

Preface to the Second Edition

We were pleasantly surprised by the success of the first edition of this book. Many of our colleagues have started to use it for teaching purposes, and feedback from industrial researchers has also shown that it is useful for practitioners. So, when Springer-Verlag approached us and asked us to revise the material for a second edition, we gladly took the opportunity to rearrange some of the existing material and to invite new authors to write two new chapters. These additional chapters cover material that has attracted considerable attention since the first edition of the book appeared. They deal with kernel methods and support vector machines on the one hand, and visualization on the other. Kernel methods represent a relatively new technology, but one which is showing great promise. Visualization methods have been around in some form or other ever since data analysis began, but are currently experiencing a renaissance in response to the increase, in numbers and size, of large data sets. In addition the chapter on rule induction has been replaced with a new version, covering this topic in much more detail.

As research continues, and new tools and methods for data analysis continue to be developed, so it becomes ever more difficult to cover all of the important techniques. Indeed, we are probably further from this goal than we were with the original edition – too many new fields have emerged over the past three years. However, we believe that this revision still provides a solid basis for anyone interested in the analysis of real data.

We are very grateful to the authors of the new chapters for working with us to an extremely tight schedule. We also would like to thank the authors of the existing chapters for spending so much time carefully revising and updating their chapters. And, again, all this would not have been possible without the help of many people, including Olfa Nasraoui, Ashley Morris, and Jim Farrand.

Once again, we owe especial thanks to Alfred Hofmann and Ingeborg Mayer of Springer-Verlag, for their continued support for this book and their patience with various delays during the preparation of this second edition.

November 2002

South San Francisco, CA, USA
London, UK

Michael R. Berthold
David J. Hand

Preface to the First Edition

The obvious question, when confronted with a book with the title of this one, is why “intelligent” data analysis? The answer is that modern data analysis uses tools developed by a wide variety of intellectual communities and that “intelligent data analysis”, or IDA, has been adopted as an overall term. It should be taken to imply the intelligent application of data analytic tools, and also the application of “intelligent” data analytic tools, computer programs which probe more deeply into structure than first generation methods. These aspects reflect the distinct influences of statistics and machine learning on the subject matter.

The importance of intelligent data analysis arises from the fact that the modern world is a data-driven world. We are surrounded by data, numerical and otherwise, which must be analysed and processed to convert it into *information* which informs, instructs, answers, or otherwise aids understanding and decision making. The quantity of such data is huge and growing, the number of sources is effectively unlimited, and the range of areas covered is vast: industrial, commercial, financial, and scientific activities are all generating such data.

The origin of this book was a wish to have a single introductory source to which we could direct students, rather than having to direct them to multiple sources. However, it soon became apparent that wider interest existed, and that potential readers other than our students would appreciate a compilation of some of the most important tools of intelligent data analysis. Such readers include people from a wide variety of backgrounds and positions who find themselves confronted by the need to make sense of data.

Given the wide range of topics we hoped to cover, we rapidly abandoned the idea of writing the entire volume ourselves, and instead decided to invite appropriate experts to contribute separate chapters. We did, however, make considerable efforts to ensure that these chapters complemented and built on each other, so that a rounded picture resulted. We are especially grateful to the authors for their patience in putting up with repeated requests for revision so as to make the chapters meld better.

In a volume such as this there are many people whose names do not explicitly appear as contributors, but without whom the work would be of substantially reduced quality. These people include Jay Diamond, Matt Easley, Sibylle Frank, Steven Greenberg, Thomas Hofmann, Joy Hollenback, Joe Iwanski, Carlo Marchesi, Roger Mitton, Vanessa Robins, Nancy Shaw, and Camille Sinanan for their painstaking proofreading and other help, as well as Stefan Wrobel, Chris Road-

VIII Preface to the First Edition

knight and Dominic Palmer-Brown for stimulating discussions and contributions which, though not appearing in print, have led to critical reassessment of how we thought some of the material should be presented.

Finally, we owe especial thanks to Alfred Hofmann from Springer-Verlag, for his enthusiasm and support for this book right from the start.

February 1999

Berkeley, California
London, United Kingdom

Michael Berthold
David J. Hand

Chapter 1

Introduction

David J. Hand
Imperial College, United Kingdom

1.1. Why “Intelligent Data Analysis”?

It must be obvious to everyone - to everyone who is reading this book, at least - that progress in computer technology is radically altering human life. Some of the changes are subtle and concealed. The microprocessors that control traffic lights or dishwashers, are examples. But others are overt and striking. The very word processor on which I am creating this chapter could not have been imagined 50 years ago; speech recognition devices, such as are now available for attachment to PCs, could have been imagined, but no-one would have had any idea of how to build such a thing.

This book is about one of those overt and striking changes: the way in which computer technology is enabling us to answer questions which would have defied an answer, perhaps even have defied a formulation, only a few decades ago. In particular, this book is about a technology which rides on top of the progress in electronic and computer hardware: the technology of data analysis.

It is fair to say that modern data analysis is a very different kind of animal from anything which existed prior to about 1950. Indeed, it is no exaggeration to say that modern data is a very different kind of animal from anything which existed before. We will discuss in some detail exactly what is meant by data in the modern world in Section 1.3 but, to get the ball rolling, it seems more convenient to begin, in this section, by briefly examining the notion of “intelligent data analysis”. Why analyse data? Why is this book concerned with “intelligent” data analysis? What is the alternative to “intelligent” data analysis? And so on. In between these two sections, in Section 1.2, we will look at the cause of all this change: the computer and its impact.

To get started, we will assume in this opening section that “data” simply comprise a collection of numerical values recording the magnitudes of various

attributes of the objects under study. Then “data analysis” describes the processing of those data. Of course, one does not set out simply to analyse data. One always has some objective in mind: one wants to answer certain questions. These questions might be high level general questions, perhaps exploratory: for example, are there any interesting structures in the data? Are any records anomalous? Can we summarise the data in a convenient way? Or the questions might be more specifically confirmatory: Is this group different from that one? Does this attribute change over time? Can we predict the value of this attribute from the measured values of these? And so on.

Orthogonal to the exploratory/confirmatory distinction, we can also distinguish between descriptive and inferential analyses. A descriptive (or summarising) analysis is aimed at making a statement about the data set to hand. This might consist of observations on the entirety of a population (all employees of a corporation, all species of beetle which live in some locality), with the aim being to answer questions about that population: what is the proportion of females? How many of the beetle species have never been observed elsewhere? In contrast, an inferential analysis is aimed at trying to draw conclusions which have more general validity. What can we say about the likely proportion of females next year? Is the number of beetle species in this locality declining? Often inferential studies are based on samples from some population, and the aim is to try to make some general statement about the broader population, most (or some) of which has not been observed. Often it is not possible to observe all of the population (indeed, this may not always be well-defined - the population of London changes minute by minute).

The sorts of tools required for exploratory and confirmatory analyses differ, just as they do for descriptive and inferential analyses. Of course, there is considerable overlap - we are, at base, analysing data. Often, moreover, a tool which appears common is used in different ways. Take something as basic as the mean of a sample as an illustration. As a description of the sample, this is fixed and accurate and is the value - assuming no errors in the computation, of course. On the other hand, as a value derived in an inferential process, it is an estimate of the parameter of some distribution. The fact that it is based on a sample - that it is an estimate - means that it is not really what we are interested in. In some sense we expect it to be incorrect, to be subject to change (if we had taken a different sample, for example, we would expect it to be different), and to have distributional properties in its own right. The single number which has emerged from the computational process of calculating the mean will be used in different ways according to whether one is interested in description or inference. The fact that the mean of sample A is larger than the mean of sample B is an observed fact - and if someone asks which sample has the larger mean we reply “A”. This may be different from what we would reply to the question “Which population has the larger mean, that from which A was drawn or that from which B was drawn?” This is an inferential question, and the variability in the data (as measured by, for example, the standard deviations of the samples) may mean we

have no confidence at all that the mean of one population is larger than that of the other.

Given the above, a possible definition of data analysis is the process of computing various summaries and derived values from the given collection of data. The word “process” is important here. There is, in some quarters, an apparent belief that data analysis simply consists of picking and applying a tool to match the presenting problem. This is a misconception, based on an artificial idealisation of the world. Indeed, the misconception has even been dignified with a name: it is called the *cookbook fallacy*, based on the mistaken idea that one simply picks an appropriate recipe from one’s collection. (And not the idea that one cooks one’s data!) There are several reasons why this is incorrect. One is that data analysis is not simply a collection of isolated tools, each completely different from the other, simply lying around waiting to be matched to the problem. Rather the tools of data analysis have complex interrelationships: analysis of variance is a linear model, as is regression analysis; linear models are a special case of generalised linear models (which generalise from straightforward linearity), and also of the general linear model (a multivariate extension); logistic regression is a generalised linear model and is also a simple form of neural network; generalised additive models generalise in a different way; nonparametric methods relax some of the assumptions of classical parametric tests, but in doing so alter the hypotheses being tested in subtle ways; and so one can go on.

A second reason that the *cookbook fallacy* is incorrect lies in its notion of matching a problem to a technique. Only very rarely is a research question stated sufficiently precisely that a single and simple application of one method will suffice. In fact, what happens in practice is that data analysis is an iterative process. One studies the data, examines it using some analytic technique, decides to look at it another way, perhaps modifying it in the process by transformation or partitioning, and then goes back to the beginning and applies another data analytic tool. This can go round and round many times. Each technique is being used to probe a slightly different aspect of the data - to ask a slightly different question of the data. Several authors have attempted to formalise this process (for example [236, 408, 427]). Often the process throws up aspects of the data that have not been considered before, so that other analytic chains are started. What is essentially being described here is a *voyage of discovery* - and it is this sense of discovery and investigation which makes modern data analysis so exciting. It was this which led a geologist colleague of mine to comment that he envied us data analysts. He and other similar experts had to spend the time and tedium collecting the data, but the data analysts were necessarily in at the kill, when the exciting structures were laid bare. Note the contrast between this notion, that modern data analysis is the most exciting of disciplines, and the lay view of statistics - that it is a dry and tedious subject suited only to those who couldn’t stand the excitement of accountancy as a profession. The explanation for the mismatch lies in the fact that the lay view is a historical view. A view based on the perception that data analysts spend their time scratching away at columns of figures (with a quill pen, no doubt!). This fails to take into account the changes

we referred to at the start of this chapter: the impact of the computer in removing the drudgery and tedium. The quill pen has been replaced by a computer. The days of mindless calculation replaced by a command to the machine - which then effortlessly, accurately, and probably effectively instantaneously carries out the calculation. We shall return to this in Section 1.2.

One reason that the word “intelligent” appears in the title of this book is also implicit in the previous paragraph: the repeated application of methods, as one attempts to tease out the structure, to understand what is going on, and to refine the questions that the researchers are seeking to answer, requires painstaking care and, above all, intelligence. “Intelligent” data analysis is not a haphazard application of statistical and machine learning tools, not a random walk through the space of analytic techniques, but a carefully planned and considered process of deciding what will be most useful and revealing.

1.2. How the Computer Is Changing Things/the Merger of Disciplines

Intelligent data analysis has its origins in various disciplines. If I were to single out two as the most important, I would choose statistics and machine learning. Of these, of course, statistics is the older - machines which can have a hope of learning have not been around for that long. But the mere fact of the youth of machine learning does not mean that it does not have its own culture, its own interests, emphases, aims, and objectives which are not always in line with those of statistics. This fact, that these two disciplines at the heart of intelligent data analysis have differences, has led to a creative tension, which has benefited the development of data analytic tools.

Statistics has its roots in mathematics. Indeed many statisticians still regard it as fundamentally a branch of mathematics. In my view this has been detrimental to the development of the discipline (see, for example [243], and other papers in that issue). Of course, it is true that statistics is a mathematical subject - just as physics, engineering, and computer science are mathematical. But this does not make it a branch of mathematics any more than it makes these other subjects branches of mathematics. At least partly because of this perception, statistics (and statisticians) have been slow to follow up promising new developments. That is, there has been an emphasis on mathematical rigour, a (perfectly reasonable) desire to establish that something is sensible on theoretical grounds before testing it in practice.

In contrast, the machine learning community has its origins very much in computer practice (and not really even in computer science, in general). This has led to a practical orientation, a willingness to test something out to see how well it performs, without waiting for a formal proof of effectiveness.

It goes without saying that both strategies can be very effective. Indeed, ideally one would apply both strategies - establish by experiment that something does seem to work and demonstrate by mathematics when and under what circumstances it does so. Thus, in principle at least, there is a great potential

for synergy between the two areas. Although, in general, I think this potential has yet to be realised, one area where it has been realised is in artificial neural network technology.

If the place given to mathematics is one of the major differences between statistics and machine learning, another is in the relative emphasis they give to models and to algorithms.

Modern statistics is almost entirely driven by the notion of a model. This is a postulated structure, or an approximation to a structure, which could have led to the data. Many different types of models exist. Indeed, recalling the comment above that the tools of data analysis have complex interrelationships, and are not simply a collection of isolated techniques, it will come as no surprise if I say that many different families of models exist. There are a few exceptional examples within statistics of schools or philosophies which are not model driven, but they are notable for their scarcity and general isolation. (The Dutch Gifi school of statistics [211], which seeks data transformations to optimise some external criterion, is an example of an algorithm driven statistical school. By applying an algorithm to a variety of data sets, an understanding emerges of the sort of behaviour one can expect when new data sets are explored.) These exceptions have more in common with approaches to data analysis developed in the machine learning community than in traditional statistics (which here is intended to include Bayesian statistics). In place of the statistical emphasis on models, machine learning tends to emphasise algorithms. This is hardly surprising - the very word “learning” contains the notion of process, an implicit algorithm.

The term “model” is very widely used - and, as often occurs when words are widely used, admits a variety of subtly different interpretations. This is perhaps rather unfortunate in data analysis since different types of models are used in different ways and different tools are used for constructing different kinds of models. Several authors have noted the distinction between empirical and mechanistic models (see, for example [81, 129, 239]). The former seek to model relationships without basing them on any underlying theory, while the latter are constructed on the basis of some supposed mechanism underlying the data generating process. Thus, for example, we could build a regression model to relate one variable to several potential explanatory variables, and perhaps obtain a very accurate predictive model, without having any claim or belief that the model in any way represented the causal mechanism; or we might believe that our model described an “underlying reality”, in which increasing one variable led to an increase in another.

We can also distinguish models designed for prediction from models designed to help understanding. Sometimes a model which is known to be a poor representation of the underlying processes (and therefore useless for “understanding”) may do better in predicting future values than one which is a good representation. For example, the so-called Independence Bayes model, a model for assigning objects to classes based on the usually false assumption of independence between the variables on which the prediction is to be based, often performs well. This sort of behaviour is now understood, but has caused confusion in the past.

Finally here (although doubtless other distinctions can be made) we can distinguish between models and patterns (and as with the distinctions above, there is, inevitably, overlap). This is an important distinction in data mining, where tools for both kinds of structure are often needed (see [244]). A model is a “large scale” structure, perhaps summarising relationships over many cases, whereas a pattern is a local structure, satisfied by a few cases or in a small region of the data space in some sense. A Box-Jenkins analysis of a time series will yield a model, but a local waveform, occurring only occasionally and irregularly, would be a pattern. Both are clearly of potential importance: we would like to detect seasonality, trend, and correlation structures in data, as well as the occasional anomaly which indicates that something peculiar has happened or is about to happen. (It is also worth noting here that the word “pattern”, as used in the phrase “pattern recognition”, has a rather different meaning. There it refers to the vector of measurements characterising a particular object - a “point” in the language of multivariate statistics.)

I commented above that the modern computer-aided model fitting process is essentially effortless. This means that a huge model space can be searched in order to find a well-fitting model. This is not without its disadvantages. The larger the set of possible models examined in order to fit a given set of data, the better a fit one is likely to obtain. This is fine if we are seeking simply to summarise the available data, but not so fine if the objective is inference. In this case we are really aiming to generalise beyond the data, essentially to other data which could have arisen by the same process (although this generalisation may be via parameters of distributions which are postulated to (approximately) underlie the data generating mechanism). When we are seeking to generalise, the data on which the model must be based will have arisen as a combination of the underlying process and the chance events which led to that particular data set being generated and chosen (sampling variability, measurement error, time variation, and so on). If the chosen model fits the data too well, then it will not merely be fitting the underlying process but will also be fitting the chance factors. Since future data will have different values for the chance factors, this will mean that our inference about future values will be poor. This phenomenon - in which the model goes too far in fitting the data - is termed overfitting. Various strategies have been developed in attempts to overcome it. Some are formal - a probability model for the chance factors is developed - while others are more ad hoc - for example, penalising the measure of how well the model fits the data, so that more complex models are penalised more heavily, or shrinking a well-fitting model. Examples are given in later chapters.

As model fitting techniques have become more refined (and quick to carry out, even on large data sets), and as massive amounts of data have accumulated, so other issues have come to light which generally, in the past, did not cause problems. In particular, subtle aspects of the underlying process can now often be detected, aspects which are so small as to be irrelevant in practice, even though they highly statistically significant are and almost certainly real. The decision as to how complex a model to choose must be based on the size of the

effects one regards as important, and not merely on the fact that a feature of the data is very unlikely to be a chance variation.

Moreover, if the data analysis is being undertaken for a colleague, or in collaboration, there are other practical bounds on how sophisticated a model should be used. The model must be comprehensible and interpretable in the terms of the discipline from which it arose.

If a significant chunk of intelligent data analysis is concerned with finding a model for data or the structures which led to the data, then another significant chunk is concerned with algorithms. These are the computational facilitators which enable us to analyse data at all. Although some basic forms of algorithm have been around since before the dawn of the computer age, others have only been developed - could only be imagined - since computers became sufficiently powerful. Computer intensive methods such as resampling, and Bayesian methods based on avoiding integration by generating random samples from arbitrary distributions, have revolutionised modern data analysis. However, to every fundamental idea, every idea which opens up a wealth of new possibilities, there are published descriptions of a hundred (Why be conservative? A thousand.) algorithms which lead to little significant progress. In fact, I would claim that there are too many algorithms being developed without any critical assessment, without any theoretical base, and without any comparison with existing methods. Often they are developed in the abstract, without any real problem in mind. I recall a suggestion being made as far back as twenty years ago, only partly tongue-in-cheek, that a moratorium should be declared on the development of new cluster analysis algorithms until a better understanding had been found for those (many) that had already been developed. This has not happened. Since then work has continued at an even more breakneck pace.

The adverse implication of this is that work is being duplicated, that effort and resources are being wasted, and that sub-optimal methods are being used. I would like to make an appeal for more critical assessment, more evaluation, more meta-analysis and synthesis of the different algorithms and methods, and more effort to place the methods in an overall context by means of their properties. I have appealed elsewhere (for example [240]) for more teaching of higher level courses in data analysis, with the emphasis being placed on the concepts and properties of the methods, rather than on the mechanical details of how to apply them. Obtaining a deeper understanding of the methods, how they behave, and why they behave the way they do, is another side of this same coin. Some institutions now give courses on data analytic consultancy work - aspects of the job other than the mechanics of how to carry out a regression, etc. - but there is still a long way to go.

Of course, because of the lack of this synthesis, what happens in practice at present is that, despite the huge wealth of methods available, the standard methods - those that are readily available in widespread data analytic packages - are the ones that get used. The others are simply ignored, even if a critical assessment might have established that they had some valuable properties, or that they were “best” in some sense under some circumstances.

Although in this section we have focused on the relationship between data analysis and the two disciplines of statistics and machine learning, we should note in passing that those two disciplines also cover other areas. This is one reason why they are not merely subdisciplines of “intelligent data analysis”. Statistics, for example, subsumes efficient design methodologies, such as experimental design and survey design, described briefly in the next section, and machine learning also covers syntactic approaches to learning. Likewise, we should also note that there are other influences on modern data analysis. The impact of the huge data sets which are being collected is one example. Modern electronics facilitates automatic data acquisition (e.g. in supermarket point of sale systems, in electronic measurement systems, in satellite photography, and so on) and some vast databases have been compiled. The new discipline of data mining has developed especially to extract valuable information from such huge data sets (see [244] for detailed discussion of such databases and ways to cope with them).

As data sets have grown in size and complexity, so there has been an inevitable shift away from direct hands-on data analysis towards indirect data analysis in which the analyst works via more complex and sophisticated tools. In a sense this is automatic data analysis. An early illustration is the use of variable selection techniques in regression. Given a clearly defined criterion (sum of squared errors, for example), one can let the computer conduct a much larger search than could have been conducted by hand. The program has become a key part of the analysis and has moved the analyst’s capabilities into realms which would be impossible unaided. Modern intelligent data analysis relies heavily on such distanced analysis. By virtue of the power it provides, it extends the analyst’s capabilities substantially (by orders of magnitude, one might say). The perspective that the analyst instructs a program to go and do the work is essentially a machine learning perspective.

1.3. The Nature of Data

This book is primarily concerned with numerical data, but other kinds exist. Examples include text data and image data. In text data the basic symbols are words rather than numbers, and they can be combined in more ways than can numbers. Two of the major challenges with text data are search and matching. These have become especially important with the advent of the World Wide Web. Note that the objects of textual data analysis are the blocks of text themselves, but the objects of numerical data analysis are really the things which have given rise to the numbers. The numbers are the result of a mapping, by means of measuring instruments, from the world being studied (be it physical, psychological, or whatever), to a convenient representation. The numerical representation is convenient because we can manipulate the numbers easily and relatively effortlessly. Directly manipulating the world which is the objective of the study is generally less convenient. (For example, to discover which of two groups of men is heavier, we could have them all stand on the pans of a giant weighing scales and see which way the scales tipped. Or we could simply add up

their weights and compare the two resulting numbers.) The gradual development of a quantitative view of the world which took place around the fourteenth and fifteenth centuries (one can argue about timescales) is what underpinned the scientific revolution and what led ultimately to our current view of the world. The development of the computer, and what it implies about data analysis means that this process is continuing.

The chapters in this book present rather idealised views on data analysis. All data, perhaps especially modern data sets which are often large and many of which relate to human beings, has the potential for being messy. A priori one should expect to find (or rather, not to find) missing values, distortions, misrecording, inadequate sampling and so on. Raw data which do not appear to show any of these problems should immediately arouse suspicion. A very real possibility is that the presented data have been cleaned up before the analyst sees them. This has all sorts of implications. Here are some illustrations.

Data may be missing for a huge variety of reasons. In particular, however, data may be missing for reasons connected with the values they would have had, had they been recorded. For example, in pain research, it would be entirely reasonable to suppose that those patients who would have had the most severe pain are precisely those who have taken an analgesic and dropped out of the study. Imagine the mistakes which would result from studying only the apparent pain scores. Clearly, to cope with this, a larger data analysis is required. Somehow one must model not only the scores which have been presented, but also the mechanism by which the missing ones went missing.

Data may be misrecorded. I can recall one incident in which the most significant digit was missed from a column of numbers because it went over the edge of the printing space. (Fortunately, the results in this case were so counter-intuitive that they prompted a detailed search for an explanation.)

Data may not be from the population they are supposed to be from. In clinical trials, for example, the patients are typically not a random sample from some well-defined population, but are typically those who happened to attend a given clinic and who also satisfied a complex of inclusion/exclusion criteria. Outliers are also a classic example here, requiring careful thought about whether they should be dropped from the analysis as anomalous, or included as genuine, if unusual, examples from the population under study.

Given that all data are contaminated, special problems can arise if one is seeking small structures in large data sets. In such cases, the distortions due to contamination may be just as large, and just as statistically significant, as the effects being sought.

In view of all this, it is very important to examine the data thoroughly before undertaking any formal analysis. Traditionally, data analysts have been taught to “familiarise themselves with their data” before beginning to model it or test it against algorithms. However, with the large size of modern data sets this is less feasible (or even entirely impossible in many cases). Here we must rely on computer programs to check the data for us. There is scope here for much research: anomalous data, or data with hidden peculiarities, can only be

shown to be such if we can tell the computer what to search for. Peculiarities which we have not imagined will slip through the net and could have all sorts of implications for the value of the conclusions one draws.

In many, perhaps most, cases a “large” data set is one which has many cases or records. Sometimes, however, the word “large” can refer to the number of variables describing each record. In bank or supermarket records, for example, details of each transaction may be retained, while in high resolution nuclear magnetic resonance spectroscopy there may be several hundred thousand variables. When there are more variables than cases, problems can arise: covariance matrices become singular, so that inversion is impossible. Even if things are not so extreme, strong correlations between variables can induce instability in parameter estimates. And even if the data are well-behaved, large numbers of variables mean that the curse of dimensionality really begins to bite. (This refers to the exponentially increasing size of sample required to obtain accurate estimates of probabilities as the number of variables increases. It manifests itself in such counterintuitive effects as the fact that most of the data in a high dimensional hypercube with uniformly distributed data will lie in a thin shell around the edge, and that the nearest sample point to any given point in a high dimensional space will be far from the given point on average.)

At the end of the previous section we commented about the role of design in collecting data. Adequate design can make the difference between a productive and an unproductive analysis. Adequate design can permit intelligent data analysis - whereas data which has been collected with little thought for how it might be analysed may not succumb to even the most intelligent of analyses. This, of course, poses problems for those concerned with secondary data analysis - the analysis of data which have been collected for some purpose other than that being addressed by the current analysis.

In scientific circles the word “experiment” describes an investigation in which (some of) the potentially influential variables can be controlled. So, for example, we might have an experiment in which we control the diet subjects receive, the distance vehicles travel, the temperature at which a reaction occurs, or the proportion of people over 40 who receive each treatment. We can then study how other variables differ between different values of those which have been controlled. The hope is that by such means one can unequivocally attribute changes to those variables which have been controlled - that one can identify causal relationships between variables.

Of course, there are many subtleties. Typically it is not possible to control all the potentially influential variables. To overcome this, subjects (or objects) are randomly assigned to the classes defined by those variables which one wishes to control. Note that this is a rather subtle point. The random assignment does not guarantee that the different groups are balanced in terms of the uncontrolled variables - it is entirely possible that a higher proportion of men (or whatever) will fall in one group than another, simply by random fluctuations. What the random assignment does do, however, is permit us to argue about the average

outcomes over the class of similar experiments, also carried out by such random allocations.

Apart from its use in eliminating bias and other distortions, so that the question of interest is really being answered, experimental design also enables one to choose an efficient data collection strategy - to find the most accurate answer to the question for given resources or the least resources required to achieve a specified accuracy. To illustrate: an obvious way to control for six factors is to use each of them at several levels (say three, for the purposes of this illustration) - but this produces 729 groups of subjects. The numbers soon mount up. Often, however, one can decide a priori that certain high order effects are unlikely to occur - perhaps the way that the effect of treatment changes according age and sex is unlikely to be affected by weight, for example (even though weight itself influences the treatment effect, and so on). In such cases it is possible to collect information on only a subset of the 729 (or however many) groups and still answer the questions of interest.

Experiments are fundamentally manipulative - by definition they require that one can control values of variables or choose objects which have particular values of variables. In contrast, surveys are fundamentally observational. We study an existing population to try to work out what is related to what. To find out how a particular population (not necessarily of people, though surveys are used very often to study groups of people) behaves one could measure every individual within it. Alternatively, one could take measurements merely on a sample. The Law of Large Numbers of statistics tells us that if we repeatedly draw samples of a given size from a population, then for larger samples the variance of the mean of the samples is less. So if we only draw a small sample from our population we might obtain only an inaccurate estimate of the mean value in which we are interested. But we can obtain an accurate estimate, to any degree of accuracy we like, by choosing a large enough sample. Interestingly enough, it is essentially the size of the sample which is of relevance here. A sample of 1000 from a population of 100,000 will have the same accuracy as one from a population of a million (if the two populations have the same distribution shape).

The way in which the sample is drawn is fundamental to survey work. If one wanted to draw conclusions about population of New York one would not merely interview people who worked in delicatessens. There is a classic example of a survey going wildly wrong in predicting people's Presidential voting intentions in the US: the survey was carried out by phone, and failed to take account of the fact that the less well-off sections of the population did not have phones. Such problems are avoided by drawing up a "sampling frame" - a list of the entire population of interest - and ensuring that the sample is randomly selected from this frame.

The idea of a simple random sample, implicit in the preceding paragraph, underlies survey work. However, more sophisticated sampling schemes have been developed - again with the objective of achieving maximum efficiency, as with experimental design. For example, in stratified sampling, the sampling frame is divided into strata according to the value of a (known) variable which is

thought to be well correlated with the target variable. Separate estimates are then obtained within each stratum, and these are combined to yield an overall estimate.

The subdisciplines of experimental and survey design have developed over the years and are now very sophisticated, with recent developments involving high level mathematics. They provide good illustrations of the effort which is necessary to ensure good and accurate data so that effective answers can be obtained in data analysis. Without such tools, if the data are distorted or have been obtained by an unknown process, then no matter how powerful one's computers and data analysis tools, one will obtain results of dubious value. The familiar computer adage "garbage in, garbage out" is particularly relevant.

1.4. Modern Data Analytic Tools

The chapters in this book illustrate the range of tools available to the modern data analyst. The opening chapters adopt a mainly statistical perspective, illustrating the modelling orientation.

Chapter 2 describes basic statistical concepts, covering such things as what 'probability means', the notions of sampling and estimates based on samples, elements of inference, as well as more recently developed tools of intelligent data analysis such as cross-validation and bootstrapping.

Chapter 3 describes some of the more important statistical model structures. Most intelligent data analysis involves issues of how variables are related, and this chapter describes such multivariate models, illustrating some of them with simple examples. The discussion includes the wide range of generalised linear models, which are a key tool in the data analyst's armoury.

Up until recently, almost all statistical practice was carried out in the 'frequentist' tradition. This is based on an objective interpretation of probability, regarding it as a real property of events. Chapter 3 assumes this approach. Recently, however, thanks to advances in computer power, an alternative approach, based on a subjective interpretation of probability as a degree of belief, has become feasible. This is the Bayesian approach. Chapter 4 provides an introductory overview of Bayesian methods.

A classical approach to supervised classification methods was to combine and transform the raw measured variables to produce 'features', defining a new data space in which the classes were linearly separable. This basic principle has been developed very substantially in the notion of support vector machines, which use some clever mathematics to permit the use of an effectively infinite number of features. Early experience suggests that methods based on these ideas produce highly effective classification algorithms. The ideas are described in Chapter 5.

Time series occupy a special place in data analysis because they are so ubiquitous. As a result of their importance, a wide variety of methods has been developed. Chapter 6 describes some of these approaches.

I remarked above, that statistics and machine learning, the two legs on which modern intelligent data analysis stands, have differences in emphasis. One of

these differences is the importance given to the interpretability of a model. For example, in both domains, recursive partitioning or tree methods have been developed. These are essentially predictive models which seek to predict the value of a response variable from one or more explanatory variables. They do this by partitioning the space of explanatory variables into discrete regions, such that a unique predicted value of the response variable is associated with each region. While there is overlap, the statistical development has been more concerned with predictive accuracy, and the machine learning development with interpretability. Tree models are closely related to (indeed, some might say are a form of) methods for rule induction, which is the subject of Chapter 7. A rule is a substructure of a model which recognises a specific pattern in the database and takes some action. From this perspective, such tools for data analysis are very much machine learning tools.

It is unlikely that anyone reading this book will not have heard the phrase “neural network”. In the context of this book an artificial neural network is a structure of simple processors with parameterised interconnections. By choosing the values of the parameters appropriately, one can use the network as a very flexible function estimator. This makes such networks powerful tools - essentially providing a very good fit to a data set and then being shrunk to avoid overfitting problems. Their flexibility means that they are less subject to the bias problems intrinsic to methods which assume a particular model form to start with (e.g. the linear form of classical regression). Artificial neural networks are important because of their power as models, but they may turn out to be just as important because of the impetus they are giving to enhanced understanding of inference and the nature of induction. Chapter 8 discusses such tools for intelligent data analysis.

Probability, and the theories of inferential statistics built on it, are the most widely accepted and used tool for handling uncertainty. However, uncertainty comes in many shapes and forms. There is, for example, stochastic uncertainty arising from the basic mechanism leading to the data, but there is also uncertainty about the values of measurements or the meaning of terms. While many - especially Bayesians - feel that this second kind can also be handled by probabilistic arguments, not everyone agrees, and other approaches have been developed. One such is the school of fuzzy reasoning and fuzzy logic. Essentially, this replaces the (acknowledged false) notion that classes are precisely known by a membership function, which allows an object to belong to more than one class, but with differing degrees. The details are given in Chapter 9, which also describes fuzzy numbers and how they may be manipulated.

One of the most exciting developments which has resulted from the growth of computer power has been the probabilistic solution of previously intractable methods by means of stochastic search and optimisation methods, such as simulated annealing and genetic algorithms. These sort of strategies are the subject of Chapter 10.

Methods for graphical display of data are as old as data itself. However, modern computational facilities have extended the scope considerably: with such

facilities, dynamic and interactive graphics are possible. Coupled with the increasingly tough demands of modern data analysis, arising from such things as huge data sets and time-dependent data sets, the field of graphical displays –or data visualization, as it is now called - has blossomed. Such developments are described in Chapter 11.

Chapters 12 and Appendix A round off the book. Chapter 12 presents some examples of real applications of the ideas in the book, ranging from relatively standard statistical applications to novel machine learning applications such as the “No hands across America” experiment. Appendix A lists and describes some of the many tools available for intelligent data analysis which now abound. The variety and range of origins of these tools indicates the interdisciplinary nature of intelligent data analysis.

1.5. Conclusion

This book is about intelligent data analysis. But if data analysis can be intelligent, then it can also be unintelligent. Unfortunately, as with good health, departures from intelligent data analysis can be in many directions. Distorted data, incorrect choice of questions, misapplication of data analytic tools, overfitting, too idealised a model, a model which goes beyond the various sources of uncertainty and ambiguity in the data, and so on, all represent possibilities for unintelligent data analysis. Because of this, it is less easy to characterise unintelligent data analysis than it is to characterise intelligent data analysis. Often only in retrospect can we see that an analysis was not such a good idea after all. This is one of the reasons why the domain of intelligent data analysis is so interesting. It is very much not a case of simply applying a directory of tools to a given problem, but rather one of critical assessment, exploration, testing, and evaluation. It is a domain which requires intelligence and care, as well as the application of knowledge and expertise about the data. It is a challenging and demanding discipline. Moreover, it is a fundamentally interdisciplinary, taking ideas from several fields of endeavour. And it is a discipline which is continuing to evolve. People sometimes speak of “new technology” as if it were a change which would happen and then finish - like the transition to decimal coinage from the old pounds, shillings, and pence in the UK in the early 1970s. But so-called “new technology” is not like that. It is really a state of permanent change. And riding on the back of it is the development of new methods of data analysis. Of course, if the problems of data analysis remain the same, then the impact of new technology will be limited - we will simply be able to do things faster and bigger. But as technology advances so the possibilities change - the frontiers of what can be achieved, what can be imagined, move back. The current state of data analysis illustrates this. Neural networks, stochastic search methods, practical Bayesian tools, all illustrate possibilities which were inconceivable not so many years ago. Moreover, new application areas present new challenges, posing new problems and requiring new solutions. Examples include financial applications and biometrics (in the sense of person identification through retina and voice

prints). Beyond this, the meaning of data is shifting: we have commented above about the growing importance of non-numeric data such as text and image data. But data about data - metadata - is also attracting growing interest (in the face of such problems as how to merge or fuse data sets which define their basic units in different ways, for example). One thing is clear, data analysis is in the most exciting period of its history. And the evidence is that the possibilities and excitement will continue to grow for many years to come.



<http://www.springer.com/978-3-540-43060-5>

Intelligent Data Analysis

An Introduction

Berthold, M.R.; Hand, D. (Eds.)

2003, XI, 515 p., Hardcover

ISBN: 978-3-540-43060-5