

Contents

1	Introduction	1
1.1	It's All About Data	1
1.2	Data Analytics, Data Mining, and Knowledge Discovery	2
	References	3
2	Data and Relations	5
2.1	The Iris Data Set	5
2.2	Data Scales	7
2.3	Set and Matrix Representations	9
2.4	Relations	10
2.5	Dissimilarity Measures	11
2.6	Similarity Measures	13
2.7	Sequence Relations	15
2.8	Sampling and Quantization	17
	Problems	19
	References	20
3	Data Preprocessing	21
3.1	Error Types	21
3.2	Error Handling	24
3.3	Filtering	25
3.4	Data Transformation	30
3.5	Data Merging	33
	Problems	33
	References	33
4	Data Visualization	35
4.1	Diagrams	35
4.2	Principal Component Analysis	37
4.3	Multidimensional Scaling	41
4.4	Sammon Mapping	44

4.5	Auto-Associator	47
4.6	Histograms	48
4.7	Spectral Analysis	50
	Problems	53
	References	54
5	Correlation	55
5.1	Linear Correlation	55
5.2	Correlation and Causality	57
5.3	Chi-Square Test for Independence	58
	Problems	60
	References	61
6	Regression	63
6.1	Linear Regression	63
6.2	Linear Regression with Nonlinear Substitution	67
6.3	Robust Regression	68
6.4	Neural Networks	68
6.5	Radial Basis Function Networks	72
6.6	Cross-Validation	74
6.7	Feature Selection	76
	Problems	77
	References	77
7	Forecasting	79
7.1	Finite State Machines	79
7.2	Recurrent Models	80
7.3	Autoregressive Models	82
	Problems	83
	References	83
8	Classification	85
8.1	Classification Criteria	85
8.2	Naive Bayes Classifier	88
8.3	Linear Discriminant Analysis	91
8.4	Support Vector Machine	93
8.5	Nearest Neighbor Classifier	95
8.6	Learning Vector Quantization	96
8.7	Decision Trees	97
	Problems	100
	References	101

9	Clustering	103
9.1	Cluster Partitions	103
9.2	Sequential Clustering	105
9.3	Prototype-Based Clustering	107
9.4	Fuzzy Clustering	109
9.5	Relational Clustering	114
9.6	Cluster Tendency Assessment	117
9.7	Cluster Validity	118
9.8	Self-Organizing Map	119
	Problems	120
	References	121
A	Brief Review of Some Optimization Methods	123
A.1	Optimization with Derivatives	123
A.2	Gradient Descent	124
A.3	Lagrange Optimization	125
	References	126
	Solutions	127
	List of Symbols	131
	Index	133