Artificial Intelligence and Statistics

Edited by

William A. Gale

Artificial Intelligence and Statistics

Edited by

William A. Gale
AT&T Bell Laboratories

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts • Menlo Park, California
Don Mills, Ontario • Wokingham, England • Amsterdam
Sydney • Singapore • Tokyo • Madrid
Bogotá • Santiago • San Juan

Library of Congress Cataloging-in-Publication Data
Main entry under title:

Artificial intelligence and statistics.

Papers prepared for the Workshop on Artificial Intelligence and Statistics, held in April 1985, in Princeton, N.J., and sponsored by AT&T Bell Laboratories.
Includes index.
1. Artificial intelligence--Congresses.
2. Statistics--Congresses. 3. Expert systems (Computer science)--Congresses. I. Gale, William A.
II. Workshop on Artificial Intelligence and Statistics (1985 : Princeton, N.J.) III. AT&T Bell Laboratories.
Q334.A764 1986 006.3 85-28631
ISBN 0-201-11569-7

Reproduced by Addison-Wesley from camera-ready copy supplied and approved by the editor.

ABCDEFGHIJ-DO-89876

Preface

This book will be of interest to statisticians or AI researchers who want to apply AI techniques in statistics. It will also be of interest to AI researchers or statisticians studying propagation of uncertainty or learning (clustering, concept formation).

Most of the papers in this book were prepared for the Workshop on Artificial Intelligence and Statistics, held in April 1985. Most of the talks presented at the Workshop are represented here by papers.

My intention in arranging the Workshop was to raise the visibility of the possible applications of AI in statistics and of statistics in AI. The goal was to bring together people interested in the various interactions of AI and statistics, giving a broader view of the total work at this point than would be possible for any one group. AT&T Bell Laboratories, which has been a leader in developing applications of AI in statistics, agreed to sponsor the Workshop.

The Workshop was a catastrophe of success, attracting four times as many requests to attend as I had planned space for, and not all could be accommodated at the Workshop. It therefore seemed that it would be worth preparing a book based on the conference to give those who could not be at the Workshop access to the papers presented there.

This book substantially increases the published literature on AI in statistics and contains some important contributions from statistics to AI. It can give you a working knowledge of current capabilities in the application of AI to statistics or some suggestions of useful research in applying statistics to AI. All the papers contain new results, and some present in-depth reviews of specific areas.

The book was typeset in Times Roman font by AT&T Bell Laboratories TROFF computer program. I thank Susan Tarczynski for a fine job of typing and correcting the manuscript.

Murray Hill, N.J. *WAG*

Contributors

Annie G. Brooking
Knowledge Based Systems
Manor House Systems
58 Clapham Corner, North Side
London SW4 - 9RZ
UNITED KINGDOM

Keith A. Butler
Advanced Technology Applications Division
Boeing Computer Services
P.O. Box 24346
Seattle, WA 98124

James E. Corter
Teachers College
Columbia University
New York, NY 10027

Thomas Ellman
Computer Science Department
450 Computer Science Building
Columbia University
New York, NY 10027

Douglas Fisher
Department of Information and Computer Science
University of California
Irvine, CA 92717

John Fox
Imperial Cancer Research Fund Laboratories
P.O. Box No. 123
Lincoln's Inn Fields
London, WC2A 3PX
UNITED KINGDOM

William A. Gale
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

David J. Hand
Institute of Psychiatry
De Crespigny Park
Denmark Hall
London SE5 8AF
UNITED KINGDOM

Stephen C. Hora
Information Systems and Quantitative Sciences
Texas Tech University
Box 4320
Lubbock, TX 79409

Peter J. Huber
Department of Statistics
Harvard University
Cambridge, MA 02138

Pat Langley
Department of Information and Computer Science
University of California
Irvine, CA 92717

P. B. Musgrove
Polytechnic of the South Bank
London
UNITED KINGDOM

R. Wayne Oldford
Massachusetts Institute of Technology
1 Amherst Street
Cambridge, MA 02138

Stephen C. Peters
Massachusetts Institute of Technology
1 Amherst Street
Cambridge, MA 02138

Robert I. Phelps
Department of Mathematics Statistics
Brunel University
Uxbridge Middlesex UB8 3PH
UNITED KINGDOM

Daryl Pregibon
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

Steven Salzberg
Applied Expert Systems, Inc.
Five Cambridge Center
Cambridge, MA 02142

David J. Spiegelhalter
MRC Biostatistics Unit
Medical Research Council Centre
Cambridge, CB2 2QH
UNITED KINGDOM

Ronald A. Thisted
The University of Chicago
Department of Statistics
5734 University Avenue
Chicago, IL 60637

John Tukey
Princeton University
Fine Hall
Washington Road
Princeton, NJ 08540

Contents

1

Overview of Artificial Intelligence and Statistics

William A. Gale

1.1 OPPORTUNITIES

Exciting applications of Artificial Intelligence (AI) technology are opening a new field of research in statistics. This new field seeks to integrate previously known tests and transformations into coherent total approaches to data analysis. This is AI in statistics. At the same time, AI researchers are turning to work by statisticians in seeking solutions to problems involving uncertainty. Representing the uncertainty of conclusions established from uncertain facts by uncertain rules is one such problem. This is statistics in AI. The chapters in this book represent the first stage of the development of AI in statistics, as well as a continuation of statistics in AI, and are together a report on AI *and* statistics.

The key use of AI in statistics has been to enable statisticians to study *strategies* of data analysis. The term strategy has been used to denote a coherent total approach to a data analytic task, following Chambers (1981). However, there is, as yet no generally accepted definition of this term. Daryl Pregibon and I (Gale and Pregibon, 1982) suggested that a definition would be the answer to questions such as

> "What do I look for?"
> "When do I look for it?"
> "How do I look for it?"
> "Why do I look for it?"
> "What do I have to do to look for it?"

In Chapter 15 of this book, Wayne Oldford and Steve Peters write "The term 'statistical strategy' will be used here to label the reasoning used by the experienced statistician in the course of the analysis of some aspect of a substantive statistical problem." David Hand, in Chapter 16, writes "*statistical strategy* has been defined as a formal description of the choices, actions, and decisions to be made while using statistical methods in the course of a study." These definitions give the general flavor of the subject matter beginning to be addressed.

We need to make statistical strategy available to more people to prevent misuse of statistical packages. Statisticians have frequently noted that current statistical packages are open to misuse, as will be clear from the background discussion that follows. The current packages provide excellent numerical processing, but the user is responsible for determining whether the processing is appropriate, and what the results mean. Statistical strategy is simply missing from the packages and is necessary to use the packages wisely. Mechanization of statistical strategy would make it widely available.

The tools developed by AI may allow us to mechanize statistical strategy. In the last decade AI researchers have demonstrated techniques for mechanizing symbolic information processing tasks that otherwise would have been assumed to be the preserve of human intelligence. These tasks include understanding the spoken or written word, interpreting still or moving images, and solving intellectual problems from playing chess to synthesis of complex molecules. In particular, the widespread success of "expert systems" in addressing problems in other fields has suggested that they might prove useful in statistics. (Fox (1985) is a readable introduction to expert systems.) AI techniques have been used in statistics for provocative demonstrations that software can now be written to do such things as translate a medical hypothesis into a statistical study, help a user select an appropriate statistical technique, and check automatically for the validity of the assumptions behind a statistical technique.

Now that several systems have demonstrated the utility of AI techniques in statistics, leading statisticians have begun to point out the opportunities for research. John Tukey (1985) recently pointed out to a meeting of statisticians "we have been unwilling to think hard about what our strategies (not just individualized tactics) really have been or should be," and "one just cannot build an expert system without thinking through a strategy." A major advantage of mechanization for statisticians is that it will support systematic testing and analysis of strategies. David Cox, one of the few statisticians to think about strategies before AI methods made them testable (Cox and Snell, 1981), spoke at the same meeting citing the need for "provision of concepts to guide the strategy of the collection and analysis of data" (Cox, 1985).

Although AI has contributed to many areas, statistics is one of relatively few disciplines that can contribute back to AI. Statistics has focused on numerical methods for estimating limits to knowledge when measurements have a random component, whereas AI has been concerned with symbolic representations of knowledge and their use. As the AI representations become deeper and represent their roots, they will always be found to rest on measurement, which always has a random component, so there is a clear need for statistics in AI.

The need for statistics in AI has so far shown itself mainly in the two areas of uncertainty and learning. In the first area, reasoning in AI systems must take into account the uncertainty of empirical relationships. The developers of the well-known Mycin system (Buchanan and Shortliffe, 1984) found uncertainty to be an essential property in representing knowledge for their domain of medical diagnosis. Their treatment of uncertainty has highlighted its importance, perhaps unduly so. The discussion of probabilistic methods in AI literature has revolved around the assumptions necessary to use Bayesian methods. Even those defending probabilistic methods in the AI literature, such as Charniak (1983), have usually been discussing the limitations of what Spiegelhalter (Chapter 2) calls "idiot Bayes." There has also been some interest among AI researchers in Shafer's (1976) theory of evidence. As Spiegelhalter points out, more sophisticated methods are available.

Statistics can also contribute to the study of learning and concept formation (or, in statistical terminology, clustering). AI is particularly concerned with finding clustering methods that can be applied to data that have categorical labels in addition to or instead of numerical attributes, whereas statisticians have long dealt with clustering on numerical attributes. The obvious area of common concern would be that with both numerical and categorical attributes. But as Fisher and Langley (Chapter 4) point out, there are strong correspondences between the purely symbolic techniques and the purely numerical techniques. Also, the AI work has not paid sufficient attention to the problem of error in the input symbolic data.

When I first thought about the workshop that led to this book, I was thinking about a workshop on "AI in Statistics." But it soon became apparent that this was a mistake, because what the statisticians need to learn is what the AI researchers know well, and vice versa. To get experts in both areas at one meeting, there must be something for both to learn, so maximum value comes from having both sides present. I believe the same holds for this book. Those particularly interested in learning from some of these chapters may well find that they could easily contribute to a discussion raised by another chapter. I hope that such will happen frequently, and that what is written here is soon outmoded.

1.2 BACKGROUND

This section presents a review of background literature on AI in statistics. I have included a thorough review of this literature because papers on AI in statistics are neither very numerous nor very accessible. I have not, however, included a similar review of statistics in AI because that literature is too extensive to be summarized readily and because it is more readily available to readers. Certain chapters of this book do include fairly extensive reviews of statistics in AI for certain key areas. David Spiegelhalter's chapter, Chapter 2, provides a review of some of the most relevant papers in propagation of uncertainty, whereas in Chapter 4 Doug Fisher and Pat Langley include a considerable review of papers in concept formation.

1.2.1 What Has Been Done

The earliest paper I have seen that explicitly called for more intelligence in statistical software is a paper by John Nelder (1977), "Intelligent Programs, the Next Stage in Statistical Computing." He motivated the discussion by pointing out that algorithmic procedures in common use in statistics packages would accept any data set of the right shape (the right lengths of vectors, or sizes of matrices). The packages did not give a warning if something was wrong, and therefore the program relied on the user to know what to check and to make the checks. He expressed the opinion, which I believe is widely shared, that "the amount of uncritical use of standard procedures is enormous." He then turned to a discussion of what might be done, using regression as an example. Nelder suggested three sources of distortion that a regression analysis should "ideally" be protected against: (1) faults in the data, (2) faults in the model - systematic part, and (3) faults in the model - random part. But, in 1977, the techniques Nelder could suggest for carrying out his ideas were not extensive and probably lacked sufficient power to accomplish the job.

John Chambers (1981) was the first to argue that expert system techniques from AI would be useful for achieving more intelligent software. He echoed Nelder in decrying reliance on "blind computational algorithms," and also argued that declining computation costs, making computers more widely available for statistics, would "precipitate much uninformed, unguided, and simply incorrect data analysis." He went on to argue that "passive solutions," such as reporting diagnostics or multiple answers, would not go far enough. He therefore suggested that expert system techniques be used to devise software playing an active role in proposing answers and actions based on the data set. He introduced the term *strategy* and mentioned regression diagnostics and robust estimation as currently available statistical tools that would be useful in a strategy.

Chambers, Pregibon, and Zayas presented a paper at Buenos Aires in 1981 reporting the result of an experiment undertaken by the three authors at AT&T Bell Laboratories in the summer of 1980. They chose regression analysis as an initial target, citing as reasons that it was a relatively small but not insignificant area, and well studied. Their work was the first to use a symbolic reasoning system to direct statistical software. They identified the broad structures needed by a consultation program as diagnostics, action rules, explanations, and dialogue. They attempted to implement an expert system with production rules using OPS4. However, as far as I could tell from an examination of the program when I started work on REX, they used very few production rules and quite a bit of Lisp programming. A program to conduct user interaction was well developed. The paper reports discussions of statistical strategy in analysis of collinearity, outliers, re-expression, and non-normality of errors. This program attempted more than REX, with fewer resources, so that it could only be described as an "initial experiment."

At the same meeting, Campbell and Woodings (1981) called for software which would "mimic the steps and checks that experienced statisticians carry out automatically." They cited inexperienced users as a source of statistical errors, pointing out that current statistical packages did little checking. They included a detailed discussion of some checks and transformations, key ingredients to a strategy, for multivariate discrimination. Their suggestions for implementation included ideas on descriptors to be kept with the data and records of actions taken using the data. Although they cited AI as a promising source of help, they did not relate the statistical needs to demonstrated AI systems.

Hájek and Ivánek (1982) described another early system. A system using 34 rules to decided among six choices of statistical test was built and run on a dozen problems. This system was not connected to a statistical package, but merely asked the user all the information it wanted, including such items determinable from the data as the number of data points. The paper did not argue the necessity or desirability of consultation systems in statistics, but suggested two types of systems based on analogies with expert systems which had been described in other domains. The two suggestions were systems to advise on the use of specific statistical packages and systems modeling the exploratory data analysis process.

Bob Blum's thesis (1982) on the RX system was slow to be recognized as an important contribution for AI in statistics. Because it was considered part of medical literature, it was overlooked by statisticians. By specializing to medicine as a ground domain, Blum was able to produce a feasibility demonstration system for translating research goals into a specific data analysis agenda. RX included hierarchical representations of medical concepts and statistical methods, together with a causal network among the medical concepts. The data available to the system were longitudinal medical records on rheumatic patients. The patients were seriously ill, with multiple diseases and therapies, which made establishing a valid study design challenging. RX was intended to (1) propose a new causal relation worth studying; (2) design a study by controlling confounding relations, finding data, and chosing an analytic method; and (3) to carry out the analysis. The first and third of these goals were not accomplished, but the second alone is a valuable example of an AI application in statistics.

Daryl Pregibon and I began work on REX late in 1981. REX gave advice on regression analysis, searching for problems in the data and proposing actions to remediate any problems found. Reading John Nelder's paper after building REX was very interesting because the three main stages of REX's analysis — checking each variable separately, establishing linearity of the model, and reviewing the residuals for problems — correspond precisely to Nelder's three sources of distortion. REX was the first system to successfully use expert system techniques to choose commands for a statistical package and to carry out complete analyses. We reported our work in Gale and Pregibon (1982) and Pregibon and Gale (1984). Chapter 8 of this book gives more detail on REX than is available in those earlier papers.

Three points stand out from John Tukey's (1982) talk at the 14th Interface Symposium. First, he emphasized the need for statistical explanations. In statistics, explanations are rarely certain. Multiple possible reasons to more or less explain what is going on will be harder to generate but more honest than recounting a single line of reasoning. Second, he made clear the importance of search in data analysis, saying, "A competent data analysis of an even moderately complex set of data is a thing of trials and retreats, of dead ends and branches." This discussion of search in statistics led me to devise the trace in Student (Chapter 10), and the work by Oldford and Peters (Chapter 15) and Huber (Chapter 12) reflects the same concern. Third, the talk proposed "cognostics," diagnostics designed for interpretation by a computer rather than by a human. We have gotten along with diagnostics off the shelf for interactive systems, but when we want to say "Do those 5000 cases *like* I am doing this one, and tell me of any *interesting* differences," then we will need more than is now available.

In 1983, Portier and Lai described a system, STATPATH, which is the only system yet to tackle the problem of a user not understanding a question or misunderstanding a question. STATPATH uses production rules to encode a binary choice tree to help the user select an analytic technique. The decisions in the choice tree are made by asking the user carefully phrased questions. If the user did not understand a question, several lines of help were available. First, the user could ask for a more fully worded question. Alternatively, the user could reply "unknown" to the question, instead of "yes" or "no," and get *both* further lines of questioning. Having followed more than one line of questioning to a recommended analytic technique, the user could then browse textual annotations to the techniques. This problem of user misunderstanding is an important one, and one for which AI techniques may be useful.