

Contents

Preface	vii
1 Introduction	1
1.1 What Is Data Mining?	2
1.2 Motivating Challenges	4
1.3 The Origins of Data Mining	6
1.4 Data Mining Tasks	7
1.5 Scope and Organization of the Book	11
1.6 Bibliographic Notes	13
1.7 Exercises	16
2 Data	19
2.1 Types of Data	22
2.1.1 Attributes and Measurement	23
2.1.2 Types of Data Sets	29
2.2 Data Quality	36
2.2.1 Measurement and Data Collection Issues	37
2.2.2 Issues Related to Applications	43
2.3 Data Preprocessing	44
2.3.1 Aggregation	45
2.3.2 Sampling	47
2.3.3 Dimensionality Reduction	50
2.3.4 Feature Subset Selection	52
2.3.5 Feature Creation	55
2.3.6 Discretization and Binarization	57
2.3.7 Variable Transformation	63
2.4 Measures of Similarity and Dissimilarity	65
2.4.1 Basics	66
2.4.2 Similarity and Dissimilarity between Simple Attributes	67
2.4.3 Dissimilarities between Data Objects	69
2.4.4 Similarities between Data Objects	72

xiv Contents

2.4.5	Examples of Proximity Measures	73
2.4.6	Issues in Proximity Calculation	80
2.4.7	Selecting the Right Proximity Measure	83
2.5	Bibliographic Notes	84
2.6	Exercises	88
3	Exploring Data	97
3.1	The Iris Data Set	98
3.2	Summary Statistics	98
3.2.1	Frequencies and the Mode	99
3.2.2	Percentiles	100
3.2.3	Measures of Location: Mean and Median	101
3.2.4	Measures of Spread: Range and Variance	102
3.2.5	Multivariate Summary Statistics	104
3.2.6	Other Ways to Summarize the Data	105
3.3	Visualization	105
3.3.1	Motivations for Visualization	105
3.3.2	General Concepts	106
3.3.3	Techniques	110
3.3.4	Visualizing Higher-Dimensional Data	124
3.3.5	Do's and Don'ts	130
3.4	OLAP and Multidimensional Data Analysis	131
3.4.1	Representing Iris Data as a Multidimensional Array	131
3.4.2	Multidimensional Data: The General Case	133
3.4.3	Analyzing Multidimensional Data	135
3.4.4	Final Comments on Multidimensional Data Analysis	139
3.5	Bibliographic Notes	139
3.6	Exercises	141
4	Classification:	
	Basic Concepts, Decision Trees, and Model Evaluation	145
4.1	Preliminaries	146
4.2	General Approach to Solving a Classification Problem	148
4.3	Decision Tree Induction	150
4.3.1	How a Decision Tree Works	150
4.3.2	How to Build a Decision Tree	151
4.3.3	Methods for Expressing Attribute Test Conditions	155
4.3.4	Measures for Selecting the Best Split	158
4.3.5	Algorithm for Decision Tree Induction	164
4.3.6	An Example: Web Robot Detection	166

4.3.7	Characteristics of Decision Tree Induction	168
4.4	Model Overfitting	172
4.4.1	Overfitting Due to Presence of Noise	175
4.4.2	Overfitting Due to Lack of Representative Samples . . .	177
4.4.3	Overfitting and the Multiple Comparison Procedure . .	178
4.4.4	Estimation of Generalization Errors	179
4.4.5	Handling Overfitting in Decision Tree Induction	184
4.5	Evaluating the Performance of a Classifier	186
4.5.1	Holdout Method	186
4.5.2	Random Subsampling	187
4.5.3	Cross-Validation	187
4.5.4	Bootstrap	188
4.6	Methods for Comparing Classifiers	188
4.6.1	Estimating a Confidence Interval for Accuracy	189
4.6.2	Comparing the Performance of Two Models	191
4.6.3	Comparing the Performance of Two Classifiers	192
4.7	Bibliographic Notes	193
4.8	Exercises	198
5	Classification: Alternative Techniques	207
5.1	Rule-Based Classifier	207
5.1.1	How a Rule-Based Classifier Works	209
5.1.2	Rule-Ordering Schemes	211
5.1.3	How to Build a Rule-Based Classifier	212
5.1.4	Direct Methods for Rule Extraction	213
5.1.5	Indirect Methods for Rule Extraction	221
5.1.6	Characteristics of Rule-Based Classifiers	223
5.2	Nearest-Neighbor classifiers	223
5.2.1	Algorithm	225
5.2.2	Characteristics of Nearest-Neighbor Classifiers	226
5.3	Bayesian Classifiers	227
5.3.1	Bayes Theorem	228
5.3.2	Using the Bayes Theorem for Classification	229
5.3.3	Naïve Bayes Classifier	231
5.3.4	Bayes Error Rate	238
5.3.5	Bayesian Belief Networks	240
5.4	Artificial Neural Network (ANN)	246
5.4.1	Perceptron	247
5.4.2	Multilayer Artificial Neural Network	251
5.4.3	Characteristics of ANN	255

xvi Contents

5.5	Support Vector Machine (SVM)	256
5.5.1	Maximum Margin Hyperplanes	256
5.5.2	Linear SVM: Separable Case	259
5.5.3	Linear SVM: Nonseparable Case	266
5.5.4	Nonlinear SVM	270
5.5.5	Characteristics of SVM	276
5.6	Ensemble Methods	276
5.6.1	Rationale for Ensemble Method	277
5.6.2	Methods for Constructing an Ensemble Classifier	278
5.6.3	Bias-Variance Decomposition	281
5.6.4	Bagging	283
5.6.5	Boosting	285
5.6.6	Random Forests	290
5.6.7	Empirical Comparison among Ensemble Methods	294
5.7	Class Imbalance Problem	294
5.7.1	Alternative Metrics	295
5.7.2	The Receiver Operating Characteristic Curve	298
5.7.3	Cost-Sensitive Learning	302
5.7.4	Sampling-Based Approaches	305
5.8	Multiclass Problem	306
5.9	Bibliographic Notes	309
5.10	Exercises	315
6	Association Analysis: Basic Concepts and Algorithms	327
6.1	Problem Definition	328
6.2	Frequent Itemset Generation	332
6.2.1	The <i>Apriori</i> Principle	333
6.2.2	Frequent Itemset Generation in the <i>Apriori</i> Algorithm	335
6.2.3	Candidate Generation and Pruning	338
6.2.4	Support Counting	342
6.2.5	Computational Complexity	345
6.3	Rule Generation	349
6.3.1	Confidence-Based Pruning	350
6.3.2	Rule Generation in <i>Apriori</i> Algorithm	350
6.3.3	An Example: Congressional Voting Records	352
6.4	Compact Representation of Frequent Itemsets	353
6.4.1	Maximal Frequent Itemsets	354
6.4.2	Closed Frequent Itemsets	355
6.5	Alternative Methods for Generating Frequent Itemsets	359
6.6	FP-Growth Algorithm	363

6.6.1	FP-Tree Representation	363
6.6.2	Frequent Itemset Generation in FP-Growth Algorithm	366
6.7	Evaluation of Association Patterns	370
6.7.1	Objective Measures of Interestingness	371
6.7.2	Measures beyond Pairs of Binary Variables	382
6.7.3	Simpson’s Paradox	384
6.8	Effect of Skewed Support Distribution	386
6.9	Bibliographic Notes	390
6.10	Exercises	404
7	Association Analysis: Advanced Concepts	415
7.1	Handling Categorical Attributes	415
7.2	Handling Continuous Attributes	418
7.2.1	Discretization-Based Methods	418
7.2.2	Statistics-Based Methods	422
7.2.3	Non-discretization Methods	424
7.3	Handling a Concept Hierarchy	426
7.4	Sequential Patterns	429
7.4.1	Problem Formulation	429
7.4.2	Sequential Pattern Discovery	431
7.4.3	Timing Constraints	436
7.4.4	Alternative Counting Schemes	439
7.5	Subgraph Patterns	442
7.5.1	Graphs and Subgraphs	443
7.5.2	Frequent Subgraph Mining	444
7.5.3	<i>Apriori</i> -like Method	447
7.5.4	Candidate Generation	448
7.5.5	Candidate Pruning	453
7.5.6	Support Counting	457
7.6	Infrequent Patterns	457
7.6.1	Negative Patterns	458
7.6.2	Negatively Correlated Patterns	458
7.6.3	Comparisons among Infrequent Patterns, Negative Pat- terns, and Negatively Correlated Patterns	460
7.6.4	Techniques for Mining Interesting Infrequent Patterns	461
7.6.5	Techniques Based on Mining Negative Patterns	463
7.6.6	Techniques Based on Support Expectation	465
7.7	Bibliographic Notes	469
7.8	Exercises	473

8	Cluster Analysis: Basic Concepts and Algorithms	487
8.1	Overview	490
8.1.1	What Is Cluster Analysis?	490
8.1.2	Different Types of Clusterings	491
8.1.3	Different Types of Clusters	493
8.2	K-means	496
8.2.1	The Basic K-means Algorithm	497
8.2.2	K-means: Additional Issues	506
8.2.3	Bisecting K-means	508
8.2.4	K-means and Different Types of Clusters	510
8.2.5	Strengths and Weaknesses	510
8.2.6	K-means as an Optimization Problem	513
8.3	Agglomerative Hierarchical Clustering	515
8.3.1	Basic Agglomerative Hierarchical Clustering Algorithm	516
8.3.2	Specific Techniques	518
8.3.3	The Lance-Williams Formula for Cluster Proximity	524
8.3.4	Key Issues in Hierarchical Clustering	524
8.3.5	Strengths and Weaknesses	526
8.4	DBSCAN	526
8.4.1	Traditional Density: Center-Based Approach	527
8.4.2	The DBSCAN Algorithm	528
8.4.3	Strengths and Weaknesses	530
8.5	Cluster Evaluation	532
8.5.1	Overview	533
8.5.2	Unsupervised Cluster Evaluation Using Cohesion and Separation	536
8.5.3	Unsupervised Cluster Evaluation Using the Proximity Matrix	542
8.5.4	Unsupervised Evaluation of Hierarchical Clustering	544
8.5.5	Determining the Correct Number of Clusters	546
8.5.6	Clustering Tendency	547
8.5.7	Supervised Measures of Cluster Validity	548
8.5.8	Assessing the Significance of Cluster Validity Measures	553
8.6	Bibliographic Notes	555
8.7	Exercises	559
9	Cluster Analysis: Additional Issues and Algorithms	569
9.1	Characteristics of Data, Clusters, and Clustering Algorithms	570
9.1.1	Example: Comparing K-means and DBSCAN	570
9.1.2	Data Characteristics	571

9.1.3	Cluster Characteristics	573
9.1.4	General Characteristics of Clustering Algorithms	575
9.2	Prototype-Based Clustering	577
9.2.1	Fuzzy Clustering	577
9.2.2	Clustering Using Mixture Models	583
9.2.3	Self-Organizing Maps (SOM)	594
9.3	Density-Based Clustering	600
9.3.1	Grid-Based Clustering	601
9.3.2	Subspace Clustering	604
9.3.3	DENCLUE: A Kernel-Based Scheme for Density-Based Clustering	608
9.4	Graph-Based Clustering	612
9.4.1	Sparsification	613
9.4.2	Minimum Spanning Tree (MST) Clustering	614
9.4.3	OPOSSUM: Optimal Partitioning of Sparse Similarities Using METIS	616
9.4.4	Chameleon: Hierarchical Clustering with Dynamic Modeling	616
9.4.5	Shared Nearest Neighbor Similarity	622
9.4.6	The Jarvis-Patrick Clustering Algorithm	625
9.4.7	SNN Density	627
9.4.8	SNN Density-Based Clustering	629
9.5	Scalable Clustering Algorithms	630
9.5.1	Scalability: General Issues and Approaches	630
9.5.2	BIRCH	633
9.5.3	CURE	635
9.6	Which Clustering Algorithm?	639
9.7	Bibliographic Notes	643
9.8	Exercises	647
10	Anomaly Detection	651
10.1	Preliminaries	653
10.1.1	Causes of Anomalies	653
10.1.2	Approaches to Anomaly Detection	654
10.1.3	The Use of Class Labels	655
10.1.4	Issues	656
10.2	Statistical Approaches	658
10.2.1	Detecting Outliers in a Univariate Normal Distribution	659
10.2.2	Outliers in a Multivariate Normal Distribution	661
10.2.3	A Mixture Model Approach for Anomaly Detection	662

xx Contents

10.2.4	Strengths and Weaknesses	665
10.3	Proximity-Based Outlier Detection	666
10.3.1	Strengths and Weaknesses	666
10.4	Density-Based Outlier Detection	668
10.4.1	Detection of Outliers Using Relative Density	669
10.4.2	Strengths and Weaknesses	670
10.5	Clustering-Based Techniques	671
10.5.1	Assessing the Extent to Which an Object Belongs to a Cluster	672
10.5.2	Impact of Outliers on the Initial Clustering	674
10.5.3	The Number of Clusters to Use	674
10.5.4	Strengths and Weaknesses	674
10.6	Bibliographic Notes	675
10.7	Exercises	680
Appendix A Linear Algebra		685
A.1	Vectors	685
A.1.1	Definition	685
A.1.2	Vector Addition and Multiplication by a Scalar	685
A.1.3	Vector Spaces	687
A.1.4	The Dot Product, Orthogonality, and Orthogonal Projections	688
A.1.5	Vectors and Data Analysis	690
A.2	Matrices	691
A.2.1	Matrices: Definitions	691
A.2.2	Matrices: Addition and Multiplication by a Scalar	692
A.2.3	Matrices: Multiplication	693
A.2.4	Linear Transformations and Inverse Matrices	695
A.2.5	Eigenvalue and Singular Value Decomposition	697
A.2.6	Matrices and Data Analysis	699
A.3	Bibliographic Notes	700
Appendix B Dimensionality Reduction		701
B.1	PCA and SVD	701
B.1.1	Principal Components Analysis (PCA)	701
B.1.2	SVD	706
B.2	Other Dimensionality Reduction Techniques	708
B.2.1	Factor Analysis	708
B.2.2	Locally Linear Embedding (LLE)	710
B.2.3	Multidimensional Scaling, FastMap, and ISOMAP	712

B.2.4	Common Issues	715
B.3	Bibliographic Notes	716
Appendix C Probability and Statistics		719
C.1	Probability	719
C.1.1	Expected Values	722
C.2	Statistics	723
C.2.1	Point Estimation	724
C.2.2	Central Limit Theorem	724
C.2.3	Interval Estimation	725
C.3	Hypothesis Testing	726
Appendix D Regression		729
D.1	Preliminaries	729
D.2	Simple Linear Regression	730
D.2.1	Least Square Method	731
D.2.2	Analyzing Regression Errors	733
D.2.3	Analyzing Goodness of Fit	735
D.3	Multivariate Linear Regression	736
D.4	Alternative Least-Square Regression Methods	737
Appendix E Optimization		739
E.1	Unconstrained Optimization	739
E.1.1	Numerical Methods	742
E.2	Constrained Optimization	746
E.2.1	Equality Constraints	746
E.2.2	Inequality Constraints	747
Author Index		750
Subject Index		758
Copyright Permissions		769