2

Classification

R. J. Henery
University of Strathclyde[1]

2.1 DEFINITION OF CLASSIFICATION

Classification has two distinct meanings. We may be given a set of observations with the aim of establishing the existence of classes or clusters in the data. Or we may know for certain that there are so many classes, and the aim is to establish a rule whereby we can classify a new observation into one of the existing classes. The former type is known as Unsupervised Learning (or Clustering), the latter as Supervised Learning. In this book when we use the term classification, we are talking of Supervised Learning. In the statistical literature, Supervised Learning is usually, but not always, referred to as discrimination, by which is meant the establishing of the classification rule from given correctly classified data.

The existence of correctly classified data presupposes that someone (the Supervisor) is able to classify without error, so the question naturally arises: why is it necessary to replace this exact classification by some approximation?

2.1.1 Rationale

There are many reasons why we may wish to set up a classification procedure, and some of these are discussed later in relation to the actual datasets used in this book. Here we outline possible reasons for the examples in Section 1.2.

1. Mechanical classification procedures may be much faster: for example, postal code reading machines may be able to sort the majority of letters, leaving the difficult cases to human readers.
2. A mail order firm must take a decision on the granting of credit purely on the basis of information supplied in the application form: human operators may well have biases, *i.e.* may make decisions on irrelevant information and may turn away good customers.

[1] *Address for correspondence*: Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, U.K.

3. In the medical field, we may wish to avoid the surgery that would be the only sure way of making an exact diagnosis, so we ask if a reliable diagnosis can be made on purely external symptoms.
4. The Supervisor (refered to above) may be the verdict of history, as in meteorology or stock-exchange transaction or investment and loan decisions. In this case the issue is one of forecasting.

2.1.2 Issues

There are also many issues of concern to the would-be classifier. We list below a few of these.

- Accuracy. There is the reliability of the rule, usually represented by the proportion of correct classifications, although it may be that some errors are more serious than others, and it may be important to control the error rate for some key class.
- Speed. In some circumstances, the speed of the classifier is a major issue. A classifier that is 90% accurate may be preferred over one that is 95% accurate if it is 100 times faster in testing (and such differences in time-scales are not uncommon in neural networks for example). Such considerations would be important for the automatic reading of postal codes, or automatic fault detection of items on a production line for example.
- Comprehensibility. If it is a human operator that must apply the classification procedure, the procedure must be easily understood else mistakes will be made in applying the rule. It is important also, that human operators believe the system. An oft-quoted example is the Three-Mile Island case, where the automatic devices correctly recommended a shutdown, but this recommendation was not acted upon by the human operators who did not believe that the recommendation was well founded. A similar story applies to the Chernobyl disaster.
- Time to Learn. Especially in a rapidly changing environment, it may be necessary to learn a classification rule quickly, or make adjustments to an existing rule in real time. "Quickly" might imply also that we need only a small number of observations to establish our rule.

At one extreme, consider the naïve 1-nearest neighbour rule, in which the training set is searched for the 'nearest' (in a defined sense) previous example, whose class is then assumed for the new case. This is very fast to learn (no time at all!), but is very slow in practice if all the data are used (although if you have a massively parallel computer you might speed up the method considerably). At the other extreme, there are cases where it is very useful to have a quick-and-dirty method, possibly for eyeball checking of data, or for providing a quick cross-checking on the results of another procedure. For example, a bank manager might know that the simple rule-of-thumb "only give credit to applicants who already have a bank account" is a fairly reliable rule. If she notices that the new assistant (or the new automated procedure) is mostly giving credit to customers who do **not** have a bank account, she would probably wish to check that the new assistant (or new procedure) was operating correctly.

2.1.3 Class definitions

An important question, that is improperly understood in many studies of classification, is the nature of the classes and the way that they are defined. We can distinguish three common cases, only the first leading to what statisticians would term classification:

1. Classes correspond to labels for different populations: membership of the various populations is not in question. For example, dogs and cats form quite separate classes or populations, and it is known, with certainty, whether an animal is a dog or a cat (or neither). Membership of a class or population is determined by an independent authority (the *Supervisor*), the allocation to a class being determined independently of any particular attributes or variables.
2. Classes result from a prediction problem. Here class is essentially an outcome that must be predicted from a knowledge of the attributes. In statistical terms, the class is a random variable. A typical example is in the prediction of interest rates. Frequently the question is put: will interest rates rise (class=1) or not (class=0).
3. Classes are pre-defined by a partition of the sample space, *i.e.* of the attributes themselves. We may say that class is a function of the attributes. Thus a manufactured item may be classed as faulty if some attributes are outside predetermined limits, and not faulty otherwise. There is a rule that has already classified the data from the attributes: the problem is to create a rule that mimics the actual rule as closely as possible. Many credit datasets are of this type.

In practice, datasets may be mixtures of these types, or may be somewhere in between.

2.1.4 Accuracy

On the question of accuracy, we should always bear in mind that accuracy as measured on the training set and accuracy as measured on unseen data (the test set) are often very different. Indeed it is not uncommon, especially in Machine Learning applications, for the training set to be perfectly fitted, but performance on the test set to be very disappointing. Usually, it is the accuracy on the unseen data, when the true classification is unknown, that is of practical importance. The generally accepted method for estimating this is to use the given data, in which we assume that all class memberships are known, as follows. Firstly, we use a substantial proportion (the training set) of the given data to train the procedure. This rule is then tested on the remaining data (the test set), and the results compared with the known classifications. The proportion correct in the test set is an unbiased estimate of the accuracy of the rule provided that the training set is randomly sampled from the given data.

2.2 EXAMPLES OF CLASSIFIERS

To illustrate the basic types of classifiers, we will use the well-known Iris dataset, which is given, in full, in Kendall & Stuart (1983). There are three varieties of Iris: Setosa, Versicolor and Virginica. The length and breadth of both petal and sepal were measured on 50 flowers of each variety. The original problem is to classify a new Iris flower into one of these three types on the basis of the four attributes (petal and sepal length and width). To keep this example simple, however, we will look for a classification rule by which the varieties can be distinguished purely on the basis of the two measurements on Petal Length and Width. We have available fifty pairs of measurements of each variety from which to learn the classification rule.

2.2.1 Fisher's linear discriminants

This is one of the oldest classification procedures, and is the most commonly implemented in computer packages. The idea is to divide sample space by a series of lines in two dimensions, planes in 3-D and, generally hyperplanes in many dimensions. The line dividing two classes is drawn to bisect the line joining the centres of those classes, the direction of the line is determined by the shape of the clusters of points. For example, to differentiate between Versicolor and Virginica, the following rule is applied:

- If Petal Width $< 3.272 - 0.3254\times$Petal Length, then Versicolor.
- If Petal Width $> 3.272 - 0.3254\times$Petal Length, then Virginica.

Fisher's linear discriminants applied to the Iris data are shown in Figure 2.1. Six of the observations would be misclassified.

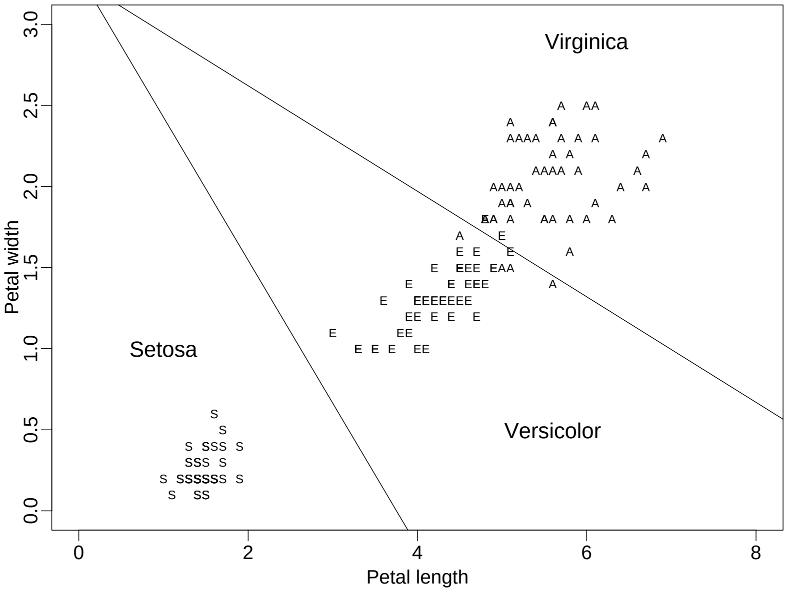

Fig. 2.1: Classification by linear discriminants: Iris data.

2.2.2 Decision tree and Rule-based methods

One class of classification procedures is based on recursive partitioning of the sample space. Space is divided into boxes, and at each stage in the procedure, each box is examined to see if it may be split into two boxes, the split usually being parallel to the coordinate axes. An example for the Iris data follows.

- If Petal Length $<$ 2.65 then Setosa.
- If Petal Length $>$ 4.95 then Virginica.

- If $2.65 <$ Petal Length < 4.95 then :

  if Petal Width < 1.65 then Versicolor;

  if Petal Width > 1.65 then Virginica.

The resulting partition is shown in Figure 2.2. Note that this classification rule has three mis-classifications.

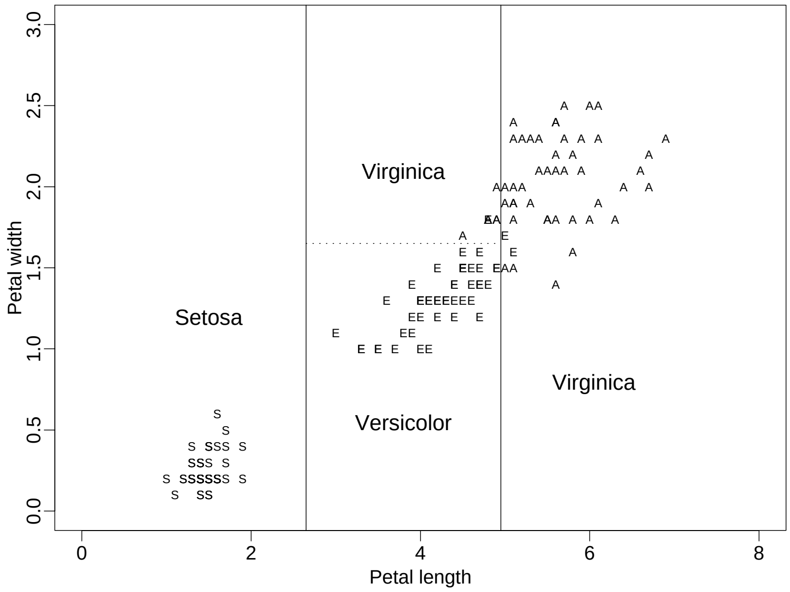

Fig. 2.2: Classification by decision tree: Iris data.

Weiss & Kapouleas (1989) give an alternative classification rule for the Iris data that is very directly related to Figure 2.2. Their rule can be obtained from Figure 2.2 by continuing the dotted line to the left, and can be stated thus:

- If Petal Length < 2.65 then Setosa.
- If Petal Length > 4.95 or Petal Width > 1.65 then Virginica.
- Otherwise Versicolor.

Notice that this rule, while equivalent to the rule illustrated in Figure 2.2, is stated more concisely, and this formulation may be preferred for this reason. Notice also that the rule is ambiguous if Petal Length < 2.65 and Petal Width > 1.65. The quoted rules may be made unambiguous by applying them in the given order, and they are then just a re-statement of the previous decision tree. The rule discussed here is an instance of a rule-based method: such methods have very close links with decision trees.

2.2.3 k-Nearest-Neighbour

We illustrate this technique on the Iris data. Suppose a new Iris is to be classified. The idea is that it is most likely to be near to observations from its own proper population. So we look at the five (say) nearest observations from all previously recorded Irises, and classify the observation according to the most frequent class among its neighbours. In Figure 2.3, the new observation is marked by a $+$, and the 5 nearest observations lie within the **circle** centred on the $+$. The apparent elliptical shape is due to the differing horizontal and vertical scales, but the proper scaling of the observations is a major difficulty of this method.

This is illustrated in Figure 2.3 , where an observation centred at $+$ would be classified as Virginica since it has 4 Virginica among its 5 nearest neighbours.

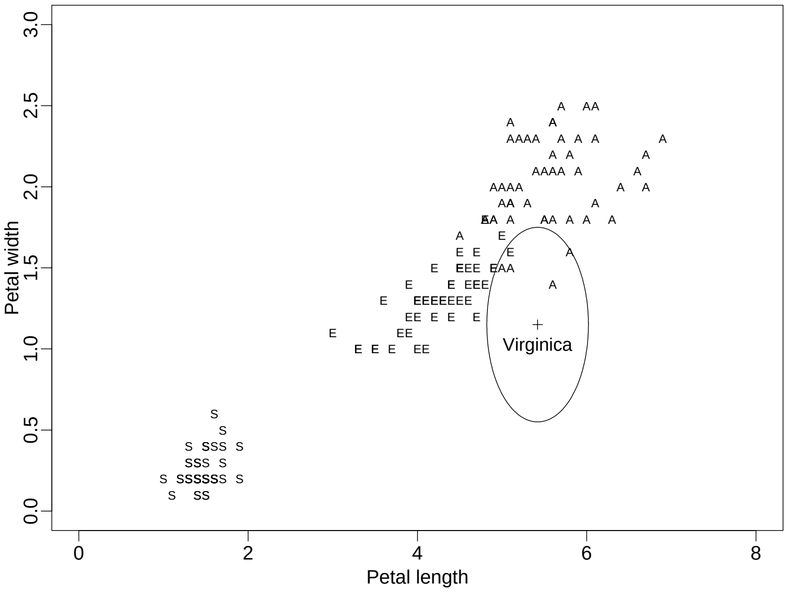

Fig. 2.3: Classification by 5-Nearest-Neighbours: Iris data.

2.3 CHOICE OF VARIABLES

As we have just pointed out in relation to k-nearest neighbour, it may be necessary to reduce the weight attached to some variables by suitable scaling. At one extreme, we might remove some variables altogether if they do not contribute usefully to the discrimination, although this is not always easy to decide. There are established procedures (for example, forward stepwise selection) for removing unnecessary variables in linear discriminants, but, for large datasets, the performance of linear discriminants is not seriously affected by including such unnecessary variables. In contrast, the presence of irrelevant variables is always a problem with k-nearest neighbour, regardless of dataset size.

2.3.1 Transformations and combinations of variables

Often problems can be simplified by a judicious transformation of variables. With statistical procedures, the aim is usually to transform the attributes so that their marginal density is approximately normal, usually by applying a monotonic transformation of the power law type. Monotonic transformations do not affect the Machine Learning methods, but they can benefit by combining variables, for example by taking ratios or differences of key variables. Background knowledge of the problem is of help in determining what transformation or combination to use. For example, in the Iris data, the product of the variables Petal Length and Petal Width gives a single attribute which has the dimensions of area, and might be labelled as Petal Area. It so happens that a decision rule based on the single variable Petal Area is a good classifier with only four errors:

- If Petal Area $<$ 2.0 then Setosa.
- If 2.0 $<$ Petal Area $<$ 7.4 then Virginica.
- If Petal Area $>$ 7.4 then Virginica.

This tree, while it has one more error than the decision tree quoted earlier, might be preferred on the grounds of conceptual simplicity as it involves only one "concept", namely Petal Area. Also, one less arbitrary constant need be remembered (*i.e.* there is one less node or cut-point in the decision trees).

2.4 CLASSIFICATION OF CLASSIFICATION PROCEDURES

The above three procedures (linear discrimination, decision-tree and rule-based, k-nearest neighbour) are prototypes for three types of classification procedure. Not surprisingly, they have been refined and extended, but they still represent the major strands in current classification practice and research. The 23 procedures investigated in this book can be directly linked to one or other of the above. However, within this book the methods have been grouped around the more traditional headings of classical statistics, modern statistical techniques, Machine Learning and neural networks. Chapters 3 – 6, respectively, are devoted to each of these. For some methods, the classification is rather abitrary.

2.4.1 Extensions to linear discrimination

We can include in this group those procedures that start from linear combinations of the measurements, even if these combinations are subsequently subjected to some non-linear transformation. There are 7 procedures of this type: Linear discriminants; logistic discriminants; quadratic discriminants; multi-layer perceptron (backprop and cascade); DIPOL92; and projection pursuit. Note that this group consists of statistical and neural network (specifically multilayer perceptron) methods only.

2.4.2 Decision trees and Rule-based methods

This is the most numerous group in the book with 9 procedures: NewID; AC^2; Cal5; CN2; C4.5; CART; IndCART; Bayes Tree; and ITrule (see Chapter 5).

2.4.3 Density estimates

This group is a little less homogeneous, but the 7 members have this in common: the procedure is intimately linked with the estimation of the local probability density at each point in sample space. The density estimate group contains: k-nearest neighbour; radial basis functions; Naive Bayes; Polytrees; Kohonen self-organising net; LVQ; and the kernel density method. This group also contains only statistical and neural net methods.

2.5 A GENERAL STRUCTURE FOR CLASSIFICATION PROBLEMS

There are three essential components to a classification problem.

1. The relative frequency with which the classes occur in the population of interest, expressed formally as the *prior* probability distribution.

2. An implicit or explicit criterion for separating the classes: we may think of an underlying input/output relation that uses observed attributes to distinguish a random individual from each class.
3. The cost associated with making a wrong classification.

Most techniques implicitly confound components and, for example, produce a classification rule that is derived conditional on a particular prior distribution and cannot easily be adapted to a change in class frequency. However, in theory each of these components may be individually studied and then the results formally combined into a classification rule. We shall describe this development below.

2.5.1 Prior probabilities and the Default rule

We need to introduce some notation. Let the classes be denoted $A_i, i = 1, \ldots, q$, and let the prior probability π_i for the class A_i be:

$$\pi_i \;=\; p(A_i)$$

It is always possible to use the no-data rule: *classify any new observation as class A_k, irrespective of the attributes of the example.* This *no-data* or *default* rule may even be adopted in practice if the cost of gathering the data is too high. Thus, banks may give credit to all their established customers for the sake of good customer relations: here the cost of gathering the data is the risk of losing customers. The default rule relies only on knowledge of the prior probabilities, and clearly the decision rule that has the greatest chance of success is to allocate every new observation to the most frequent class. However, if some classification errors are more serious than others we adopt the minimum risk (least expected cost) rule, and the class k is that with the least expected cost (see below).

2.5.2 Separating classes

Suppose we are able to observe data x on an individual, and that we know the probability distribution of x within each class A_i to be $P(x|A_i)$. Then for any two classes A_i, A_j the *likelihood ratio* $P(x|A_i)/P(x|A_j)$ provides the theoretical optimal form for discriminating the classes on the basis of data x. The majority of techniques featured in this book can be thought of as implicitly or explicitly deriving an approximate form for this likelihood ratio.

2.5.3 Misclassification costs

Suppose the cost of misclassifying a class A_i object as class A_j is $c(i, j)$. Decisions should be based on the principle that the total cost of misclassifications should be minimised: for a new observation this means minimising the expected cost of misclassification.

Let us first consider the expected cost of applying the default decision rule: *allocate all new observations to the class A_d*, using suffix d as label for the decision class. When decision A_d is made for all new examples, a cost of $c(i, d)$ is incurred for class A_i examples and these occur with probability π_i. So the expected cost C_d of making decision A_d is:

$$C_d \;=\; \sum_i \pi_i c(i, d)$$

The Bayes minimum cost rule chooses that class that has the lowest expected cost. To see the relation between the minimum error and minimum cost rules, suppose the cost of misclassifications to be the same for all errors and zero when a class is correctly identified, *i.e.* suppose that $c(i,j) = c$ for $i \neq j$ and $c(i,j) = 0$ for $i = j$.

Then the expected cost is

$$C_d = \sum_i \pi_i c(i,d) = \sum_{i \neq d} \pi_i c = c \sum_{i \neq d} \pi_i = c(1 - \pi_d)$$

and the minimum cost rule is to allocate to the class with the greatest prior probability.

Misclassification costs are very difficult to obtain in practice. Even in situations where it is very clear that there are very great inequalities in the sizes of the possible penalties or rewards for making the wrong or right decision, it is often very difficult to quantify them. Typically they may vary from individual to individual, as in the case of applications for credit of varying amounts in widely differing circumstances. In one dataset we have assumed the misclassification costs to be the same for all individuals. (In practice, credit-granting companies must assess the potential costs for each applicant, and in this case the classification algorithm usually delivers an assessment of probabilities, and the decision is left to the human operator.)

2.6 BAYES RULE GIVEN DATA x

We can now see how the three components introduced above may be combined into a classification procedure.

When we are given information x about an individual, the situation is, in principle, unchanged from the no-data situation. The difference is that all probabilities must now be interpreted as conditional on the data x. Again, the decision rule with least probability of error is to allocate to the class with the highest probability of occurrence, but now the relevant probability is the conditional probability $p(A_i|x)$ of class A_i given the data x:

$$p(A_i|x) = \text{Prob(class} A_i \text{ given } x)$$

If we wish to use a minimum cost rule, we must first calculate the expected costs of the various decisions *conditional on the given information* x.

Now, when decision A_d is made for examples with attributes x, a cost of $c(i,d)$ is incurred for class A_i examples and these occur with probability $p(A_i|x)$. As the probabilities $p(A_i|x)$ depend on x, so too will the decision rule. So too will the expected cost $C_d(x)$ of making decision A_d:

$$C_d(x) = \sum_i p(A_i|x) c(i,d)$$

In the special case of equal misclassification costs, the minimum cost rule is to allocate to the class with the greatest *posterior probability*.

When Bayes theorem is used to calculate the conditional probabilities $p(A_i|x)$ for the classes, we refer to them as the *posterior probabilities* of the classes. Then the posterior probabilities $p(A_i|x)$ are calculated from a knowledge of the prior probabilities π_i and the conditional probabilities $P(x|A_i)$ of the data for each class A_i. Thus, for class A_i suppose that the probability of observing data x is $P(x|A_i)$. Bayes theorem gives the posterior probability $p(A_i|x)$ for class A_i as:

$$p(A_i|x) = \pi_i P(x|A_i) / \sum_j \pi_j P(x|A_j)$$

The divisor is common to all classes, so we may use the fact that $p(A_i|x)$ is proportional to $\pi_i P(x|A_i)$. The class A_d with minimum expected cost (minimum risk) is therefore that for which

$$\sum_i \pi_i c(i,d) P(x|A_i)$$

is a minimum.

Assuming now that the attributes have continuous distributions, the probabilities above become probability densities. Suppose that observations drawn from population A_i have probability density function $f_i(x) = f(x \mid A_i)$ and that the prior probability that an observation belongs to class A_i is π_i. Then Bayes' theorem computes the probability that an observation x belongs to class A_i as

$$p(A_i|x) \;=\; \pi_i f_i(x) / \sum_j \pi_j f_j(x)$$

A classification rule then assigns x to the class A_d with maximal *a posteriori* probability given x:

$$p(A_d|x) = \max_i p(A_i|x)$$

As before, the class A_d with minimum expected cost (minimum risk) is that for which

$$\sum_i \pi_i c(i,d) f_i(x)$$

is a minimum.

Consider the problem of discriminating between just two classes A_i and A_j. Then assuming as before that $c(i,i) = c(j,j) = 0$, we should allocate to class i if

$$\pi_j c(i,j) f_j(x) < \pi_i c(j,i) f_i(x)$$

or equivalently

$$\frac{f_i(x)}{f_j(x)} > \frac{\pi_j}{\pi_i} \frac{c(i,j)}{c(j,i)}$$

which shows the pivotal role of the likelihood ratio, which must be greater than the ratio of prior probabilities times the relative costs of the errors. We note the symmetry in the above expression: changes in costs can be compensated in changes in prior to keep constant the threshold that defines the classification rule - this facility is exploited in some techniques, although for more than two groups this property only exists under restrictive assumptions (see Breiman *et al.*, page 112).

2.6.1 Bayes rule in statistics

Rather than deriving $p(A_i|x)$ via Bayes theorem, we could also use the empirical frequency version of Bayes rule, which, in practice, would require prohibitively large amounts of data. However, in principle, the procedure is to gather together all examples in the training set that have the same attributes (exactly) as the given example, and to find class proportions $p(A_i|x)$ among these examples. The minimum error rule is to allocate to the class A_d with highest posterior probability.

Unless the number of attributes is very small and the training dataset very large, it will be necessary to use approximations to estimate the posterior class probabilities. For example, one way of finding an approximate Bayes rule would be to use not just examples with attributes matching exactly those of the given example, but to use examples that were near the given example in some sense. The minimum error decision rule would be to allocate to the most frequent class among these matching examples. Partitioning algorithms, and decision trees in particular, divide up attribute space into regions of self-similarity: all data within a given box are treated as similar, and posterior class probabilities are constant within the box.

Decision rules based on Bayes rules are optimal - no other rule has lower expected error rate, or lower expected misclassification costs. Although unattainable in practice, they provide the logical basis for all statistical algorithms. They are unattainable because they assume complete information is known about the statistical distributions in each class. Statistical procedures try to supply the missing distributional information in a variety of ways, but there are two main lines: *parametric* and *non-parametric*. Parametric methods make assumptions about the nature of the distributions (commonly it is assumed that the distributions are Gaussian), and the problem is reduced to estimating the parameters of the distributions (means and variances in the case of Gaussians). Non-parametric methods make no assumptions about the specific distributions involved, and are therefore described, perhaps more accurately, as *distribution-free*.

2.7 REFERENCE TEXTS

There are several good textbooks that we can recommend. Weiss & Kulikowski (1991) give an overall view of classification methods in a text that is probably the most accessible to the Machine Learning community. Hand (1981), Lachenbruch & Mickey (1975) and Kendall *et al.* (1983) give the statistical approach. Breiman *et al.* (1984) describe CART, which is a partitioning algorithm developed by statisticians, and Silverman (1986) discusses density estimation methods. For neural net approaches, the book by Hertz *et al.* (1991) is probably the most comprehensive and reliable. Two excellent texts on pattern recognition are those of Fukunaga (1990) , who gives a thorough treatment of classification problems, and Devijver & Kittler (1982) who concentrate on the k-nearest neighbour approach. A thorough treatment of statistical procedures is given in McLachlan (1992), who also mentions the more important alternative approaches. A recent text dealing with pattern recognition from a variety of perspectives is Schalkoff (1992).