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.
To Hessel Flach (1923–2006)
Brief Contents

Preface
Prologue: A machine learning sampler
1  The ingredients of machine learning
2  Binary classification and related tasks
3  Beyond binary classification
4  Concept learning
5  Tree models
6  Rule models
7  Linear models
8  Distance-based models
9  Probabilistic models
10 Features
11 Model ensembles
12 Machine learning experiments
Epilogue: Where to go from here
Important points to remember
References
Index
## Contents

<table>
<thead>
<tr>
<th>Preface</th>
<th>xv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prologue: A machine learning sampler</td>
<td>1</td>
</tr>
</tbody>
</table>

### 1 The ingredients of machine learning 13

1.1 Tasks: the problems that can be solved with machine learning 14
   - Looking for structure 16
   - Evaluating performance on a task 18

1.2 Models: the output of machine learning 20
   - Geometric models 21
   - Probabilistic models 25
   - Logical models 32
   - Grouping and grading 36

1.3 Features: the workhorses of machine learning 38
   - Two uses of features 40
   - Feature construction and transformation 41
   - Interaction between features 44

1.4 Summary and outlook 46
   - What you’ll find in the rest of the book 48

### 2 Binary classification and related tasks 49

2.1 Classification 52
2.2 Scoring and ranking 61
  Assessing and visualising ranking performance 63
  Turning rankers into classifiers 69

2.3 Class probability estimation 72
  Assessing class probability estimates 73
  Turning rankers into class probability estimators 76

2.4 Binary classification and related tasks: Summary and further reading 79

3 Beyond binary classification 81
  3.1 Handling more than two classes 81
  Multi-class classification 82
  Multi-class scores and probabilities 86
  3.2 Regression 91
  3.3 Unsupervised and descriptive learning 95
    Predictive and descriptive clustering 96
    Other descriptive models 100
  3.4 Beyond binary classification: Summary and further reading 102

4 Concept learning 104
  4.1 The hypothesis space 106
    Least general generalisation 108
    Internal disjunction 110
  4.2 Paths through the hypothesis space 112
    Most general consistent hypotheses 116
    Closed concepts 116
  4.3 Beyond conjunctive concepts 119
    Using first-order logic 122
  4.4 Learnability 124
  4.5 Concept learning: Summary and further reading 127

5 Tree models 129
  5.1 Decision trees 133
  5.2 Ranking and probability estimation trees 138
    Sensitivity to skewed class distributions 143
  5.3 Tree learning as variance reduction 148
    Regression trees 148
Contents

Clustering trees ................................................. 152
5.4 Tree models: Summary and further reading ................. 155

6 Rule models .................................................. 157
6.1 Learning ordered rule lists ................................ 158
Rule lists for ranking and probability estimation .............. 164
6.2 Learning unordered rule sets ................................ 167
Rule sets for ranking and probability estimation .............. 173
A closer look at rule overlap ..................................... 174
6.3 Descriptive rule learning ................................... 176
Rule learning for subgroup discovery .......................... 178
Association rule mining ........................................... 182
6.4 First-order rule learning .................................... 189
6.5 Rule models: Summary and further reading ............... 192

7 Linear models .................................................. 194
7.1 The least-squares method .................................... 196
Multivariate linear regression ................................... 201
Regularised regression .......................................... 204
Using least-squares regression for classification .............. 205
7.2 The perceptron ................................................ 207
7.3 Support vector machines ..................................... 211
Soft margin SVM .................................................. 216
7.4 Obtaining probabilities from linear classifiers ............. 219
7.5 Going beyond linearity with kernel methods ............... 224
7.6 Linear models: Summary and further reading ............. 228

8 Distance-based models ........................................ 231
8.1 So many roads... ............................................. 231
8.2 Neighbours and exemplars ................................... 237
8.3 Nearest-neighbour classification ............................ 242
8.4 Distance-based clustering .................................... 245
\emph{K}-means algorithm .......................................... 247
Clustering around medoids ....................................... 250
Silhouettes ......................................................... 252
8.5 Hierarchical clustering ....................................... 253
8.6 From kernels to distances .................................... 258
8.7 Distance-based models: Summary and further reading .... 260
## Contents

### 9 Probabilistic models

- 9.1 The normal distribution and its geometric interpretations .................................................. 266
- 9.2 Probabilistic models for categorical data .............................................................................. 273
- Using a naive Bayes model for classification ............................................................................. 275
- Training a naive Bayes model .................................................................................................... 279
- 9.3 Discriminative learning by optimising conditional likelihood .............................................. 282
- 9.4 Probabilistic models with hidden variables .......................................................................... 286
- Expectation-Maximisation .......................................................................................................... 288
- Gaussian mixture models ........................................................................................................... 289
- 9.5 Compression-based models .................................................................................................. 292
- 9.6 Probabilistic models: Summary and further reading ............................................................. 295

### 10 Features

- 10.1 Kinds of feature ..................................................................................................................... 299
- Calculations on features .............................................................................................................. 299
- Categorical, ordinal and quantitative features ........................................................................... 304
- Structured features ...................................................................................................................... 305
- 10.2 Feature transformations ....................................................................................................... 307
- Thresholding and discretisation .................................................................................................. 308
- Normalisation and calibration ..................................................................................................... 314
- Incomplete features .................................................................................................................... 321
- 10.3 Feature construction and selection ....................................................................................... 322
- Matrix transformations and decompositions ............................................................................... 324
- 10.4 Features: Summary and further reading ............................................................................. 327

### 11 Model ensembles

- 11.1 Bagging and random forests ............................................................................................... 331
- 11.2 Boosting ............................................................................................................................... 334
- Boosted rule learning .................................................................................................................. 337
- 11.3 Mapping the ensemble landscape ....................................................................................... 338
- Bias, variance and margins ......................................................................................................... 338
- Other ensemble methods ............................................................................................................ 339
- Meta-learning .............................................................................................................................. 340
- 11.4 Model ensembles: Summary and further reading ............................................................... 341

### 12 Machine learning experiments

- 12.1 What to measure ................................................................................................................... 344
- 12.2 How to measure it ................................................................................................................ 348
Contents

12.3 How to interpret it ........................................... 351
Interpretation of results over multiple data sets ............. 354
12.4 Machine learning experiments: Summary and further reading .... 357

Epilogue: Where to go from here .......................... 360

Important points to remember ............................... 363

References ..................................................... 367

Index .......................................................... 383
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 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
Preface

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