Information Science and Statistics

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

Information Science and Statistics

Akaike and Kitagawa: The Practice of Time Series Analysis.
Cowell, Dawid, Lauritzen, and Spiegelhalter: Probabilistic Networks and Expert Systems.
Doucet, de Freitas, and Gordon: Sequential Monte Carlo Methods in Practice.
Fine: Feedforward Neural Network Methodology.
Hawkins and Olwell: Cumulative Sum Charts and Charting for Quality Improvement.
Jensen: Bayesian Networks and Decision Graphs.
Marchette: Computer Intrusion Detection and Network Monitoring: A Statistical Viewpoint.
Rubinstein and Kroese: The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation, and Machine Learning.
Studený: Probabilistic Conditional Independence Structures.
Vapnik: The Nature of Statistical Learning Theory, Second Edition.
Wallace: Statistical and Inductive Inference by Minimum Massage Length.

Vladimir Vapnik

Estimation of Dependences Based on Empirical Data

Reprint of 1982 Edition

Empirical Inference Science

Afterword of 2006

Springer

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

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

Series Editors:
Michael Jordan
Division of Computer Science and Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
USA

Jon Kleinberg
Department of Computer Science
Cornell University
Ithaca, NY 14853
USA

Bernhard Schölkopf
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen
Germany

Library of Congress Control Number: 2005938355

ISBN 978-1-4419-2158-1 ISBN 978-0-387-34239-9 (eBook)
DOI 10.1007/978-0-387-34239-9

Printed on acid-free paper.

© 2006 Springer Science+Business Media New York
Originally published by Springer Science+Business Media, Inc. in 2006
Softcover reprint of the hardcover 1st edition 2006

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 in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

Vladimir Vapnik

Estimation of Dependences Based on Empirical Data

Translated by Samuel Kotz

With 22 illustrations

Springer

To my daughter

Preface

Estimating dependences on the basis of empirical data has been, and will probably remain, a central problem in applied analysis. This problem is a mathematical interpretation of one of the basic questions of science: how to extract the existing law-like relationship from scattered data.

The simplest attack on this problem is to construct (estimate) a function from its values at certain points. Here we will formulate some general principles of estimating a functional dependence, and then develop an algorithm for the estimation using these principles.

Usually, when one seeks a general principle, intended for a solution of a wide class of problems, one focuses first upon the simplest, most basic problem. This simple version of the problem is treated theoretically with great thoroughness and the scheme obtained for a solution is then extended to all the problems of the class under consideration.

When studying the estimation of functional dependences, the functions which take only one value (i.e., constants) are usually chosen as the simplest problem. One assumes that the measurements of a constant are subject to errors. Given several such measurements, one must determine this constant. There are various ways to state this problem specifically. These are based on different models of measurements with errors. However, regardless of the model, the study of the basic problem leads to the following classical principle of estimating functional dependence based on empirical data:

> Select, from an admissible set of functions, the one which yields the best approximation of the totality of the available empirical data.

This principle is sufficiently general. It leaves the measure of the quality of the relation between the function and the empirical data undefined. Various definitions of this measure are available; for example, the amount of the mean square deviation of the functional values, the amount of the mean deviation, the maximal deviation, etc. Each definition generates its own method of estimating dependences, such as the least-squares method, the least absolute values method, etc. However, in all cases the principle of the solution (i.e., the search for a function which best approximates the data) remains unchanged.

The main content of this book deals with a study of a different, nonclassical principle of estimating dependences:

> Select, from an admissible set of functions, a function which fulfills a definite relationship between a quantity characterizing the quality of the approximation and a quantity characterizing the "complexity" of the approximating function.

This principle may need some clarification. With increasing complexity of the approximating function, one obtains successively better approximations to the available data, and may even be able to construct a function which will pass through all of the given points. This new principle, unlike the classical one, asserts that we should not strive to get close to empirical data at all costs; that is, we should not excessively complicate the approximating function. For any given amount of data, there exists a specific relationship between the complexity of the approximating function and the quality of the approximation thus obtained. By preserving this relationship, the estimated dependence most accurately characterizes the actual (unknown) dependence. Further improvements of the approximation by increasing the complexity may result in the estimated function approximating the given data better, but representing the actual function less accurately. This nonclassical principle of estimation reflects an attempt to take into account that dependence is estimated with a limited amount of data.

The idea that, with a limited amount of data, the selected function should not merely approximate empirical data but also possess some extremal properties has existed for a long time. It first received theoretical justification in the investigation of the problems of pattern recognition. The mathematical statement of pattern recognition necessarily leads to estimating a function which admits not one (as is the case in our basic problem) but two values. This additional complexity is unexpectedly of fundamental importance. The set of functions taking on two values is much more "varied" than the set of constants (i.e., functions taking on one value).

The important point is that the structure of the set of constant functions is "simple and homogeneous", while that of the set of functions taking on two values is rich and admits ordering according to its complexity. The latter is essential for estimating dependences with limited amounts of empirical data.

Thus the study of pattern recognition problems has shown that the simplest classical problem does not encompass all the problems of estimating dependences, since the class of functions associated with estimating a constant is so limited that no problem of its stratification arises.

The simplest problem of this book is the problem of pattern recognition. We use methods based on classical ideas of statistical analysis as well as those associated with the nonclassical principle of estimation for its solution. All of these methods are adopted for two other problems of estimation: regression estimation and interpretation of the results of indirect experiments.

For our new basic problem, we distinguish between two formulations: estimating functions and estimating values of a function at given points. (These two formulations coincide in the case of estimation of constants.) We distinguish between these formulations since, with a limited amount of data, there may not be enough information to estimate a function satisfactorily as a whole, but at the same time it may be enough to estimate k numbers—the values of a function at given points.

Thus this book is devoted to problems of estimating dependences with limited amounts of data. The basic idea is as follows: the attempt to take into account the fact that the amount of empirical data is limited leads us to the nonclassical principle of estimating dependences. Utilizing this principle allows us to solve delicate problems of estimation. These include determination of optimal set of features in the case of pattern recognition, determination of the structure of the approximating function in the case of regression estimation, and construction of regularizing functions for solving ill-posed problems of interpretation of indirect experiments (i.e., problems which arise due to the limited amount of data and which cannot be solved within the framework of classical setups).

The book contains ten chapters. Chapters 1 and 2 are introductory. In these, various problems of estimating dependences are considered from the common positions of minimizing the expected risk based on the empirical data and various possible approaches to minimizing risks are discussed.

Chapters 3, 4, and 5 are devoted to the study ot classical ideas of risk minimization: estimating probability density functions by means of parametric methods and utilization of this density for minimization of the risk. Chapter 3 applies these ideas to pattern recognition problems. Chapters 4 and 5 apply them to regression estimation problems. Beginning with Chapter 6 nonclassical methods of minimization of risk are studied. Chapters 6 and 7 establish the conditions for applying the method of minimization of empirical risk to solutions of problems of minimization of the expected risk for samples of limited size, while Chapters 8–10 utilize these conditions to construct a method of risk minimization based on limited data: the so-called method of structural minimization. (In Chapter 8, we consider the application of the method of structural risk minimization to the problems of pattern recognition and regression. In Chapter 9, we give an application to the solutions to ill-posed problems of interpreting results of indirect experiments. In Chapter 10, we investigate the problem of estimating values of functions at given points based on structural minimization). Finally, Addenda I and II are devoted to algorithms for structural risk minimization.

This book is intended for a wide class of readers: students in upper-level courses, graduate students, engineers, and scientists. The exposition is such that the proofs do not interfere with the basic flow of the arguments. However, all of the main assertions are proved *in toto*.

We try to avoid generalizations which are possibly important but less indicative of the basic ideas developed in this book. Therefore, in the main part of the book we consider only simple cases (such as quadratic loss functions, equally spaced observations, independent errors, etc.). As a rule, the corresponding generalizations may be achieved using standard methods. The most important of these generalizations concerning arbitrary loss functions are given at the end of the respective chapters.

The main part of the book does not require a knowledge of special branches of mathematics. However, in order to follow the proofs the reader should possess some experience in dealing with mathematical concepts.

The book is not a survey of the standard theory, and it may be biased to some extent. Nevertheless, it is our hope that the reader will find it interesting and useful.

Moscow, 1982 V. VAPNIK

Contents