

# Contents

Preface

page ix

<b>1</b>	<b>Data Mining and Analysis</b> . . . . .	<b>1</b>
1.1	Data Matrix	1
1.2	Attributes	3
1.3	Data: Algebraic and Geometric View	4
1.4	Data: Probabilistic View	14
1.5	Data Mining	25
1.6	Further Reading	30
1.7	Exercises	30
<b>PART ONE: DATA ANALYSIS FOUNDATIONS</b>		
<b>2</b>	<b>Numeric Attributes</b> . . . . .	<b>33</b>
2.1	Univariate Analysis	33
2.2	Bivariate Analysis	42
2.3	Multivariate Analysis	48
2.4	Data Normalization	52
2.5	Normal Distribution	54
2.6	Further Reading	60
2.7	Exercises	60
<b>3</b>	<b>Categorical Attributes</b> . . . . .	<b>63</b>
3.1	Univariate Analysis	63
3.2	Bivariate Analysis	72
3.3	Multivariate Analysis	82
3.4	Distance and Angle	87
3.5	Discretization	89
3.6	Further Reading	91
3.7	Exercises	91
<b>4</b>	<b>Graph Data</b> . . . . .	<b>93</b>
4.1	Graph Concepts	93
4.2	Topological Attributes	97

4.3	Centrality Analysis	102
4.4	Graph Models	112
4.5	Further Reading	132
4.6	Exercises	132
<b>5</b>	<b>Kernel Methods</b>	<b>134</b>
5.1	Kernel Matrix	138
5.2	Vector Kernels	144
5.3	Basic Kernel Operations in Feature Space	148
5.4	Kernels for Complex Objects	154
5.5	Further Reading	161
5.6	Exercises	161
<b>6</b>	<b>High-dimensional Data</b>	<b>163</b>
6.1	High-dimensional Objects	163
6.2	High-dimensional Volumes	165
6.3	Hypersphere Inscribed within Hypercube	168
6.4	Volume of Thin Hypersphere Shell	169
6.5	Diagonals in Hyperspace	171
6.6	Density of the Multivariate Normal	172
6.7	Appendix: Derivation of Hypersphere Volume	175
6.8	Further Reading	180
6.9	Exercises	180
<b>7</b>	<b>Dimensionality Reduction</b>	<b>183</b>
7.1	Background	183
7.2	Principal Component Analysis	187
7.3	Kernel Principal Component Analysis	202
7.4	Singular Value Decomposition	208
7.5	Further Reading	213
7.6	Exercises	214
<b>PART TWO: FREQUENT PATTERN MINING</b>		
<b>8</b>	<b>Itemset Mining</b>	<b>217</b>
8.1	Frequent Itemsets and Association Rules	217
8.2	Itemset Mining Algorithms	221
8.3	Generating Association Rules	234
8.4	Further Reading	236
8.5	Exercises	237
<b>9</b>	<b>Summarizing Itemsets</b>	<b>242</b>
9.1	Maximal and Closed Frequent Itemsets	242
9.2	Mining Maximal Frequent Itemsets: GenMax Algorithm	245
9.3	Mining Closed Frequent Itemsets: Charm Algorithm	248
9.4	Nonderivable Itemsets	250
9.5	Further Reading	256
9.6	Exercises	256

<b>10</b>	<b>Sequence Mining</b> . . . . .	259
10.1	Frequent Sequences	259
10.2	Mining Frequent Sequences	260
10.3	Substring Mining via Suffix Trees	267
10.4	Further Reading	277
10.5	Exercises	277
<b>11</b>	<b>Graph Pattern Mining</b> . . . . .	280
11.1	Isomorphism and Support	280
11.2	Candidate Generation	284
11.3	The gSpan Algorithm	288
11.4	Further Reading	296
11.5	Exercises	297
<b>12</b>	<b>Pattern and Rule Assessment</b> . . . . .	301
12.1	Rule and Pattern Assessment Measures	301
12.2	Significance Testing and Confidence Intervals	316
12.3	Further Reading	328
12.4	Exercises	328
<b>PART THREE: CLUSTERING</b>		
<b>13</b>	<b>Representative-based Clustering</b> . . . . .	333
13.1	K-means Algorithm	333
13.2	Kernel K-means	338
13.3	Expectation-Maximization Clustering	342
13.4	Further Reading	360
13.5	Exercises	361
<b>14</b>	<b>Hierarchical Clustering</b> . . . . .	364
14.1	Preliminaries	364
14.2	Agglomerative Hierarchical Clustering	366
14.3	Further Reading	372
14.4	Exercises and Projects	373
<b>15</b>	<b>Density-based Clustering</b> . . . . .	375
15.1	The DBSCAN Algorithm	375
15.2	Kernel Density Estimation	379
15.3	Density-based Clustering: DENCLUE	385
15.4	Further Reading	390
15.5	Exercises	391
<b>16</b>	<b>Spectral and Graph Clustering</b> . . . . .	394
16.1	Graphs and Matrices	394
16.2	Clustering as Graph Cuts	401
16.3	Markov Clustering	416
16.4	Further Reading	422
16.5	Exercises	423

<b>17</b>	<b>Clustering Validation</b> . . . . .	425
17.1	External Measures	425
17.2	Internal Measures	440
17.3	Relative Measures	448
17.4	Further Reading	461
17.5	Exercises	462
<b>PART FOUR: CLASSIFICATION</b>		
<b>18</b>	<b>Probabilistic Classification</b> . . . . .	467
18.1	Bayes Classifier	467
18.2	Naive Bayes Classifier	473
18.3	$K$ Nearest Neighbors Classifier	477
18.4	Further Reading	479
18.5	Exercises	479
<b>19</b>	<b>Decision Tree Classifier</b> . . . . .	481
19.1	Decision Trees	483
19.2	Decision Tree Algorithm	485
19.3	Further Reading	496
19.4	Exercises	496
<b>20</b>	<b>Linear Discriminant Analysis</b> . . . . .	498
20.1	Optimal Linear Discriminant	498
20.2	Kernel Discriminant Analysis	505
20.3	Further Reading	511
20.4	Exercises	512
<b>21</b>	<b>Support Vector Machines</b> . . . . .	514
21.1	Support Vectors and Margins	514
21.2	SVM: Linear and Separable Case	520
21.3	Soft Margin SVM: Linear and Nonseparable Case	524
21.4	Kernel SVM: Nonlinear Case	530
21.5	SVM Training Algorithms	534
21.6	Further Reading	545
21.7	Exercises	546
<b>22</b>	<b>Classification Assessment</b> . . . . .	548
22.1	Classification Performance Measures	548
22.2	Classifier Evaluation	562
22.3	Bias-Variance Decomposition	572
22.4	Further Reading	581
22.5	Exercises	582
<b>Index</b>		585