LEARNING FROM DATA

Concepts, Theory, and Methods

SECOND EDITION

VLADIMIR CHERKASSKY • FILIP MULIER

LEARNING FROM DATA

Concepts, Theory, and Methods

Second Edition

VLADIMIR CHERKASSKY
FILIP MULIER

IEEE PRESS

BICENTENNIAL 1807 WILEY 2007 BICENTENNIAL

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:

Cherkassky, Vladimir S.
Learning from data : concepts, theory, and methods / by Vladimir Cherkassky, Filip Mulier. – 2nd ed.
p. cm.
ISBN 978-0-471-68182-3 (cloth)
1. Adaptive signal processing. 2. Machine learning. 3. Neural networks (Computer science) 4. Fuzzy systems. I. Mulier, Filip. II. Title.
TK5102.9.C475 2007
006.3′1–dc22 2006038736

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

LEARNING FROM DATA

BICENTENNIAL 1807 WILEY 2007 BICENTENNIAL

The Wiley Bicentennial–Knowledge for Generations

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce
President and Chief Executive Officer

Peter Booth Wiley
Chairman of the Board

PREFACE

> There are two problems in modern science:
> - too many people use different terminology to solve the same problems;
> - even more people use the same terminology to address completely different issues.
>
> Anonymous

In recent years, there has been an explosive growth of methods for learning (or estimating dependencies) from data. This is not surprising given the proliferation of

- low-cost computers (for implementing such methods in software)
- low-cost sensors and database technology (for collecting and storing data)
- highly computer-literate application experts (who can pose "interesting" application problems)

A learning method is an algorithm (usually implemented in software) that estimates an unknown mapping (dependency) between a system's inputs and outputs from the available data, namely from known (input, output) samples. Once such a dependency has been accurately estimated, it can be used for prediction of future system outputs from the known input values. This book provides a unified description of principles and methods for learning dependencies from data.

Methods for estimating dependencies from data have been traditionally explored in diverse fields such as statistics (multivariate regression and classification), engineering (pattern recognition), and computer science (artificial intelligence, machine learning, and, more recently, data mining). Recent interest in learning from data has resulted in the development of biologically motivated methodologies, such as artificial neural networks, fuzzy systems, and wavelets.

Unfortunately, developments in each field are seldom related to other fields, despite the apparent commonality of issues and methods. The mere fact that hundreds of "new" methods are being proposed each year at various conferences and in numerous journals suggests a certain lack of understanding of the basic issues common to all such methods.

The premise of this book is that there are just a handful of important principles and issues in the field of learning dependencies from data. Any researcher or practitioner in this field needs to be aware of these issues in order to successfully apply a particular methodology, understand a method's limitations, or develop new techniques.

This book is an attempt to present and discuss such issues and principles (common to all methods) and then describe representative popular methods originating from statistics, neural networks, and pattern recognition. Often methods developed in different fields can be related to a common conceptual framework. This approach enables better understanding of a method's properties, and it has methodological advantages over traditional "cookbook" descriptions of various learning algorithms.

Many aspects of learning methods can be addressed under a traditional statistical framework. At the same time, many popular learning algorithms and learning methodologies have been developed outside classical statistics. This happened for several reasons:

1. Traditionally, the statistician's role has been to analyze the inferential limitations of the structural model constructed (proposed) by the application-domain expert. Consequently, the conceptual approach (adopted in statistics) is parameter estimation for model identification. For many real-life problems that require flexible estimation with finite samples, the statistical approach is fundamentally flawed. As shown in this book, learning with finite samples should be based on the framework known as risk minimization, rather than density estimation.
2. Statisticians have been late to recognize and appreciate the importance of computer-intensive approaches to data analysis. The growing use of computers has fundamentally changed the traditional boundaries between a statistician (data modeler) and a user (application expert). Nowadays, engineers and computer scientists successfully use sophisticated empirical data-modeling techniques (i.e., neural nets) to estimate complex nonlinear dependencies from the data.
3. Statistics (being part of mathematics) has developed into a closed discipline, with its own scientific jargon and academic objectives that favor analytic proofs rather than practical methods for learning from data.

Historically, we can identify three stages in the development of predictive learning methods. First, in 1985–1992 classical statistics gave way to neural networks (and other empirical methods, such as fuzzy systems) due to an early enthusiasm and naive claims that biologically inspired methods (i.e., neural nets) can achieve model-free learning not subject to statistical limitations. Even though such claims later proved to be false, this stage had a positive impact by showing the power and usefulness of flexible nonlinear modeling based on the risk minimization approach. Then in 1992–1996 came the return of statistics as the researchers and practitioners of neural networks became aware of their statistical limitations, initiating a trend toward interpretation of learning methods using a classical statistical framework. Finally, the third stage, from 1997 to present, is dominated by the wide popularity of support vector machines (SVMs) and similar margin-based approaches (such as boosting), and the growing interest in the Vapnik–Chervonenkis (VC) theoretical framework for predictive learning.

This book is intended for readers with varying interests, including researchers/practitioners in data modeling with a classical statistics background, researchers/practitioners in data modeling with a neural network background, and graduate students in engineering or computer science.

The presentation does not assume a special math background beyond a good working knowledge of probability, linear algebra, and calculus on an undergraduate level. Useful background material on optimization and linear algebra is included in Appendixes A and B, respectively. We do not provide mathematical proofs, but, whenever possible, in place of proofs we provide intuitive explanations and arguments. Likewise, mathematical formulation and discussion of the major concepts and results are provided as needed. The goal is to provide a unified treatment of diverse methodologies (i.e., statistics and neural networks), and to that end we carefully define the terminology used throughout the book. This book is not easy reading because it describes fairly complex concepts and mathematical models for solving inherently difficult (ill-posed) problems of learning with finite data. To aid the reader, each chapter starts with a brief overview of its contents. Also, each chapter is concluded with a summary containing an overview of open research issues and pointers to other (relevant) chapters.

Book chapters are conceptually organized into three parts:

- *Part I: Concepts and Theory* (Chapters 1–4). Following an introduction and motivation given in Chapter 1, we present formal specification of the inductive learning problem in Chapter 2 that also introduces major concepts and issues in learning from data. In particular, it describes an important concept called an *inductive principle*. Chapter 3 describes the regularization (or penalization) framework adopted in statistics. Chapter 4 describes Vapnik's statistical learning theory (SLT), which provides the theoretical basis for predictive learning with finite data. SLT, aka VC theory, is important for understanding various learning methods developed in neural networks, statistics, and pattern recognition, and for developing new approaches, such as SVMs (described in Chapter 9) and noninductive learning settings (described in Chapter 10).

- *Part II: Constructive Learning Methods* (Chapters 5–8). This part describes learning methods for regression, classification, and density approximation problems. The objective is to show conceptual similarity of methods originating from statistics, neural networks, and signal processing and to discuss their relative advantages and limitations. Whenever possible, we relate constructive learning methods to the conceptual framework of Part I. Chapter 5 describes nonlinear optimization strategies commonly used in various methods. Chapter 6 describes methods for density approximation, which include statistical, neural network, and signal processing techniques for data reduction and dimensionality reduction. Chapter 7 provides descriptions of statistical and neural network methods for regression. Chapter 8 describes methods for classification.
- *Part III: VC-Based Learning Methodologies* (Chapters 9 and 10). Here we describe constructive learning approaches that originate in VC theory. These include SVMs (or margin-based methods) for several inductive learning problems (in Chapter 9) and various noninductive learning formulations (described in Chapter 10).

The chapters should be followed in a sequential order, as the description of constructive learning methods is related to the conceptual framework developed in the first part of the book. A shortened sequence of Chapters 1–3 followed by Chapters 5, 6, 7 and 8 is recommended for the beginning readers who are interested only in the description of statistical and neural network methods. This sequence omits the mathematically and conceptually challenging Chapters 4 and 9. Alternatively, more advanced readers who are primarily interested in SLT and SVM methodology may adopt the sequence of Chapters 2, 3, 4, 9, and 10.

In the course of writing this book, our understanding of the field has changed. We started with the currently prevailing view of learning methods as a collection of tricks. Statisticians have their own bag of tricks (and terminology), neural networks have a different set of tricks, and so on. However, in the process of writing this book, we realized that it is possible to understand the various heuristic methods (tricks) by a sound general conceptual framework. Such a framework is provided by SLT developed mainly by Vapnik over the past 35 years. This theory combines fundamental concepts and principles related to learning with finite data, well-defined problem formulations, and rigorous mathematical theory. Although SLT is well known for its *mathematical* aspects, its *conceptual* contributions are not fully appreciated. As shown in our book, the conceptual framework provided by SLT can be used for improved understanding of various learning methods even where its mathematical results cannot be directly applied. Modern learning methods (i.e., flexible approaches using finite data) have slowly drifted away from the original problem statements posed in classical statistical decision and estimation theory. A major conceptual contribution of SLT is in revisiting the problem statement appropriate for modern data mining applications. On the very basic level, SLT makes a clear distinction between the problem formulation and a solution approach (aka inductive principle) used to solve a problem. Although this distinction appears trivial on the surface, it leads to a fundamentally new understanding of the learning problem not explained by classical theory. Although it is tempting to skip directly to constructive solutions, this book devotes enough attention to the learning problem formulation and important concepts *before* describing actual learning methods.

Over the past 10 years (since the first edition of this book), we have witnessed considerable growth of interest in SVM-related methods. Nowadays, SVM (aka kernel) methods are commonly used in data mining, statistics, signal processing, pattern recognition, genomics, and so on. In spite of such an overwhelming success and wide recognition of SVM methodology, many important VC theoretical concepts responsible for good generalization of SVMs (such as margin, VC dimension) remain rather poorly understood. For example, many recent monographs and research papers refer to SVMs as a "special case of regularization." So in this second edition, we made a special effort to emphasize the conceptual aspects of VC theory and to contrast the VC theoretical approach to learning (i.e., *system imitation*) versus the classical statistical and function approximation approach (i.e., *system identification*). Accurate interpretation of VC theoretical concepts is important for improved understanding of inductive learning algorithms, as well as for developing emerging state-of-the-art approaches based on noninductive learning settings (as discussed in Chapter 10). In this edition, we emphasize the philosophical interpretation of predictive learning, in general, and of several VC theoretical concepts, in particular. These philosophical connections appear to be quite useful for understanding recent advanced learning methods and for motivating new noninductive types of inference. Moreover, philosophical aspects of predictive learning can be immediately related to epistemology (understanding of human knowledge), as discussed in Chapter 11.

Many people have contributed directly and indirectly to this book. First and foremost, we are greatly indebted to Vladimir Vapnik of NEC Labs for his fundamental contributions to SLT and for his patience in explaining this theory to us. We would like to acknowledge many people whose constructive feedback helped improve the quality of the second edition, including Ella Bingham, John Boik, Olivier Chapelle, David Hand, Nicol Schraudolph, Simon Haykin, David Musicant, Erinija Pranckeviciene, and D. Solomatine—all of whom provided many useful comments.

This book was used in the graduate course "Predictive Learning from Data" at the University of Minnesota over the past 10 years, and we would like to thank students who took this course for their valuable feedback. In particular, we acknowledge former graduate students X. Shao, Y. Ma, T. Xiong, L. Liang, H Gao, M. Ramani, R. Singh, and Y. Kim whose research contributions are incorporated in this book in the form of several fine figures and empirical

comparisons. Finally, we would like to thank our families for their patience and support.

Vladimir Cherkassky
Filip Mulier

Minneapolis, Minnesota
March 2007

NOTATION

The following uniform notation is used throughout the book. Scalars are indicated by script letters such as a. Vectors are indicated by lowercase bold letters such as $\mathbf{w}$. Matrices are given using uppercase bold letters $\mathbf{V}$. When elements of a matrix are accessed individually, we use the corresponding lowercase script letter. For example, the (i, j) element of the matrix $\mathbf{V}$ is v_{ij}. Common notation for all chapters is as follows:

Data

n	Number of samples
d	Number of input variables
$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$	Matrix of input samples
$\mathbf{y} = [y_1, \ldots, y_n]$	Vector of output samples
$\mathbf{Z} = [\mathbf{X}, \mathbf{y}]$	Combined input–output training data or
$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$	Representation of data points in a feature space

Distribution

P	Probability
$F(\mathbf{x})$	Cumulative probability distribution function (cdf)
$p(\mathbf{x})$	Probability density function (pdf)
$p(\mathbf{x}, y)$	Joint probability density function
$p(\mathbf{x}; \omega)$	Probability density function, which is parameterized
$p(y\|\mathbf{x})$	Conditional density
$t(\mathbf{x})$	Target function

Approximating Functions

$f(\mathbf{x}, \omega)$	A class of approximating functions indexed by abstract parameter ω (ω can be a scalar, vector, or matrix). Interpretation of $f(\mathbf{x}, \omega)$ depends on the particular learning problem

$f(\mathbf{x}, \omega_0)$	The function that minimizes the expected risk (optimal solution)
$f(\mathbf{x}, \omega^*)$	Estimate of the optimal solution obtained from finite data
$f(\mathbf{x}, \mathbf{w}, \mathbf{V}) = \sum_{i=1}^{m} w_i g_i(\mathbf{x}, \mathbf{v}_i) + b$	Basis function expansion of approximating functions with bias term
$g_i(\mathbf{x}, \mathbf{v})$	Basis function in a basis function expansion
$w, \mathbf{w}, \mathbf{W}$	Parameters of approximating function
$v, \mathbf{v}, \mathbf{V}$	Basis function parameters
m	Number of basis functions
Ω	Set of parameters, as in $\mathbf{w} \in \Omega$
Δ	Margin distance
$t(\mathbf{x})$	Target function
ξ	Error between the target function and the approximating function, or error between model estimate and time output

Risk Functionals

$L(y, f(\mathbf{x}, \omega))$	Discrepancy measure or loss function
L_2	Squared discrepancy measure
$Q(\omega)$	A set of loss functions
R	Risk or average loss
$R(\omega)$	Expected risk as a function of parameters
$R_{\text{emp}}(\omega)$	Empirical risk as a function of parameters

Kernel Functions

$K(\mathbf{x}, \mathbf{x}')$	General kernel function (for kernel smothing)
$S(\mathbf{x}, \mathbf{x}')$	Equivalent kernel of a linear estimator
$H(\mathbf{x}, \mathbf{x}')$	Inner product kernel

Miscellaneous

$(\mathbf{a} \cdot \mathbf{b})$	Inner (dot) product of two vectors
$I()$	Indicator function of a Boolean argument that takes the value 1 if its argument is true and 0 otherwise. By convention, for a real-valued argument, $I(x) = 1$ for $x > 0$, and $I(x) = 0$ for $x \leq 0$
$\phi[f(\mathbf{x}, \omega)]$	Penalty functional
λ	Regularization parameter
h	VC dimension
γ_k	Learning rate for stochastic approximation at iteration step k
$[a]_+$	Positive argument, equals max $(a, 0)$
L	Lagrangian

In addition to the above notation used throughout the book, there is chapter-specific notation, which will be introduced locally in each chapter.

CONTENTS