

Luc Devroye László Györfi
Gábor Lugosi

A Probabilistic Theory of Pattern Recognition

With 99 Figures



Springer

Luc Devroye
School of Computer Science
McGill University
Montreal, Quebec, H3A 2A7
Canada

László Györfi
Gábor Lugosi
Department of Mathematics and
Computer Science
Technical University of Budapest
Budapest
Hungary

Managing Editors

I. Karatzas
Department of Statistics
Columbia University
New York, NY 10027, USA

M. Yor
CNRS, Laboratoire de Probabilités
Université Pierre et Marie Curie
4, Place Jussieu, Tour 56
F-75252 Paris Cedex 05, France

Mathematics Subject Classification (1991): 68T10, 68T05, 62G07, 62H30

Library of Congress Cataloging-in-Publication Data
Devroye, Luc.

A probabilistic theory of pattern recognition/Luc Devroye,
László Györfi, Gábor Lugosi.

p. cm.

Includes bibliographical references and index.

I. Pattern perception. 2. Probabilities. I. Györfi, László.
II. Lugosi, Gábor. III. Title.
Q327.D5 1996
003'.52'015192—dc20 95-44633

Printed on acid-free paper.

© 1996 by Springer Science+Business Media New York
Originally published by Springer-Verlag New York, Inc. in 1996
Softcover reprint of the hardcover 1st edition 1996

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 here-after 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 managed by Francine McNeill; manufacturing supervised by Jeffrey Taub.
Photocomposed copy prepared using Springer's svsing.sty macro.

9 8 7 6 5 4 3

SPIN 10830936

ISBN 978-1-4612-6877-2

ISBN 978-1-4612-0711-5 (eBook)

DOI 10.1007/978-1-4612-0711-5

Contents

Preface	v
1 Introduction	1
2 The Bayes Error	9
2.1 The Bayes Problem	9
2.2 A Simple Example	11
2.3 Another Simple Example	12
2.4 Other Formulas for the Bayes Risk	14
2.5 Plug-In Decisions	15
2.6 Bayes Error Versus Dimension	17
Problems and Exercises	18
3 Inequalities and Alternate Distance Measures	21
3.1 Measuring Discriminatory Information	21
3.2 The Kolmogorov Variational Distance	22
3.3 The Nearest Neighbor Error	22
3.4 The Bhattacharyya Affinity	23
3.5 Entropy	25
3.6 Jeffreys' Divergence	27
3.7 F -Errors	28
3.8 The Mahalanobis Distance	30
3.9 f -Divergences	31
Problems and Exercises	35

4	Linear Discrimination	39
4.1	Univariate Discrimination and Stoller Splits	40
4.2	Linear Discriminants	44
4.3	The Fisher Linear Discriminant	46
4.4	The Normal Distribution	47
4.5	Empirical Risk Minimization	49
4.6	Minimizing Other Criteria	54
	Problems and Exercises	56
5	Nearest Neighbor Rules	61
5.1	Introduction	61
5.2	Notation and Simple Asymptotics	63
5.3	Proof of Stone's Lemma	66
5.4	The Asymptotic Probability of Error	69
5.5	The Asymptotic Error Probability of Weighted Nearest Neighbor Rules	71
5.6	k -Nearest Neighbor Rules: Even k	74
5.7	Inequalities for the Probability of Error	75
5.8	Behavior When L^* Is Small	78
5.9	Nearest Neighbor Rules When $L^* = 0$	80
5.10	Admissibility of the Nearest Neighbor Rule	81
5.11	The (k, l) -Nearest Neighbor Rule	81
	Problems and Exercises	83
6	Consistency	91
6.1	Universal Consistency	91
6.2	Classification and Regression Estimation	92
6.3	Partitioning Rules	94
6.4	The Histogram Rule	95
6.5	Stone's Theorem	97
6.6	The k -Nearest Neighbor Rule	100
6.7	Classification Is Easier Than Regression Function Estimation	101
6.8	Smart Rules	106
	Problems and Exercises	107
7	Slow Rates of Convergence	111
7.1	Finite Training Sequence	111
7.2	Slow Rates	113
	Problems and Exercises	118
8	Error Estimation	121
8.1	Error Counting	121
8.2	Hoeffding's Inequality	122
8.3	Error Estimation Without Testing Data	124
8.4	Selecting Classifiers	125

8.5	Estimating the Bayes Error	128
	Problems and Exercises	129
9	The Regular Histogram Rule	133
9.1	The Method of Bounded Differences	133
9.2	Strong Universal Consistency	138
	Problems and Exercises	142
10	Kernel Rules	147
10.1	Consistency	149
10.2	Proof of the Consistency Theorem	153
10.3	Potential Function Rules	159
	Problems and Exercises	161
11	Consistency of the k-Nearest Neighbor Rule	169
11.1	Strong Consistency	170
11.2	Breaking Distance Ties	174
11.3	Recursive Methods	176
11.4	Scale-Invariant Rules	177
11.5	Weighted Nearest Neighbor Rules	178
11.6	Rotation-Invariant Rules	179
11.7	Relabeling Rules	180
	Problems and Exercises	182
12	Vapnik-Chervonenkis Theory	187
12.1	Empirical Error Minimization	187
12.2	Fingering	191
12.3	The Glivenko-Cantelli Theorem	192
12.4	Uniform Deviations of Relative Frequencies from Probabilities	196
12.5	Classifier Selection	199
12.6	Sample Complexity	201
12.7	The Zero-Error Case	202
12.8	Extensions	206
	Problems and Exercises	208
13	Combinatorial Aspects of Vapnik-Chervonenkis Theory	215
13.1	Shatter Coefficients and VC Dimension	215
13.2	Shatter Coefficients of Some Classes	219
13.3	Linear and Generalized Linear Discrimination Rules	224
13.4	Convex Sets and Monotone Layers	226
	Problems and Exercises	229
14	Lower Bounds for Empirical Classifier Selection	233
14.1	Minimax Lower Bounds	234
14.2	The Case $L_C = 0$	234
14.3	Classes with Infinite VC Dimension	238

14.4	The Case $L_C > 0$	239
14.5	Sample Complexity	245
	Problems and Exercises	247
15	The Maximum Likelihood Principle	249
15.1	Maximum Likelihood: The Formats	249
15.2	The Maximum Likelihood Method: Regression Format	250
15.3	Consistency	253
15.4	Examples	256
15.5	Classical Maximum Likelihood: Distribution Format	260
	Problems and Exercises	261
16	Parametric Classification	263
16.1	Example: Exponential Families	266
16.2	Standard Plug-In Rules	267
16.3	Minimum Distance Estimates	270
16.4	Empirical Error Minimization	275
	Problems and Exercises	276
17	Generalized Linear Discrimination	279
17.1	Fourier Series Classification	280
17.2	Generalized Linear Classification	285
	Problems and Exercises	287
18	Complexity Regularization	289
18.1	Structural Risk Minimization	290
18.2	Poor Approximation Properties of VC Classes	297
18.3	Simple Empirical Covering	297
	Problems and Exercises	300
19	Condensed and Edited Nearest Neighbor Rules	303
19.1	Condensed Nearest Neighbor Rules	303
19.2	Edited Nearest Neighbor Rules	309
19.3	Sieves and Prototypes	309
	Problems and Exercises	312
20	Tree Classifiers	315
20.1	Invariance	318
20.2	Trees with the X -Property	319
20.3	Balanced Search Trees	322
20.4	Binary Search Trees	326
20.5	The Chronological k -d Tree	328
20.6	The Deep k -d Tree	332
20.7	Quadtrees	333
20.8	Best Possible Perpendicular Splits	334
20.9	Splitting Criteria Based on Impurity Functions	336

20.10	A Consistent Splitting Criterion	340
20.11	BSP Trees	341
20.12	Primitive Selection	343
20.13	Constructing Consistent Tree Classifiers	346
20.14	A Greedy Classifier	348
	Problems and Exercises	357
21	Data-Dependent Partitioning	363
21.1	Introduction	363
21.2	A Vapnik-Chervonenkis Inequality for Partitions	364
21.3	Consistency	368
21.4	Statistically Equivalent Blocks	372
21.5	Partitioning Rules Based on Clustering	377
21.6	Data-Based Scaling	381
21.7	Classification Trees	383
	Problems and Exercises	383
22	Splitting the Data	387
22.1	The Holdout Estimate	387
22.2	Consistency and Asymptotic Optimality	389
22.3	Nearest Neighbor Rules with Automatic Scaling	391
22.4	Classification Based on Clustering	392
22.5	Statistically Equivalent Blocks	393
22.6	Binary Tree Classifiers	394
	Problems and Exercises	395
23	The Resubstitution Estimate	397
23.1	The Resubstitution Estimate	397
23.2	Histogram Rules	399
23.3	Data-Based Histograms and Rule Selection	403
	Problems and Exercises	405
24	Deleted Estimates of the Error Probability	407
24.1	A General Lower Bound	408
24.2	A General Upper Bound for Deleted Estimates	411
24.3	Nearest Neighbor Rules	413
24.4	Kernel Rules	415
24.5	Histogram Rules	417
	Problems and Exercises	419
25	Automatic Kernel Rules	423
25.1	Consistency	424
25.2	Data Splitting	428
25.3	Kernel Complexity	431
25.4	Multiparameter Kernel Rules	435

25.5	Kernels of Infinite Complexity	436
25.6	On Minimizing the Apparent Error Rate	439
25.7	Minimizing the Deleted Estimate	441
25.8	Sieve Methods	444
25.9	Squared Error Minimization	445
	Problems and Exercises	446
26	Automatic Nearest Neighbor Rules	451
26.1	Consistency	451
26.2	Data Splitting	452
26.3	Data Splitting for Weighted NN Rules	453
26.4	Reference Data and Data Splitting	454
26.5	Variable Metric NN Rules	455
26.6	Selection of k Based on the Deleted Estimate	457
	Problems and Exercises	458
27	Hypercubes and Discrete Spaces	461
27.1	Multinomial Discrimination	461
27.2	Quantization	464
27.3	Independent Components	466
27.4	Boolean Classifiers	468
27.5	Series Methods for the Hypercube	470
27.6	Maximum Likelihood	472
27.7	Kernel Methods	474
	Problems and Exercises	474
28	Epsilon Entropy and Totally Bounded Sets	479
28.1	Definitions	479
28.2	Examples: Totally Bounded Classes	480
28.3	Skeleton Estimates	482
28.4	Rate of Convergence	485
	Problems and Exercises	486
29	Uniform Laws of Large Numbers	489
29.1	Minimizing the Empirical Squared Error	489
29.2	Uniform Deviations of Averages from Expectations	490
29.3	Empirical Squared Error Minimization	493
29.4	Proof of Theorem 29.1	494
29.5	Covering Numbers and Shatter Coefficients	496
29.6	Generalized Linear Classification	501
	Problems and Exercises	505
30	Neural Networks	507
30.1	Multilayer Perceptrons	507
30.2	Arrangements	511

30.3	Approximation by Neural Networks	517
30.4	VC Dimension	521
30.5	L_1 Error Minimization	526
30.6	The Adaline and Padaline	531
30.7	Polynomial Networks	532
30.8	Kolmogorov-Lorentz Networks and Additive Models	534
30.9	Projection Pursuit	538
30.10	Radial Basis Function Networks	540
	Problems and Exercises	542
31	Other Error Estimates	549
31.1	Smoothing the Error Count	549
31.2	Posterior Probability Estimates	554
31.3	Rotation Estimate	556
31.4	Bootstrap	556
	Problems and Exercises	559
32	Feature Extraction	561
32.1	Dimensionality Reduction	561
32.2	Transformations with Small Distortion	567
32.3	Admissible and Sufficient Transformations	569
	Problems and Exercises	572
	Appendix	575
A.1	Basics of Measure Theory	575
A.2	The Lebesgue Integral	576
A.3	Denseness Results	579
A.4	Probability	581
A.5	Inequalities	582
A.6	Convergence of Random Variables	584
A.7	Conditional Expectation	585
A.8	The Binomial Distribution	586
A.9	The Hypergeometric Distribution	589
A.10	The Multinomial Distribution	589
A.11	The Exponential and Gamma Distributions	590
A.12	The Multivariate Normal Distribution	590
	Notation	591
	References	593
	Author Index	619
	Subject Index	627