

CONTENTS

Preface	v
1 Point Estimation Algorithms	1
1. Introduction	1
2. Motivation	2
3. Methods of Point Estimation	2
3.1. The Method of Moments	2
3.2. Maximum Likelihood Estimation	4
3.3. The Expectation-Maximization Algorithm	6
4. Measures of Performance	8
4.1. Bias	9
4.2. Mean Squared Error	9
4.3. Standard Error	10
4.4. Efficiency	10
4.5. Consistency	11
4.6. The Jackknife Method	11
5. Summary	13
2 Applications of Bayes Theorem	15
1. Introduction	15
2. Motivation	16
3. The Bayes Approach for Classification	17
3.1. Statistical Framework for Classification	17
3.2. Bayesian Methodology	20
4. Examples	22
4.1. Example 1: Numerical Methods	22
4.2. Example 2: Bayesian Networks	24
5. Summary	25

3	Similarity Measures	27
1.	Introduction	27
2.	Motivation	28
3.	Classic Similarity Measures	28
3.1.	Dice	30
3.2.	Overlap	30
3.3.	Jaccard	31
3.4.	Asymmetric	31
3.5.	Cosine	31
3.6.	Other Measures	32
3.7.	Dissimilarity	32
4.	Example	33
5.	Current Applications	35
5.1.	Multi-Dimensional Modeling	35
5.2.	Hierarchical Clustering	36
5.3.	Bioinformatics	37
6.	Summary	38
4	Decision Trees	39
1.	Introduction	39
2.	Motivation	41
3.	Decision Tree Algorithms	42
3.1.	ID3 Algorithm	43
3.2.	Evaluating Tests	43
3.3.	Selection of Splitting Variable	46
3.4.	Stopping Criteria	46
3.5.	Tree Pruning	47
3.6.	Stability of Decision Trees	47
4.	Example: Classification of University Students	48
5.	Applications of Decision Tree Algorithms	49
6.	Summary	50
5	Genetic Algorithms	53
1.	Introduction	53
2.	Motivation	54
3.	Fundamentals	55
3.1.	Encoding Schema and Initialization	56
3.2.	Fitness Evaluation	57

3.3.	Selection	58
3.4.	Crossover	59
3.5.	Mutation	61
3.6.	Iterative Evolution	62
4.	Example: The Traveling-Salesman	63
5.	Current and Future Applications	65
6.	Summary	66
6	Classification: Distance-based Algorithms	67
1.	Introduction	67
2.	Motivation	68
3.	Distance Functions	68
3.1.	City Block Distance	69
3.2.	Euclidean Distance	70
3.3.	Tangent Distance	70
3.4.	Other Distances	71
4.	Classification Algorithms	72
4.1.	A Simple Approach Using Mean Vector	72
4.2.	K -Nearest Neighbors	74
5.	Current Applications	76
6.	Summary	77
7	Decision Tree-based Algorithms	79
1.	Introduction	79
2.	Motivation	80
3.	ID3	80
4.	C4.5	82
5.	C5.0	83
6.	CART	84
7.	Summary	85
8	Covering (Rule-based) Algorithms	87
1.	Introduction	87
2.	Motivation	88
3.	Classification Rules	88
4.	Covering (Rule-based) Algorithms	90
4.1.	1R Algorithm	91
4.2.	PRISM Algorithm	94
4.3.	Other Algorithms	96

5. Applications of Covering Algorithms	97
6. Summary	97
9 Clustering: An Overview	99
1. Introduction	99
2. Motivation	100
3. The Clustering Process	100
3.1. Pattern Representation	101
3.2. Pattern Proximity Measures	102
3.3. Clustering Algorithms	103
3.3.1. Hierarchical Algorithms	103
3.3.2. Partitional Algorithms	105
3.4. Data Abstraction	105
3.5. Cluster Assessment	105
4. Current Applications	107
5. Summary	107
10 Clustering: Hierarchical Algorithms	109
1. Introduction	109
2. Motivation	110
3. Agglomerative Hierarchical Algorithms	111
3.1. The Single Linkage Method	112
3.2. The Complete Linkage Method	114
3.3. The Average Linkage Method	116
3.4. The Centroid Method	116
3.5. The Ward Method	117
4. Divisive Hierarchical Algorithms	118
5. Summary	120
11 Clustering: Partitional Algorithms	121
1. Introduction	121
2. Motivation	122
3. Partitional Clustering Algorithms	122
3.1. Squared Error Clustering	122
3.2. Nearest Neighbor Clustering	126
3.3. Partitioning Around Medoids	127
3.4. Self-Organizing Maps	131

4. Current Applications	132
5. Summary	132
12 Clustering: Large Databases	133
1. Introduction	133
2. Motivation	134
3. Requirements for Scalable Clustering	134
4. Major Approaches to Scalable Clustering	135
4.1. The Divide-and-Conquer Approach	135
4.2. Incremental Clustering Approach	135
4.3. Parallel Approach to Clustering	136
5. BIRCH	137
6. DBSCAN	139
7. CURE	140
8. Summary	141
13 Clustering: Categorical Attributes	143
1. Introduction	143
2. Motivation	144
3. ROCK Clustering Algorithm	145
3.1. Computation of Links	146
3.2. Goodness Measure	147
3.3. Miscellaneous Issues	148
3.4. Example	148
4. COOLCAT Clustering Algorithm	149
5. CACTUS Clustering Algorithm	151
6. Summary	152
14 Association Rules: An Overview	153
1. Introduction	153
2. Motivation	154
3. Association Rule Process	154
3.1. Terminology and Notation	154
3.2. From Data to Association Rules	157
4. Large Itemset Discovery Algorithms	158
4.1. Apriori	158
4.2. Sampling	160
4.3. Partitioning	162
5. Summary	163

15 Association Rules: Parallel and Distributed Algorithms	169
1. Introduction	169
2. Motivation	170
3. Parallel and Distributed Algorithms	171
3.1. Data Parallel Algorithms on Distributed Memory Systems	172
3.1.1. Count Distribution (CD)	172
3.2. Task Parallel Algorithms on Distributed Memory Systems	174
3.2.1. Data Distribution (DD)	174
3.2.2. Candidate Distribution (CaD)	174
3.2.3. Intelligent Data Distribution (IDD)	175
3.3. Data Parallel Algorithms on Shared Memory Systems	176
3.3.1. Common Candidate Partitioned Database (CCPD)	176
3.4. Task Parallel Algorithms on Shared Memory Systems	177
3.4.1. Asynchronous Parallel Mining (APM)	177
4. Discussion of Parallel Algorithms	177
5. Summary	179
16 Association Rules: Advanced Techniques and Measures	183
1. Introduction	183
2. Motivation	184
3. Incremental Rules	184
4. Generalized Association Rules	185
5. Quantitative Association Rules	187
6. Correlation Rules	188
7. Measuring the Quality of Association Rules	189
7.1. Lift	189
7.2. Conviction	189
7.3. Chi-Squared Test	190
8. Summary	191

17 Spatial Mining: Techniques and Algorithms	193
1. Introduction and Motivation	193
2. Concept Hierarchies and Generalization	194
3. Spatial Rules	196
4. STING	197
5. Spatial Classification	199
5.1. ID3 Extension	200
5.2. Two-Step Method	201
6. Spatial Clustering	202
6.1. CLARANS	202
6.2. GDBSCAN	203
6.3. DBCLASD	204
7. Summary	204
References	207
Index	219