

Data Analysis and Data Mining

An Introduction

ADELCHI AZZALINI

AND

BRUNO SCARPA

CONTENTS

Preface vii

Preface to the English Edition ix

1. Introduction 1

1.1. New problems and new opportunities 1

1.2. All models are wrong 9

1.3. A matter of style 12

2. A–B–C 15

2.1. Old friends: Linear models 15

2.2. Computational aspects 30

2.3. Likelihood 33

2.4. Logistic regression and GLM 40

Exercises 44

3. Optimism, Conflicts, and Trade-offs 45

3.1. Matching the conceptual frame and real life 45

3.2. A simple prototype problem 46

3.3. If we knew $f(x)$. . . 47

3.4. But as we do not know $f(x)$. . . 51

3.5. Methods for model selection 52

3.6. Reduction of dimensions and selection of most appropriate model 58

Exercises 66

4. Prediction of Quantitative Variables 68

4.1. Nonparametric estimation: Why? 68

4.2. Local regression 69

4.3. The curse of dimensionality 78

4.4. Splines 79

4.5. Additive models and GAM 89

4.6. Projection pursuit 93

4.7. Inferential aspects 94

4.8. Regression trees 98

4.9. Neural networks 106

4.10. Case studies 111

Exercises 132

5. Methods of Classification	134
5.1. Prediction of categorical variables	134
5.2. An introduction based on a marketing problem	135
5.3. Extension to several categories	142
5.4. Classification via linear regression	149
5.5. Discriminant analysis	154
5.6. Some nonparametric methods	159
5.7. Classification trees	164
5.8. Some other topics	168
5.9. Combination of classifiers	176
5.10. Case studies	183
Exercises	210
6. Methods of Internal Analysis	212
6.1. Cluster analysis	212
6.2. Associations among variables	222
6.3. Case study: Web usage mining	232
Appendix A Complements of Mathematics and Statistics	240
A.1. Concepts on linear algebra	240
A.2. Concepts of probability theory	241
A.3. Concepts of linear models	246
Appendix B Data Sets	254
B.1. Simulated data	254
B.2. Car data	254
B.3. Brazilian bank data	255
B.4. Data for telephone company customers	256
B.5. Insurance data	257
B.6. Choice of fruit juice data	258
B.7. Customer satisfaction	259
B.8. Web usage data	261
Appendix C Symbols and Acronyms	263
References	265
Author Index	269
Subject Index	271