

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 Ohwell: 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.

Studeny: Probabilistic Conditional Independence Structures.

Vapnik: The Nature of Statistical Learning Theory, Second Edition.

Wallace: Statistical and Inductive Inference by Minimum Message Length.

Vladimir Vapnik

Estimation of Dependences Based on Empirical Data

Reprint of 1982 Edition

Empirical Inference Science

Afterword of 2006

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

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 of 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

1	The Problem of Estimating Dependences from Empirical Data	1
1	The Problem of Minimizing the Expected Risk on the Basis of Empirical Data	1
2	The Problem of Pattern Recognition	4
3	The Regression Estimation Problem	5
4	The Problem of Interpreting Results of Indirect Experiments	8
5	Ill-posed Problems	10
6	Accuracy and Confidence of Risk Minimization Based on Empirical Data	13
7	The Accuracy of Estimating Dependences on the Basis of Empirical Data	15
8	Special Features of Problems of Estimating Dependences	18
	Appendix to Chapter 1. Methods for Solving Ill-posed Problems	20
A1	The Problem of Solving an Operator Equation	20
A2	Problems Well Posed in Tihonov's Sense	22
A3	The Regularization Method	23
2	Methods of Expected-Risk Minimization	27
1	Two Approaches to Expected-Risk Minimization	27
2	The Problem of Large Deviations	29
3	Prior Information in Problems of Estimating Dependences on the Basis of Empirical Data	32
4	Two Procedures for Minimizing the Expected Risk	34
5	The Problem of Estimating the Probability Density	36

6	Uniform Proximity between Empirical Means and Mathematical Expectations	39
7	A Generalization of the Glivenko–Cantelli Theorem and the Problem of Pattern Recognition	41
8	Remarks on Two Procedures for Minimizing Expected Risk on the Basis of Empirical Data	42
3	Methods of Parametric Statistics for the Pattern Recognition Problem	45
1	The Pattern Recognition Problem	45
2	Discriminant Analysis	46
3	Decision Rules in Problems of Pattern Recognition	49
4	Evaluation of Qualities of Algorithms for Density Estimation	51
5	The Bayesian Algorithm for Density Estimation	52
6	Bayesian Estimators of Discrete Probability Distributions	54
7	Bayesian Estimators for the Gaussian (Normal) Density	56
8	Unbiased Estimators	63
9	Sufficient Statistics	64
10	Computing the Best Unbiased Estimator	66
11	The Problem of Estimating the Parameters of a Density	70
12	The Maximum-Likelihood Method	73
13	Estimation of Parameters of the Probability Density Using the Maximum-Likelihood Method	76
14	Remarks on Various Methods for Density Estimation	78
4	Methods of Parametric Statistics for the Problem of Regression Estimation	81
1	The Scheme for Interpreting the Results of Direct Experiments	81
2	A Remark on the Statement of the Problem of Interpreting the Results of Direct Experiments	83
3	Density Models	84
4	Extremal Properties of Gaussian and Laplace Distributions	87
5	On Robust Methods of Estimating Location Parameters	91
6	Robust Estimation of Regression Parameters	96
7	Robustness of Gaussian and Laplace Distributions	99
8	Classes of Densities Formed by a Mixture of Densities	101
9	Densities Concentrated on an Interval	103
10	Robust Methods for Regression Estimation	105
5	Estimation of Regression Parameters	109
1	The Problem of Estimating Regression Parameters	109
2	The Theory of Normal Regression	111
3	Methods of Estimating the Normal Regression that are Uniformly Superior to the Least-Squares Method	115

4	A Theorem on Estimating the Mean Vector of a Multivariate Normal Distribution	120
5	The Gauss–Markov Theorem	125
6	Best Linear Estimators	127
7	Criteria for the Quality of Estimators	128
8	Evaluation of the Best Linear Estimators	130
9	Utilizing Prior Information	134
6	A Method of Minimizing Empirical Risk for the Problem of Pattern Recognition	139
1	A Method of Minimizing Empirical Risk	139
2	Uniform Convergence of Frequencies of Events to Their Probabilities	141
3	A Particular Case	142
4	A Deterministic Statement of the Problem	144
5	Upper Bounds on Error Probabilities	146
6	An ε -net of a Set	149
7	Necessary and Sufficient Conditions for Uniform Convergence of Frequencies to Probabilities	152
8	Properties of Growth Functions	154
9	Bounds on Deviations of Empirically Optimal Decision Rules	155
10	Remarks on the Bound on the Rate of Uniform Convergence of Frequencies to Probabilities	158
11	Remark on the General Theory of Uniform Estimating of Probabilities	159
	Appendix to Chapter 6. Theory of Uniform Convergence of Frequencies to Probabilities: Sufficient Conditions	162
A1	Sufficient Conditions for Uniform Convergence of Frequencies to Probabilities	162
A2	The Growth Function	163
A3	The Basic Lemma	168
A4	Derivation of Sufficient Conditions	170
A5	A Bound on the Quantity Γ	173
A6	A Bound on the Probability of Uniform Relative Deviation	176
7	A Method of Minimizing Empirical Risk for the Problem of Regression Estimation	181
1	Uniform Convergence of Means to Mathematical Expectations	181
2	A Particular Case	183
3	A Generalization to a Class with Infinitely Many Members	186
4	The Capacity of a Set of Arbitrary Functions	188
5	Uniform Boundedness of a Ratio of Moments	191
6	Two Theorems on Uniform Convergence	192
7	Theorem on Uniform Relative Deviation	195
8	Remarks on a General Theory of Risk Estimation	202

Appendix to Chapter 7. Theory of Uniform Convergence of Means to Their Mathematical Expectations: Necessary and Sufficient Conditions	206
A1 ε -entropy	206
A2 The Quasicube	211
A3 ε -extension of a Set	214
A4 An Auxiliary Lemma	216
A5 Necessary and Sufficient Conditions for Uniform Convergence: The Proof of Necessity	220
A6 Necessary and Sufficient Conditions for Uniform Convergence: The Proof of Sufficiency	223
A7 Corollaries	228
 8 The Method of Structural Minimization of Risk	 232
1 The Idea of the Method of Structural Risk Minimization	232
2 Moving-Control Estimators	236
3 Moving-Control Estimators in Problems of Regression Estimation	238
4 Estimating the Expected Risk for Samples of Arbitrary Size	241
5 Estimation of Indicator Functions in a Class of Linear Decision Rules	252
6 Estimation of Regression in a Class of Polynomials	254
7 Estimation of Regression in a Class of Functions Linear in Their Parameters: Moving Control Method	259
8 Estimation of Regression in a Class of Functions Linear in Their Parameters: Uniform Estimating Method	261
9 Selection of Sample	263
10 Remarks on a General Theory of Risk Minimization	265
 9 Solution of Ill-posed Problems. Interpretation of Measurements Using the Method of Structural Risk Minimization	 267
1 Ill-posed Problems of Interpreting Results of Indirect Experiments	267
2 Definitions of Convergence	268
3 Theorems on Interpreting Results of Indirect Experiments	271
4 Proofs of the Theorems	275
5 Methods of Polynomial and Piecewise Polynomial Approximations	285
6 Methods for Solving Ill-posed Measurement Problems	288
7 The Problem of Probability Density Estimation	292
8 Estimation of Smooth Densities	294
9 Density Estimation Using Parzen's Method	301
10 Density Estimation Using the Method of Structural Risk Minimization	303
 Appendix to Chapter 9. Statistical Theory of Regularization	 308

10	Estimation of Functional Values at Given Points	312
1	The Scheme of Minimizing the Overall Risk	312
2	The Method of Structural Minimization of the Overall Risk	315
3	Bounds on the Uniform Relative Deviation of Frequencies in Two Subsamples	316
4	A Bound on the Uniform Relative Deviation of Means in Two Subsamples	318
5	Estimation of Values of an Indicator Function in a Class of Linear Decision Rules	321
6	Selection of a Sample for Estimating Values of an Indicator Function	327
7	Estimation of Values of an Arbitrary Function in the Class of Functions Linear in Their Parameters	330
8	Selection of a Sample for Estimation of Values of an Arbitrary Function	332
9	Estimation of Values of an Indicator Function in the Class of Piecewise Linear Decision Rules	334
10	Estimation of Values of an Arbitrary Function in a Class of Piecewise Linear Functions	335
11	Local Algorithms for Estimating Values of Indicator Functions	336
12	Local Algorithms for Estimating Values of an Arbitrary Function	339
13	The Problem of Finding the Best Point of a Given Set	340
14	Remarks on Estimating Values of a Function	345
	Appendix to Chapter 10. Taxonomy Problems	347
A1	A Problem of Classification of Objects	347
A2	Algorithms of Taxonomy	349
	Postscript	351
	Addendum I. Algorithms for Pattern Recognition	353
1	Remarks about Algorithms	353
2	Construction of Subdividing Hyperplanes	355
3	Algorithms for Maximizing Quadratic Forms	359
4	Methods for Constructing an Optimal Separating Hyperplane	362
5	An Algorithm for External Subdivision of Values of a Feature into Gradations	364
6	An Algorithm for Constructing Separating Hyperplanes	366
	Addendum II. Algorithms for Estimating Nonindicator Functions	370
1	Remarks Concerning Algorithms	370
2	An Algorithm for Regression Estimation in a Class of Polynomials	371
3	Canonical Splines	373

4 Algorithms for Estimating Functions in a Class of Splines	379
5 Algorithms for Solving Ill-posed Problems of Interpreting Measurements	380
6 Algorithms for Estimating Multidimensional Regression in a Class of Linear Functions	381
Bibliographical Remarks	384
Chapter 1	384
Chapter 2	385
Chapter 3	386
Chapter 4	386
Chapter 5	387
Chapter 6	388
Chapter 7	388
Chapter 8	389
Chapter 9	389
Chapter 10	390
Addenda I and II	390
Bibliography	391
Index	397