

Statistics for Engineering and Information Science

Vladimir N. Vapnik

The Nature of Statistical Learning Theory

Second Edition



Springer

Vladimir N. Vapnik

The Nature of Statistical Learning Theory

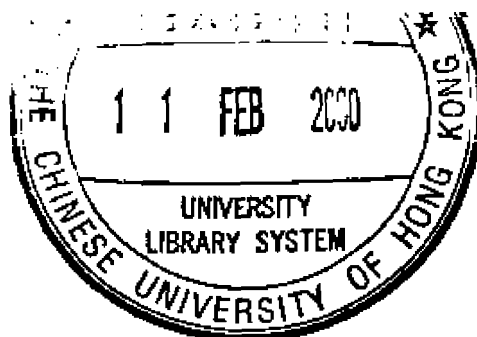
Second Edition

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 0-387-98780-0 (hbk. alk. paper)

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-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), 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.

Printed and bound by Mapte-Vail Book Manufacturing Group, York, PA.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-98780-0 Springer-Verlag New York Berlin Heidelberg 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.

ACKNOWLEDGMENTS

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	
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	A-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 Problem	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
8.3	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 Red-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**.

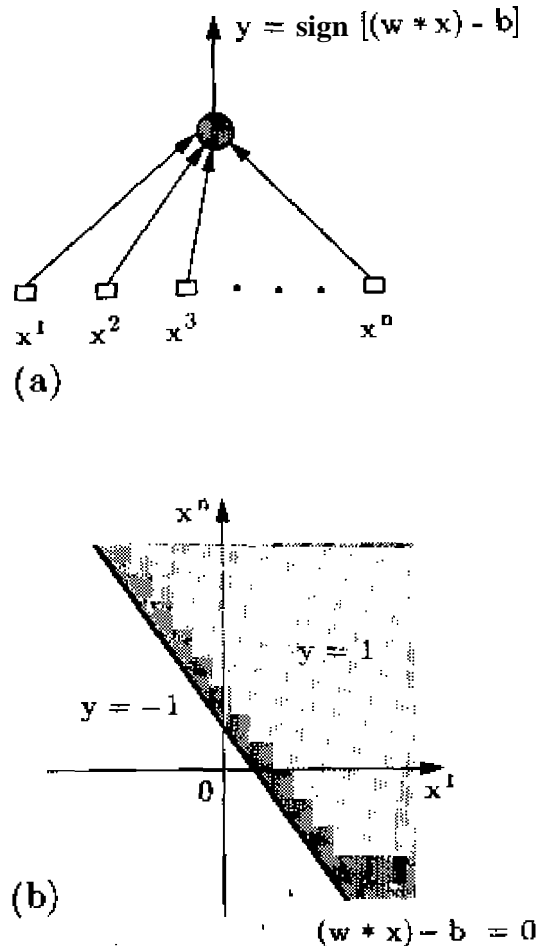


FIGURE 0.1. (a) Model of a neuron. (b) Geometrically, a neuron defines two regions in input space where it takes the values -1 and 1 . These regions are separated by the hyperplane $(w \cdot x) - b = 0$.

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 $x = (x^1, \dots, x^n) \in X \subset \mathbb{R}^n$ 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 m 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**:

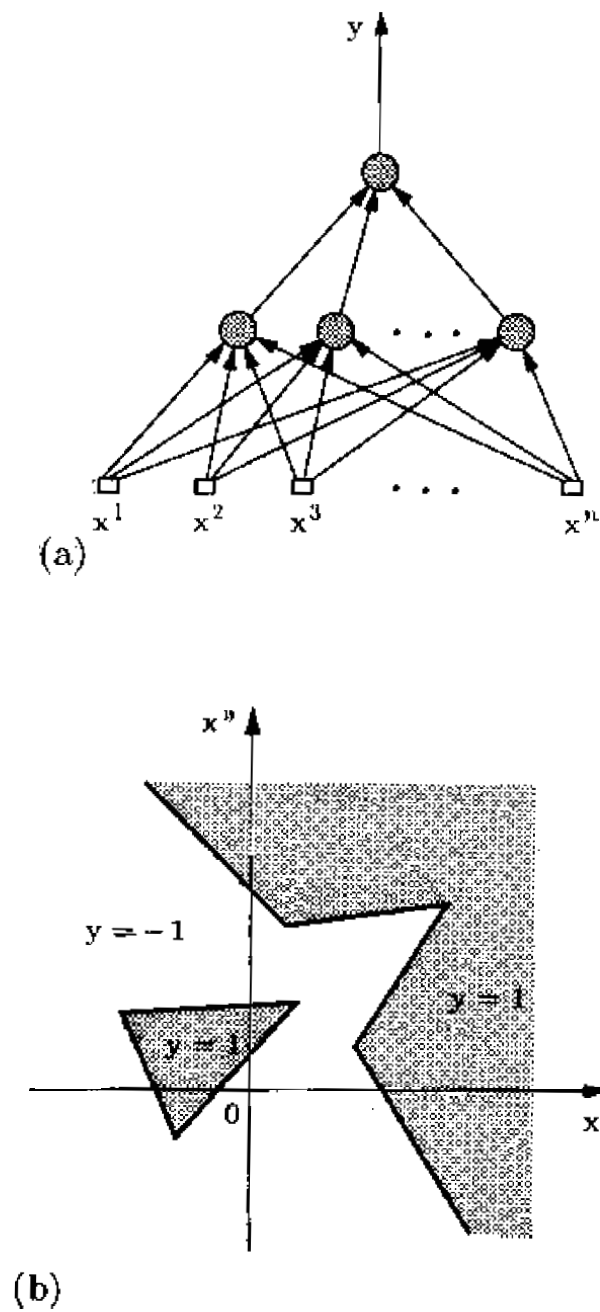


FIGURE 0.2.(a) The perceptron is a composition of several neurons. (b) Geometrically, the perceptron defines two regions in input space where it takes the values -1 and 1 . These regions are separated by a piecewise linear surface.

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

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

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.,

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

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 $\lceil R^2/\rho^2 \rceil$ 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

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

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. Vapnik, A.Ja. Tscherwonienkis, *Theorie der Zeichenerkennung*, 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 $f \in \mathcal{B}$ 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 functional) 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 Štefanyuk, 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 XXXXXXXXXX

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 $\text{sign}\{(w \cdot x) - b\}$ 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**.