the density estimation problem)

$$\int_{-\infty}^x p(x')dx' = F(x)$$

with approximation $F_\ell(x)$ on the right-hand side instead of $F(x)$.

We will solve this problem using Method P, where we consider the distance between $F(x)$ and $F_\ell(x)$ defined by the uniform metric

$$\rho_{E_2}(F(x), F_\ell(x)) = \sup_x |F(x) - F_\ell(x)| \quad (7.29)$$

and the regularization functional

$$\Omega(f) = (f, f)_H \quad (7.30)$$

defined by a norm of some reproducing kernel Hilbert space (RKHS).

To define the RKHS one has to define a symmetric positive definite kernel $K(x, y)$ and an inner product $(f, g)_H$ in Hilbert space H such that

$$(f(x), K(x, y))_H = f(y) \quad \forall f \in H \quad (7.31)$$

(the reproducing property). Note that any symmetric positive definite function $K(x, y)$ has an expansion

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y), \quad (7.32)$$

where λ_i and $\phi_i(x)$ are eigenvalues and eigenfunctions of the operator

$$Df = \int K(x, y)f(y)dy.$$

Consider the set of functions

$$f(x, c) = \sum_{i=1}^{\infty} c_i \phi_i(x), \quad (7.33)$$

for which we introduce the inner product

$$(f(x, c^*), f(x, c^{**}))_H = \sum_{i=1}^{\infty} \frac{c_i^* c_i^{**}}{\lambda_i}. \quad (7.34)$$

The kernel (7.32), inner product (7.34), and set (7.33) define an RKHS.

Indeed,

$$\begin{aligned} (f(x), K(x, y))_H &= \left(\sum_{i=1}^{\infty} c_i \phi_i(x), K(x, y) \right)_H \\ &= \left(\sum_{i=1}^{\infty} c_i \phi_i(x), \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y) \right)_H = \sum_{i=1}^{\infty} \frac{c_i \lambda_i \phi_i(y)}{\lambda_i} = f(y). \end{aligned}$$

For functions from an RKHS the functional (7.30) has the form

$$\Omega(f) = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i}, \quad (7.35)$$

where λ_i is the i th eigenvalue of the kernel $K(x, y)$. Therefore, the choice of the kernel defines smoothness requirements to the solution.

To solve the density estimation problem we use Method P with the functional defined by (7.30) and uniform metric (7.29). We choose the value of the parameter $\sigma = \sigma_\ell$ in the constraint to satisfy residual principle (7.14). Therefore, we minimize the functional

$$\Omega(f) = (f, f)_H$$

subject to the constraints

$$\sup_x \left| F_\ell(x) - \int_{-\infty}^x f(x') dx' \right| = \sigma_\ell.$$

However, for computational reasons we consider the constraints defined only at the points x_i of the training set

$$\max_i \left| F_\ell(x) - \int_{-\infty}^x f(x') dx' \right|_{x=x_i} = \sigma_\ell, \quad 1 \leq i \leq \ell.$$

We look for a solution of our equation in the form

$$f(x) = \sum_{i=1}^{\ell} \beta_i K(x_i, x), \quad (7.36)$$

where $K(x_i, x)$ is the same kernel that defines the RKHS. Taking into account (7.31) and (7.36) we rewrite functional (7.30) as follows:

$$\begin{aligned} \Omega(f) &= (f, f)_H \\ &= \left(\sum_{i=1}^{\ell} \beta_i K(x, x_i), \sum_{i=1}^{\ell} \beta_i K(x, x_i) \right)_H \\ &= \sum_{i=1}^{\ell} \beta_i \sum_{j=1}^{\ell} \beta_j (K(x, x_i), K(x, x_j))_H \end{aligned}$$

$$= \sum_{i,j=1}^{\ell} \beta_j \beta_i K(x_i, x_j). \quad (7.37)$$

To obtain the last equation we used the reproducing property (7.31).

Therefore, to solve our equation we minimize the functional

$$W(\beta) = \Omega(f, f) = \sum_{i,j=1}^{\ell} \beta_i \beta_j K(x_i, x_j) \quad (7.38)$$

subject to the constraints

$$\max_i \left| F_{\ell}(x) - \sum_{j=1}^{\ell} \beta_j \int_{-\infty}^x K(x_j, x') dx' \right|_{x=x_i} = \sigma_{\ell}, \quad 1 \leq i \leq \ell, \quad (7.39)$$

where the largest deviation defines the equality (the residual principle).

This optimization problem is closely related to the SV regression problem with an σ_{ℓ} -insensitive zone. It can be solved using the SVM technique (see Chapter 6).

To obtain the solution in the form of a mixture of densities we choose a nonnegative kernel $K(x, x_i)$ satisfying the following conditions, which we call the *condition K*:

1. The kernel has the form

$$K_{\gamma}(x, x_i) = a(\gamma) K\left(\frac{x - x_i}{\gamma}\right), \quad (7.40)$$

$$a(\gamma) \int K\left(\frac{x - x_i}{\gamma}\right) dx = 1, \quad K(0) = 1, \quad (7.41)$$

where $a(\gamma)$ is the normalization constant.

2. The value of the parameter γ affects the eigenvalues $\lambda_1(\gamma), \dots, \lambda_k(\gamma) \dots$ defined by the kernel. We consider such kernels for which the ratios $\lambda_{k+1}(\gamma)/\lambda_k(\gamma)$, $k = 1, 2, \dots$, decrease when γ increases. Examples of such functions are

$$K_{\gamma}(x, x_i) = a(\gamma) \exp\left(-\left|\frac{x - x_i}{\gamma}\right|^p\right), \quad 0 < p \leq 2. \quad (7.42)$$

Also, to obtain the solution in the form of a mixture of densities we add two more constraints:

$$\beta_i \geq 0, \quad \sum_{i=1}^{\ell} \beta_i = 1. \quad (7.43)$$

Note that our target functional also depends on the parameter γ :

$$W_\gamma(\beta) = \Omega(f) = \sum_{i,j=1}^{\ell} \beta_j \beta_i K_\gamma(x_i, x_j). \quad (7.44)$$

We call the value of the parameter γ *admissible* if for this value there exists solution of our optimization problem (the solution satisfies residual principle (7.14)).

The admissible set

$$\gamma_{\min} \leq \gamma_\ell \leq \gamma_{\max}$$

is not empty, since for Parzen's method (which also has form (7.36)) such a value does exist. Recall that the value γ_ℓ in the kernel determines the smoothness requirements on the solution: The larger the γ , the smaller the ratio λ_{k+1}/λ_k , and therefore functional (7.35) imposes stronger smoothness requirements.

For any admissible γ the SVM technique provides the unique solution with some number of elements in the mixture. We choose the solution corresponding to an admissible γ_ℓ that minimizes the functional (7.44) over both coefficients β_i and parameter γ . This choice of parameter controls the accuracy of the solution. By choosing a large admissible γ_ℓ we achieve another goal: We increase the smoothness requirements to the solution satisfying (7.14) and we select the solution with a small number of mixture elements⁶ (a small number of support vectors; see Section 6.7). One can continue to increase sparsity (by increasing σ_ℓ in (7.14)), trading sparsity for the accuracy of the solution.

7.8.1 The SVM Density Estimate: Summary

The SVM solution of the density estimation equation using Method P implements the following ideas:

1. The target functional in the optimization problem is defined by the norm of RKHS with kernel (depending on one parameter) that allows effective control of the smoothness properties of the solution.

⁶Note that we have two different descriptions of the same functional: description (7.35) in a space of functions $\phi_k(x)$ and description (7.44) in kernels $K(x, x_i)$. From (7.35) it follows that in increasing γ we require more strong filtration of the "high-frequency components" of the expansion in the space ϕ_k . It is known that one can estimate densities in a high-dimensional space using a small number of observations only if the target density is smooth (can be described by "low-frequency functions"). Therefore, in high-dimensional space the most accurate solution often corresponds to the largest admissible γ . Also, in our experiments we observed that within the admissible set the difference in accuracy obtained for solutions with different γ is not significant.

2. The solution of the equation is chosen in the form of an expansion (with nonnegative weights) on the same kernel function that defines the RKHS.
3. The distance $\rho_{E_2}(Af_\ell, F_\ell)$ defining the optimization constraints is given by the uniform metric (which allows effective use of the residual principle).
4. The solution satisfies the residual principle with the value of residual (depending only on the dimensionality and the number of observations) obtained from a Kolmogorov-Smirnov type distribution.
5. The admissible parameter γ of the kernel is chosen to control accuracy of the solution and/or sparsity of the solution.

7.8.2 Comparison of the Parzen's and the SVM methods

Note that two estimators, the Parzen's estimator

$$f_P(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} G_\gamma(x, x_i) \quad (7.45)$$

and the SVM estimator

$$f_{SVM}(x) = \sum_{i=1}^{\ell} \beta_i K_\gamma(x, x_i),$$

have the same structure. In the case where

$$G_\gamma(x, x_i) = K_\gamma(x, x_i)$$

and

$$\beta_i = \frac{1}{\ell},$$

the SVM estimator coincides with the Parzen's estimator. The solution (7.45), however, is not necessarily the solution of our optimization problem. Nevertheless, one can show that the less smooth the SVM admissible solution is, the closer it is to Parzen's solution. Indeed, the smaller is γ in the kernel function $a(\gamma)K\left(\frac{|x_j - x_i|}{\gamma}\right)$, the better the functional

$$W(\beta) = a(\gamma) \sum_{i=1}^{\ell} \beta_i^2 \quad (7.46)$$

approximates our target functional (7.38).

Parzen's type estimator is the solution for the smallest admissible γ of the following optimization problem: Minimize (over β) functional (7.46) (instead of functional (7.38)) subject to constraints (7.39) and (7.43).

Therefore, Parzen's estimator is the less sparse admissible SVM solution of this (modified) optimization problem.

Below we compare solutions obtained by Parzen's method to the solution obtained by the SVM method for different admissible values of the parameter γ . We estimated a density in the two-dimensional case defined by a mixture of two Laplacians;

$$p(x, y) = \frac{1}{8} (\exp\{-(|x - 1| + |y - 1|2)\} + \exp\{-(|x + 1|2 + |y + 1|)\}).$$

In both methods we used the same Gaussian kernels

$$G_\gamma(x, y; x', y') = K_\gamma(x, y; x', y') = \frac{1}{2\pi\gamma^2} \exp\left\{-\frac{(x - x')^2 + (y - y')^2}{2\gamma^2}\right\}$$

and defined the best parameter γ using the residual principle with $\sigma_\ell = q/\sqrt{\ell}$ and $q = 1.2$.

In both cases the density was estimated from 200 observations. The accuracy of approximation was measured in the L_1 metric

$$\Delta_\ell = \int |p_\ell(x, y) - p(x, y)| dx dy.$$

We conducted 100 such trials and constructed a distribution over the obtained values q for these trials. This distribution is presented by boxplots. The horizontal lines of the boxplot indicate 5%, 25%, 50%, 75%, and 95% quantiles of the error distribution.

Figures 7.1 and 7.2 demonstrate the trade-off between accuracy and sparsity. Figure 7.1a displays the distribution of the L_1 error, and Figure 7.1b displays the distribution of the number of terms for the Parzen's method, and for the SVM method with $\gamma_\ell = 0.9$, $\gamma_\ell = 1.1$, for the largest admissible γ_ℓ . Figure 7.2a displays the distribution of the L_1 error, and Figure 7.2b displays the distribution of the number of terms, where instead of the optimal $\sigma_\ell = q/\sqrt{\ell}$ in (9) we use $\sigma_\ell = mq/\sqrt{\ell}$ with $m = 1, 1.5, 2.1$.

7.9 CONDITIONAL PROBABILITY ESTIMATION

In this section to estimate conditional probability, we generalize the SVM density estimation method described in the previous section. Using the same ideas we solve the equation

$$\int_{-\infty}^x p(\omega|x') dF(x') = F(\omega, x) \quad (7.47)$$

when the probability distribution functions $F(x)$ and $F(x, y)$ are unknown, but data

$$(w_1, x_1), \dots, (w_\ell, x_\ell)$$

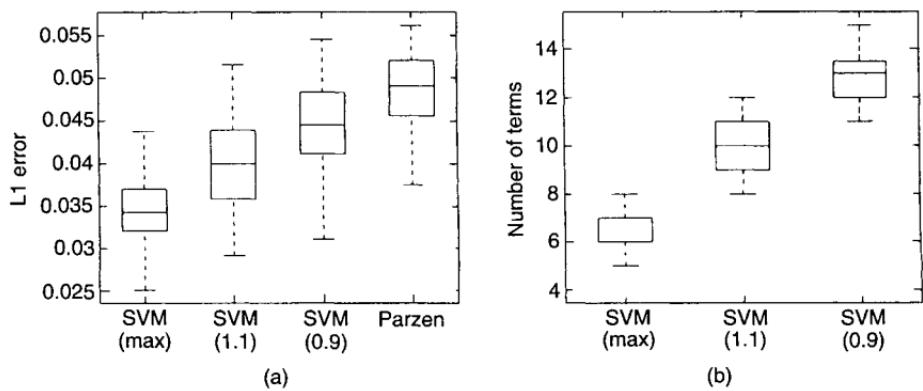


FIGURE 7.1. (a) A boxplot of the L_1 error for the SVM method with $\gamma_\ell = \gamma_{\max}$, $\gamma_\ell = 1.1$, $\gamma_\ell = 0.9$, and Parzen's method (the same result as SVM with $\gamma_\ell = \gamma_{\min}$). (b) A boxplot of the distribution on the number of terms for the corresponding cases.

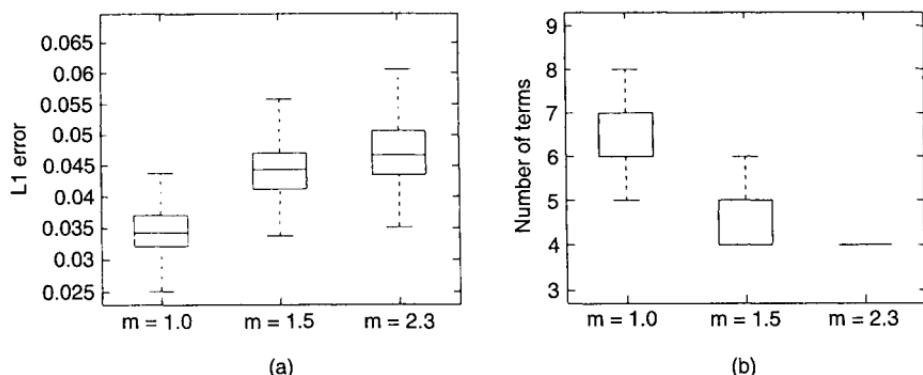


FIGURE 7.2. (a) A boxplot of the L_1 error for the SVM method with $\gamma_\ell = \gamma_{\max}$ where we use $\sigma_\ell = mq/\sqrt{\ell}$ with $m = 1, 1.5, 2.3$. (b) A boxplot of the distribution of the number of terms for the corresponding cases.

are given.

Below we first describe conditions under which one can obtain solutions of equations with both the right-hand side and the operator approximately defined, and then we describe the SVM method for conditional probability estimation.

7.9.1 Approximately Defined Operator

Consider the problem of solving the operator equation

$$Af = F$$

under the condition that (random) approximations are given not only for the function on the right-hand side of the equation but for the operator as well. We assume that instead of the exact operator A we are given a sequence of approximations A_ℓ , $\ell = 1, 2, \dots$ defined by a sequence of random continuous operators that converge in probability (below we will specify the definition of closeness of two operators) to the operator A .

As before, we consider the problem of solving the operator equation by Method T, that is, by minimizing the functional

$$W(f) = \rho_{E_2}^2(A_\ell f, F_\ell) + \gamma_\ell \Omega(f).$$

We measure the closeness of operator A and operator A_ℓ by the distance

$$\|A_\ell - A\| = \sup_f \frac{\rho_{E_2}(A_\ell f, Af)}{\Omega^{1/2}(f)}. \quad (7.48)$$

The following theorem is true [Stefanyuk, 1986].

Theorem 7.5. *For any $\varepsilon > 0$ and any constants $C_1, C_2 > 0$ there exists a value $\gamma_0 > 0$ such that for any $\gamma_\ell \leq \gamma_0$ the inequality*

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\}$$

$$\leq P\{\rho_{E_2}(F_\ell, F) > C_1 \sqrt{\gamma_\ell}\} + P\{\|A_\ell - A\| > C_2 \sqrt{\gamma_\ell}\} \quad (7.49)$$

holds true.

Corollary. From this theorem it follows that if the approximations $F_\ell(x)$ of the right-hand side of the operator equation converge in probability to the true function $F(x)$ in the metric of the space E_2 with the rate of convergence $r(\ell)$, and the approximations A_ℓ converge in probability to the true operator A in the metric defined in (7.48) with the rate of convergence $r_A(\ell)$, then there exists a function

$$r_0(\ell) = \max\{r(\ell), r_A(\ell)\} \rightarrow_{\ell \rightarrow \infty} 0$$

such that the sequence of solutions to the equation converges in probability to the desired one if

$$\frac{r_0(\ell)}{\sqrt{\gamma_\ell}} \xrightarrow{\ell \rightarrow \infty} 0$$

and γ_ℓ converges to zero with $\ell \rightarrow \infty$.

Let x belongs to some bounded support $|x| \leq C^*$. The following theorem holds true.

Theorem 7.6 *If the functional $\Omega(f)$ satisfies the condition*

$$\Omega(f) \geq C \left(\sup_x |f(x)| + \sup_x |f'(x)| \right)^2 \quad (7.50)$$

and the metric in E_2 satisfies the condition

$$\rho_{E_2}(Af_\ell, Af) \leq \sup_x |(Af_\ell)x - (Af)x|, \quad (7.51)$$

then estimation of conditional probability using Method T is consistent.

That is, if the regularization is sufficiently strong comparing to the metric $\rho_{E_2}(\cdot, \cdot)$, then the method of estimating a conditional density by solving the approximately defined integral equation is consistent.

Indeed, consider the difference

$$\begin{aligned} |(Af_\ell)(x) - (Af)(x)| &= \left| \int_{-\infty}^x f(x') d(F_\ell(x') - F(x')) \right| \\ &= \left| f(x)(F_\ell(x) - F(x)) - \int_{-\infty}^x f'(x')(F_\ell(x') - F(x')) dx' \right| \\ &\leq \sup_x |f(x)| |F_\ell(x) - F(x)| + \sup_x |f'(x)| \int_0^x |F_\ell(x) - F(x)| dx. \end{aligned}$$

Taking into account (7.51), (7.50), and the fact that vectors x belong to the bounded support we have,

$$\begin{aligned} &\|A_\ell f - Af\|_{E_2} \\ &\leq (\sup_x |f(x)| + C^* \sup_x |f'(x)|) \sup_x |F_\ell(x) - F(x)| \leq \Omega^{1/2}(f) \sup_x |F_\ell(x) - F(x)|. \end{aligned} \quad (7.52)$$

From this inequality and the definition of the norm of the operator we have

$$\|A_\ell - A\| = \sup_f \frac{\|A_\ell f - Af\|_{E_2}}{\Omega^{1/2}(f)} \leq \sup_x |F_\ell(x) - F(x)|. \quad (7.53)$$

According to Theorem 7.5 the solution f_γ of the operator equation obtained on the basis of the regularization method possesses the following

properties: For any $\varepsilon > 0$, $C_1 > 0$, $C_2 > 0$ there exists γ_0 such that for any $\gamma_\ell < \gamma_0$ the inequality

$$\begin{aligned} & P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \\ & \leq P\{\rho_{E_2}(F_\ell, F) > C_1\sqrt{\gamma_\ell}\} + P\{\|A_\ell - A\| > C_2\sqrt{\gamma_\ell}\} \\ & \leq P\{\sup_x |F_\ell(x) - F(x)| > C_1\sqrt{\gamma_\ell}\} + P\{\sup_x |F_\ell(x) - F(x)| > C_2\sqrt{\gamma_\ell}\} \end{aligned}$$

holds true. Therefore, taking into account the bounds for uniform convergence over the set of events (7.9) with VC dimension n , we obtain for sufficiently large ℓ the inequality (see bounds (3.3) and (3.23))

$$\begin{aligned} & P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \\ & \leq P\{\sup_x |F_\ell(x) - F(x)| > C_1\sqrt{\gamma_\ell}\} + P\{\sup_x |F_\ell(x) - F(x)| > C_2\sqrt{\gamma_\ell}\} \\ & \leq C(\exp\{-\gamma_\ell C_1\} + \exp\{-\gamma_\ell C_2\}). \end{aligned}$$

From this inequality we find that conditions (7.50) and (7.51) imply convergence in probability and convergence almost surely to the desired one.

7.9.2 SVM Method for Conditional Probability Estimation

Now we generalize the method obtained for solving density estimation equation to solving the conditional probability equation

$$\int_{-\infty}^x p(w|x')dF(x') = F(w, x) = p(w)F(x|w), \quad (7.54)$$

where we use the empirical distribution functions $F_\ell(x)$ and $F_\ell(x|w)$ instead of the actual distribution functions $F(x)$ and $F(x|w)$.

In our solution we follow the steps described in Section 7.8.

1. We use Method P with the target functional as a norm in RKHS defined by a kernel $K_\gamma(x, x')$ satisfying conditions \mathcal{K} (See Section 7.8):

$$\Omega(f) = (f, f)_H$$

2. We are looking for the solution in the form

$$f_w(x) = p(w|x) = p(w) \sum_{i=1}^{\ell} \beta_i K_\gamma(x, x_i) \quad (7.55)$$

with nonnegative coefficients β .

Therefore, we have to minimize the functional

$$W_\gamma(\beta) = \Omega(f) = \sum_{i,j=1}^{\ell} \beta_i \beta_j K_\gamma(x_j, x_i)$$

(see Section 7.8).

3. We define optimization constraints from the equality

$$\sup_x |(A_\ell f)x - F_\ell(w, x)| = \sigma^* p(w),$$

which for our equations has the form

$$\begin{aligned} \sup_x & \left| \int_{-\infty}^x p(w) \sum_{i=1}^{\ell} \beta_i K_\gamma(x, x_i) d \left[\frac{1}{\ell} \sum_{j=1}^{\ell} \theta(x - x_j) \right] - p(w) F_\ell(x|w) \right| \\ & = \sigma^* p(w). \end{aligned}$$

After obvious calculations we obtain the optimization constraints

$$\sup_x \left| \sum_{i=1}^{\ell} \beta_i \frac{1}{\ell} \sum_{j=1}^{\ell} K_\gamma(x_j, x_i) \theta(x - x_j) - F_\ell(x|w) \right| = \sigma^*.$$

For computational reasons we check this equality only at the points of the training set. In other words, we replace this equality with the equality

$$\max_p \left| \sum_{i=1}^{\ell} \beta_i \frac{1}{\ell} \sum_{j=1}^{\ell} K_\gamma(x_j, x_i) \theta(x_p - x_j) - F_\ell(x_p|w) \right| = \sigma^*, \quad p = 1, \dots, \ell.$$

Note that the following equality is valid

$$\int_{-\infty}^{\infty} p(w|x) dF(x) = p(w).$$

Substituting our expression (7.55) for $p(w|x)$ into the integral we obtain

$$\int_{-\infty}^{\infty} \sum_{i=1}^{\ell} \beta_i K_\gamma(x, x_i) dF(x) = 1.$$

Putting $F_\sharp(x)$ into the integral instead of $F(x)$, we obtain one more constraint:

$$\sum_{i=1}^{\ell} \beta_i \left(\frac{1}{\ell} \sum_{j=1}^{\ell} K_\gamma(x_j, x_i) \right) = 1.$$

4. Let the number of vectors belonging to class w be $\ell(w)$. Then for the residual principle we use

$$\sigma^* = \sigma_{\ell(w)} = \frac{q}{\sqrt{\ell(w)}},$$

where q is the appropriate quantile for the Kolmogorov-Smirnov type distribution. We also estimate

$$p(w) = \frac{\ell(w)}{\ell},$$

the probability of the appearance of vectors of class w .

5. We choose a γ from the admissible set

$$\gamma_{\min} \leq \gamma \leq \gamma_{\max}$$

to control the accuracy of our solution (by minimizing $W_\gamma(\beta)$) or/and the sparsity of the solution (by choosing a large γ).

7.9.3 The SVM Conditional Probability Estimate: Summary

The SVM conditional probability estimate is

$$p(w|x) = \frac{\ell(w)}{\ell} \sum_{j=1}^{\ell} K_\gamma(x, x_j) \beta_i, \quad \beta \geq 0,$$

where coefficients β_i minimize the functional

$$W_\gamma(\beta) = \sum_{j=1}^{\ell} \beta_i \beta_j K_\gamma(x_j, x_i)$$

subject to the constraints

$$\max_p \left| \sum_{i=1}^{\ell} \beta_i \frac{1}{\ell} \sum_{j=1}^{\ell} K_\gamma(x_j, x_i) \theta(x_p - x_j) - F_\ell(x_p|w) \right| = \sigma^*, \quad p = 1, \dots, \ell,$$

and the constraints

$$\beta_i \geq 0,$$

$$\sum_{i=1}^{\ell} \beta_i \left(\frac{1}{\ell} \sum_{j=1}^{\ell} K_\gamma(x_j, x_i) \right) = 1.$$

We choose γ from the admissible set

$$\gamma_{\min} \leq \gamma \leq \gamma_{\max}$$

to control the properties of our solution (accuracy and/or sparsity) minimizing $W_\gamma(\beta)$ and/or choosing a large admissible γ .

7.10 ESTIMATION OF CONDITIONAL DENSITY AND REGRESSION

To estimate the conditional density function using Method P we solve the integral equation

$$\int_{-\infty}^y \int_{-\infty}^x p(y|x) dF(x) dy = F(x, y) \quad (7.56)$$

in the situation where the probability distribution functions $F(y, x)$ and $F(x)$ are unknown but data

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

are given.

To solve this equation using the approximations

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

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

we follow exactly the same steps that we used for solving the equations for density estimation and conditional probability estimation. (See Sections 7.8, 7.9.)

1. We choose as a regularization functional the norm of the function in RKHS

$$\Omega(f) = (f(x, y), f(x, y))_H$$

defined by the kernel

$$K((x, y), (x', y')) = K_\gamma(x, x_i) K_\gamma(y, y_i)$$

satisfying the conditions \mathcal{K} .

2. We look for a solution in the form

$$p(y|x) = \sum_{i=1}^{\ell} \beta_i K_\gamma(x, x_i) K_\gamma(y, y_i), \quad \beta_i \geq 0. \quad (7.57)$$

Therefore, our target functional is

$$W_\gamma(\beta) = \Omega(f) = \sum_{i,j=1}^{\ell} \beta_i \beta_j K_\gamma(x_j, x_i) K_\gamma(y_j, y_i) \quad (7.58)$$

(see Section 7.8).

3. We obtain our optimization constraints using the uniform metric

$$\rho_{E_2}(A_\ell f, F_\ell) = \sup_{x,y} |(A_\ell f)(x, y) - F_\ell(x, y)| = \sigma_\ell.$$

For our equality we have

$$\sup_{x,y} \left| \int_{-\infty}^y \int_{-\infty}^x \sum_{i=1}^\ell \beta_i K_\gamma(x', x_i) K_\gamma(y', y_i) d \left[\frac{1}{\ell} \sum_{j=1}^\ell \theta(x' - x_j) \right] dy' - F_\ell(x, y) \right| = \sigma_\ell.$$

After simple calculations we obtain the constraint

$$\sup_{x,y} \left| \sum_{i=1}^\ell \beta_i \frac{1}{\ell} \sum_{j=1}^\ell K_\gamma(x_j, x_i) \theta(x - x_j) \int_{-\infty}^y K_\gamma(y', y_i) dy' - F_\ell(x, y) \right| = \sigma_\ell.$$

For computational reasons we check this constraint only at the training vectors

$$\max_p \left| \sum_{i=1}^\ell \beta_i \frac{1}{\ell} \sum_{j=1}^\ell K_\gamma(x_j, x_i) \theta(x_p - x_j) \int_{-\infty}^{y_p} K_\gamma(y', y_i) dy' - F_\ell(x_p, y_p) \right| = \sigma_\ell, \quad (7.59)$$

$$p = 1, \dots, \ell.$$

Note that the following equality holds true:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y|x) dF(x) dy = 1.$$

Putting expression (7.57) for $p(y|x)$ into the integral we obtain

$$\begin{aligned} & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sum_{i=1}^\ell \beta_i K_\gamma(x', x_i) K_\gamma(y', y_i) dy' dF(x) \\ &= \int_{-\infty}^{\infty} \sum_{i=1}^\ell \beta_i K_\gamma(x', x_i) dF(x) = 1. \end{aligned}$$

Using $F_\ell(x)$ instead of $F(x)$ we obtain

$$\sum_{i=1}^\ell \left(\frac{1}{\ell} \sum_{j=1}^\ell \beta_j K_\gamma(x_i, x_j) \right) = 1. \quad (7.60)$$

4. We use the residual principle with

$$\sigma_\ell = \frac{q}{\sqrt{\ell}}$$

obtained from a Kolmogorov-Smirnov type distribution and choose an admissible γ .

5. To control the properties of the solution (accuracy and/or sparsity) we choose an admissible parameter γ that minimizes the target functional and/or that is large.

Therefore, we approximate the conditional density function in the form (7.57), where the coefficients β_i are obtained from the solution of the following optimization problem: Minimize functional (5.58) subject to constraints (7.59) and constraint (5.60). Choose γ from the admissible set to control the desired properties of the solution.

To estimate the regression function

$$r(x) = \int yp(y|x)dy \quad (7.61)$$

recall that the kernel $K_\gamma(y, y_j)$ is a symmetric (density) function the integral of which is equal to 1. For such a function we have

$$\int yK_\gamma(y, y_i)dy = y_i. \quad (7.62)$$

Therefore, from (7.57), (7.61), and (7.62) we obtain the following regression function:

$$r(x) = \sum_{i=1}^{\ell} y_i \beta_i K_\gamma(x, x_i).$$

It is interesting to compare this expression with Nadaraya-Watson regression

$$r(x) = \sum_{i=1}^{\ell} y_i \left(\frac{K_\gamma(x_i, x)}{\sum_{i=1}^{\ell} K_\gamma(x_i, x)} \right), \quad (7.63)$$

where the expression in the parentheses is defined by the Parzen's estimate of density (it is the ratio of the i th term of the Parzen's density estimate to the estimate of density).

The SVM regression is smooth and has sparse representation.

7.11 REMARKS

7.11.1 Remark 1. One can use a good estimate of the unknown density.

In constructing our algorithms for estimating densities, conditional probabilities, and conditional densities we use the empirical distribution function

$F_\ell(x)$ as an approximation of the actual distribution function $F(x)$. From $F_\ell(x)$ we obtained an approximation of the density function

$$p(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \delta(x - x_i)$$

as a sum of δ -functions. In fact, this approximation of the density was used to obtain the corresponding constraints.

One can use, however, better approximations of the density, based on the (sparse) SVM estimate described in Section 7.8. Using this approximation of the density function one can obtain constraints different (perhaps more accurate) from those used. In Chapter 8 we will introduce a new principle of risk minimization that reflects this idea.

7.11.2 Remark 2. One can use both labeled (training) and unlabeled (test) data.

To estimate the conditional probability function and the conditiona density function one can use both elements of training data

$$(w_1, x_1), \dots, (w_\ell, x_\ell) \tag{7.64}$$

and elements of unlabeled (test) data

$$x^*, \dots, x_\ell^*.$$

Since according to our learning model, vectors x from the training and the test sets have the same distribution $F(x)$ generated by generator G (see Chapter 1), one can use the joint set

$$x_1, \dots, x_\ell, x_1^*, \dots, x_k^*$$

to estimate the distribution $F(x)$ (or density function $p(x)$). To estimate the distribution function $F(x|w)$ one uses the subset of vectors x from (7.64) corresponding to $w = w^*$.

7.11.3 Remark 3. Method for obtaining sparse solutions of the ill-posed problems.

The method used for density, conditional probability, and conditional density estimation is quite general. It can be applied for obtaining sparse solutions of ather operator equations.

To obtain the sparse solution one has:

- Choose the regularizer as a norm in RKHS.
- Choose L_∞ metric in E_2 .

- Use the residual principle.
- Choose the appropriate value γ from the admissible set.

Informal Reasoning and Comments — 7

7.12 THREE ELEMENTS OF A SCIENTIFIC THEORY

According to Kant any theory should contain three elements:

1. Setting the problem,
2. Resolution of the problem, and
3. Proofs.

At first glance, this remark looks obvious. However, it has a deep meaning. The crux of this remark is the idea that these three elements of theory in some sense are *independent and equally important*.

1. The precise setting of the problem provides a general point of view on 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 for searching for the solution of the problem but for justification of the solution that has already been suggested.

The first two elements of the theory reflect the understanding of the essence of the problem of interest, its philosophy. The proofs make the general (philosophical) model a scientific theory.

7.12.1 Problem of Density Estimation

In analyzing the development of the theory of density estimation one can see how profound Kant's remark is. Classical density estimation theories, both parametric and nonparametric, contained only two elements: resolution of the problem and proofs. They did not contain the setting of the problem.

In the parametric case Fisher suggested the maximum likelihood method (resolution of the problem), and later it was proved by Le Cam (1953), Ibragimov and Hasminski (1981) and others that under some (not very wide, see the example in Section 1.7.4) conditions the maximum likelihood method is consistent.

The same happened with nonparametric resolutions of the problem. First the methods were proposed: The histogram method (Rosenblatt 1956), Parzen's method (Parzen 1962), projection method (Chentsov 1963) and so on followed by proofs of their consistency. In contrast to parametric methods the nonparametric methods are consistent under very wide conditions.

The absence of the general setting of the problem made the density estimation methods look like a list of recipes. It also seems to have made heuristic efforts look like the only possible approach to improve the methods. These created a huge collection of heuristic corrections to nonparametric methods for practical applications.

The attempt to suggest the general setting of the density estimation problem was made in 1978 (Vapnik and Stefanyuk (1978)), where the density estimation problem was derived directly from the definition of the density, considered as a problem of solving an integral equation with unknown right-hand side but given data. This general (since it follows from the definition of the density) setting immediately connected density estimation theory with the fundamental theory: the theory of solving ill-posed problem.

7.12.2 Theory of Ill-Posed Problems

The theory of ill-posed problems was originally developed for solving inverse mathematical physics problems. Later, however, the general nature of this theory was discovered. It was demonstrated that one has to take into account the statements of this theory every time one faces an inverse problem, i.e., when one tries to derive the unknown causes from known consequences. In particular, the results of the theory of ill-posed problems are important for statistical inverse problems, which include the problems of density estimation, conditional probability estimation, and conditional density estimation.

The existence of ill-posed problems was discovered by Hadamard (1902). Hadamard thought that ill-posed problems are pure mathematical phenomena and that real-life problems are well-posed. Soon, however, it was

discovered that there exist important real-life problems that are ill-posed.

In 1943 A.N. Tikhonov in proving a lemma about an inverse operator, described the nature of well-posed problems and therefore discovered methods for the regularization of ill-posed problems. It took twenty years more before Phillips (1962), Ivanov (1962), and Tikhonov (1963) came to the same constructive regularization idea, described, however, in a slightly different form. The important message of regularization theory was the fact that in the problem of solving operator equations

$$Af(t) = F(x)$$

that define an ill-posed problem, the obvious resolution to the problem, minimizing the functional

$$R(f) = \|Af - F\|^2,$$

does not lead to good solutions. Instead, one should use the nonobvious resolution that suggests that one minimize the "corrupted" (regularized) functional

$$R^*(f) = \|Af - F\|^2 + \gamma\Omega(f).$$

At the beginning of the 1960s this idea was not obvious. The fact that now everybody accepts this idea as natural is evidence of the deep influence of regularization theory on the different branches of mathematical science and in particular on statistics.

7.13 STOCHASTIC ILL-POSED PROBLEMS

To construct a general theory of density estimation it was necessary to generalize the theory of solving ill-posed problem for the stochastic case.

The generalization of the theory of solving ill-posed problems introduced for the deterministic case to stochastic ill-posed problems is very straightforward. Using the same regularization techniques that were suggested for solving deterministic ill-posed problems and the same key arguments based on the lemma about inverse operators we generalized the main theorems on the regularization method (V. Vapnik and A. Stefanyuk, 1978) to a stochastic model. Later, A. Stefanyuk (1986) generalized this result for the case of an approximately defined operator,

The fact that the main problem of statistics – estimating functions from a more or less wide set of functions – is ill-posed was known to everybody. Nevertheless, the analysis of methods of solving the main statistical problems, in particular density estimation, was never considered from the formal point of view of regularization theory.⁷

⁷One possible explanation is that the theory of nonparametric methods for

Instead, in the tradition of statistics there was first the suggestion of some method for solving the problem, proving its nice properties, and then introducing some heuristic corrections to make this method useful for practical tasks (especially for multidimensional problems).

Attempts to derive new estimators from the point of view of solving stochastic ill-posed problems was started with the analysis of the various known algorithms for the density estimation problem (Aidu and Vapnik, 1989). It was observed that almost all classical algorithms (such as Parzen's method and the projection method) can be obtained on the basis of the standard regularization method of solving stochastic ill-posed problems under the condition that one chooses the empirical distribution function as an approximation to the unknown distribution function.

The attempt to construct a new algorithm at that time was inspired by the idea of constructing, a better approximation to the unknown distribution function based on the available data. Using this idea we constructed a new estimators that justify many heuristic suggestions for estimating one dimensional density functions.

In the 1980s the problem of nonparametric method density estimation was very popular among both theoretists and practitioners in statistics. The main problem was to find the law for choice of the optimal width parameter for Parzen's method. Asymptotic principles that connected the value of the width with information about smoothness properties of the actual density, properties of the kernel, and the number of observations were found.

However, for practitioners these results were insufficient for two reasons, first because they are valid only for sufficiently large data sets and second because the estimate of one free parameter was based on some unknown parameter (the smoothness parameter, say, by the number of derivatives possessed by the unknown density).

Therefore, practitioners developed their own methods for estimating the width parameter. Among these methods the leave-one-out estimate became one of the most used. There is a vast literature devoted to experimental analysis width of the parameter.

At the end of the 1980s the residual method for estimating the regularization parameter (width parameter) was proposed (Vapnik 1988). It was shown that this method is almost optimal (Vapnik et al., 1992). Also, in experiments with a wide set of one-dimensional densities it was shown that this method of choice of the width parameter outperforms many theoretical and heuristic approaches (Markovich, 1989).

density estimation had begun (in the 1950s) before the regularization methods for solving ill-posed problems were discovered. In the late 1960s and in the 1970s when the theory of ill-posed problems attracted the attention of many researchers in different branches of mathematics, the paradigm in the analysis of the density estimation problem had already been developed.

Unfortunately, most of the results in density estimation are devoted to the one-dimensional case, while the main applied interest in the density estimation problem is in the multidimensional case. For this case special methods were developed.

The most popular of these, the Gaussian mixture model method, turned out to be inconsistent (see Section 1.7.4). Nevertheless, this method is used for most high-dimensional (say 50-dimensional) problems of density estimation (for example in speech recognition).

It is known, however, that even to construct good two-dimensional density estimators one has to use new ideas.

The real challenge, however, is to find a good estimator for multidimensional densities defined on bounded support.

In this chapter we proposed a new method for multidimensional density estimation. It combines ideas from three different branches of mathematics: the theory of solving integral equations using the residual principle, the universal Kolmogorov-Smirnov distribution, which allows one to estimate the parameter for the residual principle, and the SVM technique from statistical learning theory, which was developed to approximate functions in high-dimensional spaces.

Two out of three of these ideas have been checked for solving one-dimensional density estimation problems (Vapnik 1988, Aidu and Vapnik, 1989, Vapnik et al. 1992, Markovich 1989).

The third idea, to use as the regularized functional a norm in RKHS and measure discrepancy in the L_∞ norm, is the direct result of the SVM method for function approximation using ε -insensitive loss function, described for the first time in the first edition of this book. It was partly checked for estimating one dimensional density functions.

The density estimation method described in this chapter was analyzed by Sayan Mukherjee. His experiments with estimating a density in one-, two-, and six-dimensional spaces demonstrated high accuracy and good sparsity of solutions obtained. Two of these experiments are presented in this book.

Direct solutions of the conditional probability and the conditional density estimation problems described in this chapter are a straightforward generalization of the direct density estimation method. These methods have not been checked experimentally.

Chapter 8

The Vicinal Risk Minimization Principle and the SVMs

In this chapter we introduce a new principle for minimizing the expected risk called the vicinal risk minimization (VRM) principle.¹ We use this principle for solving our main problems: pattern recognition, regression estimation, and density estimation.

We minimize the vicinal risk functional using the SVM technique and obtain solutions in the form of expansions on kernels that are different for different training points.

8.1 THE VICINAL RISK MINIMIZATION PRINCIPLE

Consider again our standard setting of the function estimation problem: In a set of functions $f(x, \alpha), \alpha \in \Lambda$, minimize the functional

$$R(\alpha) = \int L(y - f(x, \alpha))dP(x, y), \quad (8.1)$$

where $L(u)$ is a given loss function if the probability measure $P(x, y)$ is unknown but data

$$(y_1, x_1), \dots, (y_\ell, x_\ell) \quad (8.2)$$

¹With this name we would like to stress that our goal is to minimize the risk in vicinities $x \in v(x_i)$ of the training vectors x_i , $i = 1, \dots, \ell$, where (as we believe) most of points $x \in v(x_i)$ keep the same (or almost the same) value y_i as the training vector x_i , rather than to minimize the empirical risk functional defined only by the training vectors.

are given.

In the first chapters of the book in order to solve this problem we considered the empirical risk minimization principle, which suggested minimizing the functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i - f(x_i, \alpha)) \quad (8.3)$$

instead of the functional (8.1).

Later we introduced the structural risk minimization principle, where we defined a structure on a set of functions $f(x, \alpha), \alpha \in \Lambda$,

$$S_i \subset \dots \subset S_n,$$

and then we minimized functional (8.3) on the appropriately chosen element S_k of this structure.

Now we consider a new basic functional instead of the empirical risk functional (8.3) and use this functional in the structural risk minimization scheme.

Note that introduction of the empirical risk functional reflects the following reasoning: Our goal is to minimize the expected risk (8.1) when the probability measure is unknown. Let us estimate the density function from the data and then use this estimate $\hat{p}(x, y)$ in functional (8.1) to obtain the target functional

$$R_T(\alpha) = \int (L(y - f(x, \alpha)) \hat{p}(x, y) dx dy. \quad (8.4)$$

When we estimate the unknown density by the sum of δ -functions

$$\hat{p}(x, y) = \frac{1}{\ell} \sum_{i=1}^{\ell} \delta(x - x_i) \delta(y - y_i)$$

we obtain the empirical risk functional.

If we believe that both the density function and the target function are smooth, then the empirical risk functional probably is not the best approximation of the expected risk functional. The question arises as to whether there exists a better approximation of the risk functional that reflects the following two assumptions:

1. The unknown density function is smooth in a vicinity of any point x_i .
2. The function minimizing the risk functional is also smooth and symmetric in vicinity any point x_i .

Below we introduce a new target functional which we will use instead of the empirical risk functional. To introduce this functional we construct (using

data) vicinity functions $v(x_i)$ of the vectors x_i for all training vectors and then using these vicinity functions we construct the target functional. As in Section 4.5 we distinguish between two concepts of vicinity functions, hard vicinity and soft vicinity functions. Below we first introduce the concept of hard vicinity function and then consider soft vicinity function. One can also use other concepts of vicinity functions which are more appropriate for problems at hand.

8.1.1 Hard Vicinity Function

- For any x_i , $i = 1, \dots, \ell$ we define a measurable subset $v(x_i)$ of the set $X \in R^n$ (the vicinity of point x_i) with volume ν_i .

We define the vicinity of this point as the set of points that are r_i -close to $x_i = (x_i^1, \dots, x_i^n)$ (r_i depends on the point x_i)

$$v(x_i) = \{x : \|x - x_i\|_E \leq r_i\},$$

where $\|x - x_i\|_E$ is a metric in space E . For example, it can be the l_1 , the l_2 , or the l_∞ metric: l_1 metric defines the vicinity as a set

$$v(x_i) = \left\{x : \sum_{k=1}^n |x^k - x_i^k| \leq r_i\right\},$$

l_2 metric defines the vicinity as the ball of radius r_i with center at point x_i

$$v(x_i) = \left\{x : \sum_{k=1}^n |x - x_i|^2 \leq r_i^2\right\},$$

while l_∞ metric defines a cube of size $2r_i$ with a center at the point $x_i = (x_i^1, \dots, x_i^n)$

$$v(x_i) = \{x : x_i^k - r_i \leq x^k \leq x_i^k + r_i, \forall k = 1, \dots, n\}.$$

- The vicinities of different training vectors have no common points.
- We approximate the unknown density function $p(x)$ in the vicinities of vector x_i as follows. All ℓ vicinities of the training data have an equal probability measure

$$P(x \in v(x_i)) = 1/\ell.$$

The distribution of the vectors within the vicinity is uniform,

$$p(x|v(x_i)) = \frac{1}{\nu_i},$$

where ν_i is the volume of vicinity $v(x_i)$.

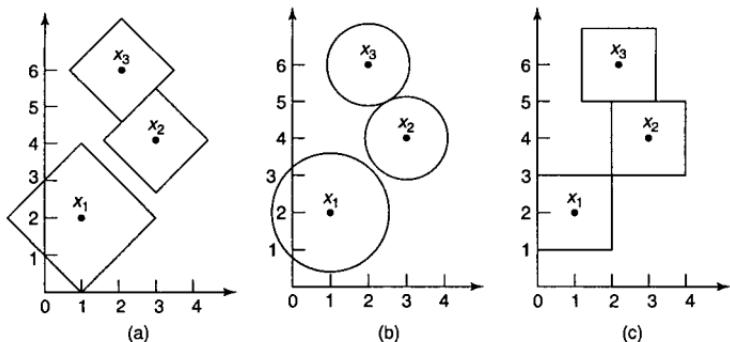


FIGURE 8.1. Vicinity of points in different metrics: (a) in the l_1 metric, (b) in the l_2 metric, and (c) in the l_∞ metric.

Figure 8.1 shows the vicinity of points in different metrics: (a) in the l_1 metric, (b) in the l_2 metric, and (c) in the l_∞ metric.

Consider the following functional, which we call the *vicinal* risk functional

$$V(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L \left(y_i - \frac{1}{\nu_i} \int_{v(x_i)} f(x, \alpha) dx \right). \quad (8.5)$$

In order to find an approximation to the function that minimizes risk functional (8.1) we are looking for the function that minimizes functional (8.5). Minimizing functional (8.5) instead of functional (8.1) we call the *vicinal risk minimization* (VRM) principle (method). Note that when $\nu_i \rightarrow 0$ the vicinal risk functional converges to the empirical risk functional.

Since the volumes of vicinities can be different for different training points, by introducing this functional we expect that the function minimizing it have different smoothness properties in the vicinities of different points.

In a sense the VRM method combines two different estimating methods: the empirical risk minimization method and 1-nearest neighbor method.

8.1.2 Soft Vicinity Function

In our definition of the vicinal method we used parameters x_i and r_i obtained from the training data to construct a *uniform* distribution function that is used in equations for VRM.

However, one can use these parameters to construct other distribution functions $p(x|x_i, r_i)$ where they define the parameters of position and width (for example, one can use the normal distribution function $p(x|x_i, r_i) = N(x_i, d_i)$). For soft vicinity functions all points of the space can belong to a vicinity of the vector x_i . However, they have different measures.

A soft vicinity function defines the following (general) form of VRM

$$\begin{aligned} V(\alpha) &= \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i - Ef(x, \alpha)) \\ &= \frac{1}{\ell} \sum_{i=1}^{\ell} L\left(y_i - \int f(x, \alpha)p(x|x_i, r_i)dx\right) \end{aligned}$$

In Section 8.3.1 we define a VRM method based on hard vicinity functions and based on soft vicinity functions.

8.2 VRM METHOD FOR THE PATTERN RECOGNITION PROBLEM

In this section we apply the VRM method to the two class $\{-1, 1\}$ pattern recognition problem. Consider the set of indicator functions

$$y = g(x, \alpha) = \text{sign}[f(x, \alpha)], \quad (8.6)$$

where $f(x, \alpha), \alpha \in \Lambda$, is a set of real-valued functions. In previous chapters we did not pay attention on the structure (8.6) of the indicator function. In order to find the function from $f(x, \alpha), \alpha \in \Lambda$, that minimizes the risk functional, we minimized the empirical functional (8.3) with the loss function $|y - f(x, \alpha)|$.

Now taking into account the structure (8.6) of indicator functions we consider another loss function

$$L(y, f(x, \alpha)) = \theta(-yf(x, \alpha)), \quad (8.7)$$

which defines the risk functional

$$R(\alpha) = \int \theta[-yf(x, \alpha)]dP(x, y), \quad (8.8)$$

where $\theta(u)$ is a step function.

To minimize this functional the VRM method suggests minimizing the functional

$$V(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta\left[-y_i \int f(x, \alpha)p(x|x_i, r_i)dx\right]. \quad (8.9)$$

For the hard vicinity function we obtain

$$V(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta\left[\frac{-y_i}{\nu_i} \int_{v(x_i)} f(x, \alpha)dx\right].$$

As in Chapter 5 we reduce this problem to the following optimization problem: Minimize the functional

$$W(f) = C \sum_{i=1}^{\ell} \xi_i + \Omega(f) \quad (8.10)$$

subject to the constraints

$$y_i \int f(x, \alpha) p(x|x_i, r_i) dx \geq 1 - \xi_i, \quad (8.11)$$

where $\Omega(f)$ is some regularization functional that we specify below.

Suppose that our set of functions is defined as follows: We map input vectors x into feature vectors z and in the feature space construct a hyperplane

$$(w, z) + b = 0$$

that separates data

$$(y_1, z_1), \dots, (y_\ell, z_\ell),$$

which are images in the feature space of our training data (8.2). (Let a kernel $K(x, x')$ defines the inner product in the feature space.)

Our goal is to find the function $f(x, \alpha)$ satisfying the constraints

$$y_i \int f(x, \alpha) p(x|x_i, r_i) dx \geq 1 - \xi_i \quad (8.12)$$

whose image in the feature space is a linear function

$$l(z) = (w^*, z) + b$$

that minimizes the functional

$$W(w) = (w, w) + C \sum_{i=1}^{\ell} \xi_i. \quad (8.13)$$

We will solve this problem using the SVM technique and call the solution the vicinal SVM solution (VSV). Note that for linear functions in the input space

$$f(x, \alpha) = (w, x) + b, \quad \alpha \in \Lambda,$$

and for vicinities where x_i is the center of mass,

$$x_i = E_{v(x_i)} x$$

the VSV solution coincides with the SVM solution. Indeed, since the target functional in the both cases is the same and

$$\int [(w, x) + b] p(x|x_i, r_i) dx = (w, x_i) + b,$$

the problems coincide.

The difference between ERM and VRM can appear in two cases, if the point x_i is not the center of mass of the vicinity $v(x_i)$ or if we consider nonlinear functions.

Let us (using the kernel $K(x, x')$) introduce two new kernels: the one-vicinal kernel

$$\mathcal{L}(x, x_i) = E_{v(x_i)} K(x, x') = \int K(x, x') p(x'|x_i, r_i) dx' \quad (8.14)$$

and the two-vicinal kernel

$$\begin{aligned} \mathcal{M}(x_i, x_j) &= E_{v(x_i)} E_{v(x_j)} K(x, x') \\ &= \int \int K(x, x') p(x|x_i, r_i) p(x'|x_j, r_j) dx dx'. \end{aligned} \quad (8.15)$$

The following theorem is true.

Theorem 8.1. *The vicinal support vector solution (VSV) has the form*

$$f(x) = \sum_{i=1}^{\ell} \beta_i \mathcal{L}(x, x_i) + b, \quad (8.16)$$

where to define coefficients β_i one has to maximize the functional

$$W(\beta) = \sum_{i=1}^{\ell} \beta_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \beta_i \beta_j \mathcal{M}(x_i, x_j) \quad (8.17)$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \beta_i = 0, \quad (8.18)$$

$$0 \leq \beta_i \leq C. \quad (8.19)$$

PROOF. Let us map input vectors x into feature vectors z . Consider samples of N points

$$x_{i_1}, \dots, x_{i_N}, \quad i = 1, \dots, \ell,$$

taken from the vicinities of points x_i , $i = 1, \dots, \ell$. Let the images of these points in feature space be

$$z_{i_1}, \dots, z_{i_N}, \quad i = 1, \dots, \ell.$$

Consider the problem of constructing the following vicinal optimal hyperplane in a feature space: Minimize the functional

$$W^*(w) = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} \xi_i \quad (8.20)$$

subject to the constraints

$$y_i \frac{1}{N} \sum_{k=1}^N [(w, z_{i_k}) + b] \geq 1 - \xi_i. \quad (8.21)$$

Note that the equivalent expression for (8.21) in the input space is

$$\frac{y_i}{N} \sum_{k=1}^N f(x_{i_k}, \alpha) \geq 1 - \xi_i. \quad (8.22)$$

As $N \rightarrow \infty$, expression (8.22) converges to (8.12). Therefore, the solution of the optimization problem defined by (8.20) and (8.21) converges to the solution of the optimization problem defined by (8.13) and (8.12).

To minimize (8.20) under constraints (8.21) we introduce the Lagrangian

$$L(w) = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} \xi_i - \sum_{i=1}^{\ell} \beta_i [(y_i \frac{1}{N} \sum_{k=1}^N (w, z_{i_k}) + b) - 1 + \xi_i] + \sum_{i=1}^{\ell} \eta_i \xi_i. \quad (8.23)$$

The solution of our optimization problem is defined by the saddle point of the Lagrangian that minimizes the functional over b , ξ_i , and w and maximizes it over β and η . As the result of minimization we obtain

$$\sum_{i=1}^{\ell} y_i \beta_i = 0, \quad (8.24)$$

$$0 \leq \beta_i \leq C, \quad (8.25)$$

and

$$w = \sum_{i=1}^{\ell} \beta_i \frac{1}{N} \sum_{k=1}^N z_{i_k}. \quad (8.26)$$

Putting (8.26) in the expression for the hyperplane we obtain

$$l(z) = (w, z) + b = \sum_{i=1}^{\ell} y_i \beta_i \frac{1}{N} \sum_{k=1}^N (z, z_{i_k}) + b. \quad (8.27)$$

Putting expression (8.26) back into the Lagrangian we obtain

$$W(\beta) = \sum_{i=1}^{\ell} \beta_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \beta_i \beta_j y_i y_j \frac{1}{N} \sum_{k=1}^N \frac{1}{N} \sum_{m=1}^N (z_{i_k}, z_{j_m}). \quad (8.28)$$

Since $(z, z') = K(x, x')$, we can rewrite expressions (8.27) and (8.28) in the form

$$f(x) = \sum_{i=1}^{\ell} y_i \beta_i \frac{1}{N} \sum_{k=1}^N K(x, x_{i_k}) + b, \quad (8.29)$$

where the coefficients β_i maximize the functional

$$W(\beta) = \sum_{i=1}^{\ell} \beta_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \beta_i \beta_j \frac{1}{N} \sum_{k=1}^N \frac{1}{N} \sum_{m=1}^N K(x_{i_k}, x_{j_m})$$

subject to constraints (8.24) and (8.25). Increasing N , we obtain

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N K(x, x_{i_k}) = \mathcal{L}(x, x_i),$$

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \frac{1}{N} \sum_{m=1}^N K(x_{i_k}, x_{j_m}) = \mathcal{M}(x_i, x_j)$$

Therefore, the VSV solution is

$$f(x) = \sum_{i=1}^{\ell} y_i \beta_i \mathcal{L}(x, x_i) + b, \quad (8.30)$$

where to define the coefficients β_i one has to maximize the functional

$$W(\beta) = \sum_{i=1}^{\ell} \beta_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \beta_i \beta_j \mathcal{M}(x_i, x_j) \quad (8.31)$$

subject to the constraints

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

$$0 \leq \beta_i \leq C.$$

8.3 EXAMPLES OF VICINAL KERNELS

In this section we give example of pairs of vicinity and kernel $K(x, y)$ that allow us to construct in the analytic form both the one-vicinal kernel $\mathcal{L}(x, x_i)$ and the two-vicinal kernel $\mathcal{M}(x_i, x_j)$. In Section 8.3.1 we introduce these kernels for hard vicinity functions and in Section 8.3.2 for soft vicinity functions.

8.3.1 Hard Vicinity Functions

We define the vicinities of points x_i , $i = 1, \dots, \ell$, using the l_∞ metric:

$$\|x - x_i\|_\infty = \sup_{1 \leq k \leq n} |x^k - x_i^k|, \quad (8.32)$$

where $x = (x^1, \dots, x^n)$ is a vector in R^n .

We define size of the vicinity of the vectors x_i , $i = 1, \dots, \ell$ from the training data

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

using the following algorithm:

1. Define the triangle matrix

$$A = \|a_{i,j}\|, \quad i > j,$$

of the pairwise distances (in the metric l_∞) of the vectors from the training set.

- 2 Define the smallest element of the matrix A (say a_{ij}).

3. Assign the value

$$d_i = \kappa a_{ij}$$

to element x_i and the value

$$d_j = \kappa a_{ij}$$

to element x_j .

Here $\kappa \leq 1/2$ is the parameter that controls the size of vicinities (usually it is reasonable to choose the maximal possible size $\kappa = 1/2$).

- 4 Choose the next smallest element a_{ms} of the matrix A . If one of the vectors (say x_m) was already assigned some value d_m , then assign the value

$$d_s = \kappa a_{ms}$$

to another vector x_s , otherwise assign this value to both vectors.

- 5 Continue this process until values d have been assigned to all vectors.

Using the value d_i we define both the vicinity of the point x_i ,

$$v(x_i) = \{x : x_i^k - d_i \leq x^k \leq x_i^k + d_i, \forall k = 1, \dots, n\}$$

and the volume

$$\nu_i = (2d_i)^n$$

of the vicinity.

Let us introduce the notation

$$v(x_i^k) = \{x^k : x_i^k - d_i \leq x^k \leq x_i^k + d_i\}.$$

Now we calculate both the one- and two-vicinal kernels for the Laplacian-type kernel

$$K(x, x') = \exp \left\{ -\frac{\|x - x'\|_{l_1}}{\Delta} \right\} = \prod_{k=1}^n \exp \left\{ -\frac{|x^k - (x')^k|}{\Delta} \right\}$$

We obtain the one-vicinal kernel

$$\begin{aligned} \mathcal{L}(x, x_i) &= \frac{1}{(2d_i)^n} \int_{v(x_i)} \exp \left\{ -\frac{|x - x'|}{\Delta} \right\} dx' \\ &= \frac{1}{2^n d_i^n} \prod_{k=1}^n \int_{v(x_i^k)} \exp \left\{ -\frac{|x^k - (x')^k|}{\Delta} \right\} d(x')^k = \prod_{k=1}^n \mathcal{L}^k(x^k, x_i^k). \end{aligned}$$

After elementary calculations we obtain

$$\begin{aligned} \mathcal{L}^k(x^k, x_i^k) &= \frac{1}{2d_i} \int_{v(x_i^k)} \exp \left\{ -\frac{|x^k - (x')^k|}{\Delta} \right\} d(x')^k \\ &= \begin{cases} \frac{\Delta}{2d_i} \left[2 - \exp \left\{ -\frac{(d_i + x^k - x_i^k)}{\Delta} \right\} - \exp \left\{ -\frac{(d_i - x^k + x_i^k)}{\Delta} \right\} \right] & \text{if } |x_i^k - x^k| \leq d_i, \\ \frac{\Delta}{2d_i} \exp \left\{ -\frac{|x^k - x_i^k|}{\Delta} \right\} (\exp \left\{ \frac{d_i}{\Delta} \right\} - \exp \left\{ -\frac{d_i}{\Delta} \right\}) & \text{if } |x_i^k - x^k| \geq d_i. \end{cases} \end{aligned}$$

The n -dimensional two-vicinal kernel is the product of one-dimensional kernels

$$\mathcal{M}(x_i, x_j) = \prod_{k=1}^n \mathcal{M}^k(x_i^k, x_j^k).$$

To calculate $\mathcal{M}^k(x_i^k, x_j^k)$ we distinguish two cases: the case where $i \neq j$ (say $i > j$) and the case where $i = j$. For the case $i \neq j$ we obtain (taking into account that different vicinities have no common points)

$$\begin{aligned} \mathcal{M}^k(x_i^k, x_j^k) &= \frac{1}{4d_i d_j} \int_{v(x_i^k)} \int_{v(x_j^k)} \exp \left\{ -\frac{|x^k - (x')^k|}{\Delta} \right\} dx' dx \\ &= \frac{\Delta^2}{4d_i d_j} \exp \left\{ -\frac{|x^k - (x')^k|}{\Delta} \right\} \left(e^{\left\{ \frac{d_i}{\Delta} \right\}} - e^{\left\{ -\frac{d_i}{\Delta} \right\}} \right) \left(e^{\left\{ \frac{d_j}{\Delta} \right\}} - e^{\left\{ -\frac{d_j}{\Delta} \right\}} \right) \end{aligned}$$

For the case $i = j$ we obtain

$$\begin{aligned}\mathcal{M}^k(x_i^k, x_i^k) &= \frac{1}{4\delta_i^2} \int_{v(x_i^k)} \int_{v(x_j^k)} \exp \left\{ -\frac{|x^k - (x')^k|}{\Delta} \right\} dx' dx \\ &= \frac{2}{4d_i^2} \int_{v(x_i^k)} dx \int_{x_i - d_i}^x \exp \left\{ -\frac{(x^k - (x')^k)}{\Delta} \right\} dx' \\ &= \frac{\Delta^2}{2d_i^2} \left(e^{\left\{ \frac{2d_i}{\Delta} \right\}} - 1 - \frac{2d_i}{\Delta} \right).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathcal{M}^k(x_i^k, x_i^k) \\ = \begin{cases} \frac{\Delta^2}{4d_i d_j} e^{\left\{ -\frac{|x_i^k - x_j^k|}{\Delta} \right\}} \left(e^{\left\{ \frac{d_i}{\Delta} \right\}} - e^{\left\{ -\frac{d_j}{\Delta} \right\}} \right) \left(e^{\left\{ \frac{d_j}{\Delta} \right\}} - e^{\left\{ -\frac{d_i}{\Delta} \right\}} \right) & \text{if } i \neq j, \\ \frac{\Delta^2}{2d_i^2} \left(e^{\left\{ \frac{2d_i}{\Delta} \right\}} - 1 - \frac{2d_i}{\Delta} \right) & \text{if } i = j. \end{cases}\end{aligned}$$

Note that when

$$\frac{d}{\Delta} \longrightarrow 0,$$

we obtain the classical SVM solution

$$\mathcal{L}(x, x_i) \longrightarrow K(x, x_i),$$

$$\mathcal{M}(x_i, x_j) \longrightarrow K(x_i, x_j).$$

Figure 8.2 shows the one-vicinal kernel obtained from the Laplacian with parameter $\Delta = 0.25$ for different values of vicinities: (a) $d = 0.02$, (b) $d = 0.5$, and (c) $d = 1$. Note that the larger the vicinity of the point x_i , the smoother the kernel approximate function in this vicinity.

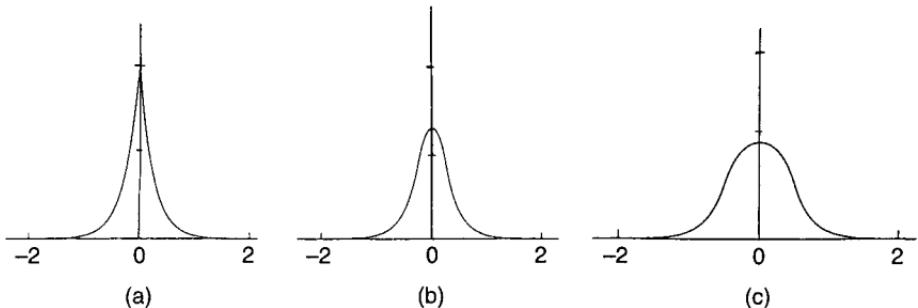


FIGURE 8.2. One-vicinal kernel obtained from Laplacian with $\Delta = 0.25$ for different values of vicinities (a) $d = 0.02$, (b) $d = 0.5$, and (c) $d = 1$.

8.3.2 Soft Vicinity Functions

To construct one- and two-vicinal kernels for the Gaussian-type kernel

$$K(x, x') = \exp \left\{ -\frac{(x - x')^2}{2\gamma^2} \right\}$$

one has make the following:

1. To define the distance between two points in the l_2 -metric.
2. To define the values d_i for all points x_i of the training data using the same algorithm that we used in the previous section.
3. To define soft vicinity functions by the normal law with parameters x_i and d_i .
4. To calculate the one- and two-vicinal functions

$$\begin{aligned} \mathcal{L}(x, x_i) &= \frac{1}{(2\pi)^{\frac{n}{2}} d_i^n} \int \exp \left\{ -\frac{(x - x')^2}{2\gamma^2} \right\} \exp \left\{ -\frac{(x' - x_i)^2}{2d_i^2} \right\} dx' \\ &= \left(1 + \frac{\sigma^2}{\gamma^2} \right)^{-\frac{n}{2}} \exp \left\{ -\frac{(x - x')^2}{2(\gamma^2 + d_i^2)} \right\}, \\ &\quad \mathcal{M}(x_j, x_i) \\ &= \frac{1}{(2\pi)^{(d_i + d_j)}} \int \int \exp \left\{ -\frac{(x - x')^2}{2\gamma^2} - \frac{(x' - x_i)^2}{2d_i^2} - \frac{(x - x_j)^2}{2d_j^2} \right\} dx dx' \\ &= \left(1 + \frac{d_i^2}{\gamma^2} + \frac{d_j^2}{\gamma^2} \right)^{-\frac{n}{2}} \exp \left\{ -\frac{(x_i - x_j)^2}{2(\gamma^2 + d_i^2 + d_j^2)} \right\}. \end{aligned}$$

8.4 NONSYMMETRIC VICINITIES

In the previous section, in order to obtain analytic expressions for vicinal kernels, we considered symmetric vicinities. This type of vicinities reflects the most simple information about problem at hand. Now our goal is to define vicinities that allow us to construct vicinal kernels reflecting some local invariants.

Below we consider the example of constructing such kernels for the digit recognition problem. However the main idea introduced in this example can be used for various function estimation problems.

It is known that any small continuous linear transformation of two dimensional images x_i can be described by six functions (Lie derivatives)

$x_{i,k}^*, k = 1, \dots, 6$ such that transformed image is

$$x = x_i + \sum_{k=1}^6 x_{i,k}^* t_k,$$

where $t_k, k = 1, \dots, 6$ are reasonable small values. Therefore different small linear transformations of image x_i are defined by six Lie derivatives of x_i and different small vectors $t = (t_1, \dots, t_6)$, say $|t| \leq c$.

Let us introduce the following vicinity of x_i

$$v_L(x_i) = \left\{ x : x = x_i + \sum_{k=1}^6 x_{i,k}^* t_k, |t| \leq c \right\}.$$

This vicinity is not necessarily symmetric. Note that if we will be able to construct one- and two-vicinal kernels

$$\mathcal{L}_L(x, x_i) = E_{v_L(x_i)} K(x, x'),$$

$$\mathcal{M}_L(x_i, x_j) = E_{v_L(x_i)} E_{v_L(x_j)} K(x, x'),$$

then the VSV solution

$$f_L(x, \alpha) = \sum_{i=1}^{\ell} y_i \alpha_i \mathcal{L}_L(x, x_i)$$

will take into account invariants with respect to small Lie transformations.

Of course it is not easy to obtain vicinal kernels in analytic form. However one can approximate these kernels by the sum

$$\mathcal{L}_L(x, x_i) = \frac{1}{N} \sum_{k=1}^N E_{v(x_k(x_i))} K(x, x') = \frac{1}{N} \sum_{k=1}^N \mathcal{L}(x, x_k(x_i))$$

$$\begin{aligned} \mathcal{M}_L(x_i, x_j) &= \frac{1}{N} \sum_{k=1}^N \frac{1}{N} \sum_{m=1}^N E_{v(x_k(x_i))} E_{v(x_m(x_j))} K(x, x') = \\ &= \frac{1}{N^2} \sum_{k=1}^N \sum_{m=1}^N \mathcal{M}(x_k(x_i), x_m(x_j)), \end{aligned}$$

where $x_k(x_i), k = 1, \dots, N$ are virtual examples obtained from x_i using small Lie transformation and $v(x_k(x_i))$ is symmetric vicinity for k -th virtual example $x_k(x_i)$ obtained from x_i .

In other words, one can use the union of symmetric vicinities of virtual examples (obtained from example x_i) to approximate a non-symmetric vicinity of example x_i .

Note that in order to obtain the state of the art performance in the digit recognition problem several authors (Y. LeCun et al. (1998), P. Simmard et al. (1998), and B. Scholkopf et al. (1996)) used virtual examples to increase the number of training examples.

In the SVM approach B. Scholkopf et al. considered the solution as expansion on the extended set of the training data

$$f(x, \alpha) = \sum_{i=1}^{\ell} y_i \sum_{k=1}^N \alpha_{i,k} K(x, x_k(x_i)), \quad (8.33)$$

where extended set included both the training data and the virtual examples obtained from the training data using Lie transformation.

In the simplified vicinal approach, where the coefficient κ that controls the vicinities $v(x_i)$ is so small that

$\mathcal{L}(x, x_i) = K(x, x_i)$, we obtain another expansion

$$f^*(x, \alpha) = \sum_{i=1}^{\ell} y_i \alpha_i \frac{1}{N} \sum_{k=1}^N K(x, x_k(x_i)), \quad (8.34)$$

where $x_k(x_i)$ is the the k th virtual example obtained from the vector x_i of the training data.

The difference between solutions $f(x, \alpha)$ and $f^*(x, \alpha)$ can be described as follows:

In $f(x, \alpha)$ one uses the following information: new (virtual) examples belong to the same class as example x_i .

In $f^*(x, \alpha)$ one uses the following information: new (virtual) examples are the same example as x_i .

The idea of constructing nonsymmetric vicinities as a union of symmetric vicinities can be used even in the case when one can not construct virtual examples. One can consider as examples from the same union a (small) cluster of examples belonging to the same class.

8.5 GENERALIZATION FOR ESTIMATION REAL-VALUED FUNCTIONS

In Chapter 6 to estimate a real-valued function from a given set of functions we used ε -insensitive loss functions

$$L(y, f(x, \alpha)) = L(|y - f(x, \alpha)|_\varepsilon).$$

For this functional we constructed the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(|y_i - f(x, \alpha)|_{\varepsilon}). \quad (8.35)$$

Now instead of functional (8.34) we will use the vicinal risk functional

$$V(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(|y_i - \int f(x, \alpha)p(x|x_i, d_i)dx|_{\varepsilon}). \quad (8.36)$$

We can rewrite the problem of minimizing (8.34) in the following form:
Minimize the functional

$$\Phi(\xi_i) = \sum_{i=1}^{\ell} L(\xi_i), \quad \xi_i \geq 0 \quad (8.37)$$

subject to the constraints

$$\begin{aligned} y_i - \int f(x, \alpha)p(x|x_i, d_i)dx &\geq -\varepsilon - \xi_i, \\ y_i - \int f(x, \alpha)p(x|x_i, d_i)dx &\leq \varepsilon + \xi_i^*. \end{aligned} \quad (8.38)$$

However, we would like to minimize the regularized functional

$$B(f) = C \sum_{i=1}^{\ell} L(\xi_i) + \Omega(f) \quad (8.39)$$

instead of (8.35), where we specify the functional $\Omega(f)$ below.

Suppose (as in Section 8.2) that our set of functions is defined as follows:
We map input vectors x into feature vectors z , and in feature space we construct a linear function

$$l(z) = (w, z) + b$$

that approximates the data

$$(y_1, z_1), \dots, (y_\ell, z_\ell),$$

which are the image of our training data (8.2) in feature space. Let the kernel $K(x, x')$ defines the inner product in feature space.

We would like to define the function that satisfies constraints (8.36) and minimizes the functional

$$\Phi = C \sum_{i=1}^{\ell} \xi_i + (w, w).$$

Consider the case where $L(u) = |u|_\varepsilon$.

The following theorem holds true.

Theorem 8.2. *The vicinal support vector solution has the form*

$$f(x) = \sum_{i=1}^{\ell} (\beta_i - \beta_i^*) \mathcal{L}(x, x_i) + b$$

where to define coefficients β_i and β^* one has to maximize the functional

$$W(\beta)$$

$$= -\varepsilon \sum_{i=1}^{\ell} (\beta_i + \beta_i^*) + \sum_{i=1}^{\ell} y(\beta_i - \beta_i^*) \sum_{i=1}^{\ell} \beta_i - \frac{1}{2} \sum_{i,j=1}^{\ell} (\beta_i - \beta_i^*)(\beta_j - \beta_j^*) \mathcal{M}(x_i, x_j)$$

subject to the constraints

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

$$0 \leq \beta_i \leq C,$$

$$0 \leq \beta_i^* \leq C,$$

where the vicinal kernels $\mathcal{L}(x, x_i)$ and $\mathcal{M}(x_i, x_j)$ are defined by equations (8.14) and (8.15).

The proof of this theorem is identical to the proof of Theorem 8.1.

One can prove analogous theorems for different loss functions $L(u) = L(|y - f(x, \alpha)|_\varepsilon)$. In particular, for the case where $L = (y - f(x, \alpha))^2$ one obtains the solution in closed form.

Theorem 8.3 *The VSV solution for the loss function*

$$L = (y - f(x, \alpha))^2$$

is

$$f(x) = Y^T \left(M + \frac{1}{C} I \right) L,$$

where

$$Y^T = (y_1, \dots, y_\ell)$$

is a $1 \times \ell$ matrix of the values y of observations,

$$M = \|\mathcal{M}(x_i, x_j)\|$$

is an $\ell \times \ell$ matrix whose elements are defined by the two-vicinal kernels,

$$L = \|\mathcal{L}(x, x_1), \dots, \mathcal{L}(x, x_\ell)\|^T$$

is an $\ell \times 1$ matrix whose elements are defined by the one-vicinal kernels $\mathcal{L}(x, x_i)$, $i = 1, \dots, \ell$, and I is the $\ell \times \ell$ identity matrix.

8.6 ESTIMATING DENSITY AND CONDITIONAL DENSITY

8.6.1 Estimating a Density Function

In Chapter 7 when we used Method P for solving the density estimation problem we reduced it to the following optimization problem: Minimize the functional

$$\Omega(f) = (f, f)_H \quad (8.40)$$

subject to the constraints

$$\sup_x \left| F_\ell(x) - \int_{-\infty}^x f(x') dx' \right| = \sigma_\ell. \quad (8.41)$$

However, for computational reasons we checked this constraint only for the ℓ points defined by the data of observations

$$\max_i \left| F_\ell(x) - \int_{-\infty}^x f(x') dx' \right|_{x=x_i} = \sigma_\ell, \quad i = 1, \dots, \ell. \quad (8.42)$$

We also considered the solution as an expansion on the kernel (that defines RKHS)

$$\begin{aligned} f(x) &= \sum_{i=1}^{\ell} \beta_i K_\gamma(x, x_i), \\ \sum_{i=1}^{\ell} \beta_i &= 1, \quad \beta \geq 0. \end{aligned} \quad (8.43)$$

Now let us look for a solution in the form

$$f^*(x) = \sum_{i=1}^{\ell} \beta_i \frac{1}{\nu_i} \int_{v(x_i)} K_\gamma(x, x') dx' = \sum_{i=1}^{\ell} \beta_i \mathcal{L}_\gamma(x, x_i). \quad (8.44)$$

For such solution we obtain (taking into account the reproducing properties of the kernel $K(x, x')$) the following optimization problem:

Minimize the functional

$$W(\beta) = \Omega(f, f) = \sum_{i,j=1}^{\ell} \beta_i \beta_j \mathcal{M}_\gamma(x_i, x_j) \quad (8.45)$$

subject to constraints (8.41) and the constraints

$$\max_i \left| F_\ell(x) - \sum_{j=1}^{\ell} \beta_j \int_{-\infty}^x \mathcal{L}_\gamma(x', x_j) dx' \right|_{x=x_i} = \sigma_\ell, \quad 1 \leq i \leq \ell, \quad (8.46)$$

where $\mathcal{L}_\gamma(x_j, x')$ and $\mathcal{M}_\gamma(x_i, x_j)$ are functions defined by equations (8.14) and (8.15), and γ is a parameter of the width of the kernel

$$K_\gamma(x, x') = a(\gamma) K\left(\frac{x - x'}{\gamma}\right).$$

As in Chapter 7 we choose γ from the admissible set to obtain the minimum (8.43) or/and sparse solution.

This estimator of the density function has an expansion on different kernels depending on $v(x_i)$.

8.6.2 Estimating a Conditional Probability Function

To use the VSV solution for conditional probability estimation we consider the analogous form of expansion as for the density estimation problem

$$p(w|x) = \frac{\ell(w)}{\ell} \sum_{i=1}^{\ell} \beta_i \mathcal{L}_\gamma(x, x_i). \quad (8.47)$$

Repeating the same reasoning as before, one shows that to find the coefficients β_i one needs to minimize the functional

$$W_\gamma(\beta) = \sum_{j=1}^{\ell} \beta_i \beta_j \mathcal{M}_\gamma(x_j, x_i) \quad (8.48)$$

subject to the constraints

$$\max_p \left| \sum_{i=1}^{\ell} \beta_i \frac{1}{\ell} \sum_{j=1}^{\ell} \mathcal{L}_\gamma(x_j, x_i) \theta(x_p - x_j) - F_\ell(x_p|w) \right| = \sigma^* \quad 1 \leq p \leq \ell \quad (8.49)$$

and the constraints

$$\sum_{i=1}^{\ell} \beta_i \left(\frac{1}{\ell} \sum_{j=1}^{\ell} \mathcal{L}_\gamma(x_j, x_i) \right) = 1, \quad (8.50)$$

$$\beta_i \geq 0. \quad (8.51)$$

We choose γ from the admissible set

$$\gamma_{\min} \leq \gamma \leq \gamma_{\max} \quad (8.52)$$

to control properties of the solution (accuracy and/or sparsity) minimizing $W_\gamma(\beta)$ and/or choosing large admissible γ .

8.6.3 Estimating a Conditional Density Function

To estimate the conditional density function we repeat the same reasoning. We use the expansion

$$p(y|x) = \sum_{i=1}^{\ell} \beta_i \mathcal{L}_{\gamma}(x, x_i) K_{\gamma}(y, y_i), \quad \beta_i \geq 0. \quad (8.53)$$

To find the coefficients β_i we minimize the functional

$$W_{\gamma}(\beta) = \sum_{i,j=1}^{\ell} \beta_i \beta_j \mathcal{M}_{\gamma}(x_j, x_i) K_{\gamma}(y_j, y_i) \quad (8.54)$$

subject to the constraints

$$\max_p \left| \sum_{i=1}^{\ell} \beta_i \frac{1}{\ell} \sum_{j=1}^{\ell} \mathcal{L}_{\gamma}(x_j, x_i) \theta(x_p - x_j) \int_{-\infty}^y K_{\gamma}(y', y_i) dy' - F_{\ell}(x_p, y_p) \right| = \sigma_{\ell}, \quad (8.55)$$

$$p = 1, \dots, \ell,$$

and the constraints

$$\sum_{i=1}^{\ell} \left(\frac{1}{\ell} \sum_{j=1}^{\ell} \beta_j \mathcal{L}_{\gamma}(x_i, x_j) \right) = 1, \quad (8.56)$$

$$\beta_i \geq 0. \quad (8.57)$$

To control the properties of the solution (accuracy and/or sparsity) we choose an admissible parameter γ that minimizes the target functional and/or that is large.

Remark. When estimating density, conditional probability, and the conditional density function we looked for a solution

$$f(x, \beta) = \sum_{i=1}^{\ell} \beta_i \mathcal{L}_{\gamma}(x, x_i)$$

that has the following singularities:

$$\beta_i \geq 0, \quad i = 1, \dots, \ell,$$

$$\mathcal{L}_{\gamma}(x, x_i) = E_{v(x_i)} K_{\gamma}(x, x'),$$

$$\mathcal{M}_{\gamma}(x_i x_j) = E_{v(x_i)} E_{v(x_j)} K_{\gamma}(x, x'),$$

where

$$K_{\gamma}(x, x_i) = a(\gamma) K \left(\frac{x - x_i}{\gamma} \right)$$

with a normalization parameter $a(\gamma)$ (see Section 7.8).

Since parameters β_i are nonnegative it is reasonable to construct solutions based on kernels $K(x, x')$ that have light tails or have finite support. In particular, one can use the kernel defined by the normal law

$$K_\Delta(x, x') = \frac{1}{\sqrt{2\pi}\gamma} \exp \left\{ -\frac{(x - x')^2}{2\gamma^2} \right\}.$$

For this kernels we have

$$\mathcal{L}_\gamma(x, x') = [2\pi(\gamma^2 + d_i^2)]^{-\frac{n}{2}} \exp \left\{ -\frac{(x - x')^2}{2(\gamma^2 + d_i^2)} \right\} \quad (8.58)$$

$$\mathcal{M}_\gamma(x_j, x_i) = [2\pi(\gamma^2 + d_i^2 + d_j^2)]^{-\frac{n}{2}} \exp \left\{ -\frac{(x - x')^2}{2(\gamma^2 + d_i^2 + d_j^2)} \right\}. \quad (8.59)$$

As a kernel $K_\gamma(x, x_i)$ defined on finite support one can consider B_n -spline

$$B_n(x, x') = \frac{1}{\gamma} \sum_{j=0}^{n+1} \frac{(-1)^j}{(n+1)!} C_j^{n+1} ((x - x') + (n-1-j)\gamma)_+^n.$$

It is known that starting with $n = 2$ a B_n -spline can be approximated by a Gaussian function

$$B_n(x, x') \approx \sqrt{\frac{6}{\pi\gamma^2(n+1)}} \exp \left\{ -\frac{6(x - x')^2}{\gamma^2(n+1)} \right\}. \quad (8.60)$$

Therefore, for one- and two-vicinal kernels constructed on the basis of kernel function defined by a B_n -spline one has either to calculate them directly or use the approximation (8.60) and expressions (5.58) and (5.59).

8.6.4 Estimating a Regression Function

To estimate the regression function

$$r(x) = \int yp(y|x)dy \quad (8.61)$$

recall that the kernel $K_\gamma(y, y_j)$ is a symmetric (density) function the integral of which is equal to 1. For such a function we have

$$\int yK_\gamma(y, y_i)dy = y_i. \quad (8.62)$$

Therefore, from (8.51), (8.56), and (7.57) we obtain the following regression function:

$$r(x) = \sum_{i=1}^{\ell} y_i \beta_i \mathcal{L}_\gamma(x, x_i)$$

Informal Reasoning and Comments — 8

The inductive principle introduced in this chapter is brand new. There remains work to properly analyze it, but the first results are good.

Sayan Mukherjee used this principle for solving the density estimation problem based on the VSV solution (so far in low-dimensional spaces). He demonstrated its advantages by comparing it to existing approaches, especially in the case where the sample size is small.

Ideas that are close to this one have appeared in the nonparametric density estimation literature. In particular, many discussions have taken place in order to modernize the Parzen's methods of density estimation. Researchers have created methods that use different values of the width at different points. It appeared that the width of the kernel at a given point should be somehow connected to the size of the vicinity of this point.

However, the realizations of this idea were too straightforward: It was proposed to choose the width of the kernel proportional to the value d_i of the vicinity of the corresponding point x_i . In other words, it was proposed to use the kernel $a(\gamma)K\left(\frac{x-x_i}{d_i\gamma}\right)$. This suggestion, however, created the following problem: When the value of the vicinity decreases, the new kernel converges to the δ -function

$$\lim_{d_i \rightarrow 0} a(d_i\gamma)K\left(\frac{x-x_i}{d_i\gamma}\right) = \delta(x - x_i).$$

In the 1980s, in constructing density estimators from various solutions of

an integral equation we observed that classical methods such as Parzen's method or the projection method are defined by different conditions for solving this integral equation with the same approximation on the right-hand side – the empirical distribution function. The idea of using a discontinuous function to approximate a continuous function in the problem of solving the integral equation that defines the derivative of the (given) right-hand side is probably not the best.

Using in the same equations the continuous approximation to the distribution function, we obtain nonclassical estimators. In particular, using a continuous piecewise linear (polygonal) approximation we obtained (in the one-dimensional case) a Parzen's-type estimator with a new kernel defined as follows (Vapnik, 1988):

$$g_{\text{new}}(x; x_i, x_{i+1}, \gamma) = \frac{a(\gamma)}{(x_{i+1} - x_i)} \int_{x_i}^{x_{i+1}} K\left(\frac{x-z}{\gamma}\right) dz,$$

where x_i, x_{i+1} are elements of the variation series of the sample and $K_\gamma(u)$ is the Parzen kernel.

This kernel converges to Parzen's kernels when $(x_{i+1} - x_i) \rightarrow 0$,

$$\lim_{(x_{i+1}-x_i) \rightarrow 0} g_{\text{new}}(x; x_i, x_{i+1}, \gamma) = a(\gamma) K\left(\frac{x-z}{\gamma}\right).$$

After the introduction of SVM methods, the (sparse) kernel approximation began to play an important role in solving various function estimation problems. As in Parzen's density estimation method, the SVM methods use the same kernel (with different values of coefficients of expansions and different support vectors). Of course, the question arises as to whether it is possible to construct different kernels for different support vectors. Using the VRM principle we obtain kernels of a new type in all the problems considered in this book.

The VRM principle was actually introduced as an attempt to understand the nature of the solutions that use different widths of kernel.

Chapter 9

Conclusion: What Is Important in Learning Theory?

9.1 WHAT IS IMPORTANT IN THE SETTING OF THE PROBLEM?

In the beginning of this book we postulated (without any discussion) that learning is a problem of *function estimation* on the basis of empirical data. To solve this problem we used a classical inductive principle – the ERM principle. Later, however, we introduced a new principle – the SRM principle. Nevertheless, the general understanding of the problem remains based on the statistics of large samples: The goal is to derive the rule that possesses the lowest risk. The goal of obtaining the “lowest risk” reflects the philosophy of large sample size statistics: The rule with low risk is good because if we use this rule for a large test set, with high probability the means of losses will be small.

Mostly, however, we face another situation. We are simultaneously given training data (pairs (x_i, y_i)) and test data (vectors x_j^*), and the goal is to use the learning machine with a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, to find the y_j^* for the *given* test data. In other words, we face the problem of estimating the *values of the unknown function at given points*.

Why should the problem of estimating the values of an unknown function at given points of interest be solved in two stages: First estimating the function and second estimating the values of the function using the estimated function? In this two-stage scheme one actually tries to solve a relatively simple problem (estimating the values of a function at given points of interest) by first solving (as an intermediate problem) a much more difficult

one (estimating the function). Recall that estimating a function requires estimating the values of the function at *all (infinite) points of the domain* where the function is defined including the points of interest. Why should one first estimate the values of the function at *all points of the domain* to estimate the values of the function at the points of interest?

It can happen that one does not have enough information (training data) to estimate the function well, but one does have enough data to estimate the values of the function at a *given finite number of points of interest*.

Moreover, in human life, decision-making problems play an important role. For learning machines these can be formulated as follows: Given the training data

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

the machine with functions $f(x, \alpha)$, $\alpha \in \Lambda$, has to find among the test data

$$x_1^*, \dots, x_k^*,$$

the one x_*^* that belongs to the first class with highest probability (decision making problem in the pattern recognition form).¹ To solve this problem one does not even need to estimate the values of the function at all given points; therefore it can be solved in situations where one does not have enough information (not enough training data) to estimate the value of a function at given points.

The key to the solution of these problems is the following observation, which for simplicity we will describe for the pattern recognition problem.

The learning machine (with a set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$) is simultaneously given two strings: the string of $\ell + k$ vectors x from the training and the test sets, and the string of ℓ values y from the training set. In pattern classification the goal of the machine is to define the string containing k values y for the test data.

For the problem of estimating the values of a function at the given points the set of functions implemented by the learning machine can be *factorized* into a finite set of equivalence classes. (Two indicator functions fall in the same equivalence class if they coincide on the string $x_1, \dots, x_{\ell+k}$). These equivalence classes can be characterized by their cardinality (how many functions they contain).

The cardinality of equivalence classes is a concept that makes the theory of estimating the function at the given points differ from the theory of estimating the function. This concept (as well as the theory of estimating the function at given points) was considered in the 1970s (Vapnik, 1979). For the set of linear functions it was found that the bound on generalization ability, in the sense of minimizing the number of errors only on the given

¹Or to find one that with the most probability possesses the largest value of y_* (decision-making in regression form).

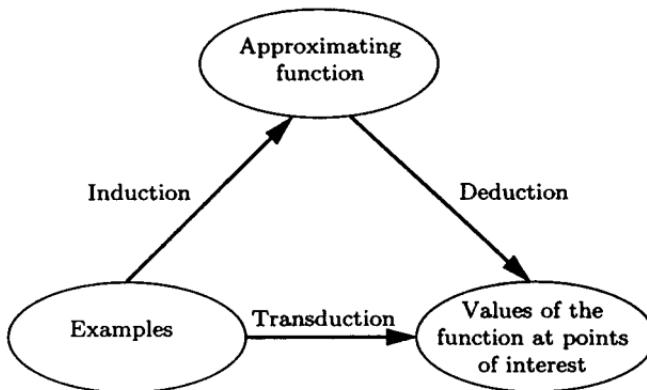


FIGURE 9.1. Different types of inference. *Induction*, deriving the function from the given data. *Deduction*, deriving the values of the given function for points of interest. *Transduction*, deriving the values of the unknown function for points of interest from the given data. The classical scheme suggests deriving S the values of the unknown function for points of interest in two steps: first using the inductive step, and then using the deduction step, rather than obtaining the direct solution in one step.

test data (along with the factors considered in this book), depends also on a new factor, the cardinality of equivalence classes. Therefore, since to minimize a risk one can minimize the obtained bound over a larger number of factors, one can find a lower minimum. Now the problem is to construct a general theory for estimating a function at the given points. This brings us to a new concept of learning.

Classical philosophy usually considers two types of inference: *deduction*, describing the movement from general to particular, and *induction*, describing the movement from particular to general.

The model of estimating the value of a function at a given point of interest describes a new concept of inference: moving *from particular to particular*. We call this type of inference *transductive inference*. (Fig. 9.1)

Note that this concept of inference appears when one would like to get the best result from a restricted amount of information. The main idea in this case was described in Section 1.9 as follows:

If you are limited to a restricted amount of information, do not solve the particular problem you need by solving a more general problem.

We used this idea for constructing a direct method of estimating the functions. Now we would like to continue developing this idea: Do not

solve the problem of estimating the values of a function at given points by estimating the entire function, and do not solve a decision-making problem by estimating the values of a function at a given points, etc.

The problem of estimating the values of a function at a given point addresses a question that has been discussed in philosophy for more than 2000 years:

What is the basis of human intelligence: knowledge of laws (rules) or the culture of direct access to the truth (intuition, adhoc inference)?

There are several different models embracing the statements of the learning problem, but from the conceptual point of view none can compare to the problem of estimating the values of the function at given points. This model can provide the strongest contribution to the 2000 years of discussions about the essence of human reason.

9.2 WHAT IS IMPORTANT IN THE THEORY OF CONSISTENCY OF LEARNING PROCESSES?

The theory of consistency of learning processes is well developed. It answers almost all questions toward understanding the conceptual model of learning processes realizing the ERM principle. The only remaining open question is that of necessary and sufficient conditions for a fast rate of convergence. In Chapter 2 we considered the sufficient condition described using annealed entropy

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda}(\ell)}{\ell} = 0$$

for the pattern recognition case. It also can be shown that the conditions

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda}(\varepsilon; \ell)}{\ell} = 0, \quad \forall \varepsilon > 0,$$

in terms of the annealed entropy $H_{\text{ann}}^{\Lambda}(\varepsilon; \ell) = \ln EN^{\Lambda}(\varepsilon; z_1, \dots, z_\ell)$ define sufficient conditions for fast convergence in the case of regression estimation.

The following question remains:

Do these equalities form the necessary conditions as well? If not, what are necessary and sufficient conditions?

Why is it important to find a concept that describes necessary and sufficient conditions for a fast rate of convergence?

As was demonstrated, this concept plays a key role in the theory of bounds. In our constructions we used the annealed entropy for finding both (nonconstructive) distribution-independent bounds and (nonconstructive)

distribution-dependent bounds. On the basis of annealed entropy, we constructed both the growth function and the generalized growth function. Proving necessity of annealed entropy for a fast rate of convergence would amount to showing that this is the best possible construction for deriving bounds on the generalization ability of learning machines. If necessary and sufficient conditions are described by another function, the constructions can be reconsidered.

9.3 WHAT IS IMPORTANT IN THE THEORY OF BOUNDS?

The theory of bounds contains two parts: the theory of nonconstructive bounds, which are obtained on the basis of the concepts of the growth function and the generalized growth function, and the theory of constructive bounds, where the main problem is estimating these functions using some constructive concept.

The main problem in the theory of bounds is in the second part. One has to introduce some constructive concept by means of which one can estimate the growth function or the generalized growth function. In 1968 we introduced the concept of the VC dimension and found the bound for the growth function (Vapnik and Chervonenkis, 1968, 1971). We proved that the value $N^\Lambda(\ell)$ is either 2^ℓ or polynomial bounded,²

$$N^\Lambda(z_1, \dots, z_\ell) \leq \left(\frac{e\ell}{h} \right)^h.$$

Note that the polynomial on the right-hand side depends on *one* free parameter h . This bound (which depends on one capacity parameter) cannot be improved (there exist examples where equality is achieved).

The challenge is to find refined concepts containing more than one parameter (say two parameters) that describe some properties of capacity (and the set of distribution functions $F(z) \in \mathcal{P}$), by means of which one can obtain better bounds.³

This is a very important question, and the answer would have immediate impact on the bounds of the generalization ability of learning machines.

²In 1972 this bound was also published by Sauer (Sauer, 1972).

³Recall the MDL bound: Even such a refined concept as the coefficient of compression provides a worse bound than one based on three (actually rough) concepts such as the value of the empirical risk, the number of observations, and the number of functions in a set.

9.4 WHAT IS IMPORTANT IN THE THEORY FOR CONTROLLING THE GENERALIZATION ABILITY OF LEARNING MACHINES?

The most important problem in the theory for controlling the generalization ability of learning machines is finding a new inductive principle for small sample sizes. In the mid-1970s, several techniques were suggested to improve the classical methods of function estimation. Among these are the various rules for choosing the degree of a polynomial in the polynomial regression problem, various regularization techniques for multidimensional regression estimation, and the regularization method for solving ill-posed problems. All these techniques are based on the same idea: to provide the set of functions with a structure and then to minimize the risk on the elements of the structure. In the 1970s the crucial role of capacity control was discovered. We call this general idea SRM to stress the importance of minimizing the risk in the element of the structures.

In SRM, one tries to control simultaneously two parameters: the value of the empirical risk and the capacity of the element of the structure.

In the 1970s the MDL principle was proposed. Using this principle, one can control the coefficient of compression.

The most important question is this:

Does there exist a new inductive principle for estimating dependency from small sample sizes?

In studies of inductive principles it is crucial to find new concepts that affect the bounds of the risk, and which therefore can be used in minimizing these bounds. To use an additional concept, we introduced a new statement of the learning problem: the local risk minimization problem. In this statement, in the framework of the SRM principle, one can control three parameters: empirical risk, capacity, and locality.

In the problem of estimating the values of a function at the given points one can use an additional concept: the cardinality of equivalence classes. This aids in controlling the generalization ability: By minimizing the bound over four parameters, one can get smaller minima than by minimizing the bound over fewer parameters. The problem is to find a new concept that can affect the upper bound of the risk. This will immediately lead to a new learning procedure, and even to a new type of reasoning (as in the case of transductive inference).

Finally, it is important to find new structures on the set of functions. It is interesting to find structures with elements containing functions that are described by large numbers of parameters, but nevertheless have low VC dimension. We have found only one such structure, and this brought us to SV machines. New structures of this kind will probably result in new types of learning machines.

9.5 WHAT IS IMPORTANT IN THE THEORY FOR CONSTRUCTING LEARNING ALGORITHMS?

The algorithms for learning should be well controlled. This means that one has to control two main parameters responsible for generalization ability: the value of the empirical risk and the VC dimension of the smallest element of the structure that contains the chosen function.

The SV technique can be considered as an effective tool for controlling these two parameters if structures are defined on the sets of linear functions in some high-dimensional feature space. This technique is not restricted only to the sets of indicator functions (for solving pattern recognition problems). At the end of Chapter 5 we described the generalization of the SV method for solving regression problems. In the framework of this generalization, using a special convolution function one can construct high-dimensional spline functions belonging to the subset of splines with a chosen VC dimension. Using different convolution functions for the inner product one can also construct different types of functions nonlinear in input space.⁴

Moreover, the SV technique goes beyond the framework of learning theory. It admits a general point of view as a new type of parameterization of sets of functions.

The matter is that in solving the function estimation problems in both computational statistics (say pattern recognition, regression, density estimation) and in computational mathematics (say, obtaining approximations to the solution to multidimensional (operator) equations of different types) the first step is describing (parameterizing) a set of functions in which one is looking for a solution.

In the first half of this century the main idea of parameterization (after the Weierstrass theorem) was polynomial series expansion. However, even in the one-dimensional case sometimes one needs a few dozen terms for accurate function approximation. To treat such a series for solving many problems the accuracy of existing computers can be insufficient.

Therefore, in the middle of the 1950s a new type of function parameterization was suggested, the so-called spline functions (piecewise polynomial functions). This type of parameterization allowed us to get an accurate

⁴Note once more that advanced estimation techniques in statistics developed in the 1980s such as projection pursuit regression, MARS, hinging hyperplanes, etc in fact consider some special approximations in the sets of functions

$$y = \sum_{j=1}^N \alpha_j K\{(x \cdot w_j)\} + b,$$

where $\alpha_1, \dots, \alpha_N$ are scalars and w_1, \dots, w_N are vectors.

solution for most one-dimensional (sometimes two-dimensional) problems. However, it often fails in, say, the four-dimensional case.

The SV parameterization of functions can be used in high-dimensional space (recall that for this parameterization the complexity of approximation depends on the number of support vectors rather than on the dimensionality of the space). By controlling the “capacity” of the set of functions one can control the “smoothness” properties of the approximation.

This type of parameterization should be taken into account whenever one considers multidimensional problems of function estimation (function approximation).

Currently we have experience only in using the SV technique for solving pattern recognition problems. However, theoretically there is no obstacle to obtain using this technique the same high level of accuracy in solving dependency estimation problems that arise in different areas of statistics (such as regression estimation, density estimation, conditional density estimation) and computational mathematics (such as solving some multidimensional linear operator equations).

One can consider the SV technique as a new type of parameterization of multidimensional functions that in many cases allows us to overcome the curse of dimensionality.⁵

9.6 WHAT IS THE MOST IMPORTANT?

The learning problem belongs to the problems of natural science: There exists a phenomenon for which one has to construct a model. In the attempts to construct this model, theoreticians can choose one of two different positions depending on which part of Hegel’s formula (describing the general philosophy of nature) they prefer:

*Whatever is real is rational, and whatever is rational is real.*⁶

The interpretation of the first part of this formula can be as follows. Somebody (say an experimenter) knows a model that describes reality, and the problem of the theoretician is to prove that this model is rational (he should define as well what it means to be rational). For example, if somebody believes and can convince the theoretician that neural networks

⁵See footnote on page 170.

⁶In Hegel’s original assertion, the meaning of the words “real” and “rational” does not coincide with the common meaning of these words. Nevertheless, according to a remark of B. Russell, the identification of the real and the rational in a common sense leads to the belief that “whatever is, is right.” Russell did not accept this idea (see B. Russell, *A History of Western Philosophy*). However, we do interpret Hegel’s formula as: “Whatever exists is right, and whatever right is exists.”

are good models of real brains, then the goal of the theoretician is to prove that this model is rational.

Suppose that the theoretician considers the model to be “rational” if it possesses some remarkable asymptotic properties. In this case, the theoretician succeeds if he or she proves (as has been done) that the learning process in neural networks asymptotically converges to local extrema and that a sufficiently large neural network can approximate well any smooth function. The conceptual part of such a theory will be complete if one can prove that the achieved local extremum is close to the global one.

The second position is a heavier burden for the theoretician: The theoretician has to define what a rational model is, then has to find this model, and finally, the must convince the experimenters to prove that this model is real (describes reality).

Probably, a rational model is one that not only has remarkable asymptotic properties but also possesses some remarkable properties in dealing with a given finite number of observations.⁷ In this case, the small sample size philosophy is a useful tool for constructing rational models.

The rational models can be so unusual that one needs to overcome prejudices of common sense in order to find them. For example, we saw that the generalization ability of learning machines depends on the VC dimension of the set of functions, rather than on the number of parameters that define the functions within a given set. Therefore, one can construct high-degree polynomials in high-dimensional input space with good generalization ability. Without the theory for controlling the generalization ability this opportunity would not be clear. Now the experimenters have to answer the question: Does generalization, as performed by real brains, include mechanisms similar to the technology of support vectors?⁸

That is why the role of theory in studies of learning processes can be more constructive than in many other branches of natural science.

This, however, depends on the choice of the general position in studies of learning phenomena. The choice of the position reflects the belief of which in this specific area of natural science is the main discoverer of truth: experiment or theory.

⁷ Maybe it has to possess additional properties. Which?

⁸ The idea that the generalization, the definition of the importance of the observed facts, and storage of the important facts, are different aspects of the same brain mechanism is very attractive.

References

Remarks on References

One of the greatest mathematicians of the century, A.N. Kolmogorov, once noted that an important difference between mathematical sciences and historical sciences is that facts once found in mathematics hold forever, while the facts found in history are reconsidered by every generation of historians.

In statistical learning theory as in mathematics the importance of results obtained depends on new facts about learning phenomena, whatever they reveal, rather than a new description of already known facts. Therefore, I tried to refer to the works that reflect the following sequence of the main events in developing the statistical learning theory described in this book:

- 1958–1962. Constructing the perceptron.
- 1962–1964. Proving the first theorems on learning processes.
- 1958–1963. Discovery of nonparametric statistics.
- 1962–1963. Discovery of the methods for solving ill-posed problems.
- 1960–1965. Discovery of the algorithmic complexity concept and its relation to inductive inference.
- 1968–1971. Discovery of the law of large numbers for the space of indicator functions and its relation to the pattern recognition problem.

- 1965–1973. Creation of a general asymptotic learning theory for stochastic approximation inductive inference.
- 1965–1972. Creation of a general nonasymptotic theory of pattern recognition for the ERM principle.
1974. Formulation of the SRM principle.
1978. Formulation of the MDL principle.
- 1974–1979. Creation of the general nonasymptotic learning theory based on both the ERM and SRM principles.
1981. Generalization of the law of large numbers for the space of real-valued functions.
1986. Construction of NN based on the back-propagation method.
1989. Discovery of necessary and sufficient conditions for consistency of the ERM principle and the ML method.
- 1989–1993. Discovery of the universality of function approximation by a sequence of superpositions of sigmoid functions.
- 1992–1995. Constructing the SV machines.

REFERENCES

- M.A. Aizerman, E.M. Braverman, and L.I. Rozonoer (1964), “Theoretical foundations of the potential function method in pattern recognition learning,” *Automation and Remote Control* **25**, pp. 821–837.
- M.A. Aizerman, E.M. Braverman, and L.I. Rozonoer (1965), “The Robbins-Monroe process and the method of potential functions,” *Automation and Remote Control*, **28**, pp. 1882–1885.
- H. Akaike (1970), “Statistical predictor identification,” *Annals of the Institute of Statistical Mathematics*, pp. 202–217.
- S. Amari (1967), “A theory of adaptive pattern classifiers,” *IEEE Trans. Elect. Comp.*, **EC-16**, pp. 299–307.
- T.W. Anderson and R.R. Bahadur (1966), “Classification into two multivariate normal distributions with different covariance matrices.” *The Annals of Mathematical Statistics* **133** (2).

- A.R. Barron (1993), "Universal approximation bounds for superpositions of a sigmoid function," *IEEE Transactions on Information Theory* **39** (3) pp. 930–945.
- J. Berger (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer.
- B. Boser, I. Guyon, and V.N. Vapnik (1992), "A training algorithm for optimal margin classifiers," *Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh ACM, pp. 144–152.
- L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Müller, E Säckinger, P. Simard, and V. Vapnik (1994), "Comparison of classifier methods: A case study in handwritten digit recognition *Proceedings 12th IAPR International Conference on Pattern Recognition*, **2**, IEEE Computer Society Press Los Alamos, California, pp. 77–83.
- L. Bottou and V. Vapnik (1992), "Local learning algorithms," *Neural Computation* **4** (6), pp. 888–901.
- L. Breiman (1993), "Hinging hyperplanes for regression, classification and function approximation," *IEEE Transactions on Information Theory* **39** (3), pp. 999–1013.
- L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone (1984), *Classification and regression trees*, Wadsworth, Belmont, CA.
- A. Bryson, W. Denham, and S. Dreyfuss (1963), "Optimal programming problem with inequality constraints. I: Necessary conditions for extremal solutions" *AIAA Journal* **1**, pp. 25–44.
- F.P. Cantelli (1933), "Sulla determinazione empirica della leggi di probabilità," *Giornale dell' Institute Italiano degli Attuari* (4).
- G.J. Chaitin (1966), "On the length of programs for computing finite binary sequences," *J. Assoc. Comput. Mach.*, **13**, pp. 547–569.
- N.N. Chentsov (1963), "Evaluation of an unknown distribution density from observations," *Soviet Math.* **4**, pp. 1559–1562.
- C. Cortes and V. Vapnik (1995), "Support Vector Networks," *Machine Learning* **20**, pp 1–25.
- R. Courant and D. Hilbert (1953), *Methods of Mathematical Physics*, J. Wiley, New York.
- G. Cybenko (1989), "Approximation by superpositions of sigmoidal function," *Mathematics of Control, Signals, and Systems* **2**, pp. 303–314.

- L. Devroye (1988), "Automatic pattern recognition: A Study of the probability of error," *IEEE Transaction on Pattern Analysis and Machine Intelligence* **10** (4), pp. 530–543.
- L. Devroye and L. Györfi (1985), *Nonparametric density estimation in L_1 view*, J. Wiley, New York.
- H. Drucker, R. Schapire, and P. Simard (1993), "Boosting performance in neural networks," *International Journal in Pattern Recognition and Artificial Intelligence* **7** (4), pp. 705–719.
- R.M. Dudley (1978), "Central limit theorems for empirical measures," *Ann. Prob.* **6** (6), pp. 899–929.
- R.M. Dudley (1984), *Course on empirical processes*, Lecture Notes in Mathematics, Vol. 1097, pp. 2–142, Springer, New York.
- R.M. Dudley (1987), "Universal Donsker classes and metric entropy," *Ann. Prob.* **15** (4), pp. 1306–1326.
- R.A. Fisher (1952), *Contributions to Mathematical Statistics*, J. Wiley, New York.
- J.H. Friedman, T. Hastie, and R. Tibshirani (1998), "Technical report," Stanford University, Statistic Department. (www-stat.stanford.edu/~ghf/#papers)
- J.H. Friedman and W. Stuetzle (1981), "Projection pursuit regression," *JASA* **76**, pp. 817–823.
- F. Girosi, and G. Anzellotti (1993), "Rate of convergence for radial basis functions and neural networks," *Artificial Neural Networks for Speech and Vision*, Chapman & Hall, pp. 97–113.
- V. I. Glivenko (1933), "Sulla determinazione empirica di probabilita'," *Giornale dell' Instituto Italiano degli Attuari* (4).
- U. Grenander (1981), *Abstract inference*, J. Wiley, New York.
- A.E. Hoerl and R.W. Kennard (1970), "Ridge regression: Biased estimation for non-orthogonal problems," *Technometrics* **12**, pp. 55–67.
- P. Huber (1964), "Robust estimation of location parameter," *Annals of Mathematical Statistics* **35** (1).
- L.K. Jones (1992), "A simple lemma on greedy approximation in Hilbert space and convergence rates for Projection Pursuit Regression," *The Annals of Statistics* **20** (1), pp. 608–613.

- I.A. Ibragimov and R.Z. Hasminskii (1981), *Statistical estimation: Asymptotic theory*, Springer, New York.
- V.V. Ivanov (1962), "On linear problems which are not well-posed," *Soviet Math. Docl.* **3** (4), pp. 981-983.
- V.V. Ivanov (1976), *The theory of approximate methods and their application to the numerical solution of singular integral equations*, Leyden, Nordhoff International.
- M. Karpinski and T. Werther (1989), "VC dimension and uniform learnability of sparse polynomials and rational functions," *SIAM J. Computing*, Preprint 8537-CS, Bonn University, 1989.
- A. N. Kolmogoroff (1933), "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell' Instituto Italiano degli Attuari* (4).
- A.N. Kolmogorov (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer. (English translation: A.N. Kolmogorov (1956), *Foundation of the Theory of Probability*, Chelsea.)
- A.N. Kolmogorov (1965), "Three approaches to the quantitative definitions of information," *Problem of Inform. Transmission* **1** (1), pp. 1-7.
- L. LeCam (1953), "On some asymptotic properties of maximum likelihood estimates and related Bayes estimate," *Univ. Calif. Public. Stat* **11**
- Y. LeCun (1986), "Learning processes in an asymmetric threshold network," *Disordered systems and biological organizations*, Les Houches, France, Springer, pp. 233-240.
- Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.J. Jackel (1990), "Handwritten digit recognition with back-propagation network," *Advances in Neural Information Processing Systems* **2** Morgan Kaufman, pp. 396-404.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner (1998), "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86**, pp. 2278-2324.
- G.G. Lorentz (1966), *Approximation of functions*, Holt-Rinehart-Winston, New York.
- G. Matheron and M. Armstrong (ed) (1987), *Geostatistical case studies (Quantitative geology and geostatistics)*, D. Reider Publishing Co.
- H.N. Mhaskar (1993), "Approximation properties of a multi-layer feed-forward artificial neural network," *Advances in Computational Mathematics* **1** pp. 61-80.

- C.A. Micchelli (1986), "Interpolation of scattered data: distance matrices and conditionally positive definite functions," *Constructive Approximation* **2** pp. 11–22.
- M.L. Miller (1990), *Subset selection in regression*, London, Chapman and Hall.
- J.J. More and G. Toraldo (1991), "On the solution of large quadratic programming problems with bound constraints," *SIAM Optimization*, **1**, (1), pp. 93–113.
- A.B.J. Novikoff (1962), "On convergence proofs on perceptrons," *Proceedings of the Symposium on the Mathematical Theory of Automata*, Polytechnic Institute of Brooklyn, Vol. XII, pp. 615–622.
- S. Paramasamy (1992), "On multivariant Kolmogorov-Smirnov distribution," *Statistics & Probability Letters* **15**, pp. 140–155.
- J.M. Parrondo and C. Van den Broeck (1993), "Vapnik-Chervonenkis bounds for generalization," *J. Phys. A*, **26**, pp. 2211–2223.
- E. Parzen (1962), "On estimation of probability function and mode." *Annals of Mathematical Statistics* **33** (3).
- D.Z. Phillips (1962), "A technique for numerical solution of certain integral equation of the first kind," *J. Assoc. Comput. Mach.* **9** pp. 84–96.
- T. Poggio and F. Girosi (1990), "Networks for Approximation and Learning," *Proceedings of the IEEE* **78** (9).
- D. Pollard (1984), *Convergence of stochastic processes*, Springer, New York.
- K. Popper (1968), *The Logic of Scientific Discovery*, 2nd ed., Harper Torch Book, New York.
- M.J.D. Powell (1992), "The theory of radial basis functions approximation in 1990," W.A. Light ed., *Advances in Numerical Analysis Volume II: Wavelets, Subdivision algorithms and radial basis functions*, Oxford University, pp. 105–210.
- J. Rissanen (1978), "Modeling by shortest data description," *Automatica*, **14**, pp. 465–471.
- J. Rissanen (1989), *Stochastic complexity and statistical inquiry*, World Scientific.
- H. Robbins and H. Monroe (1951), "A stochastic approximation method," *Annals of Mathematical Statistics* **22**, pp. 400–407.

- F. Rosenblatt (1962), *Principles of neurodynamics: Perceptron and theory of brain mechanisms*, Spartan Books, Washington D.C.
- M. Rosenblatt (1956), "Remarks on some nonparametric estimation of density functions," *Annals of Mathematical Statistics* **27**, pp. 642–669.
- D.E. Rumelhart, G.E. Hinton, and R.J. Williams (1986), Learning internal representations by error propagation. *Parallel distributed processing: Explorations in macrostructure of cognition*, Vol. I, Badford Books, Cambridge, MA., pp. 318–362.
- B. Russell (1989), *A History of Western Philosophy*, Unwin, London.
- N. Sauer (1972), "On the density of families of sets," *J. Combinatorial Theory (A)* **13** pp.145–147.
- G. Schwartz (1978), "Estimating the dimension of a model," *Annals of Statistics* **6**, pp. 461–464.
- B. Scholkopf, C. Burges, and V. Vapnik (1996) "Incorporating invariance in support vector learning machines," in book *C. von der Malsburg, W. von Seelen, J.C Vonbruggen, and S. Sendoff (eds) Artificial Neural Network – ICANN'96. Springer Lecture Notes in Computer Science* Vol. 1112, Berlin pp. 47–52.
- P.Y. Simard, Y. LeCun, and J. Denker (1993), "Efficient pattern recognition using a new transformation distance," *Neural Information Processing Systems* **5** pp. 50-58.
- P.Y. Simard, Y. LeCun, J. Denker , and B. Victorri (1998), "Transformation invariance in pattern recognition – tangent distance and tangent propagation," in the book *G.B. Orr and K. Muller (eds) Neural networks: Tricks and trade*, Springer.
- N.V. Smirnov (1970), *Theory of probability and mathematical statistics (Selected works)*, Nauka, Moscow.
- R.J. Solomonoff (1960), "A preliminary report on general theory of inductive inference," Technical Report ZTB-138, Zator Company, Cambridge, MA.
- R.J. Solomonoff (1964), "A formal theory of inductive inference," Parts 1 and 2, *Inform. Contr.*,**7**, pp. 1–22, pp. 224–254.
- R.A. Tapia and J.R. Thompson (1978), *Nonparametric probability density estimation*, The John Hopkins University Press, Baltimore.
- A.N. Tikhonov (1963), "On solving ill-posed problem and method of regularization," *Doklady Akademii Nauk USSR*, **153**, pp. 501–504.

- A.N. Tikhonov and V.Y. Arsenin (1977), *Solution of ill-posed problems*, W. H. Winston, Washington, DC.
- Ya.Z. Tsypkin (1971), *Adaptation and learning in automatic systems*, Academic Press, New York.
- Ya.Z. Tsypkin (1973), *Foundation of the theory of learning systems*, Academic Press, New York .
- V.N. Vapnik (1979), *Estimation of dependencies based on empirical Data*, (in Russian), Nauka, Moscow. (English translation: Vladimir Vapnik (1982), *Estimation of dependencies based on empirical data*, Springer, New York.)
- V.N. Vapnik (1993), "Three fundamental concepts of the capacity of learning machines," *Physica A* **200**, pp. 538–544.
- V.N. Vapnik (1988), "Inductive principles of statistics and learning theory" *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting*, **1**, Nauka, Moscow. (English translation: (1995), "Inductive principles of statistics and learning theory," in the book *Smolensky, Moser, Rumelhart, eds., Mathematical perspectives on neural networks*, Lawrence Erlbaum Associates, Inc.)
- Vladimir. Vapnik (1998), *Statistical learning theory*, J. Wiley, New York.
- V.N. Vapnik and L. Bottou (1993), "Local Algorithms for pattern recognition and dependencies estimation," *Neural Computation*, **5** (6) pp 893–908.
- V.N. Vapnik and A.Ja. Chervonenkis (1968), "On the uniform convergence of relative frequencies of events to their probabilities," *Doklady Akademii Nauk USSR* **181** (4). (English transl. Sov. Math. Dokl.)
- V.N. Vapnik and A.Ja. Chervonenkis (1971), "On the uniform convergence of relative frequencies of events to their probabilities" *Theory Probab. Appl.* **16** pp. 264–280
- V.N. Vapnik and A.Ja. Chervonenkis (1974), *Theory of Pattern Recognition* (in Russian), Nauka, Moscow. (German translation: W.N. Vapnik, A.Ja. Tscherwonenkis (1979), *Theorie der Zeichenerkennung*, Akademie, Berlin.)
- V.N. Vapnik and A.Ja. Chervonenkis (1981), "Necessary and sufficient conditions for the uniform convergence of the means to their expectations," *Theory Probab. Appl.* **26**, pp. 532–553.

V.N. Vapnik and A.Ja. Chervonenkis (1989), "The necessary and sufficient conditions for consistency of the method of empirical risk minimization" (in Russian), *Yearbook of the Academy of Sciences of the USSR* on Recognition, Classification, and Forecasting **2**, pp. 217–249, Nauka, Moscow pp 207–249. (English translation: (1991), "The necessary and sufficient conditions for consistency of the method of empirical risk minimization," *Pattern Recogn. and Image Analysis* **1** (3), pp. 284–305.)

V.N. Vapnik and A.R. Stefanyuk (1978)," Nonparametric methods for estimating probability densities," *Autom. and Remote Contr* (8).

V.V. Vasin (1970), "Relationship of several varitional methods for approximate solutions of ill-posed problems," *Math. Notes* **7**, pp. 161–166.

R.S. Wenocur and R.M. Dudley (1981), "Some special Vapnik-Chervonenkis classes," *Discrete Math.* **33**, pp. 313-318.

Index

AdaBoost algorithm 163
admissible structure 95
algorithmic complexity 10
annealed entropy 55
ANOVA decomposition 199
a posteriori information 120
a priori information 120
approximately defined operator 230
approximation rate 98
artificial intelligence 13
axioms of probability theory 60

back propagation method 126
basic problem of probability theory 62
basic problem of statistics 63
Bayesian approach 119
Bayesian inference 34
bound on the distance to the smallest risk 77
bound on the value of achieved risk 77
bounds on generalization ability of a learning machine 76

canonical separating hyperplanes 132
capacity control problem 116
cause-effect relation 9

choosing the best sparse algebraic polynomial 117
choosing the degree of a polynomial 116
classification error 19
codebook 106
complete (Popper's) nonfalsifiability 52
compression coefficient 107
conditional density estimation 228
conditional probability estimation 227
consistency of inference 36
constructive distribution-independent bound on the rate of convergence 69
convolution of inner product 140
criterion of nonfalsifiability 47

data smoothing problem 209
decisionmaking problem 296
decision trees 7
deductive inference 47
density estimation problem:
parametric (Fisher-Wald) setting 19
nonparametric setting 28
discrepancy 18

- discriminant analysis 24
- discriminant function 25
- distribution-dependent bound on the rate of convergence 69
- distribution-independent bound on the rate of convergence 69
- Δ -margin separating hyperplane 132
- empirical distribution function 28
- empirical processes 40
- empirical risk functional 20
- empirical risk minimization inductive principle 20
- ensemble of support vector machines 163
- entropy of the set of functions 42
- entropy on the set of indicator functions 42
- equivalence classes 292
- estimation of the values of a function at the given points 292
- expert systems 7
- ε -insensitivity 181
- ε -insensitive loss function 181
- feature selection problem 119
- function approximation 98
- function estimation model 17
- Gaussian 279
- generalized Glivenko–Cantelli problem 66
- generalized growth function 85
- generator of random vectors 17
- Glivenko–Cantelli problem 66
- growth function 55
- Hamming distance 106
- handwritten digit recognition 147
- hard-threshold vicinity function 103
- hard vicinity function 269
- hidden markov models 7
- hidden units 101
- Huber loss function 183
- ill-posed problems: 9
 - solution by variation method 236
 - solution by residual method 236
- solution by quasi-solution method 236
- independent trials 62
- inductive inference 55
- inner product in Hilbert space 140
- integral equations:
 - solution for exact determined equations 238
 - solution for approximately determined equations 239
- kernel function 27
- Kolmogorov–Smirnov distribution 87
- Kulback–Leibler distance 32
- Kühn–Tucker conditions 134
- Lagrange multiplier 134
- Lagrangian 134
- Laplacian 277
- law of large numbers in functional space 41
- law of large numbers 41
- law of large numbers in vector space 41
- Lie derivatives 279
- learning machine 17
- learning matrices 7
- least-squares method 21
- least-modulo method 182
- linear discriminant function 31
- linearly nonseparable case 135
- local approximation 104
- local risk minimization 103
- locality parameter 103
- loss function:
 - for AdaBoost algorithm 163
 - for density estimation 21
 - for logistic regression 156
 - for pattern recognition 21
 - for regression estimation 21
- Madaline 7
- main principle for small sample size problems 31
- maximal margin hyperplane 131
- maximum likelihood method 24
- McCulloch–Pitts neuron model 2
- measurements with the additive noise 25

- metric ε -entropy 44
- minimum description length
 - principle 104
- mixture of normal densities 26
- National Institute of Standard and Technology (NIST) digit database 173
- neural networks 126
- nontrivially consistent inference 38
- nonparametric density estimation 27
- normal discriminant function 31
- one-sided empirical process 40
- optimal separating hyperplane 131
- overfitting phenomenon 14
- parametric methods of density estimation 24
- partial nonfalsifiability 50
- Parzen's windows method 27
- pattern recognition problem 19
- perceptron 1
- perceptron's stopping rule 6
- polynomial approximation of regression 116
- polynomial machine 143
- potential nonfalsifiability 53
- probability measure 59
- probably approximately correct (PAC) model 13
- problem of demarcation 49
- pseudo-dimension 90
- quadratic programming problem 133
- quantization of parameters 110
- quasi-solution 112
- radial basis function machine 145
- random entropy 42
- random string 10
- randomness concept 10
- regression estimation problem 19
- regression function 19
- regularization theory 9
- regularized functional 9
- reproducing kernel Hilbert space 244
- residual principle 236
- rigorous (distribution-dependent) bounds 85
- risk functional 18
- risk minimization from empirical data problem 20
- robust estimators 26
- robust regression 26
- Rosenblatt's algorithm 5
- set of indicators 73
- set of unbounded functions 77
- σ -algebra 60
- sigmoid function 125
- small sample size 93
- smoothing kernel 102
- smoothness of functions 100
- soft threshold vicinity function 103
- soft vicinity function 270
- soft-margin separating hyperplane 135
- spline function:
 - with a finite number of nodes 194
 - with an infinite number of nodes 195
- stochastic approximation stopping rule 33
- stochastic ill-posed problems 113
- strong mode estimating a probability measure 63
- structural risk minimization principle 94
- structure 94
- structure of growth function 79
- supervisor 17
- support vector machines 138
- support vectors 134
- support vector ANOVA decomposition 199
- SVM_n approximation of the logistic regression 155
- SVM density estimator 247
- SVM conditional probability estimator 255
- SVM conditional density estimator 258
- tails of distribution 77
- tangent distance 150
- training set 18
- transductive inference 293

- Turing–Church thesis 177
two layer neural networks machine 145
two-sided empirical process 46
- U.S. Postal Service digit database 173
uniform one-sided convergence 39
uniform two-sided convergence 39
- VC dimension of a set of indicator functions 79
VC dimension of a set of real functions 81
VC entropy 44
VC subgraph 90
vicinal risk minimization method 267
- vicinity kernel: 273
one-vicinal kernel 273
two-vicinal kernel 273
- VRM method
for pattern recognition 273
for regression estimation 287
for density estimation 284
for conditional probability estimation 285
for conditional density estimation 286
- weak mode estimating a probability measure 63
weight decay procedure 102