



# STATISTICAL PATTERN RECOGNITION

Third Edition

Andrew R. Webb  
Keith D. Copsey

 **WILEY**



# Statistical Pattern Recognition



# Statistical Pattern Recognition

Third Edition

**Andrew R. Webb • Keith D. Copsey**

*Mathematics and Data Analysis Consultancy, Malvern, UK*



A John Wiley & Sons, Ltd., Publication

This edition first published 2011  
© 2011 John Wiley & Sons, Ltd

*Registered office*

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at [www.wiley.com](http://www.wiley.com).

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

*Library of Congress Cataloging-in-Publication Data*

Webb, A. R. (Andrew R.)

Statistical pattern recognition / Andrew R. Webb, Keith D. Copsey. – 3rd ed.  
p. cm.

Includes bibliographical references and index.

ISBN 978-0-470-68227-2 (hardback) – ISBN 978-0-470-68228-9 (paper)

I. Pattern perception–Statistical methods. I. Copsey, Keith D. II. Title.

Q327.W43 2011

006.4–dc23

2011024957

A catalogue record for this book is available from the British Library.

HB ISBN: 978-0-470-68227-2

PB ISBN: 978-0-470-68228-9

ePDF ISBN: 978-1-119-95296-1

oBook ISBN: 978-1-119-95295-4

ePub ISBN: 978-1-119-96140-6

Mobi ISBN: 978-1-119-96141-3

Typeset in 10/12pt Times by Aptara Inc., New Delhi, India

*To Rosemary,  
Samuel, Miriam, Jacob and Ethan*





# Contents

<b>Preface</b>	<b>xix</b>
<b>Notation</b>	<b>xxiii</b>
1 Introduction to Statistical Pattern Recognition	1
1.1 Statistical Pattern Recognition	1
1.1.1 Introduction	1
1.1.2 The Basic Model	2
1.2 Stages in a Pattern Recognition Problem	4
1.3 Issues	6
1.4 Approaches to Statistical Pattern Recognition	7
1.5 Elementary Decision Theory	8
1.5.1 Bayes' Decision Rule for Minimum Error	8
1.5.2 Bayes' Decision Rule for Minimum Error – Reject Option	12
1.5.3 Bayes' Decision Rule for Minimum Risk	13
1.5.4 Bayes' Decision Rule for Minimum Risk – Reject Option	15
1.5.5 Neyman–Pearson Decision Rule	15
1.5.6 Minimax Criterion	18
1.5.7 Discussion	19
1.6 Discriminant Functions	20
1.6.1 Introduction	20
1.6.2 Linear Discriminant Functions	21
1.6.3 Piecewise Linear Discriminant Functions	23
1.6.4 Generalised Linear Discriminant Function	24
1.6.5 Summary	26
1.7 Multiple Regression	27
1.8 Outline of Book	29
1.9 Notes and References	29
Exercises	31
2 Density Estimation – Parametric	33
2.1 Introduction	33

2.2	Estimating the Parameters of the Distributions	34
2.2.1	Estimative Approach	34
2.2.2	Predictive Approach	35
2.3	The Gaussian Classifier	35
2.3.1	Specification	35
2.3.2	Derivation of the Gaussian Classifier Plug-In Estimates	37
2.3.3	Example Application Study	39
2.4	Dealing with Singularities in the Gaussian Classifier	40
2.4.1	Introduction	40
2.4.2	Naïve Bayes	40
2.4.3	Projection onto a Subspace	41
2.4.4	Linear Discriminant Function	41
2.4.5	Regularised Discriminant Analysis	42
2.4.6	Example Application Study	44
2.4.7	Further Developments	45
2.4.8	Summary	46
2.5	Finite Mixture Models	46
2.5.1	Introduction	46
2.5.2	Mixture Models for Discrimination	48
2.5.3	Parameter Estimation for Normal Mixture Models	49
2.5.4	Normal Mixture Model Covariance Matrix Constraints	51
2.5.5	How Many Components?	52
2.5.6	Maximum Likelihood Estimation via EM	55
2.5.7	Example Application Study	60
2.5.8	Further Developments	62
2.5.9	Summary	63
2.6	Application Studies	63
2.7	Summary and Discussion	66
2.8	Recommendations	66
2.9	Notes and References	67
	Exercises	67
3	Density Estimation – Bayesian	70
3.1	Introduction	70
3.1.1	Basics	72
3.1.2	Recursive Calculation	72
3.1.3	Proportionality	73
3.2	Analytic Solutions	73
3.2.1	Conjugate Priors	73
3.2.2	Estimating the Mean of a Normal Distribution with Known Variance	75
3.2.3	Estimating the Mean and the Covariance Matrix of a Multivariate Normal Distribution	79
3.2.4	Unknown Prior Class Probabilities	85
3.2.5	Summary	87
3.3	Bayesian Sampling Schemes	87
3.3.1	Introduction	87

3.3.2	Summarisation	87
3.3.3	Sampling Version of the Bayesian Classifier	89
3.3.4	Rejection Sampling	89
3.3.5	Ratio of Uniforms	90
3.3.6	Importance Sampling	92
3.4	Markov Chain Monte Carlo Methods	95
3.4.1	Introduction	95
3.4.2	The Gibbs Sampler	95
3.4.3	Metropolis–Hastings Algorithm	103
3.4.4	Data Augmentation	107
3.4.5	Reversible Jump Markov Chain Monte Carlo	108
3.4.6	Slice Sampling	109
3.4.7	MCMC Example – Estimation of Noisy Sinusoids	111
3.4.8	Summary	115
3.4.9	Notes and References	116
3.5	Bayesian Approaches to Discrimination	116
3.5.1	Labelled Training Data	116
3.5.2	Unlabelled Training Data	117
3.6	Sequential Monte Carlo Samplers	119
3.6.1	Introduction	119
3.6.2	Basic Methodology	121
3.6.3	Summary	125
3.7	Variational Bayes	126
3.7.1	Introduction	126
3.7.2	Description	126
3.7.3	Factorised Variational Approximation	129
3.7.4	Simple Example	131
3.7.5	Use of the Procedure for Model Selection	135
3.7.6	Further Developments and Applications	136
3.7.7	Summary	137
3.8	Approximate Bayesian Computation	137
3.8.1	Introduction	137
3.8.2	ABC Rejection Sampling	138
3.8.3	ABC MCMC Sampling	140
3.8.4	ABC Population Monte Carlo Sampling	141
3.8.5	Model Selection	142
3.8.6	Summary	143
3.9	Example Application Study	144
3.10	Application Studies	145
3.11	Summary and Discussion	146
3.12	Recommendations	147
3.13	Notes and References	147
	Exercises	148
4	Density Estimation – Nonparametric	150
4.1	Introduction	150
4.1.1	Basic Properties of Density Estimators	150

4.2	<i>k</i> -Nearest-Neighbour Method	152
4.2.1	<i>k</i> -Nearest-Neighbour Classifier	152
4.2.2	Derivation	154
4.2.3	Choice of Distance Metric	157
4.2.4	Properties of the Nearest-Neighbour Rule	159
4.2.5	Linear Approximating and Eliminating Search Algorithm	159
4.2.6	Branch and Bound Search Algorithms: kd-Trees	163
4.2.7	Branch and Bound Search Algorithms: Ball-Trees	170
4.2.8	Editing Techniques	174
4.2.9	Example Application Study	177
4.2.10	Further Developments	178
4.2.11	Summary	179
4.3	Histogram Method	180
4.3.1	Data Adaptive Histograms	181
4.3.2	Independence Assumption (Naïve Bayes)	181
4.3.3	Lancaster Models	182
4.3.4	Maximum Weight Dependence Trees	183
4.3.5	Bayesian Networks	186
4.3.6	Example Application Study – Naïve Bayes Text Classification	190
4.3.7	Summary	193
4.4	Kernel Methods	194
4.4.1	Biasedness	197
4.4.2	Multivariate Extension	198
4.4.3	Choice of Smoothing Parameter	199
4.4.4	Choice of Kernel	201
4.4.5	Example Application Study	202
4.4.6	Further Developments	203
4.4.7	Summary	203
4.5	Expansion by Basis Functions	204
4.6	Copulas	207
4.6.1	Introduction	207
4.6.2	Mathematical Basis	207
4.6.3	Copula Functions	208
4.6.4	Estimating Copula Probability Density Functions	209
4.6.5	Simple Example	211
4.6.6	Summary	212
4.7	Application Studies	213
4.7.1	Comparative Studies	216
4.8	Summary and Discussion	216
4.9	Recommendations	217
4.10	Notes and References	217
	Exercises	218
5	Linear Discriminant Analysis	221
5.1	Introduction	221
5.2	Two-Class Algorithms	222
5.2.1	General Ideas	222

5.2.2	Perceptron Criterion	223
5.2.3	Fisher's Criterion	227
5.2.4	Least Mean-Squared-Error Procedures	228
5.2.5	Further Developments	235
5.2.6	Summary	235
5.3	Multiclass Algorithms	236
5.3.1	General Ideas	236
5.3.2	Error-Correction Procedure	237
5.3.3	Fisher's Criterion – Linear Discriminant Analysis	238
5.3.4	Least Mean-Squared-Error Procedures	241
5.3.5	Regularisation	246
5.3.6	Example Application Study	246
5.3.7	Further Developments	247
5.3.8	Summary	248
5.4	Support Vector Machines	249
5.4.1	Introduction	249
5.4.2	Linearly Separable Two-Class Data	249
5.4.3	Linearly Nonseparable Two-Class Data	253
5.4.4	Multiclass SVMs	256
5.4.5	SVMs for Regression	257
5.4.6	Implementation	259
5.4.7	Example Application Study	262
5.4.8	Summary	263
5.5	Logistic Discrimination	263
5.5.1	Two-Class Case	263
5.5.2	Maximum Likelihood Estimation	264
5.5.3	Multiclass Logistic Discrimination	266
5.5.4	Example Application Study	267
5.5.5	Further Developments	267
5.5.6	Summary	268
5.6	Application Studies	268
5.7	Summary and Discussion	268
5.8	Recommendations	269
5.9	Notes and References	270
	Exercises	270
6	Nonlinear Discriminant Analysis – Kernel and Projection Methods	274
6.1	Introduction	274
6.2	Radial Basis Functions	276
6.2.1	Introduction	276
6.2.2	Specifying the Model	278
6.2.3	Specifying the Functional Form	278
6.2.4	The Positions of the Centres	279
6.2.5	Smoothing Parameters	281
6.2.6	Calculation of the Weights	282
6.2.7	Model Order Selection	284
6.2.8	Simple RBF	285

6.2.9	Motivation	286
6.2.10	RBF Properties	288
6.2.11	Example Application Study	288
6.2.12	Further Developments	289
6.2.13	Summary	290
6.3	Nonlinear Support Vector Machines	291
6.3.1	Introduction	291
6.3.2	Binary Classification	291
6.3.3	Types of Kernel	292
6.3.4	Model Selection	293
6.3.5	Multiclass SVMs	294
6.3.6	Probability Estimates	294
6.3.7	Nonlinear Regression	296
6.3.8	Example Application Study	296
6.3.9	Further Developments	297
6.3.10	Summary	298
6.4	The Multilayer Perceptron	298
6.4.1	Introduction	298
6.4.2	Specifying the MLP Structure	299
6.4.3	Determining the MLP Weights	300
6.4.4	Modelling Capacity of the MLP	307
6.4.5	Logistic Classification	307
6.4.6	Example Application Study	310
6.4.7	Bayesian MLP Networks	311
6.4.8	Projection Pursuit	313
6.4.9	Summary	313
6.5	Application Studies	314
6.6	Summary and Discussion	316
6.7	Recommendations	317
6.8	Notes and References	318
	Exercises	318
7	Rule and Decision Tree Induction	322
7.1	Introduction	322
7.2	Decision Trees	323
7.2.1	Introduction	323
7.2.2	Decision Tree Construction	326
7.2.3	Selection of the Splitting Rule	327
7.2.4	Terminating the Splitting Procedure	330
7.2.5	Assigning Class Labels to Terminal Nodes	332
7.2.6	Decision Tree Pruning – Worked Example	332
7.2.7	Decision Tree Construction Methods	337
7.2.8	Other Issues	339
7.2.9	Example Application Study	340
7.2.10	Further Developments	341
7.2.11	Summary	342

7.3	Rule Induction	342
7.3.1	Introduction	342
7.3.2	Generating Rules from a Decision Tree	345
7.3.3	Rule Induction Using a Sequential Covering Algorithm	345
7.3.4	Example Application Study	350
7.3.5	Further Developments	351
7.3.6	Summary	351
7.4	Multivariate Adaptive Regression Splines	351
7.4.1	Introduction	351
7.4.2	Recursive Partitioning Model	351
7.4.3	Example Application Study	355
7.4.4	Further Developments	355
7.4.5	Summary	356
7.5	Application Studies	356
7.6	Summary and Discussion	358
7.7	Recommendations	358
7.8	Notes and References	359
	Exercises	359
8	Ensemble Methods	361
8.1	Introduction	361
8.2	Characterising a Classifier Combination Scheme	362
8.2.1	Feature Space	363
8.2.2	Level	366
8.2.3	Degree of Training	368
8.2.4	Form of Component Classifiers	368
8.2.5	Structure	369
8.2.6	Optimisation	369
8.3	Data Fusion	370
8.3.1	Architectures	370
8.3.2	Bayesian Approaches	371
8.3.3	Neyman–Pearson Formulation	373
8.3.4	Trainable Rules	374
8.3.5	Fixed Rules	375
8.4	Classifier Combination Methods	376
8.4.1	Product Rule	376
8.4.2	Sum Rule	377
8.4.3	Min, Max and Median Combiners	378
8.4.4	Majority Vote	379
8.4.5	Borda Count	379
8.4.6	Combiners Trained on Class Predictions	380
8.4.7	Stacked Generalisation	382
8.4.8	Mixture of Experts	382
8.4.9	Bagging	385
8.4.10	Boosting	387
8.4.11	Random Forests	389
8.4.12	Model Averaging	390

8.4.13	Summary of Methods	396
8.4.14	Example Application Study	398
8.4.15	Further Developments	399
8.5	Application Studies	399
8.6	Summary and Discussion	400
8.7	Recommendations	401
8.8	Notes and References	401
	Exercises	402
9	Performance Assessment	404
9.1	Introduction	404
9.2	Performance Assessment	405
9.2.1	Performance Measures	405
9.2.2	Discriminability	406
9.2.3	Reliability	413
9.2.4	ROC Curves for Performance Assessment	415
9.2.5	Population and Sensor Drift	419
9.2.6	Example Application Study	421
9.2.7	Further Developments	422
9.2.8	Summary	423
9.3	Comparing Classifier Performance	424
9.3.1	Which Technique is Best?	424
9.3.2	Statistical Tests	425
9.3.3	Comparing Rules When Misclassification Costs are Uncertain	426
9.3.4	Example Application Study	428
9.3.5	Further Developments	429
9.3.6	Summary	429
9.4	Application Studies	429
9.5	Summary and Discussion	430
9.6	Recommendations	430
9.7	Notes and References	430
	Exercises	431
10	Feature Selection and Extraction	433
10.1	Introduction	433
10.2	Feature Selection	435
10.2.1	Introduction	435
10.2.2	Characterisation of Feature Selection Approaches	439
10.2.3	Evaluation Measures	440
10.2.4	Search Algorithms for Feature Subset Selection	449
10.2.5	Complete Search – Branch and Bound	450
10.2.6	Sequential Search	454
10.2.7	Random Search	458
10.2.8	Markov Blanket	459
10.2.9	Stability of Feature Selection	460
10.2.10	Example Application Study	462
10.2.11	Further Developments	462
10.2.12	Summary	463



10.3	Linear Feature Extraction	463
10.3.1	Principal Components Analysis	464
10.3.2	Karhunen–Loève Transformation	475
10.3.3	Example Application Study	481
10.3.4	Further Developments	482
10.3.5	Summary	483
10.4	Multidimensional Scaling	484
10.4.1	Classical Scaling	484
10.4.2	Metric MDS	486
10.4.3	Ordinal Scaling	487
10.4.4	Algorithms	490
10.4.5	MDS for Feature Extraction	491
10.4.6	Example Application Study	492
10.4.7	Further Developments	493
10.4.8	Summary	493
10.5	Application Studies	493
10.6	Summary and Discussion	495
10.7	Recommendations	495
10.8	Notes and References	496
	Exercises	497
11	Clustering	501
11.1	Introduction	501
11.2	Hierarchical Methods	502
11.2.1	Single-Link Method	503
11.2.2	Complete-Link Method	506
11.2.3	Sum-of-Squares Method	507
11.2.4	General Agglomerative Algorithm	508
11.2.5	Properties of a Hierarchical Classification	508
11.2.6	Example Application Study	509
11.2.7	Summary	509
11.3	Quick Partitions	510
11.4	Mixture Models	511
11.4.1	Model Description	511
11.4.2	Example Application Study	512
11.5	Sum-of-Squares Methods	513
11.5.1	Clustering Criteria	514
11.5.2	Clustering Algorithms	515
11.5.3	Vector Quantisation	520
11.5.4	Example Application Study	530
11.5.5	Further Developments	530
11.5.6	Summary	531
11.6	Spectral Clustering	531
11.6.1	Elementary Graph Theory	531
11.6.2	Similarity Matrices	534
11.6.3	Application to Clustering	534
11.6.4	Spectral Clustering Algorithm	535
11.6.5	Forms of Graph Laplacian	535

11.6.6	Example Application Study	536
11.6.7	Further Developments	538
11.6.8	Summary	538
11.7	Cluster Validity	538
11.7.1	Introduction	538
11.7.2	Statistical Tests	539
11.7.3	Absence of Class Structure	540
11.7.4	Validity of Individual Clusters	541
11.7.5	Hierarchical Clustering	542
11.7.6	Validation of Individual Clusterings	542
11.7.7	Partitions	543
11.7.8	Relative Criteria	543
11.7.9	Choosing the Number of Clusters	545
11.8	Application Studies	546
11.9	Summary and Discussion	549
11.10	Recommendations	551
11.11	Notes and References	552
	Exercises	553
12	Complex Networks	555
12.1	Introduction	555
12.1.1	Characteristics	557
12.1.2	Properties	557
12.1.3	Questions to Address	559
12.1.4	Descriptive Features	560
12.1.5	Outline	560
12.2	Mathematics of Networks	561
12.2.1	Graph Matrices	561
12.2.2	Connectivity	562
12.2.3	Distance Measures	562
12.2.4	Weighted Networks	563
12.2.5	Centrality Measures	563
12.2.6	Random Graphs	564
12.3	Community Detection	565
12.3.1	Clustering Methods	565
12.3.2	Girvan–Newman Algorithm	568
12.3.3	Modularity Approaches	570
12.3.4	Local Modularity	571
12.3.5	Clique Percolation	573
12.3.6	Example Application Study	574
12.3.7	Further Developments	575
12.3.8	Summary	575
12.4	Link Prediction	575
12.4.1	Approaches to Link Prediction	576
12.4.2	Example Application Study	578
12.4.3	Further Developments	578
12.5	Application Studies	579

12.6	Summary and Discussion	579
12.7	Recommendations	580
12.8	Notes and References	580
	Exercises	580
13	Additional Topics	581
13.1	Model Selection	581
13.1.1	Separate Training and Test Sets	582
13.1.2	Cross-Validation	582
13.1.3	The Bayesian Viewpoint	583
13.1.4	Akaike's Information Criterion	583
13.1.5	Minimum Description Length	584
13.2	Missing Data	585
13.3	Outlier Detection and Robust Procedures	586
13.4	Mixed Continuous and Discrete Variables	587
13.5	Structural Risk Minimisation and the Vapnik–Chervonenkis Dimension	588
13.5.1	Bounds on the Expected Risk	588
13.5.2	The VC Dimension	589
	<b>References</b>	<b>591</b>
	<b>Index</b>	<b>637</b>



# Preface

This book provides an introduction to statistical pattern recognition theory and techniques. Most of the material presented in this book is concerned with discrimination and classification and has been drawn from a wide range of literature including that of engineering, statistics, computer science and the social sciences. The aim of the book is to provide descriptions of many of the most useful of today's pattern processing techniques including many of the recent advances in nonparametric approaches to discrimination and Bayesian computational methods developed in the statistics literature and elsewhere. Discussions provided on the motivations and theory behind these techniques will enable the practitioner to gain maximum benefit from their implementations within many of the popular software packages. The techniques are illustrated with examples of real-world applications studies. Pointers are also provided to the diverse literature base where further details on applications, comparative studies and theoretical developments may be obtained.

The book grew out of our research on the development of statistical pattern recognition methodology and its application to practical sensor data analysis problems. The book is aimed at advanced undergraduate and graduate courses. Some of the material has been presented as part of a graduate course on pattern recognition and at pattern recognition summer schools. It is also designed for practitioners in the field of pattern recognition as well as researchers in the area. A prerequisite is a knowledge of basic probability theory and linear algebra, together with basic knowledge of mathematical methods (for example, Lagrange multipliers are used to solve problems with equality and inequality constraints in some derivations). Some basic material (which was provided as appendices in the second edition) is available on the book's website.

## Scope

The book presents most of the popular methods of statistical pattern recognition. However, many of the important developments in pattern recognition are not confined to the statistics literature and have occurred where the area overlaps with research in machine learning. Therefore, where we have felt that straying beyond the traditional boundaries of statistical pattern recognition would be beneficial, we have done so. An example is the

inclusion of some rule induction methods as a complementary approach to rule discovery by decision tree induction.

Most of the methodology is generic – it is not specific to a particular type of data or application. Thus, we exclude preprocessing methods and filtering methods commonly used in signal and image processing.

## Approach

The approach in each chapter has been to introduce some of the basic concepts and algorithms and to conclude each section on a technique or a class of techniques with a practical application of the approach from the literature. The main aim has been to introduce the basic concept of an approach. Sometimes this has required some detailed mathematical description and clearly we have had to draw a line on how much depth we discuss a particular topic. Most of the topics have whole books devoted to them and so we have had to be selective in our choice of material. Therefore, the chapters conclude with a section on the key references. The exercises at the ends of the chapters vary from ‘open book’ questions to more lengthy computer projects.

## New to the third edition

Many sections have been rewritten and new material added. The new features of this edition include the following:

- A new chapter on Bayesian approaches to density estimation (Chapter 3) including expanded material on Bayesian sampling schemes and Markov chain Monte Carlo methods, and new sections on Sequential Monte Carlo samplers and Variational Bayes approaches.
- New sections on nonparametric methods of density estimation.
- Rule induction.
- New chapter on ensemble methods of classification.
- Revision of feature selection material with new section on stability.
- Spectral clustering.
- New chapter on complex networks, with relevance to the high-growth field of social and computer network analysis.

## Book outline

Chapter 1 provides an introduction to statistical pattern recognition, defining some terminology, introducing supervised and unsupervised classification. Two related approaches to supervised classification are presented: one based on the use of probability density functions

and a second based on the construction of discriminant functions. The chapter concludes with an outline of the pattern recognition cycle, putting the remaining chapters of the book into context. Chapters 2, 3 and 4 pursue the density function approach to discrimination. Chapter 2 addresses parametric approaches to density estimation, which are developed further in Chapter 3 on Bayesian methods. Chapter 4 develops classifiers based on nonparametric schemes, including the popular  $k$  nearest neighbour method, with associated efficient search algorithms.

Chapters 5–7 develop discriminant function approaches to supervised classification. Chapter 5 focuses on linear discriminant functions; much of the methodology of this chapter (including optimisation, regularisation, support vector machines) is used in some of the non-linear methods described in Chapter 6 which explores kernel-based methods, in particular, the radial basis function network and the support vector machine, and projection-based methods (the multilayer perceptron). These are commonly referred to as neural network methods. Chapter 7 considers approaches to discrimination that enable the classification function to be cast in the form of an interpretable rule, important for some applications.

Chapter 8 considers ensemble methods – combining classifiers for improved robustness. Chapter 9 considers methods of measuring the performance of a classifier.

The techniques of Chapters 10 and 11 may be described as methods of exploratory data analysis or preprocessing (and as such would usually be carried out prior to the supervised classification techniques of Chapters 5–7, although they could, on occasion, be post-processors of supervised techniques). Chapter 10 addresses feature selection and feature extraction – the procedures for obtaining a reduced set of variables characterising the original data. Such procedures are often an integral part of classifier design and it is somewhat artificial to partition the pattern recognition problem into separate processes of feature extraction and classification. However, feature extraction may provide insights into the data structure and the type of classifier to employ; thus, it is of interest in its own right. Chapter 11 considers unsupervised classification or *clustering* – the process of grouping individuals in a population to discover the presence of structure; its engineering application is to vector quantisation for image and speech coding. Chapter 12 on complex networks introduces methods for analysing data that may be represented using the mathematical concept of a graph. This has great relevance to social and computer networks.

Finally, Chapter 13 addresses some important diverse topics including model selection.

## Book website

The website [www.wiley.com/go/statistical\\_pattern\\_recognition](http://www.wiley.com/go/statistical_pattern_recognition) contains supplementary material on topics including measures of dissimilarity, estimation, linear algebra, data analysis and basic probability.

## Acknowledgements

In preparing the third edition of this book we have been helped by many people. We are especially grateful to Dr Gavin Cawley, University of East Anglia, for help and advice. We are grateful to friends and colleagues (past and present, from RSRE, DERA and QinetiQ)

who have provided encouragement and made comments on various parts of the manuscript. In particular, we would like to thank Anna Skeoch for providing figures for Chapter 12; and Richard Davies and colleagues at John Wiley for help in the final production of the manuscript. Andrew Webb is especially thankful to Rosemary for her love, support and patience.

Andrew R. Webb  
Keith D. Copsey



# Notation

Some of the more commonly used notation is given below. We have used some notational conveniences. For example, we have tended to use the same symbol for a variable as well as a measurement on that variable. The meaning should be obvious from context. Also, we denote the density function of  $x$  as  $p(x)$  and  $y$  as  $p(y)$ , even though the functions differ. A vector is denoted by a lower case quantity in bold face, and a matrix by upper case. Since pattern recognition is very much a multidisciplinary subject, it is impossible to be both consistent across all chapters and consistent with the commonly used notation in the different literatures. We have adopted the policy of maintaining consistency as far as possible within a given chapter.

$p, d$	number of variables
$C$	number of classes
$n$	number of measurements
$n_j$	number of measurements in the $j$ th class
$\omega_j$	label for class $j$
$X_1, \dots, X_p$	$p$ random variables
$x_1, \dots, x_p$	measurements on variables, $X_1, \dots, X_p$
$\mathbf{x} = (x_1, \dots, x_p)^T$	measurement vector
$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$	$n \times p$ data matrix
$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$	
$P(\mathbf{x}) = \text{prob}(X_1 \leq x_1, \dots, X_p \leq x_p)$	probability density function
$p(\mathbf{x}) = \partial P / \partial \mathbf{x}$	probability density function of class $j$
$p(\mathbf{x}   \omega_j)$	prior probability of class $j$
$p(\omega_j)$	population mean
$\mu = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$	mean of class $j, j = 1, \dots, C$
$\mu_j = \int \mathbf{x} p(\mathbf{x}   \omega_j) d\mathbf{x}$	sample mean
$\mathbf{m} = (1/n) \sum_{i=1}^n \mathbf{x}_i$	sample mean of class $j, j = 1, \dots, C; z_{ji} = 1$ if
$\mathbf{m}_j = (1/n_j) \sum_{i=1}^n z_{ji} \mathbf{x}_i$	$\mathbf{x}_i \in \omega_j, 0$ otherwise; $n_j$ -number of patterns in
	$\omega_j, n_j = \sum_{i=1}^n z_{ji}$

$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$	sample covariance matrix (maximum likelihood estimate)
$n/(n-1)\hat{\Sigma}$	sample covariance matrix (unbiased estimate)
$\hat{\Sigma}_j = \frac{1}{n_j} \sum_{i=1}^n z_{ji}(\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T$	sample covariance matrix of class $j$ (maximum likelihood estimate)
$S_j = \frac{n_j}{n_j-1} \hat{\Sigma}_j$	sample covariance matrix of class $j$ (unbiased estimate)
$S_W = \sum_{j=1}^C \frac{n_j}{n} \hat{\Sigma}_j$	pooled within class sample covariance matrix
$S = \frac{n}{n-C} S_W$	pooled within class sample covariance matrix (unbiased estimate)
$S_B = \sum \frac{n_j}{n} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T$	sample between class matrix
$S_B + S_W = \hat{\Sigma}$	
$\ \mathbf{A}\ ^2 = \sum_{ij} A_{ij}^2$	
$N(\mathbf{m}, \Sigma)$	normal (or Gaussian) distribution, mean $\mathbf{m}$ , covariance matrix $\Sigma$
$N(\mathbf{x}; \mathbf{m}, \Sigma)$	probability density function for the normal distribution, mean $\mathbf{m}$ , covariance matrix $\Sigma$ , evaluated at $\mathbf{x}$
$E[Y X]$	expectation of $Y$ given $X$
$I(\theta)$	indicator function, $I(\theta) = 1$ if $\theta = \text{true}$ else 0

# 1

# Introduction to statistical pattern recognition

Statistical pattern recognition is a term used to cover all stages of an investigation from problem formulation and data collection through to discrimination and classification, assessment of results and interpretation. Some of the basic concepts in classification are introduced and the key issues described. Two complementary approaches to discrimination are presented, namely a decision theory approach based on calculation of probability density functions and the use of Bayes theorem, and a discriminant function approach.

## 1.1 Statistical pattern recognition

### 1.1.1 Introduction

We live in a world where massive amounts of data are collected and recorded on nearly every aspect of human endeavour: for example, banking, purchasing (credit-card usage, point-of-sale data analysis), Internet transactions, performance monitoring (of schools, hospitals, equipment), and communications. The data come in a wide variety of diverse forms – numeric, textual (structured or unstructured), audio and video signals. Understanding and making sense of this vast and diverse collection of data (identifying patterns, trends, anomalies, providing summaries) requires some automated procedure to assist the analyst with this ‘data deluge’. A practical example of pattern recognition that is familiar to many people is classifying email messages (as spam/not spam) based upon message header, content and sender.

Approaches for analysing such data include those for signal processing, filtering, data summarisation, dimension reduction, variable selection, regression and classification and have been developed in several literatures (physics, mathematics, statistics, engineering, artificial intelligence, computer science and the social sciences, among others). The main focus of this book is on pattern recognition procedures, providing a description of basic techniques

together with case studies of practical applications of the techniques on real-world problems. A strong emphasis is placed on the statistical theory of discrimination, but clustering also receives some attention. Thus, the main subject matter of this book can be summed up in a single word: ‘classification’, both supervised (using class information to design a classifier – i.e. discrimination) and unsupervised (allocating to groups without class information – i.e. clustering). However, in recent years many complex datasets have been gathered (for example, ‘transactions’ between individuals – email traffic, purchases). Understanding these datasets requires additional tools in the pattern recognition toolbox. Therefore, we also examine developments such as methods for analysing data that may be represented as a graph.

Pattern recognition as a field of study developed significantly in the 1960s. It was very much an interdisciplinary subject. Some people entered the field with a real problem to solve. The large number of applications ranging from the classical ones such as automatic character recognition and medical diagnosis to the more recent ones in *data mining* (such as credit scoring, consumer sales analysis and credit card transaction analysis) have attracted considerable research effort with many methods developed and advances made. Other researchers were motivated by the development of machines with ‘brain-like’ performance, that in some way could operate giving human performance.

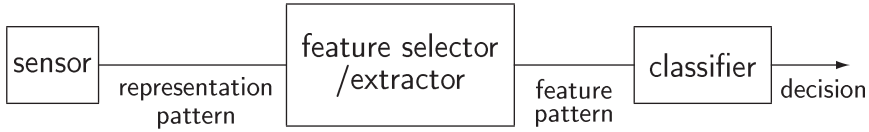
Within these areas significant progress has been made, particularly where the domain overlaps with probability and statistics, and in recent years there have been many exciting new developments, both in methodology and applications. These build on the solid foundations of earlier research and take advantage of increased computational resources readily available nowadays. These developments include, for example, kernel-based methods (including support vector machines) and Bayesian computational methods.

The topics in this book could easily have been described under the term *machine learning* that describes the study of machines that can adapt to their environment and learn from example. The machine learning emphasis is perhaps more on computationally intensive methods and less on a statistical approach, but there is strong overlap between the research areas of statistical pattern recognition and machine learning.

### 1.1.2 The basic model

Since many of the techniques we shall describe have been developed over a range of diverse disciplines, there is naturally a variety of sometimes contradictory terminology. We shall use the term ‘pattern’ to denote the  $p$ -dimensional data vector  $\mathbf{x} = (x_1, \dots, x_p)^T$  of measurements ( $^T$  denotes vector transpose), whose components  $x_i$  are measurements of the features of an object. Thus the features are the variables specified by the investigator and thought to be important for classification. In discrimination, we assume that there exist  $C$  groups or *classes*, denoted  $\omega_1, \dots, \omega_C$  and associated with each pattern  $\mathbf{x}$  is a categorical variable  $z$  that denotes the class or group membership; that is, if  $z = i$ , then the pattern belongs to  $\omega_i, i \in \{1, \dots, C\}$ .

Examples of patterns are measurements of an acoustic waveform in a speech recognition problem; measurements on a patient made in order to identify a disease (diagnosis); measurements on patients (perhaps subjective assessments) in order to predict the likely outcome (prognosis); measurements on weather variables (for forecasting or prediction); sets of financial measurements recorded over time; and a digitised image for character recognition. Therefore, we see that the term ‘pattern’, in its technical meaning, does not necessarily refer to structure within images.



**Figure 1.1** Pattern classifier.

The main topic in this book may be described by a number of terms including *pattern classifier design* or *discrimination* or *allocation rule design*. Designing the rule requires specification of the parameters of a pattern classifier, represented schematically in Figure 1.1, so that it yields the optimal (in some sense) response for a given input pattern. This response is usually an estimate of the class to which the pattern belongs. We assume that we have a set of patterns of known class  $\{(\mathbf{x}_i, z_i), i = 1, \dots, n\}$  (the *training* or *design* set) that we use to design the classifier (to set up its internal parameters). Once this has been done, we may estimate class membership for a pattern  $\mathbf{x}$  for which the class label is unknown. Learning the model from a training set is the process of *induction*; applying the trained model to patterns of unknown class is the process of *deduction*.

Thus, the uses of a pattern classifier are to provide:

- A descriptive model that explains the difference between patterns of different classes in terms of features and their measurements.
- A predictive model that predicts the class of an unlabelled pattern.

However, we might ask why do we need a predictive model? Cannot the procedure that was used to assign labels to the training set measurements also be used for the test set in classifier operation? There may be several reasons for developing an automated process:

- to remove humans from the recognition process – to make the process more reliable;
- in banking, to identify good risk applicants before making a loan;
- to make a medical diagnosis without a post mortem (or to assess the state of a piece of equipment without dismantling it) – sometimes a pattern may only be labelled through intensive examination of a subject, whether person or piece of equipment;
- to reduce cost and improve speed – gathering and labelling data can be a costly and time consuming process;
- to operate in hostile environments – the operating conditions may be dangerous or harmful to humans and the training data have been gathered under controlled conditions;
- to operate remotely – to classify crops and land use remotely without labour-intensive, time consuming, surveys.

There are many classifiers that can be constructed from a given dataset. Examples include decision trees, neural networks, support vector machines and linear discriminant functions. For a classifier of a given type, we employ a learning algorithm to search through the parameter space to find the model that best describes the relationship between the measurements and class labels for the training set. The form derived for the pattern classifier depends on a number of different factors. It depends on the distribution of the training data, and the assumptions

made concerning its distribution. Another important factor is the misclassification cost – the cost of making an incorrect decision. In many applications misclassification costs are hard to quantify, being combinations of several contributions such as monetary costs, time and other more subjective costs. For example, in a medical diagnosis problem, each treatment has different costs associated with it. These relate to the expense of different types of drugs, the suffering the patient is subjected to by each course of action and the risk of further complications.

Figure 1.1 grossly oversimplifies the pattern classification procedure. Data may undergo several separate transformation stages before a final outcome is reached. These transformations (sometimes termed preprocessing, feature selection or feature extraction) operate on the data in a way that, usually, reduces its dimension (reduces the number of features), removing redundant or irrelevant information, and transforms it to a form more appropriate for subsequent classification. The term *intrinsic dimensionality* refers to the minimum number of variables required to capture the structure within the data. In speech recognition, a preprocessing stage may be to transform the waveform to a frequency representation. This may be processed further to find formants (peaks in the spectrum). This is a *feature extraction* process (taking a possibly nonlinear combination of the original variables to form new variables). *Feature selection* is the process of selecting a subset of a given set of variables (see Chapter 10). In some problems, there is no automatic feature selection stage, with the feature selection being performed by the investigator who ‘knows’ (through experience, knowledge of previous studies and the problem domain) those variables that are important for classification. In many cases, however, it will be necessary to perform one or more transformations of the measured data.

In some pattern classifiers, each of the above stages may be present and identifiable as separate operations, while in others they may not be. Also, in some classifiers, the preliminary stages will tend to be problem specific, as in the speech example. In this book, we consider feature selection and extraction transformations that are not application specific. That is not to say the methods of feature transformation described will be suitable for any given application, however, but application-specific preprocessing must be left to the investigator who understands the application domain and method of data collection.

## 1.2 Stages in a pattern recognition problem

A pattern recognition investigation may consist of several stages enumerated below. Not all stages may be present; some may be merged together so that the distinction between two operations may not be clear, even if both are carried out; there may be some application-specific data processing that may not be regarded as one of the stages listed below. However, the points below are fairly typical.

1. Formulation of the problem: gaining a clear understanding of the aims of the investigation and planning the remaining stages.
2. Data collection: making measurements on appropriate variables and recording details of the data collection procedure (ground truth).