

Contents

1	Introduction	6
2	High-Dimensional Space	7
2.1	Properties of High-Dimensional Space	9
2.2	The High-Dimensional Sphere	10
2.2.1	The Sphere and the Cube in Higher Dimensions	10
2.2.2	Volume and Surface Area of the Unit Sphere	11
2.2.3	The Volume is Near the Equator	14
2.2.4	The Volume is in a Narrow Annulus	16
2.2.5	The Surface Area is Near the Equator	16
2.3	The High-Dimensional Cube and Chernoff Bounds	18
2.4	Volumes of Other Solids	22
2.5	Generating Points Uniformly at Random on the surface of a Sphere	24
2.6	Gaussians in High Dimension	24
2.7	Random Projection and the Johnson-Lindenstrauss Theorem	30
2.8	Bibliographic Notes	33
2.9	Exercises	34
3	Random Graphs	45
3.1	The $G(n, p)$ Model	45
3.1.1	Degree Distribution	45
3.1.2	Existence of Triangles in $G(n, d/n)$	49
3.1.3	Phase Transitions	51
3.1.4	Phase Transitions for Monotone Properties	61
3.1.5	Phase Transitions for CNF-sat	64
3.1.6	The Emerging Graph	68
3.1.7	The Giant Component	70
3.2	Branching Processes	79
3.3	Nonuniform and Growth Models of Random Graphs	84
3.3.1	Nonuniform Models	84
3.3.2	Giant Component in Random Graphs with Given Degree Distribution	85
3.4	Growth Models	86
3.4.1	Growth Model Without Preferential Attachment	86
3.4.2	A Growth Model With Preferential Attachment	93
3.5	Small World Graphs	94
3.6	Bibliographic Notes	99
3.7	Exercises	100
4	Singular Value Decomposition (SVD)	111
4.1	Singular Vectors	112
4.2	Singular Value Decomposition (SVD)	116
4.3	Best Rank k Approximations	117

4.4	Power Method for Computing the Singular Value Decomposition	119
4.5	Applications of Singular Value Decomposition	123
4.5.1	Principal Component Analysis	123
4.5.2	Clustering a Mixture of Spherical Gaussians	124
4.5.3	An Application of SVD to a Discrete Optimization Problem	128
4.5.4	SVD as a Compression Algorithm	131
4.5.5	Spectral Decomposition	131
4.5.6	Singular Vectors and ranking documents	132
4.6	Bibliographic Notes	134
4.7	Exercises	135
5	Random Walks on Graphs	144
5.1	Electrical Networks and Random Walks	144
5.2	Random Walks on Undirected Graphs	149
5.3	Random Walks in Euclidean Space	156
5.4	Random Walks on Directed Graphs	159
5.5	Finite Markov Processes	159
5.6	Markov Chain Monte Carlo	163
5.6.1	Time Reversibility	165
5.6.2	Metropolis-Hasting Algorithm	166
5.6.3	Gibbs Sampling	166
5.7	Convergence to Steady State	168
5.7.1	Using Minimum Escape Probability to Prove Convergence	174
5.8	Exercises	177
6	Learning and VC-dimension	185
6.1	Learning	185
6.2	Linear Separators, the Perceptron Algorithm and Margins	185
6.3	Nonlinear Separators, Support Vector Machines and Kernels	189
6.4	Strong and Weak Learning - Boosting	195
6.5	Number of Examples Needed for Prediction: VC-Dimension	196
6.6	Vapnik-Chervonenkis or VC-Dimension	199
6.6.1	Examples of Set Systems and Their VC-Dimension	199
6.6.2	The Shatter Function	203
6.6.3	Shatter Function for Set Systems of Bounded VC-Dimension	204
6.6.4	Intersection Systems	205
6.7	The VC Theorem	206
6.8	Priors and Bayesian Learning	210
6.9	Exercises	211
7	Algorithms for Massive Data Problems	223
7.1	Frequency Moments of Data Streams	223
7.1.1	Number of Distinct Elements in a Data Stream	224
7.1.2	Counting the Number of Occurrences of a Given Element.	227

7.1.3	Counting Frequent Elements	228
7.1.4	The Second Moment	230
7.2	Sketch of a Large Matrix	233
7.2.1	Matrix Multiplication Using Sampling	235
7.2.2	Approximating a Matrix with a Sample of Rows and Columns . . .	237
7.3	Sketches of Documents	239
7.4	Exercises	242
8	Clustering	247
8.1	Some Clustering Examples	247
8.2	A Simple Greedy Algorithm for k -clustering	249
8.3	The k -means Clustering Algorithm	250
8.4	Spectral Clustering	252
8.4.1	Significance of a Given Clustering	253
8.4.2	SVD Clusters Most Points Correctly	255
8.4.3	SVD Gives a Factor of Two Approximation to Optimal Clustering .	257
8.4.4	Convergence of the 2-means Algorithm Started on SVD Centers . .	258
8.4.5	Recursive Clustering Based on the Second Eigenvector	264
8.5	Kernel Methods	270
8.6	Agglomerative Clustering	272
8.7	Communities, Dense Submatrices	274
8.8	Flow Methods	276
8.9	Linear Programming Formulation	279
8.10	Finding a Local Cluster Without Examining the Whole graph	280
8.11	Statistical Clustering	286
8.12	Axioms for Clustering	286
8.12.1	An Impossibility Result	286
8.12.2	A Satisfiable Set of Axioms	292
8.13	Exercises	294
9	Graphical Models and Belief Propagation	301
9.1	Bayesian Networks	301
9.2	Markov Random Fields	302
9.3	Factor Graphs	304
9.4	Tree Algorithms	304
9.5	Message Passing Algorithm	305
9.6	Graphs with a Single Cycle	308
9.7	Belief Update in Networks with a Single Loop	310
9.8	Graphs with Multiple Loops	311
9.9	Clustering by Message Passing	312
9.10	Maximum Weight Matching	314
9.11	Warning Propagation	318
9.12	Correlation Between Variables	319

9.13 Exercises	323
10 Other Topics	325
10.1 Rankings	325
10.2 Hare System for Voting	327
10.3 Compressed Sensing and Sparse Vectors	328
10.3.1 Unique Reconstruction of a Sparse Vector	329
10.3.2 The Exact Reconstruction Property	331
10.3.3 Restricted Isometry Property	332
10.4 Applications	334
10.4.1 Sparse Vector in Some Coordinate Basis	334
10.4.2 A Representation Cannot be Sparse in Both Time and Frequency Domains	334
10.4.3 Biological	337
10.4.4 Finding Overlapping Cliques or Communities	338
10.4.5 Low Rank Matrices	339
10.5 Exercises	340
11 Appendix	344
11.1 Asymptotic Notation	344
11.2 Useful Inequalities	345
11.3 Sums of Series	350
11.4 Generating Functions	355
11.4.1 Generating Functions for Sequences Defined by Recurrence Rela- tionships	356
11.4.2 Exponential Generating Function	358
11.5 Probability	364
11.5.1 Sample Space, Events, Independence	364
11.5.2 Variance	367
11.5.3 Covariance	367
11.5.4 Probability Distributions	370
11.5.5 Tail Bounds	379
11.5.6 Chernoff Bounds: Bounding of Large Deviations	381
11.5.7 Hoeffding's Inequality	384
11.6 Eigenvalues and Eigenvectors	385
11.6.1 Eigenvalues and Eigenvectors	385
11.6.2 Symmetric Matrices	387
11.6.3 Extremal Properties of Eigenvalues	389
11.6.4 Eigenvalues of the Sum of Two Symmetric Matrices	390
11.6.5 Separator Theorem	392
11.6.6 Norms	392
11.6.7 Important Norms and Their Properties	394
11.6.8 Linear Algebra	397