



Computer Science and Data Analysis Series

Introduction to Machine Learning and Bioinformatics

Sushmita Mitra

Indian Statistical Institute
Kolkata, India

Sujay Datta

Texas A&M University
College Station, TX, U.S.A.

Theodore Perkins

McGill Centre for Bioinformatics
Montreal, Quebec, Canada

George Michailidis

University of Michigan
Ann Arbor, MI, U.S.A.



CRC Press

Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business
A CHAPMAN & HALL BOOK

Contents

1	Introduction	1
2	The Biology of a Living Organism	5
2.1	Cells	5
2.2	DNA and Genes	8
2.3	Proteins	12
2.4	Metabolism	15
2.5	Biological Regulation Systems: When They Go Awry	17
2.6	Measurement Technologies	19
	References	24
3	Probabilistic and Model-Based Learning	25
3.1	Introduction: Probabilistic Learning	25
3.2	Basics of Probability	27
3.3	Random Variables and Probability Distributions	40
3.4	Basics of Information Theory	56
3.5	Basics of Stochastic Processes	58
3.6	Hidden Markov Models	62
3.7	Frequentist Statistical Inference	66
3.8	Some Computational Issues	86
3.9	Bayesian Inference	89
3.10	Exercises	97
	References	100

4	Classification Techniques	101
4.1	Introduction and Problem Formulation	101
4.2	The Framework	103
4.3	Classification Methods	108
4.4	Applications of Classification Techniques to Bioinformatics Problems	124
4.5	Exercises	124
	References	125
5	Unsupervised Learning Techniques	129
5.1	Introduction	129
5.2	Principal Components Analysis	129
5.3	Multidimensional Scaling	136
5.4	Other Dimension Reduction Techniques	139
5.5	Cluster Analysis Techniques	141
5.6	Exercises	151
	References	153
6	Computational Intelligence in Bioinformatics	155
6.1	Introduction	155
6.2	Fuzzy Sets (FS)	156
6.3	Artificial Neural Networks (ANN)	161
6.4	Evolutionary Computing (EC)	167
6.5	Rough Sets (RS)	171
6.6	Hybridization	173
6.7	Application to Bioinformatics	175
6.8	Conclusion	199
6.9	Exercises	200
	References	201

7	Connections between Machine Learning and Bioinformatics	211
7.1	Sequence Analysis	211
7.2	Analysis of High-Throughput Gene Expression Data	218
7.3	Network Inference	223
7.4	Exercises	230
	References	231
8	Machine Learning in Structural Biology: Interpreting 3D Protein Images	237
8.1	Introduction	237
8.2	Background	237
8.3	ARP/WARP	247
8.4	RESOLVE	252
8.5	TEXTAL	258
8.6	ACMI	264
8.7	Conclusion	273
8.8	Acknowledgments	275
	References	275
9	Soft Computing in Biclustering	277
9.1	Introduction	277
9.2	Biclustering	278
9.3	Multi-Objective Biclustering	283
9.4	Fuzzy Possibilistic Biclustering	287
9.5	Experimental Results	291
9.6	Conclusions and Discussion	297
	References	298

10 Bayesian Machine-Learning Methods for Tumor Classification Using Gene Expression Data	303
10.1 Introduction	303
10.2 Classification Using RKHS	306
10.3 Hierarchical Classification Model	308
10.4 Likelihoods of RKHS Models	310
10.5 The Bayesian Analysis	312
10.6 Prediction and Model Choice	314
10.7 Some Examples	315
10.8 Concluding Remarks	321
10.9 Acknowledgments	322
References	322
11 Modeling and Analysis of Quantitative Proteomics Data Obtained from iTRAQ Experiments	327
11.1 Introduction	327
11.2 Statistical Modeling of iTRAQ Data	328
11.3 Data Illustration	330
11.4 Discussion and Concluding Remarks	332
11.5 Acknowledgments	334
References	334
12 Statistical Methods for Classifying Mass Spectrometry Database Search Results	339
12.1 Introduction	339
12.2 Background on Proteomics	341
12.3 Classification Methods	342
12.4 Data and Implementation	347
12.5 Results and Discussion	350
12.6 Conclusions	356
12.7 Acknowledgments	357
References	357
Index	361