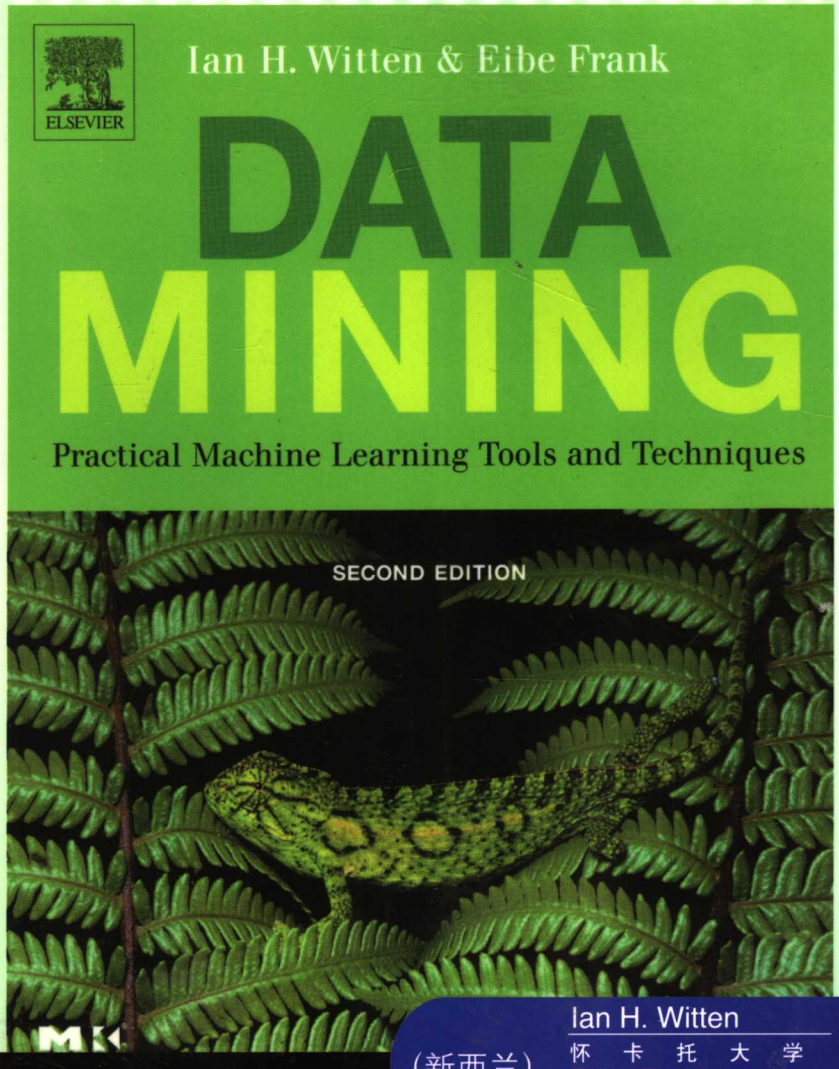
经 典 原 版 书 库

# 数据挖掘

## 实用机器学习技术

新版

（英文版 · 第2版）





Ian H. Witten
（新西兰） 怀卡托大学 著
Eibe Frank
怀卡托大学



机械工业出版社
China Machine Press

经 典 原 版 书 库

# 数据挖掘

## 实用机器学习技术

（英文版 · 第2版）

# Data Mining

## Practical Machine Learning Tools and Techniques

(Second Edition)



(新西兰) Ian H. Witten 怀卡托大学
Eibe Frank 怀卡托大学 著

机械工业出版社
China Machine Press

Ian H. Witten and Eibe Frank: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (ISBN 0-12-088407-0).



ISBN: 981-2593-15-2



**本书版权登记号：图字：01-2005-4388**

**图书在版编目（CIP）数据**

数据挖掘：实用机器学习技术（英文版·第2版）/（新西兰）威滕（Witten, I. H.）等著；-北京：机械工业出版社，2005.9

（经典原版书库）

书名原文：Data Mining: Practical Machine Learning Tools and Techniques, Second Edition

ISBN 7-111-17248-5

Ⅰ.数…　Ⅱ.威…　Ⅲ.数据采集-英文　Ⅳ.TP274

中国版本图书馆CIP数据核字（2005）第100864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：迟振春

北京牛山世兴印刷厂印刷·新华书店北京发行所发行

2005年9月第1版第1次印刷

787mm×1092mm 1/16·35印张

印数：0 001-3 000册

定价：58.00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：（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，有的还专程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: hzjsj@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）

| | | | | |
|---|---|---|---|---|
| 尤晋元 | 王　珊 | 冯博琴 | 史忠植 | 史美林 |
| 石教英 | 吕　建 | 孙玉芳 | 吴世忠 | 吴时霖 |
| 张立昂 | 李伟琴 | 李师贤 | 李建中 | 杨冬青 |
| 邵维忠 | 陆丽娜 | 陆鑫达 | 陈向群 | 周伯生 |
| 周克定 | 周傲英 | 孟小峰 | 岳丽华 | 范　明 |
| 郑国梁 | 施伯乐 | 钟玉琢 | 唐世渭 | 袁崇义 |
| 高传善 | 梅　宏 | 程　旭 | 程时端 | 谢希仁 |
| 裘宗燕 | 戴　葵 | | | |

# Foreword

Jim Gray, Series Editor
Microsoft Research

Technology now allows us to capture and store vast quantities of data. Finding patterns, trends, and anomalies in these datasets, and summarizing them with simple quantitative models, is one of the grand challenges of the information age—turning data into information and turning information into knowledge.

There has been stunning progress in data mining and machine learning. The synthesis of statistics, machine learning, information theory, and computing has created a solid science, with a firm mathematical base, and with very powerful tools. Witten and Frank present much of this progress in this book and in the companion implementation of the key algorithms. As such, this is a milestone in the synthesis of data mining, data analysis, information theory, and machine learning. If you have not been following this field for the last decade, this is a great way to catch up on this exciting progress. If you have, then Witten and Frank's presentation and the companion open-source workbench, called Weka, will be a useful addition to your toolkit.

They 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, association rules, linear models, clustering, Bayes nets, neural nets) and how to apply them in practice. With this basis, they then walk through the steps and pitfalls of various approaches. They describe how to safely scrub datasets, how to build models, and how to evaluate a model's predictive quality. Most of the book is tutorial, but Part II broadly describes how commercial systems work and gives a tour of the publicly available data mining workbench that the authors provide through a website. This Weka workbench has a graphical user interface that leads you through data mining tasks and has excellent data visualization tools that help understand the models. It is a great companion to the text and a useful and popular tool in its own right.

This book presents this new discipline in a very accessible form: as a text 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 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, this book is a great way to catch up on this exciting progress. If you have data that you want to analyze and understand, this book and the associated Weka toolkit are an excellent way to start.

# 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. In addition real data is imperfect: Some parts will be garbled, and some will be missing. Anything 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 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 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 in which the result of "learning" is an actual description of a structure that can be used to classify examples. This structural description supports explanation, understanding, and prediction. In our experience, insights gained by the applications' users 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 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 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, it is available as source code on the World Wide Web at *http://www.cs.waikato.ac.nz/ml/weka*. It is a full, industrial-strength implementation of essentially all the techniques 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 easily 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. 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 it is rarely obvious which techniques are suitable for a given situation: you need to know something about the range of possible solutions. 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. We include 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 usually 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 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 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, and curious laypeople—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 aimed at people who want to cut through to the reality that underlies the hype about machine learning and who seek a practical, nonacademic, unpretentious approach. We have avoided requiring any specific theoretical or mathematical knowledge except in some sections marked by a light gray bar in the margin. These contain optional material, often 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 and 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 to appreciate the strengths, and limitations, of the technology. However, it is not necessary for all data model 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. You 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 kinds of input and output—or *knowledge representation*—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 into 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 you to evaluate the results 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 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 implementations of machine learning schemes in Weka are written. Chapter 7 describes 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. The final chapter of Part I looks to the future.

The book describes most methods used in practical machine learning. However, it does not cover reinforcement learning, because it is rarely applied in practical data mining; genetic algorithm approaches, because these are just an optimization technique; or relational learning and inductive logic programming, because they are rarely used in mainstream data mining applications.

The data mining system that illustrates the ideas in the book is described in Part II to clearly separate conceptual material from the practical aspects of how to use it. You can skip to Part II directly from Chapter 4 if you are in a hurry to analyze your data and don't want to be bothered with the technical details.

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 postprocessing. 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, having to undergo complicated installation procedures, or—worst of all—having to change the code. 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 language 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 to slow for your application, there are compilers that translate Java programs directly into machine code, bypassing the byte-code step. This code cannot be executed on other platforms, thereby sacrificing one of Java's most important advantages.

## Updated and revised content

We finished writing the first edition of this book in 1999 and now, in April 2005, are just polishing this second edition. The areas of data mining and machine learning have matured in the intervening years. Although the core of material in this edition remains the same, we have made the most of our opportunity to update it to reflect the changes that have taken place over 5 years. There have been errors to fix, errors that we had accumulated in our publicly available errata file. Surprisingly few were found, and we hope there are even fewer in this second edition. (The errata for the second edition may be found through the book's home page at *http://www.cs.waikato.ac.nz/ml/weka/book.html.*) We have thoroughly edited the material and brought it up to date, and we practically doubled the number of references. The most enjoyable part has been adding new material. Here are the highlights.

Bowing to popular demand, we have added comprehensive information on neural networks: the perceptron and closely related Winnow algorithm in Section 4.6 and the multilayer perceptron and backpropagation algorithm

in Section 6.3. We have included more recent material on implementing nonlinear decision boundaries using both the kernel perceptron and radial basis function networks. There is a new section on Bayesian networks, again in response to readers' requests, with a description of how to learn classifiers based on these networks and how to implement them efficiently using all-dimensions trees.

The Weka machine learning workbench that accompanies the book, a widely used and popular feature of the first edition, has acquired a radical new look in the form of an interactive interface—or rather, three separate interactive interfaces—that make it far easier to use. The primary one is the Explorer, which gives access to all of Weka's facilities using menu selection and form filling. The others are the Knowledge Flow interface, which allows you to design configurations for streamed data processing, and the Experimenter, with which you set up automated experiments that run selected machine learning algorithms with different parameter settings on a corpus of datasets, collect performance statistics, and perform significance tests on the results. These interfaces lower the bar for becoming a practicing data miner, and we include a full description of how to use them. However, the book continues to stand alone, independent of Weka, and to underline this we have moved all material on the workbench into a separate Part II at the end of the book.

In addition to becoming far easier to use, Weka has grown over the last 5 years and matured enormously in its data mining capabilities. It now includes an unparalleled range of machine learning algorithms and related techniques. The growth has been partly stimulated by recent developments in the field and partly led by Weka users and driven by demand. This puts us in a position in which we know a great deal about what actual users of data mining want, and we have capitalized on this experience when deciding what to include in this new edition.

The earlier chapters, containing more general and foundational material, have suffered relatively little change. We have added more examples of fielded applications to Chapter 1, a new subsection on sparse data and a little on string attributes and date attributes to Chapter 2, and a description of interactive decision tree construction, a useful and revealing technique to help you grapple with your data using manually built decision trees, to Chapter 3.

In addition to introducing linear decision boundaries for classification, the infrastructure for neural networks, Chapter 4 includes new material on multinomial Bayes models for document classification and on logistic regression. The last 5 years have seen great interest in data mining for text, and this is reflected in our introduction to string attributes in Chapter 2, multinomial Bayes for document classification in Chapter 4, and text transformations in Chapter 7. Chapter 4 includes a great deal of new material on efficient data structures for searching the instance space: *k*D-trees and the recently invented ball trees. These

are used to find nearest neighbors efficiently and to accelerate distance-based clustering.

Chapter 5 describes the principles of statistical evaluation of machine learning, which have not changed. The main addition, apart from a note on the Kappa statistic for measuring the success of a predictor, is a more detailed treatment of cost-sensitive learning. We describe how to use a classifier, built without taking costs into consideration, to make predictions that are sensitive to cost; alternatively, we explain how to take costs into account during the training process to build a cost-sensitive model. We also cover the popular new technique of cost curves.

There are several additions to Chapter 6, apart from the previously mentioned material on neural networks and Bayesian network classifiers. More details—gory details—are given of the heuristics used in the successful RIPPER rule learner. We describe how to use model trees to generate rules for numeric prediction. We show how to apply locally weighted regression to classification problems. Finally, we describe the *X*-means clustering algorithm, which is a big improvement on traditional *k*-means.

Chapter 7 on engineering the input and output has changed most, because this is where recent developments in practical machine learning have been concentrated. We describe new attribute selection schemes such as race search and the use of support vector machines and new methods for combining models such as additive regression, additive logistic regression, logistic model trees, and option trees. We give a full account of LogitBoost (which was mentioned in the first edition but not described). There is a new section on useful transformations, including principal components analysis and transformations for text mining and time series. We also cover recent developments in using unlabeled data to improve classification, including the co-training and co-EM methods.

The final chapter of Part I on new directions and different perspectives has been reworked to keep up with the times and now includes contemporary challenges such as adversarial learning and ubiquitous data mining.

## 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 Tony Smith. Special thanks go to Mark Hall, Bernhard Pfahringer, and above all Geoff Holmes, the project leader and source of inspi-

ration. 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 made significant contributions. Since the first edition of the book the Weka team has expanded considerably: so many people have contributed that it is impossible to acknowledge everyone properly. We are grateful to Remco Bouckaert for his implementation of Bayesian networks, Dale Fletcher for many database-related aspects, Ashraf Kibriya and Richard Kirkby for contributions far too numerous to list, Niels Landwehr for logistic model trees, Abdelaziz Mahoui for the implementation of K*, Stefan Mutter for association rule mining, Gabi Schmidberger and Malcolm Ware for numerous miscellaneous contributions, Tony Voyle for least-median-of-squares regression, Yong Wang for Pace regression and the implementation of M5′, and Xin Xu for *JRip*, logistic regression, and many other contributions. Our sincere thanks go to all these people for their dedicated work and to the many contributors to Weka from outside our group at Waikato.

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. We would like to mention in particular Rob Holte, 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 2 years on many of the topics described in Chapter 7 and helped to bring us into the mainstream of machine learning.

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, and 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, YingYing Wen, and Zane Bray, who have worked with us on text mining. Colleagues Steve Jones and Malika Mahoui have also made far-reaching contributions to these and other machine learning projects. More recently we have learned much from our many visiting students from Freiburg, including Peter Reutemann and Nils Weidmann.

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 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. On his travels Eibe has benefited from interactions with Peter Turney, Joel Martin, and Berry de Bruijn in Canada and with Luc de Raedt, Christoph Helma, Kristian Kersting, Stefan Kramer, Ulrich Rückert, and Ashwin Srinivasan in Germany.

Diane Cerra and Asma Stephan of Morgan Kaufmann have worked hard to shape this book, and Lisa Royse, our production editor, has made the process go smoothly. 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 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 the first edition 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 well as the positive and helpful attitude of the long-suffering students in the machine learning course on whom we experimented.

In producing the second edition Ian was generously supported by Canada's Informatics Circle of Research Excellence and by the University of Lethbridge in southern Alberta, which gave him what all authors yearn for—a quiet space in pleasant and convivial surroundings in which to work.

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, and Immo and Ollig provided exciting diversions. Between us we hail from Canada, England, Germany, Ireland, and Samoa: New Zealand has brought us together and provided an ideal, even idyllic, place to do this work.