"Why do we not, since the phenomena are well known, build a 'knowledge refinery' as the basis of a new industry, comparable in some ways to the industry of petroleum refining, only more important in the long run? The product to be refined is codified human knowledge."
– Donald Michie

This book describes the basics of machine learning principles and algorithms used in data mining. It is suitable for advanced undergraduate and postgraduate students of computer science, researchers who want to adapt algorithms for particular data mining tasks, and advanced users of machine learning and data mining tools.

"The main strength of this book is its breadth of coverage of ML and wealth of material presented, without loss of depth. In this respect, the book is rather unique among the books on machine learning. The book covers a large majority of key methods, concepts and issues in ML. The style of writing is terse and directly to the point." – Ivan Bratko

"The authors deserve congratulations for very well organized and presented work. I also greatly appreciated the well chosen and stimulating quotations at the head of sections." – Alan McConnell-Duff

HORWOOD

ISBN 978-1-904275-21-3
9 781904 275213

KONONENKO & KUKAR

MACHINE LEARNING AND DATA MINING

HORWOOD

MACHINE LEARNING AND DATA MINING

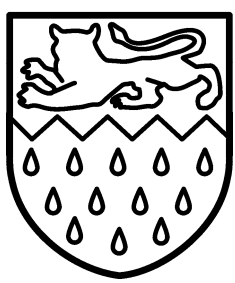
Introduction to Principles and Algorithms

IGOR KONONENKO
and
MATJAŽ KUKAR

MACHINE LEARNING AND DATA MINING:

Introduction to Principles and Algorithms

The ultimate goal of machine learning is knowledge
The ultimate goal of human learning is wisdom.



ABOUT THE AUTHORS

Igor Kononenko studied computer science at the University of Ljubljana, Slovenia, receiving his BSc in 1982, MSc in 1985 and PhD in 1990. He is now professor at the Faculty of Computer and Information Science there, teaching courses in Programming Languages, Algorithms and Data Structures; Introduction to Algorithms and Data Structures; Knowledge Engineering, Machine Learning and Knowledge Discovery in Databases. He is the head of the Laboratory for Cognitive Modelling and a member of the Artificial Intelligence Department at the same faculty. His research interests include artificial intelligence, machine learning, neural networks and cognitive modelling. He is the (co) author of 170 scientific papers in these fields and 10 textbooks. Professor Kononenko is a member of the editorial board of *Applied Intelligence* and *Informatica* journals and was also twice chair of the programme committee of the International Cognitive Conference in Ljubljana.

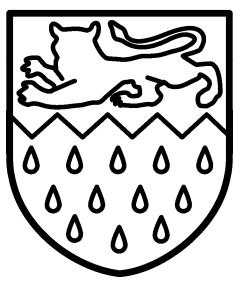
Matjaž Kukar studied computer science at the University of Ljubljana, Slovenia, receiving his BSc in 1993, MSc in 1996 and PhD in 2001. He is now the assistant professor at the Faculty of Computer and Information Science there and is also a member of the Artificial Intelligence Department at the same faculty. His research interests include knowledge discovery in databases, machine learning, artificial intelligence and statistics. Professor Kukar is the (co) author of over 50 scientific papers in these fields.

MACHINE LEARNING AND DATA MINING:

Introduction to Principles and Algorithms

Igor Kononenko
Professor of Computer and Information Science
University of Ljubljana, Slovenia

Matjaž Kukar
Assistant Professor of Computer and Information Science
University of Ljubljana, Slovenia



Horwood Publishing
Chichester, UK

HORWOOD PUBLISHING LIMITED
International Publishers in Science and Technology
Coll House, Westergate, Chichester, West Sussex, PO20 3QL, UK

First published in 2007

COPYRIGHT NOTICE
All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the permission of Horwood Publishing, Coll House, Westergate, Chichester, West Sussex, PO20 3QL, UK.

© Horwood Publishing Limited, 2007.

British Library Cataloguing in Publication Data
A catalogue record of this book is available from the British Library

ISBN – 10: 1-904275-21-4
ISBN – 13: 978-1-904275-21-3

Cover design by Jim Wilkie.
Printed and bound in the UK by Antony Rowe Limited.

Table of Contents

Foreword

In the past two decades, Machine Learning (ML) has made, among the many areas of Artificial Intelligence (AI), probably the most significant progress, and has grown into a field of its own. In this period, many of its methods found their way into practical applications, and in many application areas are now routinely used. The wide spread and visibility of ML was accelerated by powerful new techniques, and also by the appearance of effective and user-friendly implementations of ML tools, including some excellent freely available platforms, such as Weka and Orange. All this has resulted in ML, together with its related areas of Data Mining, Knowledge Discovery in Databases, and Intelligent Data Analysis, becoming an effective tool for many other disciplines as diverse as engineering, medicine, biology, ecology, biochemistry and finance. In science, it is becoming part of general scientific methodology and is now used by scientists in interpreting experimental data and in formulating new scientific theories. ML is beginning to be taught outside Computer Science and is on the way to attaining a similar status as some other subjects of general importance, in the same way as it is, for example, useful to students of medicine and biology to learn the basics of statistics.

This book is written by two scientists who have made their own original contributions to the methods and applications of ML, one of the co-authors (I. K.) having been involved since its early stages. In my view, the main strength of this book is its breadth of coverage and wealth of material presented, without loss of depth. In this respect, the book is rather unique among the books on machine learning. The book covers a large majority of key methods, concepts and issues. The style of writing is terse and directly to the point.

In comparison with other books on ML, the following chapters seem to be particularly strong. Chapter 3 systematically introduces a repertoire of basic elements that are universally used as components of ML methodology. Chapter 6 gives an outstanding review of measures for attribute selection. Chapter 8 introduces constructive induction, a topic that is rarely covered in books. In Chapters 13 and 14 on the theory of ML, the book addresses deep theoretical questions of learnability.

This book will be useful reading both for beginners to the field, and even more so to people who have already done some work in ML, but would like to broaden their knowledge and are looking for reference to learn about yet unfamiliar topics.

Ivan Bratko

Preface

In learning, each day something is gained. In following the Tao, each day something is lost.

— Lao Tse

The book describes the basics of machine learning principles and algorithms that are used in data mining and is intended for students of computer science, researchers who need to adapt various algorithms for particular data mining tasks, and for education of sophisticated users of machine learning and data mining tools.

Machine learning has exhibited tremendous progress in the last two decades. Numerous systems for machine learning are nowadays available and the number of their applications in various fields, such as industry, medicine, economics, ecology, finance, and many others, is rapidly growing. Machine learning is used for data analysis and knowledge discovery in databases, data mining, automatic generation of knowledge bases for expert systems, game playing, text classification and text mining, automatic recognition of speech, handwriting, images, etc.

The basic principle of machine learning is modeling of available data. Results of various machine learning systems may be rules, functions, relations, equation systems, probability distributions and other knowledge representations. The obtained "knowledge" aims to explain the data and can be used for forecasting, diagnostics, control, validation, and simulations.

Several excellent books have already been written on machine learning and data mining, most notably (Mitchell, 1997; Witten and Frank, 2000, 2005; Hand et al., 2001; Alpaydin, 2005). The purpose of our book is to provide a thorough and yet a comprehensible general overview of the field. Our approach differs from other textbooks by providing several different points of view to machine learning and data mining, as reflected from the structure of the book:

- Besides defining the basic notions and overviewing the machine learning methods, Chapter 1 provides the historical overview of the field and an overview of the state-of-the-art data mining tools as well as established and emerging standards.
- Chapter 2 is the least formal and introduces the psychological and philosophical issues related to machine and human learning, intelligence, and consciousness.
- Basic principles of machine learning, described in Chapter 3, are the central issue and can be considered as the core of the book.

- The next three chapters divide the machine learning process into three basic components: the representation of knowledge (Chapter 4) defines the hypothesis space, Chapter 5 describes basic search algorithms that can be used to search the hypothesis space, and the attribute quality measures (Chapter 6) are used to guide the search.

- Chapters 7 and 8 add to the whole story formal and practical guidelines for preparing, cleansing and transforming the input data in order to get the most out of it with the available machine learning algorithms. Constructive induction, described in Chapter 8, has great potentials, however, in the past not many researchers or implementations considered it worthwhile. We believe that in the future, when more research will be devoted to this area, the advantages of constructive induction will be revealed and implemented in useful practical tools.

- The following four chapters use the principles and notions, described in preceding chapters, as building blocks in order to describe particular learning algorithms. Four chapters divide the field into four main approaches. Symbolic learning (Chapter 9) includes all approaches that result in symbolic and transparent knowledge representations. Statistical learning (Chapter 10), on the other hand, results in knowledge representations that are typically hard to understand and interpret by the end user. Artificial neural networks (Chapter 11) are inspired by biological neural networks. With respect to transparency of knowledge representation, artificial neural networks are much closer to statistical learning than to symbolic learning. Cluster analysis deserves a separate chapter (Chapter 12) as the task of clustering algorithms differs a lot from other machine learning tasks.

- The last two chapters introduce formal approaches to machine learning. Chapter 13 describes a mathematically and philosophically inspired problem of identification in the limit. The results of this formal theory provide the ultimate answers to philosophical questions which, however, are of modest practical use. On the other hand, the computational learning theory, described in Chapter 14, aims to provide theoretical answers to questions of more practical value.

In order to improve readability, we provide all the references in separate sections on further reading, that also contain chapter summaries. Obvious exceptions that naturally include references are, besides this preface, also historical Sections 1.3 and 1.4. We try to keep the text free of formal derivations. Where needed, we provide separate sections for formal derivations and proofs (Sections 6.3, 7.8, and 11.7). The diamond symbol $\diamondsuit$ in the text indicates that the proof of the assertion is omitted from the text and can be found at the end of the chapter in the corresponding section on formal derivations and proofs. Obvious exceptions are the last two chapters, where proofs can be found in the literature indicated in further reading. Those sections and chapters are intended for advanced readers. In the table of contents, advanced topics are indicated with * (moderately advanced) or ** (advanced). Appendix A defines some less known formal notions, which are occasionally used in the text.

Both authors have experience in teaching various parts of the material presented in this book, to undergraduate and postgraduate students. For the undergraduate course we suggest excluding sections and chapters marked with *: Section 3.4 (comparing performance of machine learning algorithms), sections related to inductive logic programming (Sections 4.2 and 9.5), Chapter 8 (constructive induction), Section 10.5 (support vector machines), Sections 11.3 and 11.4 (Hopfield's and Bayesian neural networks), as well as all sections and chapters marked with **. Optionally, Sections 9.3 (association rules) and 9.7 (Bayesian belief networks), and/or Chapter 12 (cluster analysis) can also be kept for the postgraduate course.

The book web page can be found at the address: `mldmbook.fri.uni-lj.si`, where all information and post-press corrections will be available. The readers' comments, suggestions and indications about errors in the text are welcome, and should be sent to either author's e-mail address: `igor.kononenko@fri.uni-lj.si` or `matjaz.kukar@fri.uni-lj.si`.

Acknowledgements

The book is the result of numerous years of active research in the field of machine learning. We are grateful to Professor Ivan Bratko for his pioneering efforts in establishing an outstanding machine learning research group in Ljubljana. We are grateful to the Faculty of Computer and Information Science for support, both during research and teaching. For long term collaboration in machine learning research we are particularly grateful to Matjaž Bevk, Zoran Bosnić, Ivan Bratko, Bojan Cestnik, Janez Demšar, Aleks Jakulin, Matevž Kovačič, Nada Lavrač, Uroš Pompe, Marko Robnik Šikonja, Luka Šajn, and Blaž Zupan. We thank Luka Šajn and Alan McConnell-Duff for proofreading the text, Jožica Robnik Šikonja for proofreading Chapter 2, Ivan Bratko and the anonymous reviewer for their comments and suggestions, and Bogdan Filipič, Marko Robnik Šikonja, and Blaž Zupan for their comments on earlier drafts of the book. Useful comments on various parts of the book were provided by several people. In particular, we thank Janez Demšar for his comments on Chapter 3, Aleks Jakulin for his comments on Chapter 8, Aristidis Likas and Eduardo Hruschka for their comments on Chapter 12, and France Dacar, Matevž Kovačič and Boštjan Vilfan for their comments on Chapter 13. The author of the "spider net" graphics from the book cover is a young artist and Igor's daughter Nina Kononenko. We thank also Francesca Bonner and Ellis Horwood from Horwood Publishing for their kind assistance and patience. Finally, we are grateful to our wives and children for their loving support and patience and to whom with all our love we dedicate this book.

Ljubljana, December 2006

Igor Kononenko and Matjaž Kukar