To appear in the International Journal of Neural Systems (IJNS)

Book Review

Empirical Methods for Artificial Intelligence

Paul R. Cohen, 1995
ISBN 0-262-03225-2

Reviewed by Ron Kohavi
Silicon Graphics, Inc.
2011 N. Shoreline Blvd,
Mountain View, CA 94043-1389
ronnyk@sgi.com

1 Overview

Empirical Methods in AI gives researchers and practitioners a good introduction to exploratory data analysis (EDA), hypothesis testing, computer-intensive methods, and statistical techniques based on linear regression. The book motivates readers to use statistical techniques, which are explained through a large number of detailed examples. The basic ideas and the formulas one should use are explained, but the technical details were left to appendices and to other books. The practitioner will find cookbook-style procedures that can be easily understood and utilized, but mathematically inclined readers will not find any proofs of correctness nor a concise formal specification of the assumptions required to justify correctness of the procedures.

Cohen (1991) surveyed the Eighth National Conference on Artificial Intelligence (AAAI-90) and concluded that the methodologies used are incomplete with respect to the goals of designing and analyzing AI systems. Tichy, Lukowicz, Prechelt & Heinz (1995) showed in a very large study of over 400 articles that research papers in Computer Science are rarely validated with experimental results. More recently, and perhaps more appropriate for readers of this journal, Prechelt (1996) showed that the situation is not better in the neural network literature. Out of 190 articles published in well-known journals dedicated to neural networks, 29% did not employ even a single realistic or real learning problem. Only 8% of the articles presented results for more than one problem using real world data.

While Prechelt (1996) only looked at whether comparisons were done, Cohen went a step further and described how to design good experiments. He wrote that "books like this one encourage well-designed experiments, which, if one isn't careful, can be utterly vacuous. This is a danger, not an inevitability. Knowing the danger, we can avoid it" (p. 103).

As the title of the book implies, examples from Artificial Intelligence are used throughout the book. The common urn and coin-flip examples from introductory Probability and Statistics textbooks were replaced by AI planners, expert systems, message understanding systems, and natural language problems. We see the *Knowledge-is-power* hypothesis, Allen Newell's ideas about the *knowledge level*, and a description of the SOAR architecture. While these examples give the feeling of a familiar territory to those with broad interests in AI, they might intimidate some readers who are not familiar with the systems and concepts used. Even within the AI domain, the choice of examples is very Cohen-centric: many examples are based on work done by the author or by his students. The approach of using many different and lesser known AI examples forced the author to write many pages of details that are not necessarily relevant to the main topic of the book, thus making the text less desirable at the undergraduate level.

My recommendation is to leisurely read this book for the good motivation and the very intuitive explanations it gives to the statistical tests and procedures. Cohen wanted to write a book that was easy and accessible, so that there could be no excuse for bad methods. The book is very easy to read and it can help researchers unfamiliar with Statistics to identify the appropriate tools for their specific needs. However, while Cohen's book is easy to read, many assumptions are spread throughout the text and subtle problems are never mentioned. Readers interested in deeper understanding of the statistical procedures and the underlying assumptions required for their validity will need to read more formal texts.

2 Contents and Comments

The book is organized into nine chapters. Chapters 1-3 deal with experimental design and exploratory data analysis; hypothesis testing is left to Chapter 4. Computer intensive methods are described in Chapter 5. Chapters 6 and 7 deal with assessing performance and explaining it using analysis of variance. Chapter 8 deals with modeling, linear regression, and causality. Chapter 9 concludes with tactics for generalizing. The author gave three reasons for writing this book:

1. Computer Science has no curriculum in experimental research methods as other sciences do.

2. Systems are increasingly embedded, complex, and sophisticated; we thus need powerful research methods.

3. It is time to revise some classical views of empirical AI. We can no longer predict how a system will behave by looking at its code.

Although these reasons are important, they could be addressed by adding Statistics courses to the curriculum of AI students and increasing awareness of the importance of statistical methods. Cohen prefers to tailor the methods to AI, which does provide some advantages. He provides a diskette with statistical software called CLASP (Common Lisp Analytical Statistical Package) for $20, and an instructor's manual with homework and viewfoils.

The first two chapters stress the importance of exploratory data analysis (EDA) and the difference between EDA and hypothesis testing. EDA "finds things in haystacks, whereas statistical hypothesis testing puts them under a microscope and tells us whether they are needles and whether they are sharp" (p. 8). While the explanations are nice and intuitive, sometimes they lack a mathematical definition. For example, on page 5 we see the definitions for basic statistical terms, such as mean, median, mode, and skew. Readers unfamiliar with "skew" will have an intuitive understanding of what it is, but they will not see the formal mathematical definition.

In the preface, Cohen wrote that "the book assumes nothing about the reader, the mathematical material is light and it is developed from first principles." In Chapter 4 (Hypothesis testing and estimation), however, the author uses the term *degrees of freedom* (p. 125) with a single comment that the "number [is] closely related to sample size."

The variance is described as "the sum of squared distances between each datum and the mean, divided by the sample size *minus one*" (p. 26). The assumptions under which this statement holds are omitted; interested reader must search for them in other books. This specific estimator for the variance is unbiased for a random sample only if the mean is unknown and estimated from the sample. If, for example, the mean is known, one should divide by the sample size (and not the sample size minus one) to get an unbiased estimation of the variance. In fact, even if this was an accurate description, users would benefit from mentioning the probabilistic definition of variance and then mentioning the problem of estimating it from data. Later on (p. 127), the variance of the difference of two means is given as $\sigma^2_{x_1-x_2} = \sigma^2_{x_1} + \sigma^2_{x_2}$ with the comment that "It is interesting to consider why the variance of the differences is the sum of the individual variances." Indeed, very interesting (and false) unless the samples are independent.

Cohen has strong opinions on many topics. For example, we learn that "Ceiling effects and floor effects are due to poorly chosen test problems" (p. 81). Does this imply that machine learning researchers should not tackle problems where induction algorithms can achieve prediction accuracy close to 100%? There may be real-world domains where each error is extremely costly. Being able to improve from 99% to 99.5% means reducing the error (and cost) by half. Even though there is a ceiling effect, this does not mean that the problem is "poorly chosen."

Cohen chose to demonstrate the ceiling effect through the 1R algorithm by Holte (1993). The 1R algorithm by Holte is a simple induction algorithm that builds a decision rule based on a *single* feature. The error rate of 1R on sixteen datasets from the UCI repository (Murphy & Aha 1995) was 19.82% while that of C4.5 was 14.07%. Holte wrote that the difference was "only" 5.7%, which he thought was surprising. (I personally think that the relative error rate, 40% worse than C4.5, is the more important measure and the difference is very significant.) Both text and the tables mistakenly show the comparison between C4.5 and 1R* instead of 1R. There is a big difference between 1R and 1R*: 1R* is *not* an induction algorithm but a lower bound on the error of 1R. Holte himself wrote that "1R* is a rather optimistic upper bound [on accuracy]." Cohen does not mention this crucial point.

In the description of Efron's Bootstrap, Cohen wrote: "Bootstrap sampling distributions can be calculated for any statistic, and they assume nothing about the population, except that the sample is representative of the population" (p. 153). The term *representative* is undefined here and hides a lot more than more readers might suspect. Try to compute the probability of getting a duplicate set of measurements in the sample and you will get an extremely high estimate. Try to use the Bootstrap to estimate the accuracy of a classifier and you'll realize that things are not so simple. That is the reason the .632 bootstrap was conceived (Efron 1983), while it is totally ignored in the book. The .632 estimator is not perfect either (as no estimator can be). Recent experiments comparing it to cross-validation (Kohavi 1995) generated yet another bootstrap estimator that attempts to correct some flaws in the .632 bootstrap (Efron & Tibshirani 1995).

When describing cross-validation, the author wrote that the induction algorithm must be run on each fold separately (the memory must be cleared in his terminology) "to get k independent estimates of the effect of training" (p. 217). This is simply not true. The estimates are highly correlated as $(1 - 1/k)$ of the data is shared between folds.

Overall, I found many of the high-level explanations very good and intuitive. I believe that the many numerical examples might be useful when one tries to implement some of the tests in a program, or when one runs some statisical tests based on existing software (possibly the software provided by Cohen). I have given some examples of problems related to the lack of formal definitions and clear statements of the assumptions. There is a tradeoff between making a book accessible to a wide audience and making it precise and formal; in some cases, I personally would have liked slightly more formalism. I hope the examples mentioned here will allow you to gauge the formal level of this book and thus aid in determining whether you should read it.

References

Cohen, P. R. (1991), "A survey of the eighth national conference on artificial intelligence: Pulling together or pulling apart?", *AI Magazine* **12**(1), 17–41.

Efron, B. (1983), "Estimating the error rate of a prediction rule: improvement on cross-validation", *Journal of the American Statistical Association* **78**(382), 316–330.

Efron, B. & Tibshirani, R. (1995), Cross-validation and the bootstrap: Estimating the error rate of a prediction rule, Technical Report 477, Stanford University, Statistics department.

Holte, R. C. (1993), "Very simple classification rules perform well on most commonly used datasets", *Machine Learning* **11**, 63–90.

Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* C. S. Mellish, ed., "Proceedings of the 14th International Joint Conference on Artificial Intelligence", Morgan Kaufmann Publishers, Inc., pp. 1137–1143.

Murphy, P. M. & Aha, D. W. (1995), UCI repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html.

Prechelt, L. (1996), "A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice", *Neural Networks* **9**, –.

Tichy, W. F., Lukowicz, P., Prechelt, L. & Heinz, E. A. (1995), "Experimental evaluation in computer science: A quantitative study", *Journals of Systems and Software* **28**(1), 9–18.