Contents

Preface ix

I Linear Algebra Concepts and Matrix Decompositions

1 Vectors and Matrices in Data Mining and Pattern Recognition 3
1.1 Data Mining and Pattern Recognition 3
1.2 Vectors and Matrices 4
1.3 Purpose of the Book 7
1.4 Programming Environments 8
1.5 Floating Point Computations 8
1.6 Notation and Conventions 11

2 Vectors and Matrices 13
2.1 Matrix-Vector Multiplication 13
2.2 Matrix-Matrix Multiplication 15
2.3 Inner Product and Vector Norms 17
2.4 Matrix Norms 18
2.5 Linear Independence: Bases 20
2.6 The Rank of a Matrix 21

3 Linear Systems and Least Squares 23
3.1 LU Decomposition 23
3.2 Symmetric, Positive Definite Matrices 25
3.3 Perturbation Theory and Condition Number 26
3.4 Rounding Errors in Gaussian Elimination 27
3.5 Banded Matrices 29
3.6 The Least Squares Problem 31

4 Orthogonality 37
4.1 Orthogonal Vectors and Matrices 38
4.2 Elementary Orthogonal Matrices 40
4.3 Number of Floating Point Operations 45
4.4 Orthogonal Transformations in Floating Point Arithmetic . . . 46

5 QR Decomposition **47**
5.1 Orthogonal Transformation to Triangular Form 47
5.2 Solving the Least Squares Problem 51
5.3 Computing or Not Computing Q 52
5.4 Flop Count for QR Factorization 53
5.5 Error in the Solution of the Least Squares Problem 53
5.6 Updating the Solution of a Least Squares Problem 54

6 Singular Value Decomposition **57**
6.1 The Decomposition 57
6.2 Fundamental Subspaces 61
6.3 Matrix Approximation 63
6.4 Principal Component Analysis 66
6.5 Solving Least Squares Problems 66
6.6 Condition Number and Perturbation Theory for the Least Squares Problem 69
6.7 Rank-Deficient and Underdetermined Systems 70
6.8 Computing the SVD 72
6.9 Complete Orthogonal Decomposition 72

7 Reduced-Rank Least Squares Models **75**
7.1 Truncated SVD: Principal Component Regression 77
7.2 A Krylov Subspace Method 80

8 Tensor Decomposition **91**
8.1 Introduction 91
8.2 Basic Tensor Concepts 92
8.3 A Tensor SVD 94
8.4 Approximating a Tensor by HOSVD 96

9 Clustering and Nonnegative Matrix Factorization **101**
9.1 The k-Means Algorithm 102
9.2 Nonnegative Matrix Factorization 106

II Data Mining Applications

10 Classification of Handwritten Digits **113**
10.1 Handwritten Digits and a Simple Algorithm 113
10.2 Classification Using SVD Bases 115
10.3 Tangent Distance 122

11 Text Mining **129**
11.1 Preprocessing the Documents and Queries 130
11.2 The Vector Space Model 131
11.3 Latent Semantic Indexing 135
11.4 Clustering 139

11.5 Nonnegative Matrix Factorization 141
11.6 LGK Bidiagonalization 142
11.7 Average Performance 145

12 Page Ranking for a Web Search Engine 147
12.1 Pagerank 147
12.2 Random Walk and Markov Chains 150
12.3 The Power Method for Pagerank Computation 154
12.4 HITS 159

13 Automatic Key Word and Key Sentence Extraction 161
13.1 Saliency Score 161
13.2 Key Sentence Extraction from a Rank-k Approximation 165

14 Face Recognition Using Tensor SVD 169
14.1 Tensor Representation 169
14.2 Face Recognition 172
14.3 Face Recognition with HOSVD Compression 175

III Computing the Matrix Decompositions

15 Computing Eigenvalues and Singular Values 179
15.1 Perturbation Theory 180
15.2 The Power Method and Inverse Iteration 185
15.3 Similarity Reduction to Tridiagonal Form 187
15.4 The QR Algorithm for a Symmetric Tridiagonal Matrix 189
15.5 Computing the SVD 196
15.6 The Nonsymmetric Eigenvalue Problem 197
15.7 Sparse Matrices 198
15.8 The Arnoldi and Lanczos Methods 200
15.9 Software 207

Bibliography 209

Index 217