

Fundamentals of Pattern Recognition and Machine Learning

Ulisses Braga-Neto

Fundamentals of Pattern Recognition and Machine Learning

 Springer

Ulisses Braga-Neto
Department of Electrical
and Computer Engineering
Texas A&M University
College Station, TX, USA

ISBN 978-3-030-27655-3 ISBN 978-3-030-27656-0 (eBook)
<https://doi.org/10.1007/978-3-030-27656-0>

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Flávia

Preface

“Only the educated are free.”

–Epictetus.

The field of pattern recognition and machine learning has a long and distinguished history. In particular, there are many excellent textbooks on the topic, so the question of why a new textbook is desirable must be confronted. The goal of this book is to be a concise introduction, which combines theory and practice and is suitable to the classroom. It includes updates on recent methods and examples of applications based on the python programming language. The book does not attempt an encyclopedic treatment of pattern recognition and machine learning, which has become impossible in any case, due to how much the field has grown. A stringent selection of material is mandatory for a concise textbook, and the choice of topics made here, while dictated to a certain extent by my own experience and preferences, is believed to equip the reader with the core knowledge one must obtain to be proficient in this field. Calculus and probability at the undergraduate level are the minimum prerequisites for the book. The appendices contain short reviews of probability at the graduate level and other mathematical tools that are needed in the text.

This book has grown out of lecture notes for graduate classes on pattern recognition, bioinformatics, and materials informatics that I have taught for over a decade at Texas A&M University. The book is intended, with the proper selection of topics (as detailed below), for a one or two-semester introductory course in pattern recognition or machine learning at the graduate or advanced undergraduate level. Although the book is designed for the classroom, it can also be used effectively for self-study.

The book does not shy away from theory, since an appreciation of it is important for an education in pattern recognition and machine learning. The field is replete with classical theorems, such as the Cover-Hart Theorem, Stone’s Theorem and its corollaries, the Vapnik-Chervonenkis Theorem, and several others, which are covered in this book. Nevertheless, an effort is made in the book to strike a balance between theory and practice. In particular, examples with datasets from applications

in Bioinformatics and Materials Informatics are used throughout the book to illustrate the theory. These datasets are also used in end-of-chapter coding assignments based on python. All plots in the text were generated using python scripts, which can be downloaded from the book website. The reader is encouraged to experiment with these scripts and use them in the coding assignments. The book website also contains datasets from Bioinformatics and Materials Informatics applications, which are used in the plots and coding assignments. It has been my experience in the classroom that the understanding of the subject by students is increased significantly once they engage in assignments involving coding and data from real-world applications.

The book is organized as follows. Chapter 1 is a general introduction to motivate the topic. Chapters 2–8 concern classification. Chapters 2 and 3 on optimal and general sample-based classification are the foundational chapters on classification. Chapters 4-6 examine the three main categories of classification rules: parametric, nonparametric, and function-approximation, while Chapters 7 and 8 concern error estimation and model selection for classification. Chapter 9 on dimensionality reduction still deals with classification, but also includes material on unsupervised methods. Finally, Chapters 10 and 11 deal with clustering and regression. There is flexibility for the instructor or reader to pick topics from these chapters and use them in a different order. In particular, the “Additional Topics” sections at the end of most chapters cover miscellaneous topics, and can be included or not, without affecting continuity. In addition, for the convenience of instructors and readers, sections that contain material of a more technical nature are marked with a star. These sections could be skipped at a first reading.

The Exercises section at the end of most chapters contain problems of varying difficulty; some of them are straightforward applications of the concepts discussed in the chapter, while others introduce new concepts and extensions of the theory, some of which may be worth discussing in class. Python Assignment sections at the end of most chapters ask the reader to use python and scikit-learn to implement methods discussed in the chapter and apply them to synthetic and real data sets from Bioinformatics and Materials Informatics applications.

Based on the my experience teaching the material, I suggest that the book could be used in the classroom as follows:

1. A one-semester course focusing on classification, covering Chapters 2-9, while including the majority of the starred and additional topics sections.
2. An applications-oriented one-semester course, skipping most or all starred and additional topics sections in Chapters 2-8, covering Chapters 9-11, and emphasizing the coding assignments.
3. A two-semester sequence covering the entire book, including most or all the starred and additional topics sections.

This book is indebted to several of its predecessors. First, the classical text by Duda and Hart (1973, updated with Stork in 2001), which has been a standard reference in the area for many decades. In addition, the book by Devroye, Györfi and Lugosi (1996), which remains the gold standard in nonparametric pattern recognition. Other sources that were influential to this text are the books by McLachlan (1992), Bishop (2006), Webb (2002), and James et al. (2013).

I would like to thank all my current and past collaborators, who have helped shape my understanding of this field. Likewise, I thank all my students, both those whose research I have supervised and those who have attended my lectures, who have contributed ideas and corrections to the text. I would like to thank Ed Dougherty, Louise Strong, John Goutsias, Ascendino Dias e Silva, Roberto Lotufo, Junior Barrera, and Severino Toscano, from whom I have learned much. I thank Ed Dougherty, Don Geman, Al Hero, and Gábor Lugosi for the comments and encouragement received while writing this book. I am grateful to Caio Davi, who drew several of the figures. I appreciate very much the expert assistance provided by Paul Drougas at Springer, during difficult times in New York City. Finally, I would like to thank my wife Flávia and my children Maria Clara and Ulisses, for their patience and support during the writing of this book.

Ulisses Braga-Neto
College Station, TX
July 2020

Contents

Preface	vii
1 Introduction	1
1.1 Pattern Recognition and Machine Learning	1
1.2 Basic Mathematical Setting	2
1.3 Prediction	2
1.4 Prediction Error	4
1.5 Supervised vs. Unsupervised Learning	4
1.6 Complexity Trade-Offs	5
1.7 The Design Cycle	7
1.8 Application Examples	7
1.8.1 Bioinformatics	8
1.8.2 Materials Informatics	11
1.9 Bibliographical Notes	13
2 Optimal Classification	15
2.1 Classification without Features	15
2.2 Classification with Features	16

2.3	The Bayes Classifier	20
2.4	The Bayes Error	24
2.5	Gaussian Model	28
2.5.1	Homoskedastic Case	29
2.5.2	Heteroskedastic Case	31
2.6	Additional Topics	32
2.6.1	Minimax Classification	32
2.6.2	F-errors	34
2.6.3	Bayes Decision Theory	37
*2.6.4	Rigorous Formulation of the Classification Problem	38
2.7	Bibliographical Notes	40
2.8	Exercises	41
2.9	Python Assignments	47
3	Sample-Based Classification	51
3.1	Classification Rules	51
3.2	Classification Error Rates	54
*3.3	Consistency	55
3.4	No-Free-Lunch Theorems	59
3.5	Additional Topics	60
3.5.1	Ensemble Classification	60
3.5.2	Mixture Sampling vs. Separate Sampling	61
3.6	Bibliographical Notes	63
3.7	Exercises	63
3.8	Python Assignments	65

4 Parametric Classification	67
4.1 Parametric Plug-in Rules	67
4.2 Gaussian Discriminant Analysis	69
4.2.1 Linear Discriminant Analysis	70
4.2.2 Quadratic Discriminant Analysis	73
4.3 Logistic Classification	75
4.4 Additional Topics	77
4.4.1 Regularized Discriminant Analysis	77
*4.4.2 Consistency of Parametric Rules	79
4.4.3 Bayesian Parametric Rules	81
4.5 Bibliographical Notes	83
4.6 Exercises	84
4.7 Python Assignments	87
5 Nonparametric Classification	89
5.1 Nonparametric Plug-in Rules	89
5.2 Histogram Classification	91
5.3 Nearest-Neighbor Classification	93
5.4 Kernel Classification	95
5.5 Cover-Hart Theorem	98
*5.6 Stone's Theorem	101
5.7 Bibliographical Notes	103
5.8 Exercises	104
5.9 Python Assignments	105

6	Function-Approximation Classification	109
6.1	Support Vector Machines	109
6.1.1	Linear SVMs for Separable Data	111
6.1.2	General Linear SVMs	113
6.1.3	Nonlinear SVMs	115
6.2	Neural Networks	120
6.2.1	Backpropagation Training	126
6.2.2	Convolutional Neural Networks	129
*6.2.3	Universal Approximation Property of Neural Networks	133
*6.2.4	Universal Consistency Theorems	135
6.3	Decision Trees	136
6.4	Rank-Based Classifiers	141
6.5	Bibliographical Notes	142
6.6	Exercises	143
6.7	Python Assignments	146
7	Error Estimation for Classification	151
7.1	Error Estimation Rules	151
7.2	Error Estimation Performance	153
7.2.1	Deviation Distribution	153
7.2.2	Bias, Variance, RMS, and Tail Probabilities	153
*7.2.3	Consistency	155
7.3	Test-Set Error Estimation	157
7.4	Resubstitution	159
7.5	Cross-Validation	160

7.6	Bootstrap	163
7.7	Bolstered Error Estimation	165
7.8	Additional Topics	171
7.8.1	Convex Error Estimators	171
7.8.2	Smoothed Error Estimators	173
7.8.3	Bayesian Error Estimation	174
7.9	Bibliographical Notes	177
7.10	Exercises	179
7.11	Python Assignments	182
8	Model Selection for Classification	185
8.1	Classification Complexity	186
8.2	Vapnik-Chervonenkis Theory	189
*8.2.1	Finite Model Selection	189
8.2.2	Shatter Coefficients and VC Dimension	191
8.2.3	VC Parameters of a Few Classification Rules	192
8.2.4	Vapnik-Chervonenkis Theorem	196
8.2.5	No-Free-Lunch Theorems	197
8.3	Model Selection Methods	198
8.3.1	Validation Error Minimization	198
8.3.2	Training Error Minimization	199
8.3.3	Structural Risk Minimization	200
8.4	Bibliographical Notes	201
8.5	Exercises	202

9 Dimensionality Reduction	205
9.1 Feature Extraction for Classification	206
9.2 Feature Selection	207
9.2.1 Exhaustive Search	208
9.2.2 Univariate Greedy Search	209
9.2.3 Multivariate Greedy Search	211
9.2.4 Feature Selection and Classification Complexity	213
9.2.5 Feature Selection and Error Estimation	214
9.3 Principal Component Analysis (PCA)	216
9.4 Multidimensional Scaling (MDS)	220
9.5 Factor Analysis	222
9.6 Bibliographical Notes	224
9.7 Exercises	226
9.8 Python Assignments	228
10 Clustering	231
10.1 K-Means Algorithm	231
10.2 Gaussian Mixture Modeling	236
10.2.1 Expectation-Maximization Approach	237
10.2.2 Relationship to K -Means	243
10.3 Hierarchical Clustering	243
10.4 Self-Organizing Maps (SOM)	246
10.5 Bibliographical Notes	248
10.6 Exercises	249
10.7 Python Assignments	251

11 Regression	253
11.1 Optimal Regression	254
11.2 Sample-Based Regression	257
11.3 Parametric Regression	258
11.3.1 Linear Regression	260
11.3.2 Gauss-Markov Theorem	262
11.3.3 Penalized Least Squares	265
11.4 Nonparametric Regression	266
11.4.1 Kernel Regression	267
11.4.2 Gaussian Process Regression	267
11.5 Function-Approximation Regression	275
11.6 Error Estimation	277
11.7 Variable Selection	278
11.7.1 Wrapper Search	278
11.7.2 Statistical Testing	279
11.7.3 LASSO and Elastic Net	279
11.8 Model Selection	279
11.9 Bibliographical Notes	281
11.10 Exercises	282
11.11 Python Assignments	284
Appendix	287
A1 Probability Theory	287
A1.1 Sample Space and Events	287
A1.2 Probability Measure	289

A1.3	Conditional Probability and Independence	292
A1.4	Random Variables	293
A1.5	Joint and Conditional Distributions	298
A1.6	Expectation	299
A1.7	Vector Random Variables	305
A1.8	Convergence of Random Sequences	308
A1.9	Asymptotic Theorems	312
A2	Basic Matrix Theory	313
A3	Basic Lagrange-Multiplier Optimization	315
A4	Proof of the Cover-Hart Theorem	319
A5	Proof of Stone's Theorem	321
A6	Proof of the Vapnik-Chervonenkis Theorem	323
A7	Proof of Convergence of the EM Algorithm	327
A8	Data Sets Used in the Book	329
A8.1	Synthetic Data	329
A8.2	Dengue Fever Prognosis Data Set	330
A8.3	Breast Cancer Prognosis Data Set	330
A8.4	Stacking Fault Energy Data Set	331
A8.5	Soft Magnetic Alloy Data Set	331
A8.6	Ultrahigh Carbon Steel Data Set	331
	List of Symbols	333
	Bibliography	335
	Index	351