
Contents

1	Linear Algebra and Optimization: An Introduction	1
1.1	Introduction	1
1.2	Scalars, Vectors, and Matrices	2
1.2.1	Basic Operations with Scalars and Vectors	3
1.2.2	Basic Operations with Vectors and Matrices	8
1.2.3	Special Classes of Matrices	12
1.2.4	Matrix Powers, Polynomials, and the Inverse	14
1.2.5	The Matrix Inversion Lemma: Inverting the Sum of Matrices	17
1.2.6	Frobenius Norm, Trace, and Energy	19
1.3	Matrix Multiplication as a Decomposable Operator	21
1.3.1	Matrix Multiplication as Decomposable Row and Column Operators	21
1.3.2	Matrix Multiplication as Decomposable Geometric Operators	25
1.4	Basic Problems in Machine Learning	27
1.4.1	Matrix Factorization	27
1.4.2	Clustering	28
1.4.3	Classification and Regression Modeling	29
1.4.4	Outlier Detection	30
1.5	Optimization for Machine Learning	31
1.5.1	The Taylor Expansion for Function Simplification	31
1.5.2	Example of Optimization in Machine Learning	33
1.5.3	Optimization in Computational Graphs	34
1.6	Summary	35
1.7	Further Reading	35
1.8	Exercises	36
2	Linear Transformations and Linear Systems	41
2.1	Introduction	41
2.1.1	What Is a Linear Transform?	42
2.2	The Geometry of Matrix Multiplication	43

2.3	Vector Spaces and Their Geometry	51
2.3.1	Coordinates in a Basis System	55
2.3.2	Coordinate Transformations Between Basis Sets	57
2.3.3	Span of a Set of Vectors	59
2.3.4	Machine Learning Example: Discrete Wavelet Transform	60
2.3.5	Relationships Among Subspaces of a Vector Space	61
2.4	The Linear Algebra of Matrix Rows and Columns	63
2.5	The Row Echelon Form of a Matrix	64
2.5.1	LU Decomposition	66
2.5.2	Application: Finding a Basis Set	67
2.5.3	Application: Matrix Inversion	67
2.5.4	Application: Solving a System of Linear Equations	68
2.6	The Notion of Matrix Rank	70
2.6.1	Effect of Matrix Operations on Rank	71
2.7	Generating Orthogonal Basis Sets	73
2.7.1	Gram-Schmidt Orthogonalization and QR Decomposition	73
2.7.2	QR Decomposition	74
2.7.3	The Discrete Cosine Transform	77
2.8	An Optimization-Centric View of Linear Systems	79
2.8.1	Moore-Penrose Pseudoinverse	81
2.8.2	The Projection Matrix	82
2.9	Ill-Conditioned Matrices and Systems	85
2.10	Inner Products: A Geometric View	86
2.11	Complex Vector Spaces	87
2.11.1	The Discrete Fourier Transform	89
2.12	Summary	90
2.13	Further Reading	91
2.14	Exercises	91
3	Eigenvectors and Diagonalizable Matrices	97
3.1	Introduction	97
3.2	Determinants	98
3.3	Diagonalizable Transformations and Eigenvectors	103
3.3.1	Complex Eigenvalues	107
3.3.2	Left Eigenvectors and Right Eigenvectors	108
3.3.3	Existence and Uniqueness of Diagonalization	109
3.3.4	Existence and Uniqueness of Triangulization	111
3.3.5	Similar Matrix Families Sharing Eigenvalues	113
3.3.6	Diagonalizable Matrix Families Sharing Eigenvectors	115
3.3.7	Symmetric Matrices	115
3.3.8	Positive Semidefinite Matrices	117
3.3.9	Cholesky Factorization: Symmetric LU Decomposition	119
3.4	Machine Learning and Optimization Applications	120
3.4.1	Fast Matrix Operations in Machine Learning	121
3.4.2	Examples of Diagonalizable Matrices in Machine Learning	121
3.4.3	Symmetric Matrices in Quadratic Optimization	124
3.4.4	Diagonalization Application: Variable Separation for Optimization	128
3.4.5	Eigenvectors in Norm-Constrained Quadratic Programming	130

3.5	Numerical Algorithms for Finding Eigenvectors	131
3.5.1	The QR Method via Schur Decomposition	132
3.5.2	The Power Method for Finding Dominant Eigenvectors	133
3.6	Summary	135
3.7	Further Reading	135
3.8	Exercises	135
4	Optimization Basics: A Machine Learning View	141
4.1	Introduction	141
4.2	The Basics of Optimization	142
4.2.1	Univariate Optimization	142
4.2.1.1	Why We Need Gradient Descent	146
4.2.1.2	Convergence of Gradient Descent	147
4.2.1.3	The Divergence Problem	148
4.2.2	Bivariate Optimization	149
4.2.3	Multivariate Optimization	151
4.3	Convex Objective Functions	154
4.4	The Minutiae of Gradient Descent	159
4.4.1	Checking Gradient Correctness with Finite Differences	159
4.4.2	Learning Rate Decay and Bold Driver	159
4.4.3	Line Search	160
4.4.3.1	Binary Search	161
4.4.3.2	Golden-Section Search	161
4.4.3.3	Armijo Rule	162
4.4.4	Initialization	163
4.5	Properties of Optimization in Machine Learning	163
4.5.1	Typical Objective Functions and Additive Separability	163
4.5.2	Stochastic Gradient Descent	164
4.5.3	How Optimization in Machine Learning Is Different	165
4.5.4	Tuning Hyperparameters	168
4.5.5	The Importance of Feature Preprocessing	168
4.6	Computing Derivatives with Respect to Vectors	169
4.6.1	Matrix Calculus Notation	170
4.6.2	Useful Matrix Calculus Identities	171
4.6.2.1	Application: Unconstrained Quadratic Programming	173
4.6.2.2	Application: Derivative of Squared Norm	174
4.6.3	The Chain Rule of Calculus for Vectored Derivatives	174
4.6.3.1	Useful Examples of Vectored Derivatives	175
4.7	Linear Regression: Optimization with Numerical Targets	176
4.7.1	Tikhonov Regularization	178
4.7.1.1	Pseudoinverse and Connections to Regularization	179
4.7.2	Stochastic Gradient Descent	179
4.7.3	The Use of Bias	179
4.7.3.1	Heuristic Initialization	180
4.8	Optimization Models for Binary Targets	180
4.8.1	Least-Squares Classification: Regression on Binary Targets	181
4.8.1.1	Why Least-Squares Classification Loss Needs Repair	183

6.4.4	Optimization Algorithms for the SVM Dual	279
6.4.4.1	Gradient Descent	279
6.4.4.2	Coordinate Descent	280
6.4.5	Getting the Lagrangian Relaxation of Unconstrained Problems . . .	281
6.4.5.1	Machine Learning Application: Dual of Linear Regression	283
6.5	Penalty-Based and Primal-Dual Methods	286
6.5.1	Penalty Method with Single Constraint	286
6.5.2	Penalty Method: General Formulation	287
6.5.3	Barrier and Interior Point Methods	288
6.6	Norm-Constrained Optimization	290
6.7	Primal Versus Dual Methods	292
6.8	Summary	293
6.9	Further Reading	294
6.10	Exercises	294
7	Singular Value Decomposition	299
7.1	Introduction	299
7.2	SVD: A Linear Algebra Perspective	300
7.2.1	Singular Value Decomposition of a Square Matrix	300
7.2.2	Square SVD to Rectangular SVD via Padding	304
7.2.3	Several Definitions of Rectangular Singular Value Decomposition .	305
7.2.4	Truncated Singular Value Decomposition	307
7.2.4.1	Relating Truncation Loss to Singular Values	309
7.2.4.2	Geometry of Rank- k Truncation	311
7.2.4.3	Example of Truncated SVD	311
7.2.5	Two Interpretations of SVD	313
7.2.6	Is Singular Value Decomposition Unique?	315
7.2.7	Two-Way Versus Three-Way Decompositions	316
7.3	SVD: An Optimization Perspective	317
7.3.1	A Maximization Formulation with Basis Orthogonality	318
7.3.2	A Minimization Formulation with Residuals	319
7.3.3	Generalization to Matrix Factorization Methods	320
7.3.4	Principal Component Analysis	320
7.4	Applications of Singular Value Decomposition	323
7.4.1	Dimensionality Reduction	323
7.4.2	Noise Removal	324
7.4.3	Finding the Four Fundamental Subspaces in Linear Algebra	325
7.4.4	Moore-Penrose Pseudoinverse	325
7.4.4.1	Ill-Conditioned Square Matrices	326
7.4.5	Solving Linear Equations and Linear Regression	327
7.4.6	Feature Preprocessing and Whitening in Machine Learning	327
7.4.7	Outlier Detection	328
7.4.8	Feature Engineering	329
7.5	Numerical Algorithms for SVD	330
7.6	Summary	332
7.7	Further Reading	332
7.8	Exercises	333

8	Matrix Factorization	339
8.1	Introduction	339
8.2	Optimization-Based Matrix Factorization	341
8.2.1	Example: K-Means as Constrained Matrix Factorization	342
8.3	Unconstrained Matrix Factorization	342
8.3.1	Gradient Descent with Fully Specified Matrices	343
8.3.2	Application to Recommender Systems	346
8.3.2.1	Stochastic Gradient Descent	348
8.3.2.2	Coordinate Descent	348
8.3.2.3	Block Coordinate Descent: Alternating Least Squares	349
8.4	Nonnegative Matrix Factorization	350
8.4.1	Optimization Problem with Frobenius Norm	350
8.4.1.1	Projected Gradient Descent with Box Constraints	351
8.4.2	Solution Using Duality	351
8.4.3	Interpretability of Nonnegative Matrix Factorization	353
8.4.4	Example of Nonnegative Matrix Factorization	353
8.4.5	The I-Divergence Objective Function	356
8.5	Weighted Matrix Factorization	356
8.5.1	Practical Use Cases of Nonnegative and Sparse Matrices	357
8.5.2	Stochastic Gradient Descent	359
8.5.2.1	Why Negative Sampling Is Important	360
8.5.3	Application: Recommendations with Implicit Feedback Data	360
8.5.4	Application: Link Prediction in Adjacency Matrices	360
8.5.5	Application: Word-Word Context Embedding with GloVe	361
8.6	Nonlinear Matrix Factorizations	362
8.6.1	Logistic Matrix Factorization	362
8.6.1.1	Gradient Descent Steps for Logistic Matrix Factorization	363
8.6.2	Maximum Margin Matrix Factorization	364
8.7	Generalized Low-Rank Models	365
8.7.1	Handling Categorical Entries	367
8.7.2	Handling Ordinal Entries	367
8.8	Shared Matrix Factorization	369
8.8.1	Gradient Descent Steps for Shared Factorization	370
8.8.2	How to Set Up Shared Models in Arbitrary Scenarios	370
8.9	Factorization Machines	371
8.10	Summary	375
8.11	Further Reading	375
8.12	Exercises	375
9	The Linear Algebra of Similarity	379
9.1	Introduction	379
9.2	Equivalence of Data and Similarity Matrices	379
9.2.1	From Data Matrix to Similarity Matrix and Back	380
9.2.2	When Is Data Recovery from a Similarity Matrix Useful?	381
9.2.3	What Types of Similarity Matrices Are “Valid”?	382
9.2.4	Symmetric Matrix Factorization as an Optimization Model	383
9.2.5	Kernel Methods: The Machine Learning Terminology	383

9.3	Efficient Data Recovery from Similarity Matrices	385
9.3.1	Nyström Sampling	385
9.3.2	Matrix Factorization with Stochastic Gradient Descent	386
9.3.3	Asymmetric Similarity Decompositions	388
9.4	Linear Algebra Operations on Similarity Matrices	389
9.4.1	Energy of Similarity Matrix and Unit Ball Normalization	390
9.4.2	Norm of the Mean and Variance	390
9.4.3	Centering a Similarity Matrix	391
9.4.3.1	Application: Kernel PCA	391
9.4.4	From Similarity Matrix to Distance Matrix and Back	392
9.4.4.1	Application: ISOMAP	393
9.5	Machine Learning with Similarity Matrices	394
9.5.1	Feature Engineering from Similarity Matrix	395
9.5.1.1	Kernel Clustering	395
9.5.1.2	Kernel Outlier Detection	396
9.5.1.3	Kernel Classification	396
9.5.2	Direct Use of Similarity Matrix	397
9.5.2.1	Kernel K-Means	397
9.5.2.2	Kernel SVM	398
9.6	The Linear Algebra of the Representer Theorem	399
9.7	Similarity Matrices and Linear Separability	403
9.7.1	Transformations That Preserve Positive Semi-definiteness	405
9.8	Summary	407
9.9	Further Reading	407
9.10	Exercises	407
10	The Linear Algebra of Graphs	411
10.1	Introduction	411
10.2	Graph Basics and Adjacency Matrices	411
10.3	Powers of Adjacency Matrices	416
10.4	The Perron-Frobenius Theorem	419
10.5	The Right Eigenvectors of Graph Matrices	423
10.5.1	The Kernel View of Spectral Clustering	423
10.5.1.1	Relating Shi-Malik and Ng-Jordan-Weiss Embeddings	425
10.5.2	The Laplacian View of Spectral Clustering	426
10.5.2.1	Graph Laplacian	426
10.5.2.2	Optimization Model with Laplacian	428
10.5.3	The Matrix Factorization View of Spectral Clustering	430
10.5.3.1	Machine Learning Application: Directed Link Prediction	430
10.5.4	Which View of Spectral Clustering Is Most Informative?	431
10.6	The Left Eigenvectors of Graph Matrices	431
10.6.1	PageRank as Left Eigenvector of Transition Matrix	433
10.6.2	Related Measures of Prestige and Centrality	434
10.6.3	Application of Left Eigenvectors to Link Prediction	435
10.7	Eigenvectors of Reducible Matrices	436
10.7.1	Undirected Graphs	436
10.7.2	Directed Graphs	436

10.8	Machine Learning Applications	439
10.8.1	Application to Vertex Classification	440
10.8.2	Applications to Multidimensional Data	442
10.9	Summary	443
10.10	Further Reading	443
10.11	Exercises	444
11	Optimization in Computational Graphs	447
11.1	Introduction	447
11.2	The Basics of Computational Graphs	448
11.2.1	Neural Networks as Directed Computational Graphs	451
11.3	Optimization in Directed Acyclic Graphs	453
11.3.1	The Challenge of Computational Graphs	453
11.3.2	The Broad Framework for Gradient Computation	455
11.3.3	Computing Node-to-Node Derivatives Using Brute Force	456
11.3.4	Dynamic Programming for Computing Node-to-Node Derivatives	459
11.3.4.1	Example of Computing Node-to-Node Derivatives	461
11.3.5	Converting Node-to-Node Derivatives into Loss-to-Weight Derivatives	464
11.3.5.1	Example of Computing Loss-to-Weight Derivatives	465
11.3.6	Computational Graphs with Vector Variables	466
11.4	Application: Backpropagation in Neural Networks	468
11.4.1	Derivatives of Common Activation Functions	470
11.4.2	Vector-Centric Backpropagation	471
11.4.3	Example of Vector-Centric Backpropagation	473
11.5	A General View of Computational Graphs	475
11.6	Summary	478
11.7	Further Reading	478
11.8	Exercises	478
	Bibliography	483
	Index	491