Wiley Series on Methods and Applications in Data Mining

Daniel T. Larose, Series Editor

Second Edition

DISCOVERING KNOWLEDGE IN DATA

An Introduction to Data Mining

Daniel T. Larose • Chantal D. Larose

IEEE computer society

WILEY

WILEY SERIES ON METHODS AND APPLICATIONS IN DATA MINING

Series Editor: **Daniel T. Larose**

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition • Daniel T. Larose and Chantal D. Larose

Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data • Darius M. Dziuda

Knowledge Discovery with Support Vector Machines • Lutz Hamel

Data-Mining on the Web: Uncovering Patterns in Web Content, Structure, and Usage • Zdravko Markov and Daniel Larose

Data Mining Methods and Models • Daniel Larose

Practical Text Mining with Perl • Roger Bilisoly

SECOND EDITION

DISCOVERING KNOWLEDGE IN DATA

An Introduction to Data Mining

DANIEL T. LAROSE
CHANTAL D. LAROSE

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Larose, Daniel T.
Discovering knowledge in data : an introduction to data mining / Daniel T. Larose and Chantal D. Larose. – Second edition.
pages cm
Includes index.
ISBN 978-0-470-90874-7 (hardback)
1. Data mining. I. Larose, Chantal D. II. Title.
QA76.9.D343L38 2014
006.3'12–dc23
2013046021

CONTENTS

List of Tables

List of Illustrations

Preface

What is Data Mining?

According to the Gartner Group,

> Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.

Today, there are a variety of terms used to describe this process, including *analytics*, *predictive analytics*, *big data*, *machine learning*, and *knowledge discovery in databases*. But these terms all share in common the objective of mining actionable nuggets of knowledge from large data sets. We shall therefore use the term *data mining* to represent this process throughout this text.

Why is This Book Needed?

Humans are inundated with data in most fields. Unfortunately, these valuable data, which cost firms millions to collect and collate, are languishing in warehouses and repositories. *The problem is that there are not enough trained human analysts available who are skilled at translating all of these data into knowledge*, and thence up the taxonomy tree into wisdom. This is why this book is needed.

The McKinsey Global Institute reports:[1]

> There will be a shortage of talent necessary for organizations to take advantage of big data. A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data…. We project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions. … In addition, we project a need for 1.5 million additional managers and analysts in the United States who can ask the right questions and consume the results of the analysis of big data effectively.

This book is an attempt to help alleviate this critical shortage of data analysts. *Discovering Knowledge in Data: An Introduction to Data Mining* provides readers with:

- The models and techniques to uncover hidden nuggets of information,
- The insight into how the data mining algorithms really work, and

- The experience of actually performing data mining on large data sets.

Data mining is becoming more widespread everyday, because it empowers companies to uncover profitable patterns and trends from their existing databases. Companies and institutions have spent millions of dollars to collect megabytes and terabytes of data, but are not taking advantage of the valuable and actionable information hidden deep within their data repositories. However, as the practice of data mining becomes more widespread, companies which do not apply these techniques are in danger of falling behind, and losing market share, because their competitors are applying data mining, and thereby gaining the competitive edge.

In *Discovering Knowledge in Data*, the step-by-step, hands-on solutions of real-world business problems, using widely available data mining techniques applied to real-world data sets, will appeal to managers, CIOs, CEOs, CFOs, and others who need to keep abreast of the latest methods for enhancing return-on-investment.

What's New for the Second Edition?

The second edition of *Discovery Knowledge in Data* is enhanced with an abundance of new material and useful features, including:

- Nearly 100 pages of new material.
- Three new chapters:
  - Chapter 5: *Multivariate Statistical Analysis* covers the hypothesis tests used for verifying whether data partitions are valid, along with analysis of variance, multiple regression, and other topics.