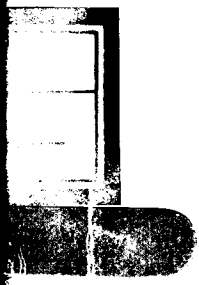
Classification Algorithms

Mike James



Classification Algorithms

Mike James

COLLINS
8 Grafton Street, London W1

Collins Professional and Technical Books
William Collins Sons & Co. Ltd
8 Grafton Street, London W1X 3LA

First published in Great Britain by
Collins Professional and Technical Books 1985

British Library Cataloguing in Publication Data
James, M.
Classification algorithms.
1. Classification
I. Title
001'.01'2 Z696

ISBN 0 00-383054-3

Typeset by V & M Graphics Ltd, Aylesbury, Bucks
Printed and bound in Great Britain by
Mackays of Chatham, Kent

Preface

Classification, the assignment of an object to one of a number of predetermined groups, is of fundamental importance in many areas of science and technology. Its very importance has forced each discipline to develop its own methods and terminology and as a result classification has become a collection of islands with little communication and cross fertilisation. Even the term 'classification' is far from standard and it is not difficult to find disciplines where the terms 'pattern recognition', 'discriminant analysis' and 'supervised learning' are far more common. I have chosen the title 'classification' rather than any of the others because it is the least partisan and the most directly descriptive of the task in hand – I can only hope that specialists in subjects where other terminology reigns will not feel too neglected.

Classification is a subject where intuition can be used to construct practical procedures that work very well and this has sometimes resulted in the loss of a sound theoretical basis for the work. While this may sometimes be unimportant – a classification rule that works is all that can be hoped for – this lack of theory can mean that more difficult problems remain unsolved. Not only can a lack of theory restrict the range of problems that be solved, it can often result in an over-optimistic opinion that a problem has been solved. For example, many methods of classification have been criticised for appearing to work well during the design and testing phase but being inadequate when used 'for real'. In this case the fault usually lies with the inadequate but intuitively obvious test procedures used to estimate the success that the rule will achieve. The theory of classification described in this book is essentially concerned with finding an optimal classification rule, ways of assessing its performance and ways of understanding how it achieves its results. Most of this theory is traditional statistics but much of it comes from the newer areas of pattern recognition and artificial intelligence. So important are these two areas of study that Chapter 10 is devoted to a description of some of their very individual approaches to classification.

Although the theory of classification is explained in the following

chapters the practice of classification has not been forgotten! Wherever possible any methods that are described are also accompanied by programs written in BASIC. This not only allows the reader to try the methods, it also clears up any ambiguity about how a method is to be applied. To further enable the reader to become familiar with the methods a data generation program is given in Appendix 2 and this combined with the other programs can be used to explore the ideas and methods before confronting any real data. The analysis programs themselves have been written to be compact so as to make their entry via a keyboard easier but they are suitable for practical application. (Disk versions of these programs are available, see Appendix 2 for details.)

Classification theory still contains many unanswered questions but we are now at a stage where many classification tasks can be automated. Much of the practical work that remains to be done is the application of known methods to create computer programs that are easy to use and fit in with our human ways of working. The future of classification is therefore a joint enterprise in which statisticians, computer scientists and practitioners in many diverse fields all have a contribution to make.

My grateful thanks are due to Bob Mays for the impetus to start writing this book, to Annette Harris and Jane Patience for help in the preparation of the manuscript, to Bernard Watson, Richard Miles and Paul Stringer of Collins Professional and Technical Books for its final production, and also to my wife Sue for her interest and encouragement throughout the project.

Mike James

Notation

$\|A\|$	Determinant of **A**.
$d_k(x)$	Quadratic discriminant for group k.
D^2	Sample Mahalanobis distance squared.
e	Total number of cases misclassified.
e_i	Eigenvalue i.
E(x)	Mean or average of x.
E(x|A)	Mean or average of x given A.
E_1	Estimate of error rate using independent samples and unknown apriori probabilities.
E_2	Estimate of error rate using independent samples and known apriori probabilities.
E_A	Apparent error rate.
E_L	Leaving-one-out error rate.
E_{ij}	Number of cases from group i classified as group j.
E_i^*	Number of cases misclassified from group i.
f_k	Linear discriminant function for group k.
f_p	Population linear discriminant function.
f_s	Sample linear discriminant function.
G	Number of groups.
ln	Natural log.
M,m	Total sample size.
m_k	Number of cases in group k.
m_{ia}	Number of cases actually in group i.
m_{ic}	Number of cases classified to group i.
N,n	Number of variables.
P(A)	Probability of A.
P(A|B)	Probability of A given B.
SE(x)	Standard error of x.
S_k	Sample covariance matrix for group k.
s_{ij}^k	Element ij of group k's covariance matrix.
U_i	Unbiased discriminant function.

VAR(x)	Variance of x.
w(x)	Two group linear discriminant.
w_p	Population two group linear discriminant.
w_s	Sample two group linear discriminant.
$\mathbf{x}$	Vector of measurements.
$\bar{\mathbf{x}}$	Sample mean.
x_i	Variable i.
$\bar{x}_i$	Sample mean of variable i.
x_{ij}	jth measurement on variable i.
δ^2	Population Mahalanobis distance squared.
ϵ	Population error rate.
ϵ_{ij}	Population error rate for group i classified as group j.
λ_i	Eigenvalue i.
$\boldsymbol{\mu}_k$	Population mean vector in group k.
μ_i	Population mean of variable i.
σ_{ij}	Population covariance between i and j.
$\boldsymbol{\Sigma}$	Population covariance matrix.
$\sum_i$	Sum over i.
χ^2	Chi squared statistic.

Contents

Chapter One
Classification

The need to decide which of a number of possible groups an object belongs to is a surprisingly general problem. For the most part, unless the classification is obvious and trivial we still depend on human expertise to classify on the basis of observations. For example a doctor diagnoses diseases using years of medical training and practice. In a similar way a botanist identifies plant species, a psychologist recognises personality types and so on. As the computer has become more and more accessible so it has become attractive to try and use it either to replace the experts or at the very least to guide and help them.

Even before computer use became common statisticians and others developed fairly simple methods of objective classification based on standard probability theory. However, as the classification problem has proved so important in so many different fields of application it has suffered from being re-solved very many times. Each time a discipline has re-invented the subject of classification it has introduced its own jargon, its own notation and its own favourite methods. For example classification is known as pattern recognition, discriminate analysis, decision theory, assignment analysis etc. Perhaps the most recent and most important re-use of classification analysis is in the area of 'expert systems', programs which seek directly to replace expert reasoning using AI (artifical intelligence) techniques. This proliferation of nomenclature and techniques has caused a great deal of unnecessary confusion in what is both a simple and important subject. The aim of this book is to explain the general theory and specific practice of classification analysis without exclusive reference to any particular field of application. However, to make the subject seem less theoretical, BASIC programs are provided wherever possible so that the reader can explore the methods discussed.

The range of classification

Although classification has a very wide application it is often difficult for a

specialist to see that the task in hand is indeed a classification problem. The trouble lies in the application of the idea of assigning an object to a group that it originates from. At its most simple the problem presents itself in terms of objects and groups. For example, if you need to find a way of naming a particular variety of plant then it is obvious that the problem is to find and use a set of measurements that will allow you to assign any specimen to the correct species, i.e. group of plants. In other situations the idea of a group and assignment may be difficult to see. For example, a doctor diagnosing an illness may think of his problem as one of deciding which disease the patient has, but a moment's thought will show that diagnosis is equivalent to assigning the patient to one of a number of possible groups of diseases based on observation of his or her symptoms. In the same way any problem that involves making a decision can be recast as one of assignment to a number of groups – each group representing a possible decision. As an extreme case consider the problem of delivering a judgement of guilt or innocence. This can be considered either as a decision based on the evidence available or as an assignment of the accused into a group of guilty people or a group of innocent people. In other words much that is found in the classical statistical subject of decision theory can be thought of as classification analysis [1].

In the same way predicting which of a number of outcomes will be realised can also be cast as a classification problem. For example the question 'will it rain today?' can either be thought of as prediction or it can be thought of as assigning 'today' to one of the the two possible groups 'rainy' or 'dry'. As prediction of a continuous variable such as temperature is usually associated with regression analysis [2] it will come as no surprise to learn that there are connections between classification and regression. However, it is not profitable to pursue classification as if it were nothing more than a branch of regression analysis because this obscures too much of the simple basis of the subject. It is much better to follow a logical development motivated by the need to classify objects efficiently than to try to find obscure links between classification and other traditional statistical methods.

Discriminant analysis

Many readers will be wondering why the title of this book is not 'Discriminant Analysis' or something similar. If you already know something of the subject of classification then no matter what your specialism you will almost certainly have come across the term 'discriminant analysis'. However, although the term is certainly the one most used to describe any sort of activity connected with the task of classifying unknown objects into groups, its exact meaning depends very much on who is using it!

There are so many different forms of 'discriminant analysis' and so many different usages of the term 'discriminant functions' that it is advisable to steer clear of the term as much as possible. Later chapters will point out which methods are known as discriminant analysis and by whom but it has to be admitted that excessive usage of the term leads to much confusion in practice!

What classification is not!

One of the potential hazards of learning any new approach to a problem is the tendency to see everything in terms of it. This sort of 'blanket application' of a method quickly brings it into disrepute and the inevitable backlash of underuse follows! To avoid this general effect and to avoid the specific misuse of classification analysis it is worth giving instances of its applicability and its non-applicability.

As already stated a number of times, classification analysis addresses itself to the problem of assigning an object to one of a number of possible groups on the basis of observations made on the object. There are ways in which this brief can be elaborated but this is the essence of the problem. Notice that there is nothing said about how we know the existence of the groups to which the object is to be assigned. In classification analysis the existence and structure of the groups themselves is of secondary importance it is the assignment of new cases that concerns us.

There are two main statistical methods that classification analysis is often confused with – 'cluster analysis' and 'analysis of variance'. If you have a mass of currently undifferentiated data and you are curious as to whether or not it has any natural group structure then the method that you should employ is cluster analysis [3]. Cluster analysis attempts to identify any possible tendency for data to 'clump' together to form groups. It is most definitely not concerned with the problem of classifying new objects into existing groups. And to emphasise this point, classification analysis is not concerned with identifying any possible groupings that might be contained within a mass of data.

The second possibility is that you suspect that a particular grouping exists within some data and would like statistical proof of this conjecture. This is the situation most often confused with classification analysis. The reasoning used is that if you can successfully classify new cases to the hypothesised groups with a reasonable accuracy then this is evidence that the groups are more than a figment of the imagination. While this reasoning is true to a certain extent there are more efficient and accurate statistical techniques for testing hypotheses about the existence of differences between groups. These

techniques are usually grouped under the heading of analysis of variance or ANOVA and are very well known to most users of statistics [4]. The ideas of ANOVA are in fact often used in classification analysis to determine whether or not the groups involved are sufficiently different to make accurate assignment a possibility and in this connection they are covered in Chapter 7 and 8. However, it is very important not to think of using the efficiency of classification as proof that groups are indeed different in some way.

As an example of the appropriateness of these three different methods of analysis consider the following three situations with respect to the same data. Suppose a psychologist administers a test to a community of patients, then:

(1) If the aim is to discover if there are a number of different groups of illness then cluster analysis is appropriate. The data would be submitted to a cluster analysis program and the output would be a suggested grouping of the patients according to the test results.

(2) If the patients are already 'labelled' in some way, i.e. some of them have one form of schizophrenia and the rest have another form, then the objective might be to test the hypothesis that the two groups perform differently on the psychological test. In this case the appropriate analysis would be ANOVA or, more likely, Multivariate ANOVA (MANOVA). Submitting the data, including an indication of which type of schizophrenia each patient had, to a MANOVA program would produce a test statistic and a significance level that would indicate the likelihood that the observed differences in performance of the two groups came about by chance alone. From this information the psychologist could conclude either that there was a real difference between the two groups on test, or there was not.

(3) Finally, if the objective is, later on, to assign a new patient to one of the two schizophrenia groups, then the appropriate method is classification analysis. In this case the data would be submitted to a classification program, once again along with an indication of the group to which each patient belonged, and this would be used to construct, and possibly test, the performance of a 'classification rule'. This rule would be the main objective of the exercise and if it was effective it would be used subsequently to assign new patients to the correct group of schizophrenics – that is it would be used for diagnosis.

In practice this neat division of methods into cluster analysis to investigate natural groupings, ANOVA to prove that groups are different and classification to assign new objects to the groups is less than clear cut. In particular, classification analysis is often accompanied by a need to discover

that characteristics make the groups different. For example if the two groups of schizophrenics do perform differently on the tests the psychologist may want to know the nature of this difference. Finding characteristics that distinguish groups is usually considered to be a branch of classification analysis known as 'feature selection' (see Chapters 7 and 8). However it uses too many of the ideas of ANOVA to be considered as a completely isolated subject. So, as with most statistical techniques, the reason for using classification analysis may be a little more complicated than suggested by the three examples described above but it is important to realise the overall objective of any method and use that which is most appropriate.

Prerequisites

This book is essentially about the application of the theories of probability and statistics to a practical problem. However the emphasis does not fall on theory. It is important to understand how the methods work but this does not entail understanding the formal proofs of a method's correctness or its derivation. Most of the methods can be understood by imagining what is happening in terms of geometry. In subsequent chapters you will find that wherever possible the mathematics is introduced either as a final summary, an additional clarification or as a practice algorithm. In other words mathematics is not relied on to convey an idea without some support! The use of mathematics as the sole explanation of an idea is often adequate but unless it is interpreted and illustrated it is rarely complete. However, to get the most value from this book it would be helpful to know a little probability theory [5], a little statistics [6] and a little matrix algebra [7].

As well as explanations of methods you will also find a number of BASIC programs in the following chapters. Some of these allow you to analyse sample data sets so that you can gain experience with the methods of classification before you approach any real data! Others are straightforward implementations of the methods described and serve both as programs that can be used on real data and to clarify any points of procedure that are difficult to describe in the text. All of the programs are written in a standard form of Microsoft BASIC and they should be easy to convert to any other dialect. (To be exact they are written in Microsoft V5 and will run on almost any CP/M machine.) A particular problem with writing programs that are intended to run on any machine is 'file handling'. Most dialects of BASIC have their own way of opening, reading, writing and closing files but it would be far too restricted to expect data to be input from the keyboard each time for the sake of compatibility! The solution adopted here is to use separate subroutines for each file operation and expect that the reader will

supply appropriate modifications to make them work on any given machine. As well as the modifications necessary to make the programs work there is also a great deal of scope for adding graphics and for generally improving the output of all of the programs but of course this is an optional extra!

References

1. Ferguson, T.S. (1967) *Mathematical Statistics - A Decision Theoretic Approach*. Academic Press.
2. Draper, N.R. and Smith, H. (1966) *Applied Regression Analysis*. Wiley.
3. Everitt, B.S. (1974) *Cluster Analysis*. Heinemann.
4. Sheffe, H. (1959) *The Analysis of Variance*. Wiley.
5. Meyer, P.L. (1965) *Introductory Probability and Statistical Applications*. Addison–Wesley, Massachusetts.
6. Wetherill, G.B. (1967) *Elementary Statistical Methods*. Methuen.
7. Searle, S.R. (1966) *Matrix Algebra for the Biological Sciences*. Wiley.