1

Introduction and Examples

This book is primarily about *pattern recognition*, which covers a wide range of activities from many walks of life. It is something which we humans are particularly good at; we receive data from our senses and are often able, immediately and without conscious effort, to identify the source of the data. For example, many of us can

recognize faces we have not seen for many years, even in disguise,
recognize voices over a poor telephone line,
as babies recognize our mothers by smell,
distinguish the grapes used to make a wine, and sometimes
even recognize the vineyard and year,
identify thousands of species of flowers and
spot an approaching storm.

Science, technology and business has brought to us many similar tasks, including

diagnosing diseases,
detecting abnormal cells in cervical smears,
recognizing dangerous driving conditions,
identifying types of car, aeroplane, . . . ,
identifying suspected criminals by fingerprints and DNA profiles,
reading Zip codes (US postal codes) on envelopes,
reading hand-written symbols (on a penpad computer),
reading maps and circuit diagrams,
classifying galaxies by shape,
picking an optimal move or strategy in a game such as chess,
identifying incoming missiles from radar or sonar signals,
detecting shoals of fish by sonar,
checking packets of frozen peas for 'foreign bodies',
spotting fake 'antique' furniture,

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

deciding which customers will be good credit risks and
spotting good opportunities on the financial markets.

Humans can (and do) do some of the tasks quite well, but the technological pressure is to build machines which can perform such tasks more accurately or faster or more cheaply than humans, or even to release humans from drudgery. There are also purely technological tasks such as reading bar codes at which humans are poor. *Pattern recognition* is the discipline of building such machines:

> 'It is felt that the decision-making processes of a human being are somewhat related to the recognition of patterns; for example the next move in a chess game is based upon the present position on the board, and buying or selling stocks is decided by a complex pattern of information. The goal of pattern recognition research is to clarify these complicated mechanisms of decision-making processes and to automate these functions using computers. However, because of the complex nature of the problem, most pattern recognition research has been concentrated on more realistic problems, such as the recognition of Latin characters and the classification of waveforms.'
> (Fukunaga, 1990, p. 1)

Since the best humans can perform many of these tasks very well, even better than the best machines, it has been of great interest to understand how we do so, and this is of independent scientific interest. So there has for many years been an interchange of ideas between engineers building pattern recognition systems and psychologists and physiologists studying human and animal brains. Twice this has led to great enthusiasm about machines influenced by ideas from psychology and biology. The first was in the late 1950s with the *perceptron*, the second in the mid 1980s over *neural networks*. Both rapidly left their biological roots, and were studied by mathematical techniques against engineering performance goals as pattern recognizers. This book is not about the impact of the study of neural networks as models of animal brains, but discusses what are more accurately (but rarely) called *artificial* neural networks which have been developed by a community which was originally biologically motivated (although many 'neural network' methods were not). Thus for the purposes of this book, a neural network is a method which arose or was popularized by the neural network community and has been or could be used for pattern recognition. Many of the originators of the current wave of interest were more careful in their terminology; whereas Hopfield (1982) did talk about neural networks, Rumelhart & McClelland (1986) used the term 'parallel distributed processing', and 'connectionist' has also been popular (for example, see Hinton, 1989a).

Marginal notes such as this replace footnotes and offer explanation, sidelines, and opinion.

Many of the ideas had arisen earlier in the pattern recognition context, but without the seductive titles had made little impact.

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

One characteristic of human pattern recognition is that it is mainly *learnt*. We cannot describe the rules we use to recognize a particular face, and will probably be unable to describe it well enough for anyone else to use the description for recognition. On the other hand, botanists can give the rules they use to identify flowering plants.

Gooseberries are the fruits of the species *Ribes grossularia*.

We should never underestimate the power of simply remembering some or all of the examples and comparing test examples with our memory.

Most learning involves a *teacher*. If we try enough different wines from unlabelled bottles, we may well discover that there are common groupings, and that one group has the aroma of gooseberries (if the latter have been experienced). But we will need a teacher to tell us that the common factor is that they were made (in part) from the *sauvignon blanc* grape. The discovery of new groupings is called *unsupervised* pattern recognition. A more common mode of learning both for us and for machines is to be given a collection of labelled examples, known as the *training set*, and from these to distil the essence of the grouping. This is *supervised* pattern recognition and is used to classify future examples into one of the same set of classes (or say it is none of these).

There is a subject known as *machine learning* which has emerged from the artificial intelligence and computer science communities. It too is concerned with distilling structure from labelled examples, although the labels are usually ‘true’ and ‘false’.

> ‘Machine Learning is generally taken to encompass automatic learning procedures based on logical or binary operations, that learn a task from a series of examples.’
>
> ‘Machine Learning aims to generate classifying expressions simple enough to be understood easily by humans. They must mimic human reasoning sufficiently well to provide insight into the decision process. Like statistical approaches, background knowledge may be exploited in development, but operation is assumed without human intervention.’ (Michie *et al.*, 1994, p. 2)

This stresses the need for a comprehensible explanation, which is needed in some but not all pattern recognition tasks. We have already noted that we cannot explain our identification of faces, and to recognize Zip codes no explanation is needed, just speed and accuracy.

This quotation mentions statistical approaches, and statistics is the oldest of the disciplines concerned with automatically finding structure in examples. As in the quotation, statistics is often thought of as being less automatic than the other disciplines, but this is largely an artefact of its greater age; its current research frontiers are very much concerned with replacing the human choice of methods by computation. Furthermore, statistics encompasses what the community of statisticians do, of whom your author is one!

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

1.1 How do neural methods differ?

Assertions are often made that neural networks provide a new approach to computing, involving analog (real-valued) rather than digital signals and massively parallel computation. For example, Haykin (1994, p. 2) offers a definition of a neural network adapted from Aleksander & Morton (1990):

> 'A neural network is a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:
>
> 1. Knowledge is acquired by the network through a learning process.
> 2. Interneuron connection strengths known as synaptic weights are used to store the knowledge.'

Many neural networks are excluded by this definition, including those of Kohonen. One could ask how a machine comes to have 'natural' properties.

In practice the vast majority of neural network applications are run on single-processor digital computers, although specialist parallel hardware is being developed (if not yet massively parallel). However, all the other methods we consider use real signals and can be parallelized to a considerable extent; it is far from clear that neural network methods will have an advantage as parallel computation becomes common, although they are frequently so slow that they need a speed-up. (Parallelization on real hardware has proved to be non-trivial; see Pitas, 1993 and Przytula & Prasanna, 1993.) We will argue that a large speed-up can be achieved by designing better learning algorithms using experience borrowed from other fields.

The traditional methods of statistics and pattern recognition are either *parametric* based on a family of models with a small number of parameters, or *non-parametric* in which the models used are totally flexible. One of the impacts of neural network methods on pattern recognition has been to emphasize the need in large-scale practical problems for something in between, families of models with large but not unlimited flexibility given by a large number of parameters. The two most widely used neural network architectures, *multi-layer perceptrons* and *radial basis functions* (RBFs), provide two such families (and several others already existed in statistics).

The name 'multi-layer perceptrons' is confusing; they are not multiple layers of perceptrons. We call them feed-forward neural nets.

Another difference in emphasis is on *'on-line'* methods, in which the data are not stored except through the changes the learning algorithm has made. The theory of such algorithms is studied for a very long stream of examples, but the practical distinction is less clear, as this stream is made up either by repeatedly cycling through the training set or by sampling the training examples (with replacement). In contrast, methods which use all the examples together are called *'batch'* methods.

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

It is often forgotten that there are intermediate positions, such as using small batches chosen from the training set.

1.2 The pattern recognition task

Except in Chapter 9 we will be exclusively concerned with supervised pattern recognition. Thus we are given a set of K pre-determined classes, and assume (in theory) the existence of an oracle that could correctly label each example which might be presented to us. When we receive an example, some measurements are made, known as *features*, and these data are fed into the pattern recognition machine, known as the *classifier*. This is allowed to report

Someone else may have made the measurements for us.

'this example is from class ℓ' or
'this example is from none of these classes' or
'this example is too hard for me'.

The second category are called *outliers* and the third *rejects* or *'doubt'* reports. Both can have great importance in applications. Suppose we have a medical diagnosis aid. We would want it to report any patient who apparently had an unknown disease, and we would also want it to ask the opinion of a senior doctor if there was real doubt. Often rejects are referred to a more expensive second tier of classification, perhaps a human or (as in Zip code recognition) a slower but more powerful method or even (as in analytical chemistry) for more expensive measurements to be made. Many pattern recognition systems always make a firm classification, but this seems to us more often to be bad design than a conscious decision that a firm decision was necessary.

It may help to know which classes are plausible.

The primary assessment of a system will be by its performance; a Zip code recognition system might be required to reject less than 2% of the examples and mis-read less than 0.5% of the remainder. In medical diagnosis we will be more interested in some errors than others, in particular in missing a disease, so the errors will need to be weighted. There may be a cost trade-off between rejection and error rate.

This might be unrealistic for hand-written addresses, and is well beyond current performance levels.

The other aspect of performance stressed in the quote from Michie *et al.* (1994, p. 2) is the power of explanation. Users need to have confidence in the system before it will be adopted. No one really cares if an odd letter is mis-routed, but patients do care if they are mis-diagnosed, and when a civilian airliner is mistaken for an enemy aircraft, questions are raised. So for some tasks 'black boxes' are unacceptable whatever their performance advantage (possibly even if they appear perfect on test). The methods of Chapters 7 and 8 are often found to be more acceptable for such tasks.

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

Some tasks are slightly different. We (and medics) often think of medical diagnosis as deciding which disease a patient has, but this ignores the possibility of two or more concurrent diseases; what we should really be asking is whether the patient has this disease for each of a range of diseases. This can be thought of as a compound decision, the classes each being a subset of the diseases, but it is normally helpful to make use of special structure within the classes.

Design issues

Although most of this book is about designing the pattern recognition machine, often the most important aspect of design is to choose the right features. If the wrong things are measured (or, more often these days with digital data, if the data are condensed too much) the task may be unachievable. Much of the enhanced success of Zip-code recognition systems has come from better features (for example, Simard *et al.*, 1993) rather than through more complicated classifiers. Sometimes good features can be found by training a classifier on a large number of features and extracting good ones (for example, by the methods of Chapters 9 and 10), but most often problem-specific insights are used.

In a few problem domains very specific rules are known which can be used to design a classifier; as an extreme example compilers can classify C programs as correct or invalid without needing to see any previous programs. Such information is often in the form of a formal *grammar*, and systems based on specifying such grammars are often called *syntactic* pattern recognition systems (Fu, 1982; Gonzalez & Thomason, 1978), but are of very restricted application. Allowing stochastic grammars in which the structure is given but the probabilities are learnt allows a little more flexibility. Chou (1989) gives an example of recognizing typeset mathematical expressions using a stochastic grammar.

In the vast majority of applications no structural assumptions are made, all the structure in the classifier being learnt from data. In the pattern recognition literature this is known as *statistical pattern recognition*. The training set is regarded as a sample from a population of possible examples, and the statistical similarities of each class extracted, or more precisely the significant differences between classes are found. A parametric or non-parametric model is constructed for the distribution of features for examples from each class, and statistical decision theory used to find an optimal classification. This is sometimes known (Dawid, 1976) as the *sampling paradigm*.

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

Another view, the *diagnostic paradigm*, goes back in the statistical literature at least to Cox (1958), and was developed in medical applications by Jerome Cornfield. This said that we were not interested in what the classes looked like, but only given an example in what the distribution over classes is *for similar examples*. The main method of this approach became known as *logistic discrimination* (Anderson, 1982), but was never widely known even in statistics and (as far as we could ascertain) appears in no pattern recognition text. This is the main approach of the neural network school.

When humans are learning concepts, we are often able to ask questions or to seek the classifications of examples which we synthesize (this being a paradigm of experimental science). Alternatively, we may describe our understanding to an expert, who will then supply a counter-example. Can we allow our machines to do the same? The idea has occurred in machine learning (Angluin, 1987, 1988, 1993), but apparently only for learning logical concepts.

We will sometimes have qualitative knowledge about the task in hand; we might know that only the sign of one of the features was material, or that the probability of a positive outcome was increasing in some continuous feature. Of course we should design the classifier to agree with such information, which Abu-Mostafa (1990, 1993, 1995a, b, c) calls 'hints'. Sometimes this is easy (just use the sign of the feature) but it can be very difficult (as in monotonicity). Generally hints (if true) help to avoid over-fitting to the training set, and this seems to be the real explanation of the gains in exchange-rate performance observed by Abu-Mostafa (1995a).

Method tuning and checking

All methods have some knobs which can be tweaked. Sometimes taking the class of the nearest training-set example is regarded as a fully automatic method, but we need to specify the metric used to find the nearest. (If the answer is 'use Euclidean distance' we still have to specify the units of measurement.)

How should those knobs be set? The most obvious way is to choose them to maximize performance. One thing we should *not* do is to evaluate the performance on a *test set* and choose the best-performing classifier, since we will then have no way to measure the true performance. We can keep back another test set, called a *validation set*, and use the performance on that to set the knobs. However, to obtain a sensitive measure of the performance, the validation set will

Note that this is not the procedure called *cross-validation*, despite the misuse of that term in the neural networks literature.

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

need to be very large, and this is data which could otherwise be used for learning.

This problem has been ignored for a long time, but now methods to use the training set for both learning and knob-setting are beginning to be used. These are discussed in Chapter 2 and illustrated on the quite small running examples that we chose.

To see why this is a real issue, consider Figure 1.1. Without knowing the true curve, it is hard to tell which of plots (b) and (c) is closer to the truth.

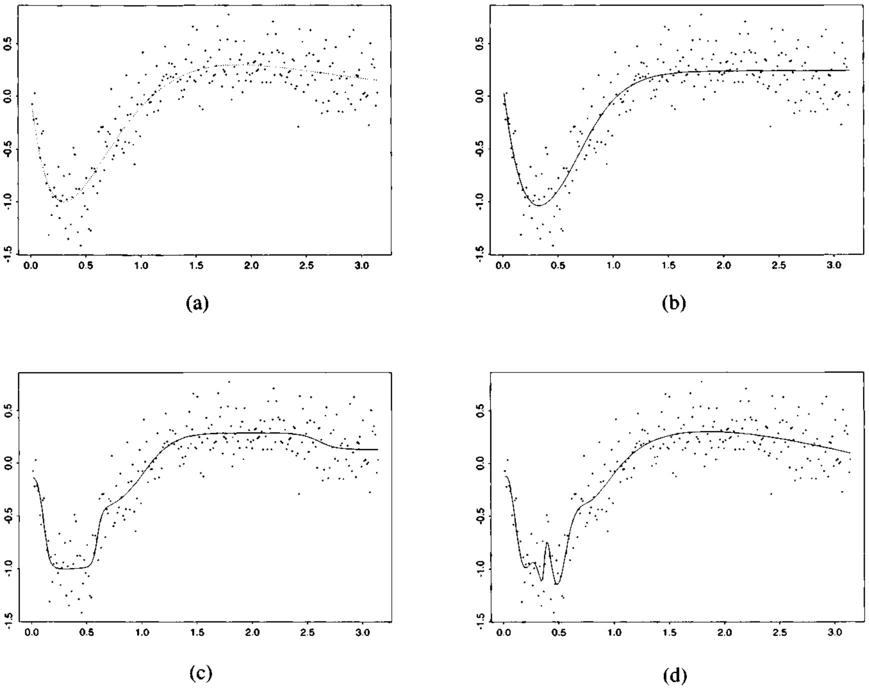

Figure 1.1: An illustration of model selection. Plot (a) shows 250 points generated by from the curve shown plus random noise, and plots (b–d) show fits by a single-hidden-layer neural network with 2, 4 and 8 hidden units.

Performance assessment

We will often want to choose between different candidate classifiers, and it will be usual to check that the performance targets are likely to be met. This needs an experimental test of the classifiers on some unseen examples. Such experiments are often (usually?) very poorly designed, and slanted towards a favourite method. The reader is urged to consult a good book on experimental design (such as Box *et al.*, 1978) before conducting such experiments.

Many of the experiments reported in the literature are designed to compare methods, when there is even more scope for confusion. In

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

medicine, methods (treatments) are compared in *double-blind* trials so there can be no preferential treatment, and in pure science experiments must be repeatable. (The large-scale trial of the StatLog project reported in Michie *et al.*, 1994, was designed to be run in these ways.) One source of confusion is that such trials may confuse the merits of the methods with the expertise of the experimenter in using them; this is a particular difficulty when the experimenter's own invention is in the trial. Two cases are of interest. One is where every method is used by a real expert and so assesses the best attainable performance. The other is when all methods are used by typical (or even new) users, which might provide a basis for recommendations to such users.

Prechelt (1994) surveyed two leading neural network journals for 1993 and half of 1994. He deemed an evaluation of an algorithm acceptable if it used two or more realistic or real problems and compared at least one alternative algorithm. Only 18% passed—in his words 'sad, but true'. Note that this book is not about evaluating algorithms, but we have used real examples to explore the merits and limitations of the methods. Amazingly, almost all books on pattern recognition or neural networks include no real or realistic examples.

This test does not consider experimental biases nor if an evaluation of the significance of the results was made.

1.3 Overview of the remaining chapters

Our approach to building a classifier will be based on statistical decision theory. In Chapter 2 we consider the Bayes rule, the best possible classifier if we knew everything about the population of examples, and then various approximations we can make if we have to learn from a training set. This includes several ways to use parametric models (which we assume to be false but perhaps convenient approximations); these sections include the classic methods based on the multidimensional normal distribution but also some improvements which are much less well known.

The next questions are: how complicated do our models need to be, and how well do they perform? These are discussed in Sections 2.6 and 2.7. There is a trade-off between adapting well to complexity of the real structure in the examples and fitting the structure of our particular training set (Figure 1.1). This explains why we are not interested in the usual asymptotics of mathematical statistics; as we receive more data we will want to choose more complicated models, and only limit the model complexity to avoid over-fitting the current training set. Another view of the effect of model flexibility on over-fitting is the study of *generalization* in Section 2.8.

Chapters 3 to 5 make weaker assumptions than standard parametric models. In Chapter 3 we study how we could use linear methods. Both Chapters 4 and 5 discuss how to apply flexible families of functions from the feature space $\mathcal{X}$ to d-dimensional Euclidean space $\mathbb{R}^d$, building on the linear methods, and consider the commonest such families, neural networks and radial basis functions, as well as splines and their generalizations.

The sixth chapter is on (nearly) non-parametric methods, where minimal assumptions are made about the classes. Most of these methods are based on looking at the classes of nearby examples, in some methods after designing a set of representative examples to replace the training set. That chapter also includes the use of mixtures of densities to model very general distributions.

Chapter 7 is about a rather different class of methods that partition the feature space $\mathcal{X}$ into regions and assign a class to each. This is done by splitting along a feature at a time, and then subdividing each subregion recursively. *Classification trees* have been considered in both statistics and machine learning; they are often easy to interpret but not amongst the highest performers.

Belief networks, also known as causal probability networks and Bayes networks, are not primarily designed for classification, but to explain the relationships between all of the observations. They are the subject of Chapter 8. They are very good for explanation, but may be less good for classification (as the finite amount of training data has to be used to learn more structure than just the relationship of the class to their features). Their strength is that they can incorporate qualitative knowledge about causal relationships amongst the features (an earlier and more sophisticated use of 'hints'). Also included in that chapter are the methods of Boltzmann machines and hierarchical mixtures of experts which can be considered within the framework of belief nets.

. . . and many more names beside

Chapters 9 and 10 are concerned with finding good features and choosing which features to use.

The appendix discusses a number of complements; some are statistical background and some explore issues a little further than is needed for pattern recognition.

1.4 Examples

The examples have been chosen to illustrate the properties of the methods we describe; not every method is used on each.

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

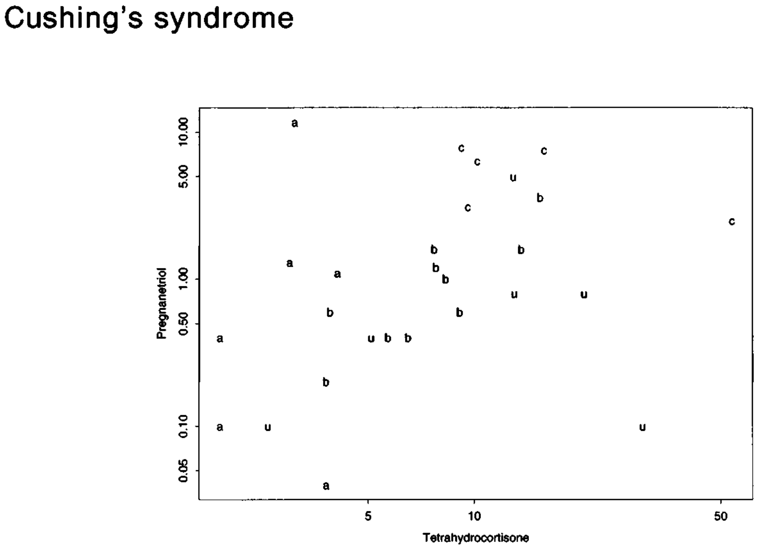

Figure 1.2: Results of two diagnostic tests on patients with Cushing's syndrome.

These data are taken from Aitchison & Dunsmore (1975, Tables 11.1–3) on diagnostic tests on patients with Cushing's syndrome, a hypersensitive disorder associated with over-secretion of cortisol by the adrenal gland. This dataset has three recognized types of the syndrome represented as a, b, c. (These encode 'adenoma', 'bilateral hyperplasia' and 'carcinoma', and represent the underlying cause of over-secretion. This can only be determined histopathologically.) The observations are urinary excretion rates (mg/24h) of the steroid metabolites tetrahydrocortisone and pregnanetriol, and are considered on log scale.

One of the patients of unknown type (marked u) was later found to be of a fourth type, and another was measured faultily.

Titterington (1976) discusses a different dataset which had 87 patients, five types, and fifteen measurements per patient, which suggests the current dataset is an abstraction of the full problem.

Synthetic two-class problem

This is a 'realistic' problem from Ripley (1994a), used there (and here) to illustrate how methods work. There are two features and two classes; each class has a bimodal distribution as should be clear from Figure 1.3. The class distributions were chosen to allow a best-possible error rate of about 8%, and are in fact equal mixtures of two normal distributions. The component normal distributions have a common covariance matrix.

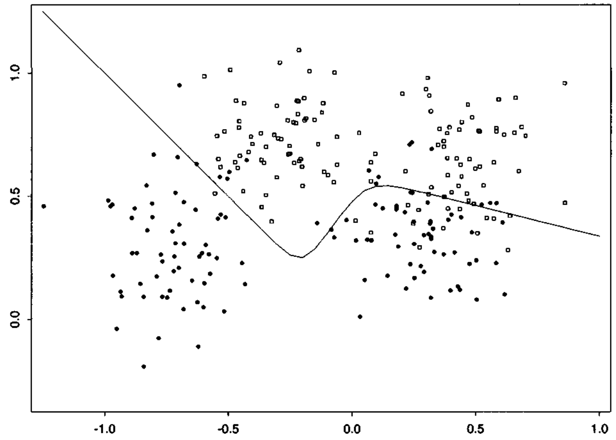

Figure 1.3: Two-class synthetic data from Ripley (1994a). The two classes are shown by solid circles and open squares: there are 125 points in each class.

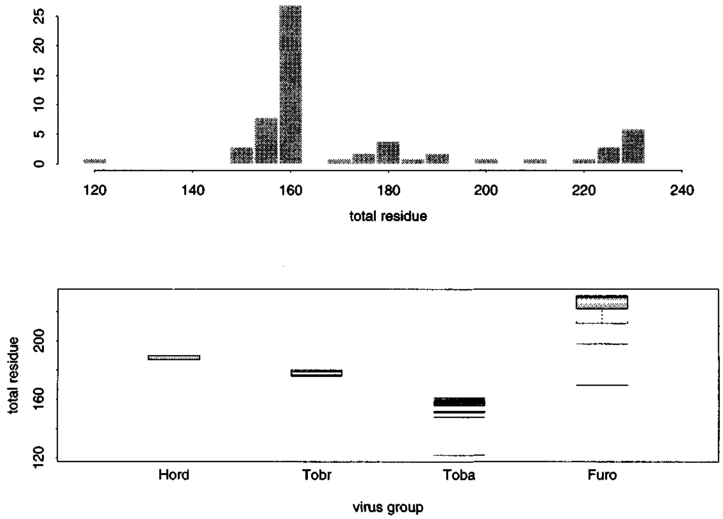

Figure 1.4: Histogram and boxplot by group of the viruses dataset. A boxplot is a representation of the distribution; the central grey box shows the middle 50% of the data, with median as a white bar. 'Whiskers' go out to plausible extremes, with outliers marked by bars.

Viruses

This is a dataset on 61 viruses with rod-shaped particles affecting various crops (tobacco, tomato, cucumber and others) described by Fauquet *et al.* (1988) and analysed by Eslava-Gómez (1989). There are 18 measurements on each virus, the number of amino acid residues per molecule of coat protein; the data come from a total of 26 sources. There is an existing classification by the number of RNA molecules and mode of transmission, into

No experimental details are provided in the source.

39 *Tobamoviruses* with monopartite genomes spread by contact,

6 *Tobraviruses* with bipartite genomes spread by nematodes,

3 *Hordeiviruses* with tripartite genomes, transmission mode unknown and

13 'furoviruses', 12 of which are known to be spread fungally.

The question of interest to Fauquet *et al.* was whether the furoviruses form a distinct group, and they performed various multivariate analyses.

One initial question with this dataset is whether the numbers of residues are absolute or relative. The data are counts from 0 to 32, with the totals per virus varying from 122 to 231. The average numbers for each amino acid range from 1.4 to 20.3. As a classification problem, this is very easy as Figure 1.4 shows. The histogram shows a multimodal distribution, and the boxplots show an almost complete separation by virus type. The only exceptional value is one virus in the furovirus group with a total of 170; this is the only virus in that group whose mode of transmission is unknown and Fauquet *et al.* (1988) suggest it has been tentatively classified as a *Tobamovirus*. The other outlier in that group (with a total of 198) is the only beet virus. The conclusions of Fauquet *et al.* may be drawn from the totals alone.

It is interesting to see if there are subgroups within the groups, so we will only use this dataset in Chapter 9, principally to investigate further the largest group (the *Tobamoviruses*). There are two viruses with identical scores, of which only one is included in the analyses. (No analysis of these data could differentiate between the two.)

Leptograpsus crabs

Campbell & Mahon (1974) studied rock crabs of the genus *Leptograpsus*. One species, *L. variegatus*, had been split into two new species, previously grouped by colour form, orange and blue. Preserved specimens lose their colour, so it was hoped that morphological differences would enable museum material to be classified.

Data are available on 50 specimens of each sex of each species, collected on sight at Fremantle, Western Australia. Each specimen has measurements on the width of the frontal lip `FL`, the rear width `RW`, and length along the midline `CL` and the maximum width `CW` of the carapace, and the body depth `BD` in mm.

Forensic glass

Our next example comes from forensic testing of glass collected by B. German on 214 fragments of glass, and taken from Murphy & Aha (1995). Each case has a measured refractive index and composition (weight percent of oxides of Na, Mg, Al, Si, K, Ca, Ba and Fe). The fragments were originally classed as seven types, one of which was absent in this dataset. The categories which occur are window float glass (70), window non-float glass (76), vehicle window glass (17), containers

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

(13), tableware (9) and vehicle headlamps (29). The composition sums to around 100%; what is not anything else is sand.

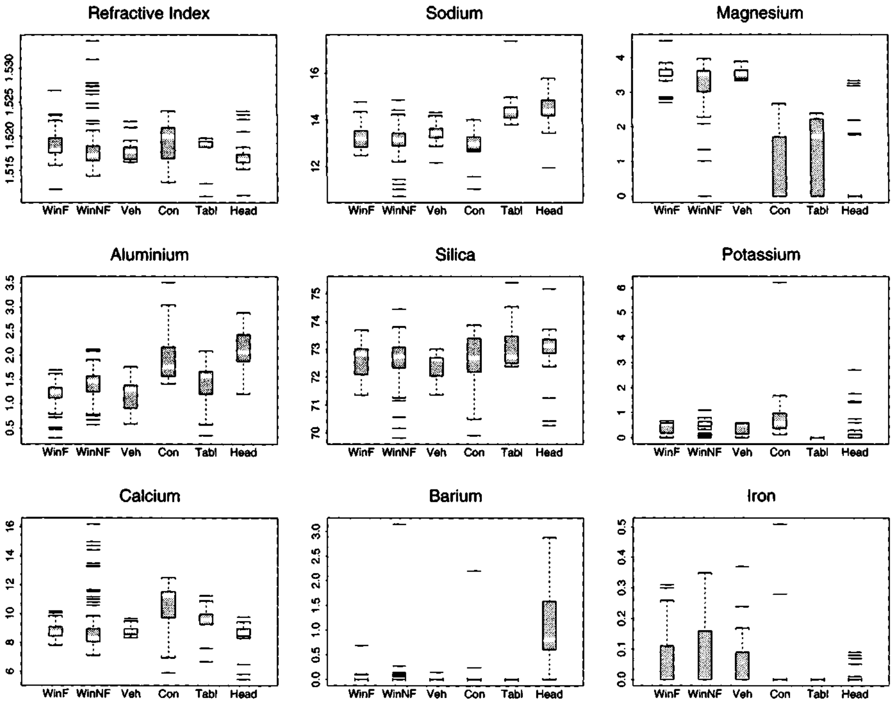

Figure 1.5: Boxplots of the features of the forensic glass data.

Figure 1.5 shows boxplots of the features. Some discrimination between glass types is apparent even from single features; for example headlamp glass is high in barium (although some examples have none), high in sodium and aluminium and low in iron. The three types of window glass appear similar, with one exceptional fragment of window non-float glass having a high refractive index, high barium and calcium and low magnesium and sodium. The containers group also contains a couple of exceptions. Characterizing populations with exceptions (especially 2 out of 13) can be difficult, and it may be easier to remove the exceptions in the training phase.

This example is really too small to divide, so methods have been assessed by 10-fold cross-validation using the same random partition for each method. The best methods have an estimated error rate of about 24%.

This is discussed in Section 2.7.

Diabetes in Pima Indians

A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

by the US National Institute of Diabetes and Digestive and Kidney Diseases, and are available from Murphy & Aha (1995). A previous report by Smith *et al.* (1988) found an error rate of about 24%. The reported variables are

number of pregnancies
plasma glucose concentration in an oral glucose tolerance test
diastolic blood pressure (mm Hg)
triceps skin fold thickness (mm)
serum insulin (μU/ml)
body mass index (weight in kg/(height in m)^2)
diabetes pedigree function
age in years

Many of these had zero values where these were impossible, so are taken to be missing values. Of the 768 records, 376 were incomplete (most prevalently in serum insulin). Most of our illustrations omit serum insulin and use the 532 complete records on the remaining variables. These were randomly split into a training set of size 200 and a test set of size 332. Methods which can deal with missing values were given 100 of the incomplete cases as part of the training set.

Note that 33% of the population were reported to have diabetes, so an error rate of 33% can be achieved by declaring all test cases to be non-diabetic. Our best methods reduce this to about 20%.

Some aspects of this dataset were considered by Wahba *et al.* (1995).

Data availability

All these datasets are available by anonymous ftp from the Internet site

`ftp.stats.ox.ac.uk`   IP address `163.1.20.1`

in directory `/pub/PRNN`. The datasets and other material are available by pointing your World Wide Web browser at

`http://www.stats.ox.ac.uk/~ripley/PRbook/`

1.5 Literature

The classic books on pattern recognition are Duda & Hart (1973), Devijver & Kittler (1982) and Fukunaga (2nd edn 1990), all of which pre-date the impact of neural networks on the subject. There are a small number of introductory texts (James, 1988; Therrien, 1989; Schalkoff, 1992) and two specialist monographs on kernel methods

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press

(Hand, 1982; Coomans & Broeckaert, 1986). Some conference proceedings, for example Devijver & Kittler (1987), provide a good overview of applications.

Classical statistical techniques are discussed in most texts on *multivariate analysis* such as T. W. Anderson (1984) and Mardia *et al.* (1979) and in slightly more specialized books by Lachenbruch (1975), Goldstein & Dillon (1978), Hand (1981) and McLachlan (1992).

There are now very many books on neural networks, particularly on parts of the subject not discussed here. Approaches to modelling memory from the point of view of statistical physics are covered by Amit (1989), Peretto (1992) and Hertz *et al.* (1991). Haykin (1994) is modern, comprehensive but unselective (and untroubled by real applications). Amari (1993) and Ripley (1993) give two statistical views of the neural network field, and Bishop (1995a) is slanted towards pattern recognition. Arbib (1995) provides many short sketches of topics over a very wide range of neural networks. One important area of neural network methods which we do not consider is the prediction of time series, the subject of a competition analysed by Weigend & Gershenfeld (1993), including expository papers.

There is now one text on general machine learning, Langley (1996), and it appears in some artificial intelligence texts (for example, Winston, 1992; Russell & Norvig, 1995). There are many more aspects than we shall consider, including incorporating domain knowledge as illustrated by King *et al.* (1992). Langley & Simon (1995) and Bratko & Muggleton (1995) discuss applications of machine learning with claimed real-world benefits.

Books which cover more than one of these three areas are rare. Krishnaiah & Kanal (1982) was a very good overview at its time; the recent edited volumes by Cherkassky *et al.* (1994) and Michie *et al.* (1994) contain several good overviews.

Face recognition is a popular application of pattern recognition surveyed by Samal & Iyengar (1992). Golomb *et al.* (1991) and Flocchini *et al.* (1992) give two example systems.

There is a very large literature on character recognition, and non-European alphabets with at least hundreds of classes provide a severe test of pattern recognition methods. The articles by Baird (1993), Cohen *et al.* (1991), Le Cun *et al.* (1989, 1990a), Gader *et al.* (1991), Guyon *et al.* (1992), Impedovo *et al.* (1991), Knerr *et al.* (1991), Lee (1991), Martin & Pitman (1990, 1991), Pavlidis (1993), Simard *et al.* (1993), Singer & Tishby (1994), Suen *et al.* (1992, 1993) and Wakahara (1993) provide some flavours.

https://doi.org/10.1017/CBO9780511812651.002 Published online by Cambridge University Press