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

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