Statistics for Engineering
and Information Science

Series Editors
M. Jordan, S.L. Lauritzen, J.F. Lawless, V. Nair

Springer Science+Business Media, LLC

Statistics for Engineering and Information Science

Akaike and Kitagawa: The Practice of Time Series Analysis.
Cowell, Dawid, Lauritzen, and Spiegelhalter: Probabilistic Networks and Expert Systems.
Fine: Feedforward Neural Network Methodology.
Hawkins and Olwell: Cumulative Sum Charts and Charting for Quality Improvement.
Vapnik: The Nature of Statistical Learning Theory, Second Edition.

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

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

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

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.

Production managed by Frank McGuckin; manufacturing supervised by Erica Bresler.
Photocomposed copy prepared from the author's LaTeX 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 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 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