
Data Analytics

Lizenz zum Wissen.




Sichern Sie sich umfassendes Technikwissen mit Sofortzugriff auf tausende Fachbücher und Fachzeitschriften aus den Bereichen: Automobiltechnik, Maschinenbau, Energie + Umwelt, E-Technik, Informatik + IT und Bauwesen.

Exklusiv für Leser von Springer-Fachbüchern: Testen Sie Springer für Professionals 30 Tage unverbindlich. Nutzen Sie dazu im Bestellverlauf Ihren persönlichen Aktionscode **C0005406** auf www.springerprofessional.de/buchaktion/



Springer für Professionals.

Digitale Fachbibliothek. Themen-Scout. Knowledge-Manager.

-  Zugriff auf tausende von Fachbüchern und Fachzeitschriften
-  Selektion, Komprimierung und Verknüpfung relevanter Themen durch Fachredaktionen
-  Tools zur persönlichen Wissensorganisation und Vernetzung

www.entschieden-intelligenter.de

Springer für Professionals

 Springer

Thomas A. Runkler

Data Analytics

Models and Algorithms for
Intelligent Data Analysis

2nd Edition

 **Springer** Vieweg

Thomas A. Runkler
Siemens AG
München, Germany

ISBN: 978-3-658-14074-8 ISBN: 978-3-658-14075-5 (eBook)
DOI 10.1007/978-3-658-14075-5

Library of Congress Control Number: 2016942272

Springer Vieweg

© Springer Fachmedien Wiesbaden 2012, 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer Vieweg is a brand of Springer Nature
The registered company is Springer Fachmedien Wiesbaden GmbH

Preface

The information in the world doubles every 20 months. Important data sources are business and industrial processes, text and structured databases, images and videos, and physical and biomedical data. Data analytics allows to find relevant information, structures, and patterns, to gain new insights, to identify causes and effects, to predict future developments, or to suggest optimal decisions. We need models and algorithms to collect, preprocess, analyze, and evaluate data, from various fields such as statistics, machine learning, pattern recognition, system theory, operations research, or artificial intelligence. With this book, you will learn about the most important methods and algorithms for data analytics. You will be able to choose appropriate methods for specific tasks and apply these in your own data analytics projects. You will understand the basic concepts of the growing field of data analytics, which will allow you to keep pace and to actively contribute to the advancement of the field.

This text is designed for undergraduate and graduate courses on data analytics for engineering, computer science, and math students. It is also suitable for practitioners working on data analytics projects. The book is structured according to typical practical data analytics projects. Only basic mathematics is required. This material has been used for more than ten years in numerous courses at the Technical University of Munich, Germany, in short courses at several other universities, and in tutorials at international scientific conferences. Much of the content is based on the results of industrial research and development projects at Siemens.

History of the book versions:

- Data Analytics, second edition, 2016, English
- Data Mining, second edition, 2015, German
- Data Analytics, 2012, English
- Data Mining, 2010, German
- Information Mining, 2000, German

Thanks go to everybody who has contributed to this material, in particular to the reviewers and my students for suggesting improvements and pointing out errors and to the editorial and publisher team for their professional collaboration.

Munich, Germany
April 2016

Thomas A. Runkler

Contents

1	Introduction	1
1.1	It's All About Data	1
1.2	Data Analytics, Data Mining, and Knowledge Discovery	2
	References	3
2	Data and Relations	5
2.1	The Iris Data Set	5
2.2	Data Scales	8
2.3	Set and Matrix Representations	10
2.4	Relations	11
2.5	Dissimilarity Measures	12
2.6	Similarity Measures	14
2.7	Sequence Relations	16
2.8	Sampling and Quantization	18
	Problems	21
	References	22
3	Data Preprocessing	23
3.1	Error Types	23
3.2	Error Handling	26
3.3	Filtering	27
3.4	Data Transformation	32
3.5	Data Integration	35
	Problems	35
	References	36
4	Data Visualization	37
4.1	Diagrams	37
4.2	Principal Component Analysis	39
4.3	Multidimensional Scaling	43
4.4	Sammon Mapping	47

4.5	Auto-Associator	51
4.6	Histograms	51
4.7	Spectral Analysis	54
	Problems	57
	References	57
5	Correlation	59
5.1	Linear Correlation	59
5.2	Correlation and Causality	61
5.3	Chi-Square Test for Independence	62
	Problems	65
	References	65
6	Regression	67
6.1	Linear Regression	67
6.2	Linear Regression with Nonlinear Substitution	71
6.3	Robust Regression	72
6.4	Neural Networks	73
6.5	Radial Basis Function Networks	77
6.6	Cross-Validation	79
6.7	Feature Selection	81
	Problems	82
	References	82
7	Forecasting	85
7.1	Finite State Machines	85
7.2	Recurrent Models	86
7.3	Autoregressive Models	88
	Problems	89
	References	89
8	Classification	91
8.1	Classification Criteria	91
8.2	Naive Bayes Classifier	95
8.3	Linear Discriminant Analysis	97
8.4	Support Vector Machine	99
8.5	Nearest Neighbor Classifier	102
8.6	Learning Vector Quantization	103
8.7	Decision Trees	104
	Problems	107
	References	108

9 Clustering	111
9.1 Cluster Partitions	111
9.2 Sequential Clustering	113
9.3 Prototype-Based Clustering	115
9.4 Fuzzy Clustering	117
9.5 Relational Clustering	123
9.6 Cluster Tendency Assessment	127
9.7 Cluster Validity	128
9.8 Self-Organizing Map	129
Problems	130
References	131
Appendix A Brief Review of Some Optimization Methods	133
A.1 Optimization with Derivatives	133
A.2 Gradient Descent	134
A.3 Lagrange Optimization	135
References	137
Solutions	139
Index	143

List of Symbols

$\forall x \in X$	for each x in X
$\exists x \in X$	there exists an x in X
\Rightarrow	if ... then ...
\Leftrightarrow	if and only if
$\int_a^b f dx$	integral of f from $x = a$ to $x = b$
$\frac{\partial f}{\partial x}$	partial derivative of f with respect to x
\wedge	conjunction
\vee	disjunction
\cap	intersection
\cup	union
\neg	complement
\setminus	set difference
\subset, \subseteq	inclusion
\cdot	product, inner product
\times	Cartesian product, vector product
$\{\}$	empty set
$[x, y]$	closed interval from x to y
$(x, y], [x, y)$	half-bounded intervals from x to y
(x, y)	open interval from x to y
$ x $	absolute value of x
$ X $	cardinality of the set X
$\ x\ $	norm of vector x
$\lfloor x \rfloor$	smallest integer $a \geq x$
$\lceil x \rceil$	largest integer $a \leq x$
$\binom{n}{m}$	vector with the components n and m , binomial coefficient
∞	infinity
$a \ll b$	a is much less than b
$a \gg b$	a is much greater than b
$\alpha(t)$	time-variant learning rate

$\operatorname{argmin} X$	index of the minimum of X
$\operatorname{argmax} X$	index of the maximum of X
$\arctan x$	arctangent of x
$\operatorname{artanh} x$	inverse hyperbolic tangent of x
c_{ij}	covariance between features i and j
$CE(U)$	classification entropy of U
$\operatorname{cov} X$	covariance matrix of X
$d(a, b)$	distance between a and b
$\operatorname{eig} X$	eigenvectors and eigenvalues of X
F_c	Fourier cosine transform
F_s	Fourier sine transform
$h(X)$	Hopkins index of X
$H(a, b)$	Hamming distance between a and b
$H(Z)$	minimal hypercube or entropy of Z
$H(Z \mid a)$	entropy of Z given a
$\inf X$	infimum of X
λ	eigenvalue, Lagrange variable
$L(a, b)$	edit distance between a and b
$\lim_{x \rightarrow y}$	limit as x approaches y
$\log_b x$	logarithm of x to base b
$\max X$	maximum of X
$\min X$	minimum of X
$a \bmod b$	a modulo b
$N(\mu, \sigma)$	Gaussian distribution with mean μ and standard deviation σ
NaN	undefined (not a number)
$PC(U)$	partition coefficient of U
\mathbb{R}	set of real numbers
\mathbb{R}^+	set of positive real numbers
r	radius
s	standard deviation
s_{ij}	correlation between features i and j
$\sup X$	supremum of X
$\tanh x$	hyperbolic tangent of x
u_{ik}	membership of the k th vector in the i th cluster
X	set or matrix X
\bar{x}	average of X
X^T, x^T	transpose of the matrix X , or the vector x
x_k	k th vector of X
$x^{(i)}$	i th component of X
$x_k^{(i)}$	i th component of the k th vector of X
x	scalar or vector x
$x(t)$	time signal
$x(j2\pi f)$	spectrum