Statistical Modeling: The Two Cultures by Leo Breiman

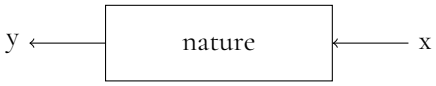
Leda Liang, Yash Nair, Zitong Yang

1 Background

Professor Leo Breiman was an incredibly distinguished statistician with a unique career. Breiman is most well known for his work on developing random forests [1], however he initially tried to pursue more theoretical work. After being convinced to study math over philosophy, Professor Breiman's first academic position was teaching probability at UCLA, which is as far as one can get from applied work within the field of statistics.

This paper on the two cultures of statistical modeling published during Professor Breiman's retirement was motivated by applied work and problems he encountered during his sabbatical [2]. In the time he spent working in industry, Professor Breiman noticed the increasing complexity and size of modern datasets and problems. This led to his frustrations with his statistical colleague favoring use of simpler data models with nice theoretical results over more practical algorithmic models with better performance.

2 Two Cultures



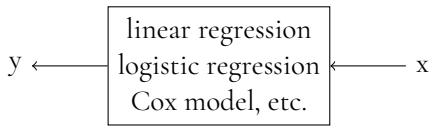
In statistics, there are two goals of learning from data:

1. Prediction - Learning how to predict for future inputs
2. Information - Learning about the nature of the underlying process

Staticians can be divided into two approaches:

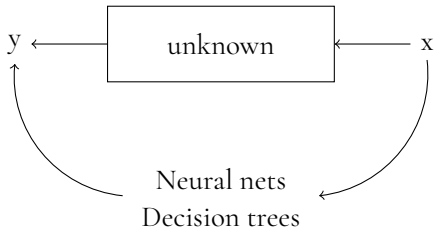
1. Data modeling
2. Algorithmic modeling

2.1 Culture 1: Data Modeling



In the early 2000's when this paper was written, Professor Breiman claims that 98% of statisticians fall under the data modeling approach. In data modeling, it is assumed that the data is generated by some model plus some noise and the goal is to estimate parameters of the model. The model assumptions are validated by a yes or no decision based on goodness of fit tests and residual examination.

2.2 Culture 2: Algorithmic Modeling



In algorithmic modeling, it is assumed that nature is complex and data does not come from a model that can be simply described. In this approach, the scientists goal is to use whichever algorithm appears to best predict the output and evaluation is based on measuring predictive accuracy.

2.3 Shifting Toward Algorithmic Modeling

In modern problems, the size of data collected keeps growing. The data modeling approach is more motivated by an academic setting. In reality, the data modeling assumptions become restrictive. In comparison, the flexibility of algorithmic modeling makes it more useful for consulting and practical applications.

2.3.1 Example 1: The Ozone Project

The goal of the ozone project was to predict ozone levels to warn the public of days with hazardous air quality conditions in Los Angeles. Ideally this would encourage people to reduce driving and time spent outside on such days.

The dataset consists of measurements from dozens of weather stations across the west coast and different layers of the atmosphere. Each weather station reports hourly readings of hundreds of variables including temperature, humidity, and wind speed. When researchers tried using the data modeling approach by using large linear regression models, too many false alarms were made.

Example 2: The Chlorine Project

The goal of the chlorine project was to predict the toxicity of a chemical compound from its mass spectra. The chemical structure of chemical compounds can tell scientists about its toxicity. Although mass spectra can be cheaply obtained, it requires expensive labor to analyze manually.

The dataset consists of 30,000 compounds with known chemical structure and mass spectra where the mass spectra consists of frequencies of each molecular weight. Linear regression, one of the most popular data models, failed in this application because the dimension was too large. However, decision tress had 95% accuracy.

3 Cons of Data Modeling

Breiman, of course, takes issue with the data modeling approach. He summarizes his primary complaint as "The conclusions are about the model's mechanism, and not about nature's mechanism." In other words, inferences made about a model's parameters, say, answer questions only relating to those parameters and will not, in general, answer questions about the way nature works (since the model may not accurately describe nature's mechanism).

3.1 A Case Study: the Linear Model

This philosophy is illustrated by a consideration of the linear model:

$$y = b_0 + \sum_{m=1}^{M} b_m x_m + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

While there is plenty of elegant theory that can be derived from this model (such as a variety of tests, confidence intervals, and prediction intervals), the issue is that the model often may not accurately describe nature.

Typically, this issue is addressed by performing a goodness of fit test. Often, this is done by computing the R^2 multiple correlation coefficient, and deciding that the model is accurate so long as the value is sufficiently close to one. Breiman illustrates this through an example of an experiment designed to test gender discrimination in salary among statistics faculty. A t-test was performed and significance detected at the 5% level, however, the analysis raises many questions. Concerns were raised about whether or not the data available is sufficient to answer the statistical question posed and whether or not the model accurately describes nature. Perhaps more covariates are needed so as to (better) satisfy the linear model assumption. Perhaps there is heteroskedasticity in the data and the simple (homoskedastic) linear model is not an accurate description of the data-generating process. In either case, the conclusions of the study are at best questionable.

3.2 Goodness of Fit Testing and Residual Analysis

While many papers do check their modeling assumptions by using goodness of fit tests, Breiman argues that even this is not enough to justify the use of the model.

The main qualm that Breiman has with goodness of fit testing is that most tests lack power. In particular, most tests are "omnibus" (i.e., they are designed to at least have some power against a wide variety of alternatives); these tests, however, lack power against particular alternatives. The consequence of this is that these goodness of fit tests will often fail to reject, leading the research to think that his or her model accurately describes the data when it really does not. Perhaps it is more fruitful to view Breiman's qualm here with the relative seriousness of errors of Type I and II. In particular, it seems that in his view, a Type II error (declaring that the model is correct, when it actually is not) is much more serious than a Type I error (declaring that the model is incorrect, when it actually is) since in the latter case, at least the researcher's confidence in his or her model/results will not be falsely bolstered. Maybe a test which controls the Type II error rate would at least, in part, address Breiman's qualms with the current standard of goodness of fit testing.

Residual analysis is another common way of assessing goodness of fit. However, according to Breiman, William Cleveland (a father of residual analysis), admitted in a talk that it lacks power in situations when there are more than four of five covariates.

3.3 Multiplicity of Data Models: Rashomon

The final problem Breiman mentions regarding data modelling is that there is a multiplicity of data models that could equally well describe the data. In particular, consider two statisticians who fit two *different* models to the same data and each of these two models passes the goodness of fit tests that each statistician runs. Which statistician is correct? Whose model should we believe? Whose inferences should we trust? Goodness of fit tests, beyond their lack of power, are incapable of answering this questions: they provide only "yes–no" answers to whether a single model accurately describes the data; they do not give a principled comparison between the two different models themselves, saying which is better than the other. Breiman illustrates this issue by referencing the medical field, in which it is commonplace to use the Cox model, even though there may be other models that fit the data (according to a goodness of fit test, e.g.,) just as well.

3.4 Predictive Accuracy

As a solution (and preliminary introduction to the algorithmic modeling approach), Breiman discusses predictive accuracy as a means by which different models can be compared. In details, Breiman suggests first to compute the predictive accuracy of each model (e.g., $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ in regression problems where y_i is the ground truth and $\hat{y}_i$ is the prediction for the i^{th} unit) and then to select the model with the lower value. It is possible that there is significant noise in the true response y_i given the covariates x_i in which case the predictive accuracy of even "good" models will be small. However, from the viewpoint of comparison this is not a problem: all that matters it the *relative* accuracy of two different models, even if the predictive accuracy of the "better" model is small it will still (typically) be greater than that of a "worse" model.

An import observation is that if this measure of predictive accuracy is computed on the data on which the model was trained, then the estimate of predictive accuracy can be heavily biased. In such a case, Breiman suggests two commonly used approaches in the machine learning community for debiasing the estimate: 1. evaluating the predictive accuracy/loss on a held-out test set, or 2. computing the cross-validation estimate of accuracy/loss. Using these tools, one can more accurately assess the true predictive accuracy of various models, and hence compare them more readily.

4 Pros and Cons of Algorithmic Modeling

Algorithmic modeling emerged in the mid-1980s, marked by the development of powerful algorithms such as random forests and neural networks. This era saw the formation of a vibrant new research community comprising young computer scientists, physicists, engineers, and a few pioneering statisticians. Publications primarily appeared in venues like *Neural Information Processing* and *Journal of Machine Learning Research*, marking a shift in the discourse surrounding statistical modeling.

4.1 Theoretical Aspects of Algorithmic Modeling

Algorithmic modeling, especially supervised learning, is intrinsically a statistical phenomenon. A key insight is that low training error, combined with a dataset size exceeding the number of 'degrees of freedom', tends to yield low test error. This relationship is crucial for understanding the effectiveness of machine learning models.

The following equation encapsulates this idea:

$$\Pr_{S\sim D^{|S|}}\left[|\text{Test}_D(f) - \text{Train}_S(f)| \leq \sqrt{\frac{\log|F| + \log\frac{1}{\delta}}{|S|}} \text{ for all } f \in F\right] > 1-\delta$$

Discussion on the $p > n$ Scenario: There has been significant discourse around the failure of these principles in scenarios where feature dimensions exceed sample size (the $p > n$ scenario), particularly in deep learning. However, these challenges can be mitigated with appropriate techniques.

4.2 Three Lessons from Algorithmic Modeling

- Rashomon: The Rashomon effect emphasizes the existence of multiple good models leading to similar outcomes, highlighting the complexity and richness of modeling.
- Occam: Occam's Razor poses a conflict between simplicity and accuracy in model selection. Simplicity often comes at the cost of reduced accuracy.
- Bellman: Bellman's principle touches on the dimensionality issue in data modeling. While high dimensionality can enhance prediction, it also complicates the model.

4.3 Statistical Modeling: A Third Culture

The third culture in statistical modeling deviates from previous ones by emphasizing the functional capabilities of models, such as reasoning and problem-solving, rather than solely focusing on predictive accuracy. For instance, in language modeling, despite the abundance of data, what matters more is the model's ability to solve complex problems like the Riemann hypothesis.

In conclusion, the evolution of statistical modeling into these three distinct cultures highlights the dynamic nature of the field. As data and computational capabilities expand, our understanding and utilization of statistical models also evolve, leading us to rethink fundamental concepts and approaches.

5 Reflection

The key perceptions to takeaway from this paper is that statisticians should not let the restrictive models of the data modeling approach to prevent them from working exciting new problems. Statisticians should search for the model that gives the best solution – either algorithmic or data models.

Professor Cox's comment on this paper is particularly important to keep in mind. This paper has made a caricature of the field of statistics and the divide has been portrayed very dramatically. Another key criticism is that the argument against data modeling is mostly in the prediction setting and data models provide many useful theoretical results for other purposes.

References

[1] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[2] L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.