

The Nature of Statistical Learning Theory

Springer Science+Business Media, LLC

Vladimir N. Vapnik

The Nature of Statistical Learning Theory

With 33 Illustrations



Springer

Vladimir N. Vapnik
AT&T Bell Laboratories
101 Crawfords Corner Road
Holmdel, NJ 07733 USA

Library of Congress Cataloging-in-Publication Data

Vapnik, Vladimir Naumovich.

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

p. cm.

Includes bibliographical references and index.

ISBN 978-1-4757-2442-4

ISBN 978-1-4757-2440-0 (eBook)

DOI 10.1007/978-1-4757-2440-0

Softcover reprint of the hardcover 1st edition 1995

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

Q325.7.V37 1995

006.3'1'015195—dc20

95-24205

Printed on acid-free paper.

© 1995 Springer Science+Business Media New York

Originally published by Springer-Verlag New York, Inc in 1995.

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 coordinated by Bill Imbornoni; manufacturing supervised by Joseph Quatela.
Photocomposed copy prepared from the author's LaTeX file.

9 8 7 6 5 4 3 2 1

In memory of my mother

Preface

Between 1960 and 1980 a revolution in statistics occurred: Fisher's paradigm introduced in the 1920–1930s 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 the 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 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 are new methods of inference for various statistical problems.

To develop these methods (that 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 the 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 forthcoming book *Statistical Learning Theory*, J. Wiley, 1996). However to describe the ideas without simplifications I needed to introduce new concepts (new mathematical constructions) some of which are non-trivial.

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 process.

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 consider 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 repetitions 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 support of Larry Jackel, the head of 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, 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, 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.

Vladimir N. Vapnik

*AT&T Bell Laboratories,
Holmdel, March 1995*

Contents

Preface	vii
Introduction: Four Periods in the Research of the Learning Problem	1
0.1 Rosenblatt's Perceptron (The 1960s)	1
0.2 Construction of the Fundamentals of Learning Theory (The 1960–1970s)	7
0.3 Neural Networks (The 1980s)	11
0.4 Returning to the Origin (The 1990s)	14
Chapter 1 Setting of the Learning Problem	15
1.1 Function Estimation Model	15
1.2 The Problem of Risk Minimization	16
1.3 Three Main Learning Problems	16
1.3.1 Pattern Recognition	17
1.3.2 Regression Estimation	17
1.3.3 Density Estimation (Fisher–Wald Setting)	17
1.4 The General Setting of the Learning Problem	18
1.5 The Empirical Risk Minimization (ERM) Inductive Principle	18
1.6 The Four Parts of Learning Theory	19
Informal Reasoning and Comments — 1	21
1.7 The Classical Paradigm of Solving Learning Problems	21

1.7.1	Density Estimation Problem (Maximum Likelihood Method)	22
1.7.2	Pattern Recognition (Discriminant Analysis) Problem	22
1.7.3	Regression Estimation Model	23
1.7.4	Narrowness of the ML Method	24
1.8	Nonparametric Methods of Density Estimation	25
1.8.1	Parzen's Windows	25
1.8.2	The Problem of Density Estimation is Ill-Posed	26
1.9	Main Principle for Solving Problems Using a Restricted Amount of Information	28
1.10	Model Minimization of the Risk Based on Empirical Data	29
1.10.1	Pattern Recognition	29
1.10.2	Regression Estimation	29
1.10.3	Density Estimation	30
1.11	Stochastic Approximation Inference	31
Chapter 2	Consistency of Learning Processes	33
2.1	The Classical Definition of Consistency and the Concept of Nontrivial Consistency	34
2.2	The Key Theorem of Learning Theory	36
2.2.1	Remark on the ML Method	37
2.3	Necessary and Sufficient Conditions for Uniform Two-Sided Convergence	38
2.3.1	Remark on Law of Large Numbers and its Generalization	39
2.3.2	Entropy of the Set of Indicator Functions	40
2.3.3	Entropy of the Set of Real Functions	41
2.3.4	Conditions for Uniform Two-Sided Convergence	43
2.4	Necessary and Sufficient Conditions for Uniform One-Sided Convergence	44
2.5	Theory of Nonfalsifiability	45
2.5.1	Kant's Problem of Demarcation and Popper's Theory of Nonfalsifiability	45
2.6	Theorems about Nonfalsifiability	47
2.6.1	Case of Complete (Popper's) Nonfalsifiability	48
2.6.2	Theorem about Partial Nonfalsifiability	48
2.6.3	Theorem about Potential Nonfalsifiability	50
2.7	Three Milestones in Learning Theory	52
Informal Reasoning and Comments — 2		55
2.8	The Basic Problems of Probability Theory and Statistics	56
2.8.1	Axioms of Probability Theory	56
2.9	Two Modes of Estimating a Probability Measure	59
2.10	Strong Mode Estimation of Probability Measures and the Density Estimation Problem	61
2.11	The Glivenko–Cantelli Theorem and its Generalization	62

2.12 Mathematical Theory of Induction	63
Chapter 3 Bounds on the Rate of Convergence of Learning Processes	65
3.1 The Basic Inequalities	66
3.2 Generalization for the Set of Real Functions	68
3.3 The Main Distribution Independent Bounds	71
3.4 Bounds on the Generalization Ability of Learning Machines	72
3.5 The Structure of the Growth Function	75
3.6 The VC Dimension of a Set of Functions	76
3.7 Constructive Distribution-Independent Bounds	79
3.8 The Problem of Constructing Rigorous (Distribution-Dependent) Bounds	81
Informal Reasoning and Comments — 3	83
3.9 Kolmogorov–Smirnov Distributions	83
3.10 Racing for the Constant	85
3.11 Bounds on Empirical Processes	86
Chapter 4 Controlling the Generalization Ability of Learning Processes	89
4.1 Structural Risk Minimization (SRM) Inductive Principle . .	90
4.2 Asymptotic Analysis of the Rate of Convergence	92
4.3 The Problem of Function Approximation in Learning Theory	94
4.4 Examples of Structures for Neural Nets	97
4.5 The Problem of Local Function Estimation	98
4.6 The Minimum Description Length (MDL) and SRM Principles	100
4.6.1 The MDL Principle	102
4.6.2 Bounds for the MDL Principle	103
4.6.3 The SRM and the MDL Principles	104
4.6.4 A Weak Point of the MDL Principle	106
Informal Reasoning and Comments — 4	107
4.7 Methods for Solving Ill-Posed Problems	108
4.8 Stochastic Ill-Posed Problems and the Problem of Density Estimation	109
4.9 The Problem of Polynomial Approximation of the Regression	111
4.10 The Problem of Capacity Control	112
4.10.1 Choosing the Degree of the Polynomial	112
4.10.2 Choosing the Best Sparse Algebraic Polynomial . . .	113
4.10.3 Structures on the Set of Trigonometric Polynomials	114
4.10.4 The Problem of Features Selection	115
4.11 The Problem of Capacity Control and Bayesian Inference .	115
4.11.1 The Bayesian Approach in Learning Theory	115

4.11.2 Discussion of the Bayesian Approach and Capacity Control Methods	117
Chapter 5 Constructing Learning Algorithms	119
5.1 Why can Learning Machines Generalize?	119
5.2 Sigmoid Approximation of Indicator Functions	121
5.3 Neural Networks	122
5.3.1 The Back-Propagation Method	122
5.3.2 The Back-Propagation Algorithm	125
5.3.3 Neural Networks for the Regression Estimation Problem	126
5.3.4 Remarks on the Back-Propagation Method	126
5.4 The Optimal Separating Hyperplane	127
5.4.1 The Optimal Hyperplane	127
5.4.2 The Structure of Canonical Hyperplanes	128
5.5 Constructing the Optimal Hyperplane	129
5.5.1 Generalization for the Nonseparable Case	131
5.6 Support Vector (SV) Machines	133
5.6.1 Generalization in High-Dimensional Space	135
5.6.2 Convolution of the Inner Product	135
5.6.3 Constructing SV Machines	136
5.6.4 Examples of SV Machines	137
5.7 Experiments with SV Machines	141
5.7.1 Example in the Plane	142
5.7.2 Handwritten Digit Recognition	142
5.7.3 Some Important Details	146
5.8 Remarks on SV Machines	149
5.9. SV Machines for the Regression Estimation Problem	151
5.9.1 ε -Insensitive Loss-Function	151
5.9.2 Minimizing the Risk Using Convex Optimization Procedure	152
5.9.3 SV Machine with Convolved Inner Product	155
Informal Reasoning and Comments — 5	157
5.10 The Art of Engineering Versus Formal Inference	157
5.11 Wisdom of Statistical Models	160
5.12 What Can One Learn from Digit Recognition Experiments?	162
5.12.1 Influence of the Type of Structures and Accuracy of Capacity Control	162
5.12.2 SRM Principle and the Problem of Feature Construction	164
5.12.3 Is the Set of Support Vectors a Robust Characteristic of the Data?	165
Conclusion: What is Important in Learning Theory?	167
6.1 What is Important in the Setting of the Problem?	167

6.2	What is Important in the Theory of Consistency of Learning Processes?	170
6.3	What is Important in the Theory of Bounds?	171
6.4	What is Important in the Theory for Controlling the Generalization Ability of Learning Machines?	172
6.5	What is Important in the Theory for Constructing Learning Algorithms?	173
6.6	What is the Most Important?	174
References		177
	Remarks on References	177
	References	178
Index		185