

Cambridge University Press

978-0-521-85267-8 - Introduction to Clustering Large and High-Dimensional Data

Jacob Kogan

Frontmatter

[More information](#)

Introduction to Clustering Large and High-Dimensional Data

There is a growing need for a more automated system of partitioning data sets into groups, or clusters. For example, as digital libraries and the World Wide Web continue to grow exponentially, the ability to find useful information increasingly depends on the indexing infrastructure or search engine. Clustering techniques can be used to discover natural groups in data sets and to identify abstract structures that might reside there, without having any background knowledge of the characteristics of the data. Clustering has been used in a variety of areas, including computer vision, VLSI design, data mining, bioinformatics (gene expression analysis), and information retrieval, to name just a few.

This book focuses on a few of the most important clustering algorithms, providing a detailed account of these major models in an information retrieval context. The beginning chapters introduce the classic algorithms in detail, while the later chapters describe clustering through divergences and show recent research for more advanced audiences.

Jacob Kogan is an Associate Professor in the Department of Mathematics and Statistics at the University of Maryland, Baltimore County. Dr. Kogan received his Ph.D. in Mathematics from the Weizmann Institute of Science and has held teaching and research positions at the University of Toronto and Purdue University, as well as a Fulbright Fellowship to Israel. His research interests include text and data mining, optimization, calculus of variations, optimal control theory, and robust stability of control systems. Dr. Kogan is the author of *Bifurcations of Extremals in Optimal Control and Robust Stability and Convexity: An Introduction* and coeditor of *Grouping Multidimensional Data: Recent Advances in Clustering*.

Cambridge University Press

978-0-521-85267-8 - Introduction to Clustering Large and High-Dimensional Data

Jacob Kogan

Frontmatter

[More information](#)

Introduction to Clustering Large and High-Dimensional Data

JACOB KOGAN

University of Maryland, Baltimore County



Cambridge University Press

978-0-521-85267-8 - Introduction to Clustering Large and High-Dimensional Data

Jacob Kogan

Frontmatter

[More information](#)

CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press

32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.orgInformation on this title: www.cambridge.org/9780521852678

© Jacob Kogan 2007

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2007

Printed in the United States of America

*A catalog record for this publication is available from the British Library.**Library of Congress Cataloging in Publication Data*

Kogan, Jacob, 1954–

Introduction to clustering large and high-dimensional data / Jacob Kogan.

p. cm.

Includes bibliographical references and index.

ISBN-13: 978-0-521-85267-8 (hardback)

ISBN-10: 0-521-85267-6 (hardback)

ISBN-13: 978-0-521-61793-2 (pbk.)

ISBN-10: 0-521-61793-6 (pbk.)

1. Cluster analysis – Data processing. 2. Cluster analysis – Computer programs. 3. Computer algorithms. 4. Dimensional analysis – Data processing. 5. Dimensional analysis – Computer programs. I. Title.

QA278.K594 2007

519.5'3 – dc22

2006024381

ISBN-13 978-0-521-85267-8 hardback

ISBN-10 0-521-85267-6 hardback

ISBN-13 978-0-521-61793-2 paperback

ISBN-10 0-521-61793-6 paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

Cambridge University Press

978-0-521-85267-8 - Introduction to Clustering Large and High-Dimensional Data

Jacob Kogan

Frontmatter

[More information](#)

God is a comedian playing to an audience
too afraid to laugh.

– Voltaire (1694–1778)

Cambridge University Press

978-0-521-85267-8 - Introduction to Clustering Large and High-Dimensional Data

Jacob Kogan

Frontmatter

[More information](#)

Contents

<i>Foreword by Michael W. Berry</i>	<i>page xi</i>
<i>Preface</i>	xiii
1 Introduction and motivation	1
1.1 A way to embed ASCII documents into a finite dimensional Euclidean space	3
1.2 Clustering and this book	5
1.3 Bibliographic notes	6
2 Quadratic k-means algorithm	9
2.1 Classical batch k -means algorithm	10
2.1.1 Quadratic distance and centroids	12
2.1.2 Batch k -means clustering algorithm	13
2.1.3 Batch k -means: advantages and deficiencies	14
2.2 Incremental algorithm	21
2.2.1 Quadratic functions	21
2.2.2 Incremental k -means algorithm	25
2.3 Quadratic k -means: summary	29
2.3.1 Numerical experiments with quadratic k -means	29
2.3.2 Stable partitions	31
2.3.3 Quadratic k -means	35
2.4 Spectral relaxation	37
2.5 Bibliographic notes	38
3 BIRCH	41
3.1 Balanced iterative reducing and clustering algorithm	41

3.2	BIRCH-like k -means	44
3.3	Bibliographic notes	49
4	Spherical k-means algorithm	51
4.1	Spherical batch k -means algorithm	51
4.1.1	Spherical batch k -means: advantages and deficiencies	53
4.1.2	Computational considerations	55
4.2	Spherical two-cluster partition of one-dimensional data	57
4.2.1	One-dimensional line vs. the unit circle	57
4.2.2	Optimal two cluster partition on the unit circle	60
4.3	Spherical batch and incremental clustering algorithms	64
4.3.1	First variation for spherical k -means	65
4.3.2	Spherical incremental iterations–computations complexity	68
4.3.3	The “ping-pong” algorithm	69
4.3.4	Quadratic and spherical k -means	71
4.4	Bibliographic notes	72
5	Linear algebra techniques	73
5.1	Two approximation problems	73
5.2	Nearest line	74
5.3	Principal directions divisive partitioning	77
5.3.1	Principal direction divisive partitioning (PDDP)	77
5.3.2	Spherical principal directions divisive partitioning (sPDDP)	80
5.3.3	Clustering with PDDP and sPDDP	82
5.4	Largest eigenvector	87
5.4.1	Power method	88
5.4.2	An application: hubs and authorities	88
5.5	Bibliographic notes	89
6	Information theoretic clustering	91
6.1	Kullback–Leibler divergence	91
6.2	k -means with Kullback–Leibler divergence	94
6.3	Numerical experiments	96
6.4	Distance between partitions	98
6.5	Bibliographic notes	99

Contents	ix
7 Clustering with optimization techniques	101
7.1 Optimization framework	102
7.2 Smoothing k -means algorithm	103
7.3 Convergence	109
7.4 Numerical experiments	114
7.5 Bibliographic notes	122
8 k-means clustering with divergences	125
8.1 Bregman distance	125
8.2 φ -divergences	128
8.3 Clustering with entropy-like distances	132
8.4 BIRCH-type clustering with entropy-like distances	135
8.5 Numerical experiments with (ν, μ) k -means	140
8.6 Smoothing with entropy-like distances	144
8.7 Numerical experiments with (ν, μ) smoka	146
8.8 Bibliographic notes	152
9 Assessment of clustering results	155
9.1 Internal criteria	155
9.2 External criteria	156
9.3 Bibliographic notes	160
10 Appendix: Optimization and linear algebra background	161
10.1 Eigenvalues of a symmetric matrix	161
10.2 Lagrange multipliers	163
10.3 Elements of convex analysis	164
10.3.1 Conjugate functions	166
10.3.2 Asymptotic cones	169
10.3.3 Asymptotic functions	173
10.3.4 Smoothing	176
10.4 Bibliographic notes	178
11 Solutions to selected problems	179
<i>Bibliography</i>	189
<i>Index</i>	203

Foreword

Clustering is a fundamental data analysis task with broad seemingly distant applications that include psychology, biology, control and signal processing, information theory, and data mining, to name just a few. The term “clustering” has been used in a variety of contexts over the last 50 years. The recent dramatic increase in computing power brings renewed interest to this fascinating research discipline. The production of a comprehensive survey would be a monumental task given the extensive literature in this area.

This book is motivated by information retrieval applications that are typically characterized by large, sparse, and high-dimensional data. Rather than covering as many clustering techniques as possible, the book provides a more detailed account of a few important clustering algorithms. The exposition varies from an excellent coverage of introductory (more elementary) material for novice readers to more detailed discussions of recent research results involving sophisticated mathematics for graduate students and research experts.

The book focuses on k -means clustering, which is by far the most popular partitioning algorithm widely used in applications. A detailed and elementary description of the classical quadratic k -means clustering algorithm provides a gentle introduction to clustering for undergraduate science majors, including engineering majors. Spherical k -means and information-theoretical k -means are introduced and connections between various versions of the algorithm are discussed. The

relationship between the quadratic k -means clustering algorithm and deterministic annealing is described.

A unified approach to the creation of k -means algorithms through divergences is covered, along with a treatment of clustering problems as *continuous* optimization problems. This is in contrast to traditional search methods for the “best” partition of a finite set into a predefined number of groups.

The BIRCH and PDDP clustering algorithms are discussed in great detail. PDDP is designed to handle large and sparse data while BIRCH is capable of handling large and more general data sets. Kogan demonstrates how both algorithms can be used to generate initial partitions for k -means. A version of BIRCH with divergences (rather than with the squared Euclidean norm) is also described in the book.

At the end of each chapter, pointers to references relevant to the material of that chapter are provided. The bibliography contains several references to clustering results not covered in the book and should prove to be quite valuable for anyone interested in learning about or contributing to clustering-based research.

Michael W. Berry
University of Tennessee

Preface

Clustering is a fundamental problem that has numerous applications in many disciplines. Clustering techniques are used to discover natural groups in data sets and to identify abstract structures that might reside there, without having any background knowledge of the characteristics of the data. They have been used in a variety of areas, including bioinformatics; computer vision; VLSI design; data mining; gene expression analysis; image segmentation; information retrieval; information theory; machine learning, object, character, and pattern recognition; signal compression; text mining; and Web page clustering.

While grouping, or clustering, is a building block in a wide range of applications, this book is motivated by the document clustering problem. The problem is characterized by very high-dimensional and sparse data. This book illustrates in depth applications of mathematical techniques for clustering large, sparse, and high-dimensional data.

The book is based on a one-semester introductory course given at the University of Maryland, Baltimore County, in the fall of 2001 and repeated in the spring of 2004. The course introduces a mixed population of advanced undergraduate and graduate students to basic results in this research area. In the fall of 2005, the course was modified and offered to a class of undergraduate computer science students at Ort Braude Engineering College, Israel. A special effort has been made to keep the exposition as simple as possible.

The classical k -means clustering algorithm is introduced first with the squared Euclidean distance. Ability to work with quadratic functions of one scalar variable is the only mathematical skill needed to comprehend this material. Later in the book the algorithm with Bregman and Csiszar divergences is introduced, thus providing the instructor with an opportunity to calibrate the level of generality that fits a specific student population. Problems inserted in the body of the book facilitate understanding theoretical material;¹ suggested projects require minimal programming skills and help the student to grasp the reality of large high-dimensional data sets. Each chapter is concluded by a brief bibliography section. These sections attempt to direct an interested reader to references relevant to the material of the corresponding chapters.

Clustering is used in a number of traditionally distant fields to describe methods for grouping of unlabeled data. Different research communities have different terminologies for clustering and the context in which clustering techniques are used. In selecting the material the author has been following his own research interests. Time limitations of one-semester courses do not allow one to cover many important relevant results. The author believes in the following results due to G. Leitmann [95]:

Theorem. *There does not exist a best method, that is, one which is superior to all other methods, for solving all problems in a given class of problems.*

Proof: By contradiction. □

Corollary. *Those who believe, or claim, that their method is the best one suffer from that alliterative affliction, ignorance/arrogance.*

Proof: By observation. □

¹ While some of the problems are straightforward, others may require investment of time and effort.

Cambridge University Press

978-0-521-85267-8 - Introduction to Clustering Large and High-Dimensional Data

Jacob Kogan

Frontmatter

[More information](#)

Preface

xv

Clustering has attracted research attention for more than 50 years, and many important results available in the literature are not covered in this book. A partial list of excellent publications on the subject is provided in Section 1.3 and [79]. This book's references stretch from the 1956 work of Steinhaus [125] to the recent work of Teboulle [128] (who brought [125] and a number of additional relevant references to the author's attention). Undoubtedly, many important relevant contributions are missing.

In spite of my efforts, many typos, errors, and mistakes surely remain in the book. While it is my unfortunate duty to accept full responsibility for them, I would like to encourage readers to email suggestions and comments to kogan@umbc.edu.

Acknowledgments

My first thanks go to Efim Gendler and Igal Lichtman, who introduced me to the fascinating area of text mining. I am indebted to Charles Nicholas for his support and collaboration.

I am grateful to Pavel Berkhin, Mike Berry, Dan Boley, Inderjit Dhillon, Joydeep Ghosh, and Zeev Volkovich for their help and influence. Special thanks go to Marc Teboulle who, over many years, has tirelessly tried to attract my attention to the beauty of optimization theory and its powerful techniques.

The productive environment at Ort Braude College developed by Shmaryahu Rozner, Rosa Azhari, and Zeev Barzilay has fostered a fruitful interaction between teaching and research that has been important to the development of this book. Special thanks go to Baruch Filhawari whose constant help made my stay at Ort Braude a pleasant experience and work on the book so productive.

Many thanks go to my students at UMBC and Ort Braude for their help in developing this book. Of the many fine people whom I have been associated with at Cambridge University Press, I would especially like to thank Lauren Cowles for her patience and assistance. Last but

Cambridge University Press

978-0-521-85267-8 - Introduction to Clustering Large and High-Dimensional Data

Jacob Kogan

Frontmatter

[More information](#)

xvi

Preface

not least, I would like to express my thanks to Rouben Rostamian for his help with L^AT_EX-related issues.

I have also been aided by the research support of Northrop Grumman Mission Systems, the U.S. Department of Defense, the United States–Israel Binational Science Foundation, and the J. William Fulbright Foreign Scholarship Board. This support is gratefully acknowledged.