

Principles of Data Mining

David Hand
Heikki Mannila
Padhraic Smyth

Contents

<i>List of Tables</i>	xvii
<i>List of Figures</i>	xix
<i>Series Foreword</i>	xxv
<i>Preface</i>	xxvii
1 Introduction	1
1.1 Introduction to Data Mining	1
1.2 The Nature of Data Sets	4
1.3 Types of Structure: Models and Patterns	9
1.4 Data Mining Tasks	11
1.5 Components of Data Mining Algorithms	15
1.5.1 Score Functions	16
1.5.2 Optimization and Search Methods	16
1.5.3 Data Management Strategies	17
1.6 The Interacting Roles of Statistics and Data Mining	18
1.7 Data Mining: Dredging, Snooping, and Fishing	22
1.8 Summary	23
1.9 Further Reading	24
2 Measurement and Data	25
2.1 Introduction	25
2.2 Types of Measurement	26
2.3 Distance Measures	31
2.4 Transforming Data	38

<i>Contents</i>	
2.5 The Form of Data	41
2.6 Data Quality for Individual Measurements	44
2.7 Data Quality for Collections of Data	47
2.8 Conclusion	52
2.9 Further Reading	52
3 Visualizing and Exploring Data	53
3.1 Introduction	53
3.2 Summarizing Data: Some Simple Examples	55
3.3 Tools for Displaying Single Variables	57
3.4 Tools for Displaying Relationships between Two Variables	62
3.5 Tools for Displaying More Than Two Variables	70
3.6 Principal Components Analysis	74
3.7 Multidimensional Scaling	84
3.8 Further Reading	90
4 Data Analysis and Uncertainty	93
4.1 Introduction	93
4.2 Dealing with Uncertainty	94
4.3 Random Variables and Their Relationships	97
4.3.1 Multivariate Random Variables	97
4.4 Samples and Statistical Inference	102
4.5 Estimation	105
4.5.1 Desirable Properties of Estimators	106
4.5.2 Maximum Likelihood Estimation	108
4.5.3 Bayesian Estimation	116
4.6 Hypothesis Testing	124
4.6.1 Classical Hypothesis Testing	124
4.6.2 Hypothesis Testing in Context	130
4.7 Sampling Methods	132
4.8 Conclusion	138
4.9 Further Reading	139
5 A Systematic Overview of Data Mining Algorithms	141
5.1 Introduction	141
5.2 An Example: The CART Algorithm for Building Tree Classifiers	145
5.3 The Reductionist Viewpoint on Data Mining Algorithms	151

<i>Contents</i>		<i>xi</i>
5.3.1 Multilayer Perceptrons for Regression and Classification	153	
5.3.2 The A Priori Algorithm for Association Rule Learning	157	
5.3.3 Vector-Space Algorithms for Text Retrieval	160	
5.4 Discussion	162	
5.5 Further Reading	164	
6 Models and Patterns	165	
6.1 Introduction	165	
6.2 Fundamentals of Modeling	167	
6.3 Model Structures for Prediction	168	
6.3.1 Regression Models with Linear Structure	169	
6.3.2 Local Piecewise Model Structures for Regression	174	
6.3.3 Nonparametric "Memory-Based" Local Models	175	
6.3.4 Stochastic Components of Model Structures	178	
6.3.5 Predictive Models for Classification	180	
6.3.6 An Aside: Selecting a Model of Appropriate Complexity	183	
6.4 Models for Probability Distributions and Density Functions	184	
6.4.1 General Concepts	184	
6.4.2 Mixtures of Parametric Models	185	
6.4.3 Joint Distributions for Unordered Categorical Data	188	
6.4.4 Factorization and Independence in High Dimensions	188	
6.5 The Curse of Dimensionality	193	
6.5.1 Variable Selection for High-Dimensional Data	194	
6.5.2 Transformations for High-Dimensional Data	195	
6.6 Models for Structured Data	197	
6.7 Pattern Structures	203	
6.7.1 Patterns in Data Matrices	203	
6.7.2 Patterns for Strings	206	
6.8 Further Reading	208	
7 Score Functions for Data Mining Algorithms	211	
7.1 Introduction	211	
7.2 Scoring Patterns	212	
7.3 Predictive versus Descriptive Score Functions	215	
7.3.1 Score Functions for Predictive Models	215	
7.3.2 Score Functions for Descriptive Models	217	

	Contents	
7.4	Scoring Models with Different Complexities	220
7.4.1	General Concepts in Comparing Models	220
7.4.2	Bias-Variance Again	221
7.4.3	Score Functions That Penalize Complexity	224
7.4.4	Score Functions Using External Validation	227
7.5	Evaluation of Models and Patterns	229
7.6	Robust Methods	231
7.7	Further Reading	232
8	Search and Optimization Methods	235
8.1	Introduction	235
8.2	Searching for Models and Patterns	238
8.2.1	Background on Search	238
8.2.2	The State-Space Formulation for Search in Data Mining	241
8.2.3	A Simple Greedy Search Algorithm	243
8.2.4	Systematic Search and Search Heuristics	244
8.2.5	Branch-and-Bound	246
8.3	Parameter Optimization Methods	247
8.3.1	Parameter Optimization: Background	247
8.3.2	Closed Form and Linear Algebra Methods	249
8.3.3	Gradient-Based Methods for Optimizing Smooth Functions	250
8.3.4	Univariate Parameter Optimization	251
8.3.5	Multivariate Parameter Optimization	255
8.3.6	Constrained Optimization	259
8.4	Optimization with Missing Data: The EM Algorithm	260
8.5	Online and Single-Scan Algorithms	265
8.6	Stochastic Search and Optimization Techniques	266
8.7	Further Reading	268
9	Descriptive Modeling	271
9.1	Introduction	271
9.2	Describing Data by Probability Distributions and Densities	272
9.2.1	Introduction	272
9.2.2	Score Functions for Estimating Probability Distributions and Densities	274
9.2.3	Parametric Density Models	275
9.2.4	Mixture Distributions and Densities	279

	Contents	xiii
9.2.5	The EM Algorithm for Mixture Models	281
9.2.6	Nonparametric Density Estimation	284
9.2.7	Joint Distributions for Categorical Data	287
9.3	Background on Cluster Analysis	293
9.4	Partition-Based Clustering Algorithms	296
9.4.1	Score Functions for Partition-Based Clustering	297
9.4.2	Basic Algorithms for Partition-Based Clustering	302
9.5	Hierarchical Clustering	308
9.5.1	Agglomerative Methods	311
9.5.2	Divisive Methods	314
9.6	Probabilistic Model-Based Clustering Using Mixture Models	315
9.7	Further Reading	324
10	Predictive Modeling for Classification	327
10.1	A Brief Overview of Predictive Modeling	327
10.2	Introduction to Classification Modeling	329
10.2.1	Discriminative Classification and Decision Boundaries	330
10.2.2	Probabilistic Models for Classification	331
10.2.3	Building Real Classifiers	335
10.3	The Perceptron	339
10.4	Linear Discriminants	341
10.5	Tree Models	343
10.6	Nearest Neighbor Methods	347
10.7	Logistic Discriminant Analysis	352
10.8	The Naive Bayes Model	353
10.9	Other Methods	356
10.10	Evaluating and Comparing Classifiers	359
10.11	Feature Selection for Classification in High Dimensions	362
10.12	Further Reading	363
11	Predictive Modeling for Regression	367
11.1	Introduction	367
11.2	Linear Models and Least Squares Fitting	368
11.2.1	Computational Issues in Fitting the Model	370
11.2.2	A Probabilistic Interpretation of Linear Regression	372
11.2.3	Interpreting the Fitted Model	375
11.2.4	Inference and Generalization	377
11.2.5	Model Search and Model Building	378

	<i>Contents</i>
11.2.6 Diagnostics and Model Inspection	381
11.3 Generalized Linear Models	384
11.4 Artificial Neural Networks	391
11.5 Other Highly Parameterized Models	393
11.5.1 Generalized Additive Models	393
11.5.2 Projection Pursuit Regression	395
11.6 Further Reading	397
12 Data Organization and Databases	399
12.1 Introduction	399
12.2 Memory Hierarchy	400
12.3 Index Structures	402
12.3.1 B-trees	402
12.3.2 Hash Indices	403
12.4 Multidimensional Indexing	404
12.5 Relational Databases	405
12.6 Manipulating Tables	409
12.7 The Structured Query Language (SQL)	413
12.8 Query Execution and Optimization	415
12.9 Data Warehousing and Online Analytical Processing (OLAP)	417
12.10 Data Structures for OLAP	419
12.11 String Databases	420
12.12 Massive Data Sets, Data Management, and Data Mining	421
12.12.1 Force the Data into Main Memory	422
12.12.2 Scalable Versions of Data Mining Algorithms	423
12.12.3 Special-Purpose Algorithms for Disk Access	424
12.12.4 Pseudo Data Sets and Sufficient Statistics	425
12.13 Further Reading	426
13 Finding Patterns and Rules	427
13.1 Introduction	427
13.2 Rule Representations	428
13.3 Frequent Itemsets and Association Rules	429
13.3.1 Introduction	429
13.3.2 Finding Frequent Sets and Association Rules	433
13.4 Generalizations	435
13.5 Finding Episodes from Sequences	436

	<i>Contents</i>	<i>xv</i>
13.6 Selective Discovery of Patterns and Rules	438	
13.6.1 Introduction	438	
13.6.2 Heuristic Search for Finding Patterns	439	
13.6.3 Criteria for Interestingness	440	
13.7 From Local Patterns to Global Models	442	
13.8 Predictive Rule Induction	443	
13.9 Further Reading	447	
14 Retrieval by Content	449	
14.1 Introduction	449	
14.2 Evaluation of Retrieval Systems	452	
14.2.1 The Difficulty of Evaluating Retrieval Performance	452	
14.2.2 Precision versus Recall	453	
14.2.3 Precision and Recall in Practice	456	
14.3 Text Retrieval	456	
14.3.1 Representation of Text	457	
14.3.2 Matching Queries and Documents	461	
14.3.3 Latent Semantic Indexing	465	
14.3.4 Document and Text Classification	469	
14.4 Modeling Individual Preferences	470	
14.4.1 Relevance Feedback	470	
14.4.2 Automated Recommender Systems	471	
14.5 Image Retrieval	472	
14.5.1 Image Understanding	473	
14.5.2 Image Representation	473	
14.5.3 Image Queries	474	
14.5.4 Image Invariants	475	
14.5.5 Generalizations of Image Retrieval	476	
14.6 Time Series and Sequence Retrieval	476	
14.6.1 Global Models for Time Series Data	478	
14.6.2 Structure and Shape in Time Series	480	
14.7 Summary	481	
14.8 Further Reading	482	
Appendix: Random Variables	485	
References	491	
Index	525	