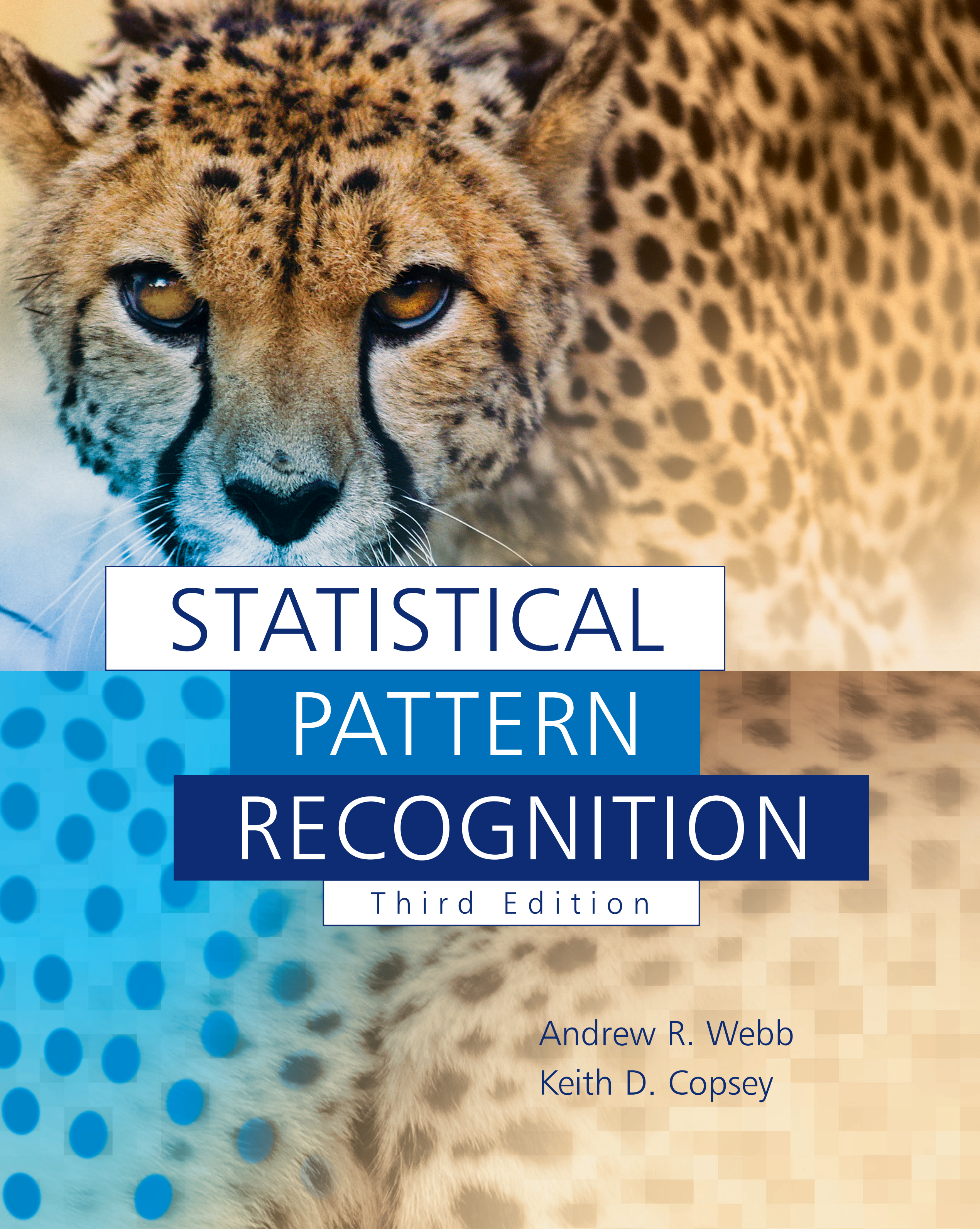

STATISTICAL PATTERN RECOGNITION

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

WILEY
A John Wiley & Sons, Ltd., Publication

This edition first published 2011

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.

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)
1. 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

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

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

Statistical Pattern Recognition, Third Edition. Andrew R. Webb and Keith D. Copsey.

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 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ 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 $\omega_1, \ldots, \omega_C$ and associated with each pattern $\boldsymbol{x}$ is a categorical variable z that denotes the class or group membership; that is, if $z = i$, then the pattern belongs to ω_i, $i \in \{1, \ldots, C\}$.

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.

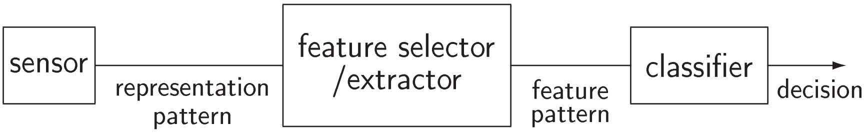

Figure 1.1 Pattern classifier.

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 $\{(\boldsymbol{x}_i, z_i), i = 1, \ldots, n\}$ (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 $\boldsymbol{x}$ 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).