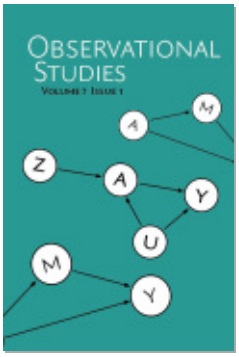
Comments on Breiman: Statistical Modelling: The Two Cultures and Commentaries



Peter Bickel

Observational Studies, Volume 7, Issue 1, 2021, pp. 17-20 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0018*

Observational Studies Vol. 7.1 (2021): Published July 2021

Comments on Breiman: Statistical modelling: the two cultures and commentaries

Peter Bickel **bickel@stat.berkeley.edu**
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA

Abstract

In a challenging paper 20 years ago, Leo Breiman challenged the statistical culture of his time. Some perceptive comments by David Cox and Brad Efron appeared with it. In this paper I try to look at this work in the light of modern culture and find much to agree but also much to disagree with. It's still a pleasure to read.

Keywords: Statistics, machine learning, philosophy

It's a pleasure to reread Leo's wonderful polemic and the commentaries on it. When written and, even more In retrospect, his critiques of our profession ring true as did the much earlier ones of J.W.Tukey in "The future of data analysis" (Tukey, 1962).

Differences, then and now

However, I disagree with Leo's simple dichotomy between data and algorithmic modeling. First, I agree with D.R.Cox in his commentary on this paper, that before data usually come some questions about the world as we sense it. This is clear from Leo's examples. Data in some forms bearing on these questions usually come next. We then integrate what we know, or think we know, with the data in some way to arrive at "answers" to our questions or rather "explanations" which help us understand what we have sensed better. How we do this integration and what we produce, as Cox implicitly makes clear, cannot be easily categorized. We, typically, abstract, using what we think we know and simplifying assumptions, to obtain possible approximations to the processes which we are interested in and which lead to the data. As statisticians we think of the data being generated by some probability mechanism. Think of this as a story. A model is a class of stories we believe the truth belongs to or is close enough to for our purposes. Any member story in the model is characterized by unknowns of different kinds we call parameters. A method is a way we devise of using the data to get closer to the truth, in a way we care about, by estimating some of these parameters. What we think of as getting closer to the truth is, I think, the domain of decision theory. We next devise a method to estimate these parameters. This method is implemented through an algorithm. There are many possible algorithms for any given method. The method can be related to the model structure as, for example, maximum likelihood or Bayes. Or the method and algorithm can come from

other considerations, as in classification using neural nets or random forests. For any given model there are many methods, and a method can be applicable to many models. I believe that what Leo calls an algorithmic model is in fact an algorithm implementing a method. Underlying the algorithms he discusses is I claim, a probability model in Leo's and most current examples in machine learning, the story is that what one is observing is an iid sample, observations generated in the same probabilistic way, from a high dimensional universe possessing an observable property (class,response,..). Nothing further is assumed. The goal is to predict the value of the property for a new observation when the property of interest is not observed. Success is measured, for example,by the probability of correct classification or closeness between the true and predicted responses. Leo makes a good case that the methods for nonparametric classification and regression he applies and the approach to variable selection he advocates in his examples do much better at prediction than methods such as logistic regression, classical at the time of his article. He correctly goes further to argue that "variable importance" is identified more accurately by his approach of using the drop in predictive power when a variable is omitted than with classical variable selection. By now as we know this paradigm embodied by machine learning is wildly successful.in prediction for activities ranging from distinguishing image types and hence face recognition, to ad placement, to disease diagnosis.

Not the cure to all the world's ills

Successful applications typically occur when there are both large numbers of observations in a training set and large numbers of potential predictors. As Leo pointed out then, the number of such situations has increased dramatically over the last 20 years. This is the situation he refers to as the "Blessing of dimensionality". But if the number of observations is small compared to the ratio of within class variability to between class variablity the blessing disappears. The algorithmic model can fail badly, as do others, when assumptions are grossly violated. Say if the observation to be classified comes from a distribution very different from that of the training sample objects. A neural net trained to distinguish between airplanes and fishes might fail badly when presented with a Zeppelin.

Rashomon and Occam

Under the heading "Rashomon" Leo points to the multiplicity of "good" generating probability mechanisms. If the sample size and the model are big enough this situation can arise naturally. In Leo's and machine learning's usual model for prediction and a method for estimating conditional expectations nonparametrically, implemented by a good algorithm, this is not a problem for prediction. Of course, there may also be many poor solutions, which can be present if the size of the model greatly exceeds that of the data. In practice,this is dealt with by algorithms which reflect sparsity assumptions,in essence a much smaller model. In any case,this poses problems for "Occam", Leo's second criterion of simplicity. It is certainly the Achilles heel of neural nets and other main machine learning algorithms Occam is the main pointer in Leo's paper to most of the other situations in which statistics is used.These are what Tukey referred to as the "exploratory" and "confirmatory" parts of statistics. That is, the back and forth processes by which we try to use what we have

learned to provide partial answers to our questions, relate what we have learned to our other knowledge, refine our model in the light of the data, assess our confidence in the validity of what we have learned and perhaps plan to gather more or other data. These are the aspects of statistics which algorithmic models do not relate to explicitly.

Algorithmic models and probability models

There is a sense in which data models and probability models can be reconciled in the context of prediction. Whatever prediction method we develop on a set of data we are interested in making predictions for the future. The assumption implicitly underlying most of machine learning, is that one is dealing with a sample of of objects from some unknown population P (probability distribution on some space) so that one is observing $X_1, ..., X_n$ distributed iid from P.A classification rule based on a particular sample should be well defined for any other sample and necessarily is a parameter, that is a function, of the empirical distribution. In all sorts of concrete situations ranging from applying least squares to neural nets any such parameter can be extended to a population parameter, that is a function of P. On the other hand, standard quantities such as Maximum likelihood estimates, in non convex situations are defined as the limits of algorithms, whose iterates may be thought of as population parameters specialized to the empirical distribution of the data.

Where are we 20 years later?

Technically,through the united efforts of computer scientists, probabilists and statisticians we have gained a much better and more rigorous understanding of the interaction of complexity of statistical models and sample size. What Leo has called algorithmic models have moved into other parts of probability and statistics. More broadly they are an integral part of what Jordan calls "intelligence augmentation" where "computation and data are used to augment human intelligence and creativity" (Jordan, 2019). Machine learning and statistics have to some extent merged under the rubric of data science. For instance, Markov decision processes have become reinforcement learning. Good prediction, when possible, has rightly become a necessary though not sufficient ingredient of statistical analysis. A rounded view of data analysis more generally can be found in Yu and Kumbier (2020). What is still not resolved is how, when algorithmic models are applied to data, their output leads to an increase in human understanding beyond prediction. Trying to decide causation, which has had a resurgence in modern statistical work, requires interpretability.which in turn relies on domain knowledge (as should the model specification when more than the minimum is available).

Interpretability

What is interpretability? The only rigorous definition in Wikipedia, due to Alfred Tarski, is framed in the language of mathematical logic and seems not useful for our ordinary understanding. In statistics I like to think of it as relating a few simple primitive notions whose interactions we know to the model to which the data has been fit. As such it seems akin to simplicity and governed by Occam's razor. As Leo pointed out in his response

to comments on his paper, current models of nature, such as quantum physics are not simple. In general,the primitives of interpretability are always relative to the actor. The interpretability of strategies in a chess game differs greatly from beginner to chess master. The conclusions that Einstein or Newton drew from known data were very different from Everyman's. Yet such conclusions are based, as is Conway's Game of Life, on a series of steps each of which is eventually separately interpretable and can be linked to each other in a clear way. That is not yet the case for the algorithmic models successful in large scale prediction. Whether they will be is akin to the question of whether artificial intelligence will supersede biological intelligence, an issue discussed in Jordan (2019) and many other places.

References

Michael I Jordan. Artificial intelligence—the revolution hasn't happened yet. *Harvard Data Science Review*, 1(1), 2019.

John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1): 1–67, 1962.

Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.