MACHINE LEARNING

The Art and Science of Algorithms that Make Sense of Data

As one of the most comprehensive machine learning texts around, this book does justice to the field's incredible richness, but without losing sight of the unifying principles.

Peter Flach's clear, example-based approach begins by discussing how a spam filter works, which gives an immediate introduction to machine learning in action, with a minimum of technical fuss. He covers a wide range of logical, geometric and statistical models, and state-of-the-art topics such as matrix factorisation and ROC analysis. Particular attention is paid to the central role played by features.

Machine Learning will set a new standard as an introductory textbook:

- The Prologue and Chapter 1 are freely available on-line, providing an accessible first step into machine learning.
- The use of established terminology is balanced with the introduction of new and useful concepts.
- Well-chosen examples and illustrations form an integral part of the text.
- Boxes summarise relevant background material and provide pointers for revision.
- Each chapter concludes with a summary and suggestions for further reading.
- A list of 'Important points to remember' is included at the back of the book together with an extensive index to help readers navigate through the material.

MACHINE LEARNING

The Art and Science of Algorithms that Make Sense of Data

PETER FLACH

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107096394

First published 2012
12th printing 2017

Printed in Singapore by Markono Print Media Pte Ltd

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-09639-4 Hardback
ISBN 978-1-107-42222-3 Paperback

Additional resources for this publication at www.cs.bris.ac.uk/home/flach/mlbook

To Hessel Flach (1923–2006)

Brief Contents

Contents

Preface

This book started life in the Summer of 2008, when my employer, the University of Bristol, awarded me a one-year research fellowship. I decided to embark on writing a general introduction to machine learning, for two reasons. One was that there was scope for such a book, to complement the many more specialist texts that are available; the other was that through writing I would learn new things – after all, the best way to learn is to teach.

The challenge facing anyone attempting to write an introductory machine learning text is to do justice to the incredible richness of the machine learning field without losing sight of its unifying principles. Put too much emphasis on the diversity of the discipline and you risk ending up with a 'cookbook' without much coherence; stress your favourite paradigm too much and you may leave out too much of the other interesting stuff. Partly through a process of trial and error, I arrived at the approach embodied in the book, which is is to emphasise both unity and diversity: unity by separate treatment of *tasks* and *features*, both of which are common across any machine learning approach but are often taken for granted; and diversity through coverage of a wide range of logical, geometric and probabilistic models.

Clearly, one cannot hope to cover all of machine learning to any reasonable depth within the confines of 400 pages. In the Epilogue I list some important areas for further study which I decided not to include. In my view, machine learning is a marriage of statistics and knowledge representation, and the subject matter of the book was chosen to reinforce that view. Thus, ample space has been reserved for tree and rule learning, before moving on to the more statistically-oriented material. Throughout the book I have placed particular emphasis on intuitions, hopefully amplified by a generous use

of examples and graphical illustrations, many of which derive from my work on the use of ROC analysis in machine learning.

How to read the book

The printed book is a linear medium and the material has therefore been organised in such a way that it can be read from cover to cover. However, this is not to say that one couldn't pick and mix, as I have tried to organise things in a modular fashion.

For example, someone who wants to read about his or her first learning algorithm as soon as possible could start with Section 2.1, which explains binary classification, and then fast-forward to Chapter 5 and read about learning decision trees without serious continuity problems. After reading Section 5.1 that same person could skip to the first two sections of Chapter 6 to learn about rule-based classifiers.

Alternatively, someone who is interested in linear models could proceed to Section 3.2 on regression tasks after Section 2.1, and then skip to Chapter 7 which starts with linear regression. There is a certain logic in the order of Chapters 4–9 on logical, geometric and probabilistic models, but they can mostly be read independently; similar for the material in Chapters 10–12 on features, model ensembles and machine learning experiments.

I should also mention that the Prologue and Chapter 1 are introductory and reasonably self-contained: the Prologue does contain some technical detail but should be understandable even at pre-University level, while Chapter 1 gives a condensed, high-level overview of most of the material covered in the book. Both chapters are freely available for download from the book's web site at `www.cs.bris.ac.uk/~flach/mlbook`; over time, other material will be added, such as lecture slides. As a book of this scope will inevitably contain small errors, the web site also has a form for letting me know of any errors you spotted and a list of errata.

Acknowledgements

Writing a single-authored book is always going to be a solitary business, but I have been fortunate to receive help and encouragement from many colleagues and friends. Tim Kovacs in Bristol, Luc De Raedt in Leuven and Carla Brodley in Boston organised reading groups which produced very useful feedback. I also received helpful comments from Hendrik Blockeel, Nathalie Japkowicz, Nicolas Lachiche, Martijn van Otterlo, Fabrizio Riguzzi and Mohak Shah. Many other people have provided input in one way or another: thank you.

José Hernández-Orallo went well beyond the call of duty by carefully reading my manuscript and providing an extensive critique with many excellent suggestions for improvement, which I have incorporated so far as time allowed. José: I will buy you a free lunch one day.

Many thanks to my Bristol colleagues and collaborators Tarek Abudawood, Rafal Bogacz, Tilo Burghardt, Nello Cristianini, Tijl De Bie, Bruno Golénia, Simon Price, Oliver Ray and Sebastian Spiegler for joint work and enlightening discussions. Many thanks also to my international collaborators Johannes Fürnkranz, Cèsar Ferri, Thomas Gärtner, José Hernández-Orallo, Nicolas Lachiche, John Lloyd, Edson Matsubara and Ronaldo Prati, as some of our joint work has found its way into the book, or otherwise inspired bits of it. At times when the project needed a push forward my disappearance to a quiet place was kindly facilitated by Kerry, Paul and David, Renée, and Trijntje.

David Tranah from Cambridge University Press was instrumental in getting the process off the ground, and suggested the pointillistic metaphor for 'making sense of data' that gave rise to the cover design (which, according to David, is 'just a canonical silhouette' not depicting anyone in particular – in case you were wondering...). Mairi Sutherland provided careful copy-editing.

I dedicate this book to my late father, who would certainly have opened a bottle of champagne on learning that 'the book' was finally finished. His version of the problem of induction was thought-provoking if somewhat morbid: the same hand that feeds the chicken every day eventually wrings its neck (with apologies to my vegetarian readers). I am grateful to both my parents for providing me with everything I needed to find my own way in life.

Finally, more gratitude than words can convey is due to my wife Lisa. I started writing this book soon after we got married – little did we both know that it would take me nearly four years to finish it. Hindsight is a wonderful thing: for example, it allows one to establish beyond reasonable doubt that trying to finish a book while organising an international conference and overseeing a major house refurbishment is really not a good idea. It is testament to Lisa's support, encouragement and quiet suffering that all three things are nevertheless now coming to full fruition. Dank je wel, meisje!

Peter Flach, Bristol