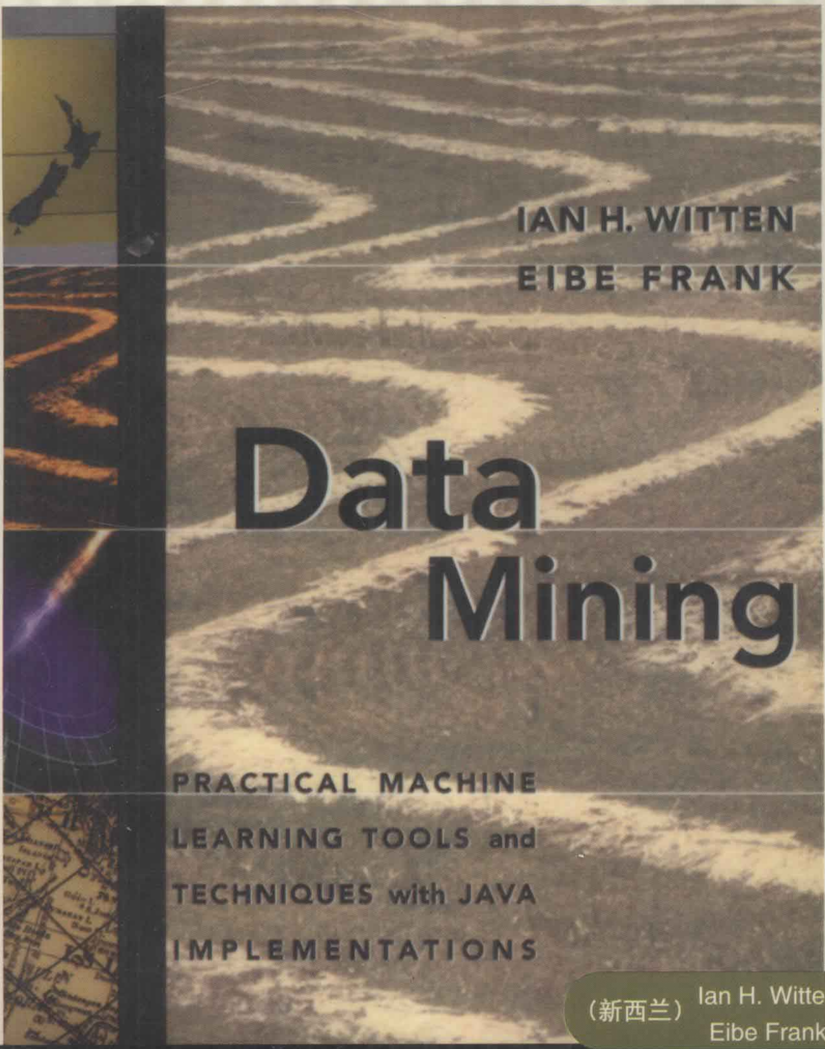
HZ BOOKS
华章教育

经典原版书库

数据挖掘

实用机器学习技术及 Java 实现

（英文版）



IAN H. WITTEN
EIBE FRANK

Data Mining

PRACTICAL MACHINE LEARNING TOOLS and TECHNIQUES with JAVA IMPLEMENTATIONS

（新西兰） Ian H. Witten
Eibe Frank 著

械工业出版社
na Machine Press

ELSEVIER

经典原版书库

数据挖掘

实用机器学习技术及 Java 实现

（英文版）

Data Mining

Practical Machine Learning Tools and Techniques with Java Implementations

（新西兰） Ian H. Witten
Eibe Frank 著

机械工业出版社
China Machine Press

本书版权登记号：图字：01-2003-5835

图书在版编目（CIP）数据

数据挖掘：实用机器学习技术及Java实现（英文版）/（新西兰）威腾（Witten, I. H. ）等著. -北京：机械工业出版社，2003.9

（经典原版书库）

书名原文：Data Mining: Practical Learning Tools and Techniques with Java Implementations

ISBN 7-111-12769-2

Ⅰ. 数…　Ⅱ. 威…　Ⅲ. 数学采集－英文　Ⅳ. TP274

中国版本图书馆CIP数据核字（2003）第068443号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）
责任编辑：朱　劼
北京昌平奔腾印刷厂印刷 · 新华书店北京发行所发行
2005年1月第1版第2次印刷
787mm × 1092mm 1/16 · 25印张
印数：3 001-4 000册
定价：40.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线电话（010）68326294

出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭橥了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短、从业人员较少的现状下，美国等发达国家在其计算机科学发展的几十年间积淀的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章图文信息有限公司较早意识到“出版要为教育服务”。自1998年开始，华章公司就将工作重点放在了遴选、移译国外优秀教材上。经过几年的不懈努力，我们与Prentice Hall，Addison-Wesley，McGraw-Hill，Morgan Kaufmann等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出Tanenbaum，Stroustrup，Kernighan，Jim Gray等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及庋藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专诚为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍，为进一步推广与发展打下了坚实的基础。

随着学科建设的初步完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都步入一个新的阶段。为此，华章公司将加大引进教材的力度，在“华章教育”的总规划之下出版三个系列的计算机教材：除“计算机科学丛书”之外，对影印版的教材，则单独开辟出“经典原版书库”；同时，引进全美通行的教学辅导书“Schaum’s Outlines”系列组成“全美经典学习指导系列”。为了保证这三套丛书的权威性，同时也为了更好地为学校和老师们服务，华章公司聘请了中国科学院、北京大学、清华大学、国防科技大学、复旦大学、上海交通大学、南京大学、浙江大学、中国科技大学、哈尔滨工业大学、西安交通大学、中国人民大学、北京航空航天大学、北京邮电大学、中山大学、解放军理工大学、郑州大学、湖北工学院、中国国

家信息安全测评认证中心等国内重点大学和科研机构在计算机的各个领域的著名学者组成“专家指导委员会”，为我们提供选题意见和出版监督。

这三套丛书是响应教育部提出的使用外版教材的号召，为国内高校的计算机及相关专业的教学度身订造的。其中许多教材均已为M. I. T.，Stanford，U.C. Berkeley，C. M. U. 等世界名牌大学所采用。不仅涵盖了程序设计、数据结构、操作系统、计算机体系结构、数据库、编译原理、软件工程、图形学、通信与网络、离散数学等国内大学计算机专业普遍开设的核心课程，而且各具特色——有的出自语言设计者之手、有的历经三十年而不衰、有的已被全世界的几百所高校采用。在这些圆熟通博的名师大作的指引之下，读者必将在计算机科学的宫殿中由登堂而入室。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证，但我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。教材的出版只是我们的后续服务的起点。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

电子邮件：hzedu@hzbook.com
联系电话：（010）68995264
联系地址：北京市西城区百万庄南街1号
邮政编码：100037

专家指导委员会

（按姓氏笔画顺序）

Foreword

Jim Gray, Series Editor
Microsoft Research

Technology now allows us to capture and store vast quantities of data. Finding and summarizing the patterns, trends, and anomalies in these data sets is one of the grand challenges of the information age.

There has been stunning progress in data mining and machine learning in the last decade. The marriage of statistics, machine learning, information theory, and computing has created a solid science with a firm mathematical base and very powerful tools. Witten and Frank present much of this progress in their book and in the Java implementations of the key algorithms. This is a milestone in the synthesis of data mining, data analysis, information theory, and machine learning.

The authors present the basic theory of automatically extracting models from data and then validating those models. The book does an excellent job of explaining the various models (decision trees, rules, and linear models) and how to apply them in practice. With this foundation, the book then walks the reader through the steps and pitfalls of various approaches. Most of the book is tutorial, but one chapter broadly describes how commercial systems work, and another does a walkthrough of the Java tools that the authors provide through a website.

This book presents this new discipline in a very accessible form: suitable both to train the next generation of practitioners and researchers and to inform lifelong learners like myself. Witten and Frank have a passion for simple and elegant solutions. They approach each topic with this mindset, grounding all concepts in concrete examples and urging the reader to consider the simple techniques first and then to progress to the more sophisticated ones if the simple ones prove inadequate.

If you are interested in databases and have not been following the machine learning field for the last decade, this book is a great way to catch up on this exciting progress.

About the Authors

Ian Witten is professor of computer science at the University of Waikato in Hamilton, New Zealand. He has taught at Essex University and at the University of Calgary, where he was head of computer science from 1982 to 1985. He holds degrees in mathematics from Cambridge University, computer science from the University of Calgary, and a Ph.D. in electrical engineering from Essex University, England. He has published extensively in academic conferences and journals on machine learning.

The underlying theme of his current research is the exploitation of information about a user's past behavior to expedite interaction in the future. In pursuit of this theme, he has been drawn into machine learning, which seeks ways to summarize, restructure, and generalize past experience; adaptive text compression, that is, using information about past text to encode upcoming characters; and user modeling, which is the general area of characterizing user behavior.

He directs a large project at Waikato on machine learning and its application to agriculture and has also been active recently in the area of document compression, indexing, and retrieval. He has also written many books over the last 15 years, the most recent of which is *Managing Gigabytes: Compressing and Indexing Documents and Images, second edition* (Morgan Kaufmann 1999) with A. Moffat and T. C. Bell.

Eibe Frank is a Ph.D. candidate in computer science at the University of Waikato. His research focus is machine learning. He holds a degree in Computer Science from the University of Karlstruhe in Germany and is the author of several papers presented at machine learning conferences and published in journals.

Preface

The convergence of computing and communication has produced a society that feeds on information. Yet most of the information is in its raw form: data. If *data* is characterized as recorded facts, then *information* is the set of patterns, or expectations, that underlie the data. There is a huge amount of information locked up in databases—information that is potentially important but has not yet been discovered or articulated. Our mission is to bring it forth.

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data. Of course, there will be problems. Many patterns will be banal and uninteresting. Others will be spurious, contingent on accidental coincidences in the particular dataset used. And real data is imperfect: some parts are garbled, some missing. Anything that is discovered will be inexact: there will be exceptions to every rule and cases not covered by any rule. Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful.

Machine learning provides the technical basis of data mining. It is used to extract information from the raw data in databases—information that is expressed in a comprehensible form and can be used for a variety of purposes. The process is one of abstraction: taking the data, warts and all, and inferring whatever structure underlies it. This book is about the tools and techniques of machine learning that are used in practical data mining for finding, and describing, structural patterns in data.

As with any burgeoning new technology that enjoys intense commercial attention, the use of data mining is surrounded by a great deal of hype in the technical—and sometimes the popular—press. Exaggerated reports appear of the secrets that can be uncovered by setting learning algorithms loose on oceans of data. But there is no magic in machine learning, no hidden power, no alchemy. Instead there is an identifiable body of simple and practical techniques that can often extract useful information from raw data. This book describes these techniques and shows how they work.

We interpret machine learning as the acquisition of structural descriptions from examples. The kind of descriptions that are found can be used for prediction, explanation, and understanding. Some data mining applications focus on prediction: forecasting what will happen in new situations from data that describe what happened in the past, often by guessing the classification of new examples. But we are equally—perhaps more—interested in applications where the result of "learning" is an actual description of a structure that can be used to classify examples. This structural description supports explanation and understanding as well as prediction. In our experience, insights gained by the user are of most interest in the majority of practical data mining applications; indeed, this is one of machine learning's major advantages over classical statistical modeling.

The book explains a wide variety of machine learning methods. Some are pedagogically motivated: simple schemes designed to explain clearly how the basic ideas work. Others are practical: real systems that are used in applications today. Many are contemporary and have been developed only in the last few years.

A comprehensive software resource, written in the Java language, has been created to illustrate the ideas in the book. Called the Waikato Environment for Knowledge Analysis, or Weka[1] for short, this utility is available as source code on the World Wide Web via *www.mkp.com/datamining* or at *www.cs.waikato.ac.nz/ml/weka*. It is a full, industrial-strength implementation of essentially all the techniques that are covered in this book. It includes illustrative code and working implementations of machine learning methods. It offers clean, spare implementations of the simplest techniques, designed to aid understanding of the mechanisms involved. It also provides a workbench that includes full, working, state-of-the-art implementations of many popular learning schemes that can be used for practical data mining or for research. Finally, it contains a framework, in the form of a Java class library, that supports applications that use embedded machine learning and even the implementation of new learning schemes.

The objective of this book is to introduce the tools and techniques for machine learning that are used in data mining. After reading it, you will understand what these techniques are and appreciate their strengths and applicability. If you wish to experiment with your own data, you will be able to do this with the Weka software.

[1] Found only on the islands of New Zealand, the *weka* (pronounced to rhyme with "Mecca") is a flightless bird with an inquisitive nature.

The book spans the gulf between the intensely practical approach taken by trade books that provide case studies on data mining and the more theoretical, principle-driven exposition found in current textbooks on machine learning. (A brief description of these books appears in the *Further reading* section at the end of Chapter 1.) This gulf is rather wide. In order to apply machine learning techniques productively, you need to understand something about how they work; this is not a technology that you can apply blindly and expect to get good results. Different problems yield to different techniques, but these are early days for data mining, and it is never clear which techniques are suitable for a given situation: you need to know something about the range of possible solutions. And we cover an extremely wide range of techniques. We can do this because, unlike many trade books, this volume does not promote any particular commercial software or approach. The book contains a large number of examples, but they use illustrative datasets that are small enough to allow you to follow what is going on. Real datasets are far too large to show this (and in any case are invariably company confidential). Our datasets are chosen not to illustrate actual large-scale practical problems, but to help you understand what the different techniques do, how they work, and what their range of application is.

The book is aimed at the technically aware general reader who is interested in the principles and ideas underlying the current practice of data mining. It will also be of interest to information professionals who need to become acquainted with this new data mining technology, and to all those who wish to gain a detailed technical understanding of what machine learning involves. It is written for an eclectic audience of information systems practitioners, programmers, consultants, developers, information technology managers, specification writers, patent examiners, curious lay people—as well as students and professors—who need an easy-to-read book with lots of illustrations that describes what the major machine learning techniques are, what they do, how they are used, and how they work. It is practically oriented, with a strong "how to" flavor, and includes algorithms, code, and implementations. All those involved in practical data mining will benefit directly from the techniques described. The book is also aimed at people who want to cut through to the reality that underlies the hype about machine learning and who seek a practical, non-academic, unpretentious approach. We have avoided requiring any specific theoretical or mathematical knowledge, except in some sections that are marked by a light gray bar in the margin. These passages contain material for the more technical or theoretically inclined reader and may be skipped without loss of continuity.

The book is organized in layers that make the ideas accessible to readers who are interested in grasping the basics, as well as to those who would like more depth of treatment, along with full details on the techniques covered. We believe that consumers of machine learning need to have some idea of how the algorithms they use work. It is often observed that data models are only as good as the person who interprets them, and that person needs to know something about how the models are produced in order to appreciate the strengths, and limitations, of the technology. However, it is not necessary for all users to have a deep understanding of the finer details of the algorithms.

We address this situation by describing machine learning methods at successive levels of detail. The reader will learn the basic ideas, the topmost level, by reading the first three chapters. Chapter 1 describes, through examples, what machine learning is and where it can be used; it also provides actual practical applications. Chapters 2 and 3 cover the different kinds of input and output—or *knowledge representation*—that are involved. Different kinds of output dictate different styles of algorithm, and at the next level, Chapter 4 describes the basic methods of machine learning, simplified to make them easy to comprehend. Here the principles involved are conveyed in a variety of algorithms without getting bogged down in intricate details or tricky implementation issues. To make progress in the application of machine learning techniques to particular data mining problems, it is essential to be able to measure how well you are doing. Chapter 5, which can be read out of sequence, equips the reader to evaluate the results that are obtained from machine learning, addressing the sometimes complex issues involved in performance evaluation.

At the lowest and most detailed level, Chapter 6 exposes in naked detail the nitty-gritty issues of implementing a spectrum of machine learning algorithms, including the complexities that are necessary for them to work well in practice. Although many readers may want to ignore this detailed information, it is at this level that the full, working, tested Java implementations of machine learning schemes are written. Chapter 7 discusses practical topics involved with engineering the input to machine learning—for example, selecting and discretizing attributes—and covers several more advanced techniques for refining and combining the output from different learning techniques. Chapter 8 describes the Java code that accompanies the book. You can skip to this chapter directly from Chapter 4 if you are in a hurry to get on with analyzing your data and don't want to be bothered with the technical details. Finally, Chapter 9 looks to the future.

The book does not cover all machine learning methods. In particular, we do not discuss neural nets because this technique produces predictions rather than structural descriptions; also, it is well described in some recent books on data mining. Nor do we cover reinforcement learning since it is rarely applied in practical data mining; nor genetic algorithm approaches since these are really just an optimization technique; nor Bayesian networks because algorithms for learning them are not yet robust enough to be deployed; nor relational learning and inductive logic programming since they are rarely used in mainstream data mining applications.

Java has been chosen for the implementations of machine learning techniques that accompany this book because, as an object-oriented programming language, it allows a uniform interface to learning schemes and methods for pre- and post-processing. We have chosen Java instead of C++, Smalltalk, or other object-oriented languages because programs written in Java can be run on almost any computer without having to be recompiled, or having to go through complicated installation procedures, or—worst of all—having to change the code itself. A Java program is compiled into byte-code that can be executed on any computer equipped with an appropriate interpreter. This interpreter is called the *Java virtual machine.* Java virtual machines—and, for that matter, Java compilers—are freely available for all important platforms.

Like all widely used programming languages, Java has received its share of criticism. Although this is not the place to elaborate on such issues, in several cases the critics are clearly right. However, of all currently available programming languages that are widely supported, standardized, and extensively documented, Java seems to be the best choice for the purpose of this book. Its main disadvantage is speed of execution—or lack of it. Executing a Java program is several times slower than running a corresponding program written in C because the virtual machine has to translate the byte-code into machine code before it can be executed. In our experience the difference is a factor of three to five if the virtual machine uses a *just-in-time compiler.* Instead of translating each byte-code individually, a just-in-time compiler translates whole chunks of byte-code into machine code, thereby achieving significant speedup. However, if this is still too slow for your application, there are compilers that translate Java programs directly into machine code, bypassing the byte-code step. Of course, this code cannot be executed on other platforms, thereby sacrificing one of Java's most important advantages.

Teaching materials on the Web

There are teaching materials available online at *www.mkp.com/books_catalog/bookpage1.asp#Extras* consisting of

- Powerpoint slides

  Powerpoint presentation containing all the figures from the book that can be downloaded
- exams

  an exam and the corresponding answers; available for instructors only
- assignments

  Assignment 1
  Assignment 2
  Assignment 3

Assignment 4
Assignment 5
Assignment 6

- quizzes

Quiz 1
Quiz 2
Quiz 3
Quiz 4
Quiz 5
Quiz 6
Quiz 7
Quiz 8
Quiz 9

Acknowledgments

Writing the acknowledgments is always the nicest part! A lot of people have helped us, and we relish this opportunity to thank them. This book has arisen out of the machine learning research project in the Computer Science Department at the University of Waikato, New Zealand. We have received generous encouragement and assistance from the academic staff members on that project: John Cleary, Sally Jo Cunningham, Matt Humphrey, Lyn Hunt, Bob McQueen, Lloyd Smith, and above all, Geoff Holmes, the project leader and source of inspiration. All who have worked on the machine learning project here have contributed to our thinking: we would particularly like to mention Steve Garner, Stuart Inglis, and Craig Nevill-Manning for helping us to get the project off the ground in the beginning when success was less certain and things were more difficult.

The Weka system that illustrates the ideas in this book forms a crucial component of it. It was conceived by the authors and designed and implemented by Eibe Frank, along with Len Trigg and Mark Hall. Many people in the machine learning laboratory at Waikato have made significant contributions, especially Yong Wang, who provided the implementation of M5′.

Tucked away as we are in a remote (but very pretty) corner of the southern hemisphere, we greatly appreciate the visitors to our department who play a crucial role in acting as sounding boards and helping us to develop our thinking in diverse ways. We would like to mention in particular Rob Holte, Bernhard Pfahringer, Carl Gutwin, and Russell Beale, each of whom visited us for several months; David Aha, who although he only came for a few days did so at an early and fragile stage of the project and performed a great service by his enthusiasm

and encouragement; and Kai Ming Ting, who worked with us for two years on many of the topics discussed in Chapter 7 and helped to bring us into the mainstream of machine learning.

Past students at Waikato have played a significant role in the development of the project. Jamie Littin worked on ripple-down rules and relational learning. Brent Martin explored instance-based learning and nested instance-based representations. Murray Fife slaved over relational learning. Nadeeka Madapathage investigated the use of functional languages for expressing machine learning algorithms. Other graduate students have influenced us in numerous ways, particularly Gordon Paynter, who has been deeply involved with the keyphrase extraction scheme described in Section 9.4, and Zane Bray, who worked on the text mining scheme outlined in the same section. Colleagues Steve Jones, Tony Smith, and Malika Mahoui have also made great and far-reaching contributions to these and other machine learning projects.

Ian Witten would like to acknowledge the formative role of his former students at Calgary, particularly Brent Krawchuk, Dave Maulsby, Thong Phan, and Tanja Mitrovic, all of whom helped him develop his early ideas in machine learning, as did faculty members Bruce MacDonald, Brian Gaines, and David Hill at Calgary, and John Andreae at the University of Canterbury. Eibe Frank is particularly indebted to his former supervisor at the University of Karlsruhe, Klaus-Peter Huber (now with SAS Institute), who infected him with the fascination of machines that learn.

Diane Cerra and Belinda Breyer of Morgan Kaufmann have worked hard to shape this book, and Elisabeth Beller, our production editor, has made the process go very smoothly for us. Bronwyn Webster has provided excellent support at the Waikato end.

We gratefully acknowledge the unsung efforts of the anonymous reviewers, one of whom in particular made a great number of pertinent and constructive comments that helped us to improve this book very significantly. In addition, we would like to thank the librarians of the Repository of Machine Learning Databases at the University of California, Irvine, whose carefully collected datasets have been invaluable in our research.

Our research has been funded by the New Zealand Foundation for Research, Science, and Technology and the Royal Society of New Zealand Marsden Fund. The Department of Computer Science at the University of Waikato has generously supported us in all sorts of ways, and we owe a particular debt of gratitude to Mark Apperley for his enlightened leadership and warm encouragement. Part of this book was written while both authors were visiting the University of Calgary, Canada, and the support of the Computer Science department there is gratefully acknowledged—as is the positive and helpful attitude of the long-suffering students in the machine learning course, on whom we experimented.

Last, and most of all, we are grateful to our families and partners. Pam, Anna, and Nikki were all too well aware of the implications of having an author in the house ("not again!") but let Ian go ahead and write the book anyway. Julie was always supportive, even when Eibe had to burn the midnight oil in the machine learning lab. The six of us hail from Canada, England, Germany, Ireland, and Samoa: New Zealand has brought us together and provided an ideal, even idyllic, place to write this book.

Contents