Principles of Data Mining (draft version)

David Hand Heikki Mannila Padhraic Smyth

Table of Contents

Chapter 1.
Introduction

1.1 Introduction to Data Mining
1.2 The Nature of Data Sets
1.3 Types of Structure: Models and Patterns
1.4 Data Mining Tasks
1.5 Components of Data Mining Algorithms
1.6 The Interacting Roles of Statistics and Data Mining
1.7 Data Mining: Dredging, Snooping, and Fishing
1.8 Summary
1.9 Further Reading

Chapter 2.
Measurement and Data

2.1 Introduction
2.2 Types of Measurement
2.3 Distance Measures
2.4 Transforming Data
2.5 The Form of Data
2.6 Data Quality for Individual Measurements
2.7 Data Quality for Collections of Data
2.8 Conclusion
2.9 Further Reading

Chapter 3.
Visualizing and Exploring Data

3.1 Introduction
3.2 Summarizing Data: Some Simple Examples
3.3 Tools for Displaying Single Variables
3.4 Tools for Displaying Relationships Between Two Variables
3.5 Tools for Displaying More Than Two Variables
3.6 Principal Components Analysis
3.7 Multidimensional Scaling
3.8 Further Reading

Chapter 4.
Data Analysis and Uncertainty

4.1 Introduction
4.2 Dealing with Uncertainty
4.3 Random Variables and their Relationships
4.4 Samples and Statistical Inference
4.5 Estimation
4.6 Hypothesis Testing
4.7 Sampling Methods
4.8 Conclusion
4.9 Further Reading

Chapter 5
A Systematic Overview of Data Mining Algorithms

5.1 Introduction
5.2 An Example: The CART Algorithm for Building Tree Classifiers
5.3 The Reductionist Viewpoint on Data Mining Algorithms
5.3.1 Multilayer Perceptrons for Regression and Classification
5.3.2 The A Priori Algorithm for Association Rule Learning
5.3.3 Vector-Space Algorithms for Text Retrieval
5.4 Discussion
5.5 Further Reading
5.6 Bibliography

Chapter 6
Models and Patterns

6.1 Introduction
6.2 Fundamentals of Modeling
6.3 Model Structures for Prediction
6.3.1 Regression Models with Linear Structure
6.3.2 Local Piecewise Model Structures for Regression
6.3.3 Nonparametric "Memory-Based" Local Models
6.3.4 Stochastic Components of Model Structures
6.3.5 Predictive Models for Classification
6.3.6 An Aside: Selecting a Model of Appropriate Complexity
6.4 Models for Probability Distributions and Density Functions
6.4.1 General Concepts
6.4.2 Mixtures of Parametric Models
6.4.3 Joint Distributions for Unordered Categorical Data

6.4.4 Factorization and Independence in High Dimensions
6.5 The Curse of Dimensionality
6.5.1 Variable Selection for High-Dimensional Data
6.5.2 Transformations for High-Dimensional Data
6.6 Models for Structured Data
6.7 Pattern Structures
6.7.1 Patterns in Data Matrices
6.7.2 Patterns for Strings
6.8 Further Reading
References

Chapter 7
Score Functions for Data Mining Algorithms

7.1 Introduction
7.2 Scoring Patterns
7.3 Predictive versus Descriptive Score Functions
7.3.1 Score Functions for Predictive Models
7.3.2 Score Functions for Descriptive Models
7.4 Scoring Models with Different Complexities
7.4.1 General Concepts in Comparing Models
7.4.2 Bias-Variance Again
7.4.3 Score Functions That Penalize Complexity
7.4.4 Score Functions Using External Validation
7.5 Evaluation of Models and Patterns
7.6 Robust Methods
7.7 Further Reading
7.8 Bibliography

Chapter 8
Search and Optimization Methods

8.1 Introduction
8.2 Searching for Models and Patterns
8.2.1 Background on Search
8.2.2 The State-Space Formulation for Search in Data Mining
8.2.3 A Simple Greedy Search Algorithm
8.2.4 Systematic Search and Search Heuristics
8.2.5 Branch-and-Bound
8.3 Parameter Optimization Methods
8.3.1 Parameter Optimization: Background
8.3.2 Closed Form and Linear Algebra Methods

8.3.3 Gradient-Based Methods for Optimizing Smooth Functions
8.3.4 Univariate Parameter Optimization
8.3.5 Multivariate Parameter Optimization
8.3.6 Constrained Optimization
8.4 Optimization with Missing Data: The EM Algorithm
8.5 Online and Single-Scan Algorithms
8.6 Stochastic Search and Optimization Techniques
8.7 Further Reading

Chapter 9
Descriptive Modeling

9.1 Introduction
9.2 Describing Data by Probability Distributions and Densities
9.2.1 Introduction
9.2.2 Score Functions for Estimating Probability Distributions and Densities
9.2.3 Parametric Density Models
9.2.4 Mixture Distributions and Densities
9.2.5 The EM Algorithm for Mixture Models
9.2.6 Nonparametric Density Estimation
9.2.7 Joint Distributions for Categorical Data
9.3 Background on Cluster Analysis
9.4 Partition-Based Clustering Algorithms
9.4.1 Score Functions for Partition-Based Clustering
9.4.2 Basic Algorithms for Partition-Based Clustering
9.5 Hierarchical Clustering
9.5.1 Agglomerative Methods
9.5.2 Divisive Methods
9.6 Probabilistic Model-Based Clustering Using Mixture Models
9.7 Further Reading
References

Chapter 10
Predictive Modeling for Classification

10.1 A Brief Overview of Predictive Modeling
10.2 Introduction to Classification Modeling
10.2.1 Discriminative Classification and Decision Boundaries
10.2.2 Probabilistic Models for Classification
10.2.3 Building Real Classifiers
10.3 The Perceptron
10.4 Linear Discriminants
10.5 Tree Models
10.6 Nearest Neighbor Methods
10.7 Logistic Discriminant Analysis
10.8 The Naive Bayes Model
10.9 Other Methods
10.10 Evaluating and Comparing Classifiers
10.11 Feature Selection for Classification in High Dimensions
10.12 Further Reading
References

Chapter 11
Predictive Modeling for Regression

11.1 Introduction
11.2 Linear Models and Least Squares Fitting
11.2.1 Computational Issues in Fitting the Model
11.2.2 A Probabilistic Interpretation of Linear Regression
11.2.3 Interpreting the Fitted Model
11.2.4 Inference and Generalization
11.2.5 Model Search and Model Building
11.2.6 Diagnostics and Model Inspection
11.3 Generalized Linear Models
11.4 Artificial Neural Networks
11.5 Other Highly Parameterized Models
11.5.1 Generalized Additive Models
11.5.2 Projection Pursuit Regression
11.6 Further Reading
References

Chapter 12
Data Organization and Databases

12.1 Introduction
12.2 Memory Hierarchy
12.3 Index Structures
12.3.1 B-trees
12.3.2 Hash Indices
12.4 Multidimensional Indexing
12.5 Relational Databases
12.6 Relational Algebra
12.7 The Structured Query Language (SQL)
12.8 Query Execution and Optimization
12.9 Data Warehousing and On-Line Analytical Processing (OLAP)
12.10 Data Structures for OLAP
12.11 String Databases
12.12 Very Large Data Sets, Data Management, and Data Mining
12.12.1 Force the Data into Main Memory
12.12.2 Scalable Versions of Data Mining Algorithms
12.12.3 Special-Purpose Algorithms for Disk Access
12.12.4 Pseudo Data Sets and Sufficient Statistics
12.13 Further Reading
References

Chapter 13
Finding Patterns and Rules

13.1 Introduction
13.2 Rule Representations
13.3 Frequent Itemsets and Association Rules
13.3.1 Introduction
13.3.2 Finding Frequent Sets and Association Rules
13.4 Generalizations
13.5 Finding Episodes from Sequences
13.6 Selective Discovery of Patterns and Rules
13.6.1 Introduction
13.6.2 Heuristic Search for Finding Patterns
13.6.3 Criteria for Interestingness
13.7 From Local Patterns to Global Models
13.8 Predictive Rule Induction
13.9 Further Reading
References

Chapter 14
Retrieval by Content

14.1 Introduction
14.2 Evaluation of Retrieval Systems
14.2.1 The Difficulty of Evaluating Retrieval Performance
14.2.2 Precision versus Recall
14.2.3 Precision and Recall in Practice
14.3 Text Retrieval
14.3.1 Representation of Text
14.3.2 Matching Queries and Documents
14.3.3 Latent Semantic Indexing
14.3.4 Relevance Feedback
14.4 Automated Recommender Systems
14.5 Document and Text Classification
14.6 Image Retrieval
14.6.1 Image Understanding
14.6.2 Image Representation
14.6.3 Image Queries
14.6.4 Image Invariants
14.6.5 Generalizations of Image Retrieval
14.7 Time Series and Sequence Retrieval
14.7.1 Global Models for Time Series Data
14.7.2 Structure and Shape in Time Series
14.8 Summary
14.9 Further Reading
References