

CONTENTS

PREFACE

xi

CHAPTER 1 AN INTRODUCTION TO DATA MINING

1

- 1.1 What is Data Mining? 1
- 1.2 Wanted: Data Miners 2
- 1.3 The Need for Human Direction of Data Mining 3
- 1.4 The Cross-Industry Standard Practice for Data Mining 4
 - 1.4.1 Crisp-DM: The Six Phases 5
- 1.5 Fallacies of Data Mining 6
- 1.6 What Tasks Can Data Mining Accomplish? 8
 - 1.6.1 Description 8
 - 1.6.2 Estimation 8
 - 1.6.3 Prediction 10
 - 1.6.4 Classification 10
 - 1.6.5 Clustering 12
 - 1.6.6 Association 14
- References 14
- Exercises 15

CHAPTER 2 DATA PREPROCESSING

16

- 2.1 Why do We Need to Preprocess the Data? 17
- 2.2 Data Cleaning 17
- 2.3 Handling Missing Data 19
- 2.4 Identifying Misclassifications 22
- 2.5 Graphical Methods for Identifying Outliers 22
- 2.6 Measures of Center and Spread 23
- 2.7 Data Transformation 26
- 2.8 Min-Max Normalization 26
- 2.9 Z-Score Standardization 27
- 2.10 Decimal Scaling 28
- 2.11 Transformations to Achieve Normality 28
- 2.12 Numerical Methods for Identifying Outliers 35
- 2.13 Flag Variables 36
- 2.14 Transforming Categorical Variables into Numerical Variables 37
- 2.15 Binning Numerical Variables 38
- 2.16 Reclassifying Categorical Variables 39
- 2.17 Adding an Index Field 39
- 2.18 Removing Variables that are Not Useful 39
- 2.19 Variables that Should Probably Not Be Removed 40
- 2.20 Removal of Duplicate Records 41