Data Mining and Data Warehousing

This textbook is written to cater to the needs of undergraduate students of computer science, engineering, and information technology for a course on data mining and data warehousing. It brings together fundamental concepts of data mining and data warehousing in a single volume. Important topics including information theory, decision tree, Naïve Bayes classifier, distance metrics, partitioning clustering, associate mining, data marts and operational data store are discussed comprehensively. The text simplifies the understanding of the concepts through exercises and practical examples. Chapters such as classification, associate mining and cluster analysis are discussed in detail with their practical implementation using Weka and R language data mining tools. Advanced topics including big data analytics, relational data models, and NoSQL are discussed in detail. Unsolved problems and multiple-choice questions are interspersed throughout the book for better understanding.

Parteek Bhatia is Associate Professor in the Department of Computer Science and Engineering at the Thapar Institute of Engineering and Technology, Patiala, India. He has more than twenty years' teaching experience. His current research includes natural language processing, machine learning, and human–computer interface. He has taught courses including, data mining and data warehousing, big data analysis, and database management systems, at undergraduate and graduate levels.

Data Mining and Data Warehousing

Principles and Practical Techniques

Parteek Bhatia

CAMBRIDGE UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108727747

First published 2019

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

ISBN 978-1-108-72774-7 Paperback

To
my parents, Mr Ved Kumar and Mrs Jagdish Bhatia
my supportive wife, Dr Sanmeet Kaur
loving sons, Rahat and Rishan

Contents

Figures

Tables

Preface

In the modern age of artificial intelligence and business analytics, data is considered as the oil of this cyber world. The mining of data has huge potential to improve business outcomes, and to carry out the mining of data there is a growing demand for database mining experts. This book intends training learners to fill this gap.

This book will give learners sufficient information to acquire mastery over the subject. It covers the practical aspects of data mining, data warehousing, and machine learning in a simplified manner without compromising on the details of the subject. The main strength of the book is the illustration of concepts with practical examples so that the learners can grasp the contents easily. Another important feature of the book is illustration of data mining algorithms with practical hands-on sessions on Weka and R language (a major data mining tool and language, respectively). In this book, every concept has been illustrated through a step-by-step approach in tutorial form for self-practice in Weka and R. This textbook includes many pedagogical features such as chapter wise summary, exercises including probable problems, question bank, and relevant references, to provide sound knowledge to learners. It provides the students a platform to obtain expertise on technology, for better placements.

Video sessions on data mining, machine learning, big data and DBMS are also available on my YouTube channel. Learners are requested to subscribe to this channel https://www.youtube.com/user/parteekbhatia to get the latest updates through video sessions on these topics.

Your suggestions for further improvements to the book are always welcome. Kindly e-mail your suggestions to parteek.bhatia@gmail.com.

I hope you enjoy learning from this book as much as I enjoyed writing it.

Acknowledgments

Writing the acknowledgments is the most emotional part of book writing. It provides an opportunity to pay gratitude to all those who matter in your life and have helped you achieve your dream and aspirations. With the grace of God and three years of effort, I have reached this stage.

I would like to express my gratitude to the many people who saw me through this book, who motivated me directly or indirectly to write this book, to all those who provided support, talked things over, read, wrote, offered comments, and assisted in the editing, proofreading, and design.

Writing a textbook is an enormous task and it requires a great deal of motivation. I appreciate the writings of great authors like Dr A. P. J. Abdul Kalam, Mr Robin Sharma, Mr Shiv Kehra and Mr Jack Canfield, who have inspired me to contribute to the education of our young generation by writing simplified content without compromising on the depth of the subject.

Writing a book is not possible without the support and motivation of one's family. I feel blessed to have Dr Sanmeet Kaur as my wife; she has always been there to support and encourage me, despite all the time it took me, on this project. Since we both belong to the same field and same profession, having long discussions with her on different aspects of the subject is the most beautiful part of learning. These discussions helped me a long way in shaping the contents of the book. Secondly, she has always been there to take care of our whole family during my engagement with this book.

I am blessed to be born into a family of teachers. My parents, Mr Ved Kumar and Mrs Jagdish Bhatia have always provided a guiding path for excellence in life. Their life journey, in itself, is a learning path for me. I thank the almighty for providing me two loving sons, Rahat and Rishan, who filled our life with love and happiness. I thank my parents-in-laws, Mr Dalip Singh and Mrs Joginder Kaur whose daughter Sanmeet filled our home with love and affection. I thank my elder brother Mr Suneet Kumar and *bhabi ji* Mrs Dimple Bhatia, for always showering their love and blessings on me.

I am blessed to have mentors like M. L. Aeri, former Principal, DAV College, Amritsar; Mr O. P. Bhardwaj, former Head, Department of Computer Science, DAV College, Amritsar; Dr R. K. Sharma Professor, DCSE, TIET; Dr Seema Bawa, Professor DCSE, TIET; Dr Maninder Singh, Head CSED, TIET, and Dr Deepak Garg, former Head CSED, TIET, who groomed me as a teacher. I wish to thank my friends Dr Amardeep Gupta and Mr V. P. Singh, who always lend their ears to my thoughts and aspirations. I would like to thank my colleagues at TIET who motivate and provide an excellent environment for growth.

The production of this book involved a lot of help from my team of students consisting of Ms Sujata Singla, Mr Divanshu Singh, Ms Suhandhi, Mr Aditya and Ms Sawinder Kaur, who read the whole manuscript and helped me in editing and refining the text. I acknowledge the contribution of Ms Sujata in implementing the data mining algorithms in R and her assistance in finalizing the contents. There were great insights form Mr Divanshu Singh, who provided feedback and helped me refine the contents in the portions on web mining and search engine.

I would like to express my gratitude to my students at Thapar Institute of Engineering and Technology, Patiala, for their curiosity and zeal for learning which motivated me to write on this topic. I also want to thank the students at other institutes with whom I had the opportunity to interact during my 'invited talks'. They provided much-needed motivation, without which this book would have been impossible.

I want to acknowledge Dr Mark Polczynski, former Director of the MS at Marquette University, USA; Dr Saed Sayad, an Adjunct Professor at the University of Toronto; Mr Martin Fowler, ThoughtWorks, USA, for granting permission to use content from some of their published works.

I thank my publisher, Cambridge University Press, for publishing this book. I thank Mr Gauravjeet Singh of Cambridge University Press and his team for editing, refining and designing the contents, thereby providing life to the manuscript in the form of a book.