

Contents

1	Introduction	1
2	Concepts and Terminology	3
2.1	Machine Learning and Data Mining	3
2.2	Bioinformatics: Connections to Data Types	5
2.3	Six Types of Data	6
2.3.1	Vectors	7
2.3.2	Sets	7
2.3.3	Sequences and Strings	8
2.3.4	Trees	9
2.3.5	Graphs	10
2.3.6	Nodes in a Graph	11
2.3.7	Notes on Data Types	12
2.4	Structure of This Book	13
3	Learning Vectors	15
3.1	Unsupervised Learning	16
3.1.1	Clustering – Objectives	16
3.1.2	Clustering – K -means	17
3.1.3	Clustering – Constrained K -means	20
3.1.4	Clustering – Finite Mixture Model	23
3.1.5	Clustering – Hierarchical Clustering	27
3.1.6	Biclustering and Visualization	32
3.1.7	Probabilistic Model (Generative Model)	32
3.1.8	Matrix Factorization	34
3.2	Supervised Learning	42
3.2.1	K -nearest Neighbors	42
3.2.2	Decision Stump	44
3.2.3	Decision Tree	47
3.2.4	Naive Bayes Classifier and Bayesian Belief Network Classifier	50
3.2.5	Linear (Ridge) Regression	53
3.2.6	Logistic Regression	57
3.2.7	Layered Neural Network and Deep Learning	64
3.2.8	Ensemble Learning via Sampling	73

3.2.9	Ensemble Learning with Three Hypotheses	75
3.2.10	Ensemble Learning: AdaBoost	79
3.2.11	Support Vector Machine	84
3.3	Feature Learning	91
3.3.1	Feature Selection	92
3.3.2	Principal Component Analysis	94
3.3.3	Laplacian Eigenmaps	96
3.3.4	Sparse Learning	98
3.4	Kernel Learning	100
3.4.1	Kernel K -means	101
3.4.2	Kernel Ridge Regression	103
3.4.3	Kernel Principal Component Analysis	107
3.4.4	Summary of Kernel Learning	109
3.5	Applications to Life Sciences	110
3.5.1	Unsupervised Learning: Biclustering	110
3.5.2	Supervised Learning	115
4	Learning Sets	119
4.1	Frequent Pattern Mining	120
4.1.1	Terms, Concepts and Problem Setting	120
4.1.2	Downward Closure Property	121
4.1.3	Two Types of Approaches	122
4.1.4	Apriori Algorithm	122
4.1.5	FP-growth Algorithm	123
4.2	Probabilistic Model	126
4.2.1	Topic Model	126
4.3	Kernel Learning	135
4.3.1	Kernels for Sets	135
4.4	Applications to Life Sciences	143
4.4.1	Biclustering by Frequent Pattern Mining	143
5	Learning Sequences	145
5.1	Frequent Subsequence Mining	146
5.1.1	Preparation	146
5.1.2	Two Types of Approaches	146
5.1.3	Generalized Sequential Patterns	147
5.1.4	PrefixSpan	147
5.1.5	Notes on PrefixSpan	151
5.2	Probabilistic Models for Sequences	152
5.2.1	Mixture Markov Model	154
5.2.2	Hidden Markov Model (HMM)	156
5.2.3	Various Extensions of HMM for Sequences	163
5.3	Kernel Learning	169
5.3.1	Spectrum Kernel	170
5.3.2	All Subsequence Kernel	170
5.3.3	Fast Computation by Suffix Tree	172

5.4	Applications to Life Sciences	178
5.4.1	Frequent Subsequence Mining for Sequence Motifs	178
5.4.2	Mixture Markov Model for Metabolic Network	179
5.4.3	Multiple Sequence Alignment by HMM	180
5.4.4	Predicting β -sheet by Stochastic Tree Grammar	187
5.4.5	Identifying Timing Difference of Gene Expression by HMM	188
5.4.6	Sequence Analysis by String Kernel	190
6	Learning Trees	191
6.1	Probabilistic Models	192
6.1.1	Hidden Tree Markov Model	193
6.1.2	Ordered Tree Markov Model	197
6.1.3	Summary of Probabilistic Models for Trees	206
6.2	Kernel Learning	207
6.2.1	All Subtree Kernel	207
6.2.2	Searching Isomorphic Subtrees: Transforming a Tree Following a Lexicographical Order	208
6.2.3	All Subtree Kernel Computation Reduced to All Subsequence Kernel Computation	209
6.2.4	Algorithm	209
6.2.5	Time Complexity	210
6.3	Frequent Subtree Mining	210
6.3.1	Problem Setting	210
6.3.2	Solutions	211
6.4	Applications to Life Sciences	212
6.4.1	Glycans: Typical Biological Molecules as Trees	212
6.4.2	Application of Probabilistic Models to Glycans	213
6.4.3	Application of Frequent Subtree Mining to Glycan	216
7	Learning Graphs	219
7.1	Frequent Subgraph Mining	220
7.1.1	Problem Setting	220
7.1.2	gSpan Algorithm	221
7.1.3	Reverse Search	225
7.1.4	Removing Redundant Frequent Subgraphs	227
7.2	Kernel Learning	229
7.2.1	Notes on Kernel Learning	234
7.3	Applications to Life Sciences	235
8	Learning Nodes in a Graph	237
8.1	Unsupervised Learning (Single Graph)	240
8.1.1	Spectral Clustering	240
8.2	Unsupervised Learning (Multiple Graphs)	248
8.2.1	Spectral Clustering	248
8.2.2	Matrix Factorization	249
8.3	Label Propagation: Semi-Supervised Learning (Single Graph)	250

8.3.1	Problem Setting	250
8.3.2	Basic Label Propagation	251
8.3.3	Flexible Label Propagation	252
8.4	Label Propagation: Semi-Supervised Learning (Multiple Graphs) .	254
8.4.1	Input Data: Multiple Graphs	254
8.4.2	Problem Setting 1: Weights over Graphs	254
8.4.3	Problem Setting 2: Weights over Localized Information . .	259
9	Data Integrative Learning	261
9.1	Integrative Learning for Multiple Vectors	262
9.1.1	Kernel learning	263
9.1.2	Canonical Correlation Analysis	265
9.1.3	Kernel Canonical Correlation Analysis	268
9.2	Integrative Learning for Vectors and Graphs	272
9.2.1	Kernel Learning	274
9.2.2	Frequent Pattern Mining	275
9.3	Integrative Learning for Vectors and Nodes in a Graph	279
9.3.1	Semi-supervised Classification	280
9.3.2	Semi-supervised Clustering	285
9.3.3	Collaborative Matrix Factorization	286
9.4	Integrative Learning for Nodes in Multiple Graphs	289
9.5	Multiple Kernel Learning (MKL)	290
9.6	Applications to Life Sciences	292
9.6.1	Integrative Learning of Vectors and Graphs	292
9.6.2	Integrative Learning of Vectors and Nodes in a Graph . . .	293
10	Prediction Results Evaluation	295
10.1	Supervised Learning and Binary Labels	298
10.1.1	Prediction without Scores	298
10.1.2	Prediction with Scores	305
10.2	Supervised Learning and Continuous Labels	312
10.2.1	Root Mean Square Deviation	312
10.2.2	Transforming into Binary Labels	312
10.3	Unsupervised Learning (Clustering)	312
10.3.1	RI (Rand Index)	313
10.3.2	Adjusted Rand Index (ARI)	314
10.3.3	Normalized Mutual Information (NMI)	315
10.3.4	Other Measures for Clustering	317
10.4	Ranking	318
10.4.1	Binary Labels Relevant to the Query	318
10.4.2	Multiple Labels Relevant to Query	320
10.5	Summary	322

Appendix A Basics and Derivations in the Main Text	323
A.1 Sampling	323
A.2 Statistics: One Variable (Univariate)	323
A.2.1 Basic Statistics	324
A.2.2 Entropy	325
A.2.3 Sigmoid function	325
A.2.4 Probability Distributions	326
A.2.5 Discrete Probability Distribution (Probability Mass Function (PMF))	326
A.2.6 Continuous Probability Distribution (Probability Density Function (PDF))	327
A.3 Statistics: Two Variables	329
A.3.1 Basic Statistics	329
A.4 Matrix	331
A.4.1 Basics	331
A.4.2 Singular Matrix	332
A.4.3 Eigenvalue Problem	333
A.4.4 Generalized Eigenvalue Problem	334
A.4.5 Singular Value Decomposition (SVD)	334
A.4.6 Positive Semidefinite Matrix	335
A.5 Kernel Function	336
A.5.1 Inner Product Space	336
A.5.2 Definition of Kernel Functions	337
A.5.3 Properties of Kernel Functions	338
A.5.4 Examples of Kernel Functions	339
A.6 Norm	339
A.6.1 Vectors	340
A.6.2 Matrix	341
A.6.3 Computation of Matrix Norm	343
A.6.4 Regularization – Constraints by Norm	343
A.7 Nodes in a Graph	344
A.7.1 Graph Laplacian	344
A.8 Optimization	346
A.8.1 Convex Optimization	346
A.8.2 Least Squares	347
A.8.3 Alternating Least Squares (ALS)	347
A.8.4 Steepest Descent	347
A.8.5 Method of Lagrange Multipliers	348
A.8.6 Karush-Kuhn-Tucker (KKT) conditions	349
A.8.7 Rayleigh Quotient	350
A.8.8 Maximum Likelihood Estimation	351
A.8.9 Expectation-Maximization (EM) Algorithm	352
A.8.10 Bayes Estimation (Bayes Learning)	353
A.8.11 Gibbs Sampling	354
A.9 Derivations in the Main Text	355
A.9.1 Derivation of (3.9) in K -means	355

A.9.2	Derivation of (3.51) and (3.52) in Matrix Factorization . . .	356
A.9.3	Derivation of (3.118) in Logistic Regression	357
A.9.4	Derivation of (3.148) and (3.152) in Layered Neural Network	358
A.9.5	Derivation of (3.209) in Support Vector Machine	359
A.9.6	Derivation of (3.253) in Kernel K -means	360
A.9.7	Derivation of (3.267) in Kernel Ridge Regression	360
A.9.8	Derivation of (A.11) of Variance	361
A.9.9	Derivation of (A.40) in Covariance	363