

CONTENTS

PREFACE

xi

1 DIMENSION REDUCTION METHODS

1

Need for Dimension Reduction in Data Mining

1

Principal Components Analysis

2

Applying Principal Components Analysis to the *Houses* Data Set

5

How Many Components Should We Extract?

9

Profiling the Principal Components

13

Communalities

15

Validation of the Principal Components

17

Factor Analysis

18

Applying Factor Analysis to the *Adult* Data Set

18

Factor Rotation

20

User-Defined Composites

23

Example of a User-Defined Composite

24

Summary

25

References

28

Exercises

28

2 REGRESSION MODELING

33

Example of Simple Linear Regression

34

Least-Squares Estimates

36

Coefficient of Determination

39

Standard Error of the Estimate

43

Correlation Coefficient

45

ANOVA Table

46

Outliers, High Leverage Points, and Influential Observations

48

Regression Model

55

Inference in Regression

57

t -Test for the Relationship Between x and y

58

Confidence Interval for the Slope of the Regression Line

60

Confidence Interval for the Mean Value of y Given x

60

Prediction Interval for a Randomly Chosen Value of y Given x

61

Verifying the Regression Assumptions

63

Example: *Baseball* Data Set

68

Example: *California* Data Set

74

Transformations to Achieve Linearity

79

Box-Cox Transformations

83

Summary

84

References

86

Exercises

86

3	MULTIPLE REGRESSION AND MODEL BUILDING	93
	Example of Multiple Regression	93
	Multiple Regression Model	99
	Inference in Multiple Regression	100
	t -Test for the Relationship Between y and x_i	101
	F -Test for the Significance of the Overall Regression Model	102
	Confidence Interval for a Particular Coefficient	104
	Confidence Interval for the Mean Value of y Given $x_1, x_2, \dots, x_m$	105
	Prediction Interval for a Randomly Chosen Value of y Given $x_1, x_2, \dots, x_m$	105
	Regression with Categorical Predictors	105
	Adjusting R^2 : Penalizing Models for Including Predictors That Are Not Useful	113
	Sequential Sums of Squares	115
	Multicollinearity	116
	Variable Selection Methods	123
	Partial F -Test	123
	Forward Selection Procedure	125
	Backward Elimination Procedure	125
	Stepwise Procedure	126
	Best Subsets Procedure	126
	All-Possible-Subsets Procedure	126
	Application of the Variable Selection Methods	127
	Forward Selection Procedure Applied to the <i>Cereals</i> Data Set	127
	Backward Elimination Procedure Applied to the <i>Cereals</i> Data Set	129
	Stepwise Selection Procedure Applied to the <i>Cereals</i> Data Set	131
	Best Subsets Procedure Applied to the <i>Cereals</i> Data Set	131
	Mallows' C_p Statistic	131
	Variable Selection Criteria	135
	Using the Principal Components as Predictors	142
	Summary	147
	References	149
	Exercises	149
4	LOGISTIC REGRESSION	155
	Simple Example of Logistic Regression	156
	Maximum Likelihood Estimation	158
	Interpreting Logistic Regression Output	159
	Inference: Are the Predictors Significant?	160
	Interpreting a Logistic Regression Model	162
	Interpreting a Model for a Dichotomous Predictor	163
	Interpreting a Model for a Polychotomous Predictor	166
	Interpreting a Model for a Continuous Predictor	170
	Assumption of Linearity	174
	Zero-Cell Problem	177
	Multiple Logistic Regression	179
	Introducing Higher-Order Terms to Handle Nonlinearity	183
	Validating the Logistic Regression Model	189
	WEKA: Hands-on Analysis Using Logistic Regression	194
	Summary	197

References	199
Exercises	199
5 NAIVE BAYES ESTIMATION AND BAYESIAN NETWORKS	204
Bayesian Approach	204
Maximum a Posteriori Classification	206
Posterior Odds Ratio	210
Balancing the Data	212
Naive Bayes Classification	215
Numeric Predictors	219
WEKA: Hands-on Analysis Using Naive Bayes	223
Bayesian Belief Networks	227
Clothing Purchase Example	227
Using the Bayesian Network to Find Probabilities	229
WEKA: Hands-On Analysis Using the Bayes Net Classifier	232
Summary	234
References	236
Exercises	237
6 GENETIC ALGORITHMS	240
Introduction to Genetic Algorithms	240
Basic Framework of a Genetic Algorithm	241
Simple Example of a Genetic Algorithm at Work	243
Modifications and Enhancements: Selection	245
Modifications and Enhancements: Crossover	247
Multipoint Crossover	247
Uniform Crossover	247
Genetic Algorithms for Real-Valued Variables	248
Single Arithmetic Crossover	248
Simple Arithmetic Crossover	248
Whole Arithmetic Crossover	249
Discrete Crossover	249
Normally Distributed Mutation	249
Using Genetic Algorithms to Train a Neural Network	249
WEKA: Hands-on Analysis Using Genetic Algorithms	252
Summary	261
References	262
Exercises	263
7 CASE STUDY: MODELING RESPONSE TO DIRECT MAIL MARKETING	265
Cross-Industry Standard Process for Data Mining	265
Business Understanding Phase	267
Direct Mail Marketing Response Problem	267
Building the Cost/Benefit Table	267
Data Understanding and Data Preparation Phases	270
Clothing Store Data Set	270
Transformations to Achieve Normality or Symmetry	272
Standardization and Flag Variables	276

Deriving New Variables	277
Exploring the Relationships Between the Predictors and the Response	278
Investigating the Correlation Structure Among the Predictors	286
Modeling and Evaluation Phases	289
Principal Components Analysis	292
Cluster Analysis: BIRCH Clustering Algorithm	294
Balancing the Training Data Set	298
Establishing the Baseline Model Performance	299
Model Collection A: Using the Principal Components	300
Overbalancing as a Surrogate for Misclassification Costs	302
Combining Models: Voting	304
Model Collection B: Non-PCA Models	306
Combining Models Using the Mean Response Probabilities	308
Summary	312
References	316

INDEX