# 1 Introduction

## 1.1 Background

Searching for patterns in biological data has a long history, perhaps best typified by the taxonomists who arranged species into groups based on their similarities and differences. As computer resources improved there was a growth in the scope and availability of new computational methods. Some of the first biologists to exploit these methods worked in disciplines such as vegetation ecology. Vegetation ecologists often collect species data from many quadrats and they require methods that will organise the data so that relevant structures are revealed. The increase in computer power, and the parallel growth in software, allowed analyses that were previously impractical for other than small data sets.

However, the greatest data analysis challenges are undoubtedly more recent. Advances in experimental methods have generated, and continue to generate, enormous volumes of genetic information that present significant storage, retrieval and analysis challenges see Slonim (2002) for a useful review. Simultaneously there has been a growth in non-biological commercial databases and the belief that they contain information which can be used to improve company profits. One consequence of these challenges is that there is now a wide, and increasing, variety of analysis tools that have the potential to extract important information from biological data.

The analysis tools that extract information from data can be placed into two broad and overlapping categories: cluster and classification methods. This book

examines a wide range of techniques from both categories to illustrate their potential for biological research. For example, they can be used to:

1. find patterns in gene expression profiles or biodiversity survey data;
2. identify class membership from some predictor variables (e.g. type of cancer, sex of a specimen, used or unused habitat);
3. identify features that either contribute to class membership or can be used to predict class membership.

## 1.2 Book structure

This chapter introduces the main themes and defines some terms that are used throughout the rest of the book. Chapter 2 outlines the important roles for exploratory data analysis prior to formal clustering and classification analyses. The main part of this chapter introduces multivariate approaches, such as principal components analysis and multi-dimensional scaling, which can be used to reduce the dimensionality of a dataset. They also have a role in the clustering of data. Chapter 3 examines different approaches to the clustering of biological data. The next three chapters, 4 to 6, are concerned with classification, i.e. placing objects into known classes. Chapter 4 outlines some issues that are common to all classification methods while Chapters 5 and 6 examine traditional statistical approaches and a variety of more recent, 'non-statistical', methods. Chapter 7 is related to the classification chapters because it examines the difficulties of assessing the accuracy of a set of predictions.

Although the methods covered in the book are capable of analysing very large data sets, only small data sets are used for illustrative purposes. Most data sets are described and listed in separate appendices. Others are freely available from a range of online sources.

## 1.3 Classification

Classification is the placing of objects into predefined groups, the archetypal biological example being identifying a biological specimen (assigning a species name). This type of process is also called 'pattern recognition', which is defined by Wikipedia (2005) as 'a field within the area of machine learning and can be defined as ''the act of taking in raw data and taking an action based on the category of the data. As such, it is a collection of methods for supervised learning'' '. The terms 'machine learning' and 'supervised learning' are defined in Section 1.6. However, classification is used here in a more general sense to include an element of prediction. For example, do we expect to find a species

in a particular habitat (species distribution modelling), or given a particular genetic profile do we expect an individual to belong to a disease-free or disease-prone class?

Despite this increased breadth for the definition, it excludes the type of statistical predictions generated by, for example, a linear regression. Techniques such as linear regression predict a continuous, quantitative value, e.g. the expected yield of a crop given a certain fertiliser application. In this text prediction is applied almost exclusively to the identification of class membership, a categorical variable. Techniques, such as logistic regression and artificial neural networks, which produce a real-valued output (often between 0 and 1) that is subsequently split into discrete categories by the application of some threshold value, are included. Throughout this text the term classifier is used to describe a general predictive technique which allocates cases to a small number of predefined classes.

## 1.4 Clustering

Clustering is related to classification. Both techniques place objects into groups or classes. The important difference is that the classes are not predefined in a cluster analysis. Instead objects are placed into 'natural' groups. The term natural is a little misleading since the groupings are almost entirely dependent on the protocols that were established before starting the clustering. Clustering and classification approaches differ in the type of 'learning' method used. Classification methods use supervised methods while clustering algorithms are unsupervised (see Section 1.6). It is also important to realise that there is no concept of accuracy for an unsupervised classification, the only way of judging the outcome is by its 'usefulness'. For example, is the classification 'fit for purpose'? If it is not then one should rethink the analysis.

There is some unfortunate confusion in the use of terms. For example, Gordon's (1981, 1999) excellent books have 'Classification' as the title, but they are entirely concerned with clustering of data. In this text classification is restricted to supervised classification methods.

## 1.5 Structures in data

### *1.5.1 Structure in tables*

A common starting point for all clustering and classification algorithms is a table in which rows represent cases (objects) and columns are the variables. Once data are in this format their original source, for example microarrays,

quadrats, etc., is largely irrelevant. The hope is that we can find some structure (signals) in these data that enable us to test or develop biological hypotheses. If the data contain one or more signals there will be some structure within the table. The role of a cluster or classification analysis is to recognise the structure.

In the simplest case this structure may be masked by something as simple as the current row and column ordering. Table 1.1 is a simple example in which there is no obvious structure.

A clear structure becomes apparent once the rows and columns are reordered (Table 1.2). Unfortunately, analysing real data is rarely that simple. Although few clustering and classification algorithms explicitly re-order the table, they will generate information which reveals the presence of some structure. This topic is covered again in Chapter 2.

### 1.5.2 *Graphical identification of structure*

Looking for structures in data tables is only feasible when the table dimensions are relatively small. As tables become larger graphical summaries become more useful. Principal components analysis (Chapter 2) is a commonly used method that can reveal structures in interval data. As an example the data from Table 1.1 were subjected to an analysis. At the end of the analysis each case has a score for two new variables, PC1 and PC2. If these scores are plotted against each other it is obvious that there is some structure to these data (Figure 1.1). Note that the points are arranged in the same order as the re-ordered table (Table 1.2). No further explanation or interpretation is presented here but it is obvious that this approach can reveal structures in data tables.

Table 1.1. *A simple dataset with ten cases (1–10) and nine variables (a–i)*

| Case | a | b | c | d | e | f | g | h | i |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 0 | 1 | 4 | 0 | 5 | 3 | 0 | 0 |
| 2 | 2 | 4 | 3 | 0 | 5 | 0 | 1 | 4 | 3 |
| 3 | 1 | 3 | 2 | 0 | 4 | 0 | 0 | 5 | 4 |
| 4 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 3 | 4 |
| 5 | 4 | 4 | 5 | 2 | 3 | 1 | 3 | 2 | 1 |
| 6 | 3 | 5 | 4 | 1 | 4 | 0 | 2 | 3 | 2 |
| 7 | 5 | 3 | 4 | 3 | 2 | 2 | 4 | 1 | 0 |
| 8 | 3 | 1 | 2 | 5 | 0 | 4 | 4 | 0 | 0 |
| 9 | 4 | 2 | 3 | 4 | 1 | 3 | 5 | 0 | 0 |
| 10 | 0 | 2 | 1 | 0 | 3 | 0 | 0 | 4 | 5 |

Table 1.2. *The dataset from Table 1.1 with re-ordered rows and columns. Zeros have been removed to highlight the structure*

| Case | f | d | g | a | c | b | e | h | i |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 4 | 3 | 2 | 1 | | | | |
| 8 | 4 | 5 | 4 | 3 | 2 | 1 | | | |
| 9 | 3 | 4 | 5 | 4 | 3 | 2 | 1 | | |
| 7 | 2 | 3 | 4 | 5 | 4 | 3 | 2 | 1 | |
| 5 | 1 | 2 | 3 | 4 | 5 | 4 | 3 | 2 | 1 |
| 6 | | 1 | 2 | 3 | 4 | 5 | 4 | 3 | 2 |
| 2 | | | 1 | 2 | 3 | 4 | 5 | 4 | 3 |
| 3 | | | | 1 | 2 | 3 | 4 | 5 | 4 |
| 10 | | | | | 1 | 2 | 3 | 4 | 5 |
| 4 | | | | | | 1 | 2 | 3 | 4 |

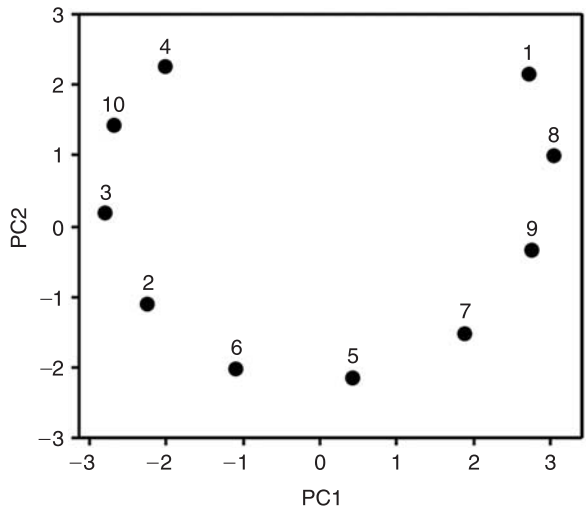


**Figure 1.1** Principal component score plot for the data from Table 1.1. Points are labelled with the case numbers.

## 1.6 Glossary

The fields of clustering and classification are awash with terms and acronyms that may be unfamiliar. This is not helped by redundancy and overlap in their use. For example, some methods have more than one name while the same term may be used for completely different methods. The following is a short glossary of some of the more important terms that appear throughout the book.

### 1.6.1 *Algorithm*

An algorithm is a finite set of unambiguous instructions that describe each step needed to accomplish a task. Normally, the term refers to the design of a computer program. An algorithm differs from an heuristic (chance-based) approach because, given some initial state, it always produces the same result. For example, given a shuffled deck of playing cards, the following simple algorithm will always produce a pile of red cards on the left and a pile of black cards on the right. The order of the cards in each pile will be random.

Repeat
  Turn over the top card.
  If it is red place it to the left of the shuffled deck.
  If it is black place it to the right of the shuffled deck.
Until no cards remain

### 1.6.2 *Bias*

Bias is the difference between the estimate of a value and its true value. A low bias is a desirable property for all classifiers. Unfortunately this may be difficult to achieve given the need to trade off bias against variance. Bias and variance are two parameters that feature significantly in the evaluation of classifier performance.

### 1.6.3 *Deviance*

The deviance is a measure of the variation unexplained by a model and is, therefore, a guide to the fit of data to a statistical model. It is calculated as twice the log-likelihood of the best model minus the log-likelihood of the current model. The best model is generally a saturated model (see maximum likelihood estimation).

### 1.6.4 *Learning*

Supervised learning

The use of the term 'learning' is potentially fraught with problems when biologists are involved (see Stevens-Wood (1999) for a review of 'real learning' within the context of machine learning algorithms). However, learning, as applied to classifier development, refers to the gradual reduction in error as training cases are presented to the classifier. This is an iterative process with the amount of error reducing through each iteration as the classifier 'learns' from the training cases. Supervised learning applies to algorithms in which a 'teacher' continually assesses, or supervises, the classifier's performance during training by marking predictions as correct or incorrect. Naturally, this assumes that each case has a valid class label. Learning in this sense relates to the way

in which a set of parameter values are adjusted to improve the classifier's performance. However, as Hand (1997, p. 41) points out 'one may ask what is the difference between ''learning'' and the less glamorous sounding ''recursive parameter estimation'' '. The answer is, of course, none!

Unsupervised learning

In unsupervised learning there is no teacher and the classifier is left to find its own classes (also known as groups or clusters). The assumption is that there are 'natural' classes that the classifier will identify, albeit with some post-classification processing by the user. A typical example would be finding habitat classes in a satellite image or patterns in gene expression levels that may represent a particular cell state. The biggest gains from an unsupervised classification are likely in knowledge-poor environments, particularly when there are large amounts of unlabelled data. In these situations unsupervised learning can be used as a type of exploratory data analysis (Chapter 2) which may provide some evidence about the underlying structures in the data. As such it should be viewed as a means of generating, rather than testing, hypotheses.

Machine learning

Machine learning (ML) is a subset of artificial intelligence (AI) and incorporates a large number of statistical ideas and methods and uses algorithms that 'learn' from data. Learning is based around finding the relationship between a set of features and class labels in a set of training examples (supervised learning).

Induction and deduction are two types of reasoning. Induction, the leap from the particular to the general ('seeing the woods through the trees'), can be used to generate classification rules from examples by establishing 'cause–effect' relationships between existing facts (Giraud-Carrier and Martinez, 1995). For example, an inductive system may be able to recognise patterns in chess games without having any knowledge of the rules (Quinlan, 1983). Perhaps a similar system could recognise patterns in gene expression without understanding the rules that control them. Deduction is concerned with reasoning using existing knowledge. Thus, in the chess example it would be necessary to know the rules of chess before reasoning about the patterns. The requirement for previously acquired knowledge is one of the main differences between deduction and induction. Because deduction can be used to predict the consequences of events or to determine the prerequisite conditions for an observed event (Giraud-Carrier and Martinez, 1995) it is at the heart of most expert systems.

Although most ML systems are based on induction there are potential problems with this approach to learning. The first is that learning by induction

requires many examples (a large sample size), although this may be outweighed by the necessity for little background knowledge (Dutton and Conroy, 1996). In a knowledge-poor discipline, such as ecology, this can be a major advantage. Another problem with inductive learning systems is that they are not guaranteed to be 'future-proof', they only make use of what happened in the 'past'. Consequently noisy or contradictory data may create problems. These difficulties are not restricted to ML programs; for example, it is well known that relationships modelled by a regression analysis should not be extrapolated beyond the limits of the original data. Noise does not have consistent effects on ML algorithms; for example, artificial neural networks, genetic algorithms and instance-based methods all use 'distributed' representations of the data while others, such as decision trees and logic, do not (Quinlan, 1993). This is important because distributed representations should be less susceptible to small alterations (noise); hence they have the potential to be more robust classifiers (Dutton and Conroy, 1996).

Carbonell (1990) placed ML techniques into four major paradigms: inductive learning; analytical learning; connectionist methods; and genetic methods. The first uses inductive techniques to derive a general class description from a database of labelled cases (the training or learning set). The analytical paradigm, including methods such as case-based reasoning, uses deduction and induction to learn from exemplars (which may be a single case). The last two are rather loosely based on biological concepts. The connectionist paradigm is an approximation of a real biological neural network while the final paradigm uses simulated selection and evolution to converge towards a solution.

### 1.6.5 *Maximum likelihood estimation*

The aim of maximum likelihood estimation is to find the values for parameters that maximise the probability of obtaining a particular set of data. If the parameters of, for example, a probability distribution are known it is a relatively simple task to determine the probability of obtaining a sample with particular characteristics. However, it is common to have a sample of values obtained from a population whose parameters are unknown. The sample could have come from many different probability distributions: maximum likelihood methods find the parameters that maximise the probability of drawing that sample.

The frequent references to log-likelihoods in connection with maximum likelihood estimation are a computational short-cut, rather than a theoretical necessity. This is because maximum likelihood estimates involve the multiplication of many small probabilities which, because of rounding errors, may

get truncated to zero. If logarithms are used the multiplications become additions and the truncation problem is reduced. Because probabilities are less than one they have negative logarithms which become larger as the probability declines.

*1.6.6 Microarray*

A microarray, or microchip, is an important bioinformatics tool that generates large amounts of data that can be subsequently processed by clustering and classification methods. Typically a microarray will have several thousand, very small DNA spots arranged in a grid on glass or other solid material. Each spot has a fragment of a DNA or RNA molecule which corresponds to part of a known gene. Messenger RNA is extracted from biological samples and reverse-transcribed into cDNA, which is then amplified by a PCR. During the PCR amplification radioactive or fluorescent nucleotides are incorporated which are later used to detect the amount of hybridisation with the DNA fragments on the chip. It is assumed that the amount of cDNA for a particular gene reflects its activity (expression level) within the biological sample. After measurement and post-processing a table of expression values is obtained for each gene within each biological sample. These are the data that are subjected to the subsequent clustering and classification procedures to identify patterns that may provide biological insights. Quackenbush's (2001) review of computational analysis methods for microarray data is a good starting point to link clustering and classification algorithms to microarray data.

*1.6.7 Multicollinearity*

This is an important topic in regression-type analyses. Multicollinearity occurs when predictor variables are highly correlated, or one predictor is a function of two or more of the other predictors, for example $x_1 = x_2 + x_3$. Multicollinearity is a problem because it may prevent calculations from being completed or it can lead to instability in the regression coefficients. The instability is often manifested in coefficients having the 'wrong' sign. Thus, although we may expect $y$ to increase as $x$ increases ($b > 0$), multicollinearity can lead to a zero or negative value for $b$. This obviously makes interpretation very difficult.

*1.6.8 Occam's razor*

Put simply Occam's razor is a philosophy that says if a choice must be made between a number of options the simplest is the best. In the context of clustering and, particularly, classification analyses this tends to relate to the number of variables or other design decisions such as the number of weights in the neural network. In general simpler models should be preferred to more

complex ones. However, Domingos (1999) discusses the role of Occam's razor in classification models and argues that choosing the simplest model may not always be desirable.

#### 1.6.9 *Ordination*

Ordination is a term that is normally applied to a particular class of multivariate techniques when they are applied to ecological data. Generally they are geometrical projection methods that attempt to present multivariate data in fewer dimensions. Palmer's 'Ordination Methods for Ecologists' website has more details (http://ordination.okstate.edu/).

## 1.7 Recommended reading and other resources

#### 1.7.1 *Books*

There are many good books that include the topics covered in this book. However, most of them have a restricted coverage, although this is usually combined with a greater depth. I have avoided the temptation to recommend subject-specific texts because the terminology and examples often obscure their explanations for readers from other disciplines. For example, Legendre and Legendre (1998) is an excellent, comprehensive book but non-ecologists may find some of the detail unhelpful. Dudoit and Fridlyand (2003) surveyed the use of microarray data in classification problems. The book by Davis (1986), which is aimed at geologists, has the clearest description of eigen values and eigen vectors that I have ever read. Unsurprisingly there are many books which fall within the machine learning and data mining spheres that may be useful. Most of these are listed in the relevant chapters but there are three that are particularly useful. First is the excellent book by Duda *et al*. (2001) which has very clear explanations for much of the theoretical background. Anyone interested in applying clustering and classification methods to their data will benefit from spending time reading this book. Second is Hand's (1997) excellent book on classification rules, which is particularly strong on assessing the performance of classifiers. The final book is Michie *et al*. (1994) which draws together the results from a series of comparative analyses. It has the added advantage of being available in a free download format.

#### 1.7.2 *Software*

There are many commercial packages but often people would like to try out analyses using free software. Fortunately there are several free, but professional, packages available. The two packages listed below are general purpose, and there are many other packages that are restricted to a small number of analyses