Matrix Methods in Data Mining and Pattern Recognition

Fundamentals of Algorithms

The SIAM series on Fundamentals of Algorithms is a collection of short user-oriented books on state-of-the-art numerical methods. Written by experts, the books provide readers with sufficient knowledge to choose an appropriate method for an application and to understand the method's strengths and limitations. The books cover a range of topics drawn from numerical analysis and scientific computing. The intended audiences are researchers and practitioners using the methods and upper level undergraduates in mathematics, engineering, and computational science.

Books in this series not only provide the mathematical background for a method or class of methods used in solving a specific problem but also explain how the method can be developed into an algorithm and translated into software. The books describe the range of applicability of a method and give guidance on troubleshooting solvers and interpreting results. The theory is presented at a level accessible to the practitioner. MATLAB® software is the preferred language for codes presented since it can be used across a wide variety of platforms and is an excellent environment for prototyping, testing, and problem solving.

The series is intended to provide guides to numerical algorithms that are readily accessible, contain practical advice not easily found elsewhere, and include understandable codes that implement the algorithms.

Series Volumes

Eldén, L., *Matrix Methods in Data Mining and Pattern Recognition*
Hansen, P. C., Nagy, J. G., and O'Leary, D. P., *Deblurring Images: Matrices, Spectra, and Filtering*
Davis, T. A., *Direct Methods for Sparse Linear Systems*
Kelley, C. T., *Solving Nonlinear Equations with Newton's Method*

Lars Eldén
Linköping University
Linköping, Sweden

Matrix Methods in Data Mining and Pattern Recognition

siam
Society for Industrial and Applied Mathematics
Philadelphia

10 9 8 7 6 5 4 3 2 1

Figures 6.2, 10.1, 10.7, 10.9, 10.11, 11.1, and 11.3 are from L. Eldén, Numerical linear algebra in data mining, *Acta Numer.*, 15:327–384, 2006. Reprinted with the permission of Cambridge University Press.

Figures 14.1, 14.3, and 14.4 were constructed by the author from images appearing in P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:711–720, 1997.

Library of Congress Cataloging-in-Publication Data

Eldén, Lars, 1944-
Matrix methods in data mining and pattern recognition / Lars Eldén.
p. cm. — (Fundamentals of algorithms ; 04)
Includes bibliographical references and index.
ISBN 978-0-898716-26-9 (pbk. : alk. paper)
1. Data mining. 2. Pattern recognition systems—Mathematical models. 3. Algebras, Linear. I. Title.

QA76.9.D343E52 2007
05.74—dc20 2006041348

Contents

Preface

The first version of this book was a set of lecture notes for a graduate course on data mining and applications in science and technology organized by the Swedish National Graduate School in Scientific Computing (NGSSC). Since then the material has been used and further developed for an undergraduate course on numerical algorithms for data mining and IT at Linköping University. This is a second course in scientific computing for computer science students.

The book is intended primarily for undergraduate students who have previously taken an introductory scientific computing/numerical analysis course. It may also be useful for early graduate students in various data mining and pattern recognition areas who need an introduction to linear algebra techniques.

The purpose of the book is to demonstrate that there are several very powerful numerical linear algebra techniques for solving problems in different areas of data mining and pattern recognition. To achieve this goal, it is necessary to present material that goes beyond what is normally covered in a first course in scientific computing (numerical analysis) at a Swedish university. On the other hand, since the book is application oriented, it is not possible to give a comprehensive treatment of the mathematical and numerical aspects of the linear algebra algorithms used.

The book has three parts. After a short introduction to a couple of areas of data mining and pattern recognition, linear algebra concepts and matrix decompositions are presented. I hope that this is enough for the student to use matrix decompositions in problem-solving environments such as MATLAB®. Some mathematical proofs are given, but the emphasis is on the existence and properties of the matrix decompositions rather than on how they are computed. In Part II, the linear algebra techniques are applied to data mining problems. Naturally, the data mining and pattern recognition repertoire is quite limited: I have chosen problem areas that are well suited for linear algebra techniques. In order to use intelligently the powerful software for computing matrix decompositions available in MATLAB, etc., some understanding of the underlying algorithms is necessary. A very short introduction to eigenvalue and singular value algorithms is given in Part III.

I have not had the ambition to write a book of recipes: "given a certain problem, here is an algorithm for its solution." That would be difficult, as the area is far too diverse to give clear-cut and simple solutions. Instead, my intention has been to give the student a set of tools that may be tried as they are but, more likely, that will need to be modified to be useful for a particular application. Some of the methods in the book are described using MATLAB scripts. They should not

be considered as serious algorithms but rather as pseudocodes given for illustration purposes.

A collection of exercises and computer assignments are available at the book's Web page: www.siam.org/books/fa04.

The support from NGSSC for producing the original lecture notes is gratefully acknowledged. The lecture notes have been used by a couple of colleagues. Thanks are due to Gene Golub and Saara Hyvönen for helpful comments. Several of my own students have helped me to improve the presentation by pointing out inconsistencies and asking questions. I am indebted to Berkant Savas for letting me use results from his master's thesis in Chapter 10. Three anonymous referees read earlier versions of the book and made suggestions for improvements. Finally, I would like to thank Nick Higham, series editor at SIAM, for carefully reading the manuscript. His thoughtful advice helped me improve the contents and the presentation considerably.

Lars Eldén
Linköping, October 2006