Gavin Whitesitt

Professor Lee

STAT 517

September 20, 2018

Don’t Put Yourself in a Box

The general idea of the paper: “Statistical Modeling: The Two Cultures” is that in the scientific community there are generally two ways that scientists can apply statistical modeling. One way is by assuming data is generated by a stochastic data model, the other is by using algorithmic models and treating the data mechanism as unknown. The author Leo Breiman asserts that the former is used almost exclusively and as a result has kept statisticians from working on a wide set of interesting and contemporary problems as well as leading to questionable conclusion in many papers.

In our complex and nuanced world that we live in it is not surprising that we there are many things that we do not fully understand. Nature operates on some vector of X variables and produces some output Y. In science our main goals are prediction (To be able to determine response given a set of input variables) and information (To understand some understanding of how nature associates response variables to input variables… the mechanism). Throughout Breiman’s paper he makes the argument that statisticians have boxed themselves in by only focusing on the latter.

The data modeling culture can be summarized by assuming that there is a model that explains the data and then validating this model based on the goodness-of-fit tests and residual examination. The author estimates that at the time of the paper being written (2001) 98% of all statisticians primarily use this approach. Common models used in this culture are linear regression, logistic regression and Cox models.

By contrast roughly only 2% of statisticians are part of the algorithmic modeling culture. The algorithmic culture does not try to determine causation, they instead try to simply figure out how x gets to y. The model validation is measured simply by the predictive accuracy. Common models used in the algorithmic modeling culture include decision tress, neural nets, support vector machines and k nearest neighbors classifiers and regressors. Breiman thinks by ignoring the algorithmic approach irrelevant theory and questionable scientific conclusions have been made.

Breiman then talks about his experiences as a consultant that have informed his opinions about the dogma associated with the data modeling culture and the upside of the algorithmic one. The failures of creating a model that assumes the data is generated by a stochastic data model can be illustrated by a few examples in Breiman’s work as a consultant. The Ozone project; an attempt to predict ozone levels 12 hours in advance, was a large failure and an example of a when using traditional methods was the wrong approach. The problem: in the 1970’s ozone levels would at some points due to a complex chemical reaction that was determined by the weather and local pollutants become dangerously high. Breiman was tasked with determining the risk level 12 hours than the prediction was previously capable of. If this task was accomplished the warnings could be much more effective potentially preventing a major health threat in Los Angeles.

Because the traffic in Los Angeles Basin was regular the variation at any given time varied only a few percent so the total emissions were roughly constant. Therefore, this problem was largely one of accurate weather prediction. 450 meteorological variables for a period of seven years were available. The training set was the first 5 years and the test set were the last two years. First large linear regression was used. This was followed by variable selection and again by addition of quadratic terms followed by more linear regression. In the 1970’s the computing power obviously was much lower than it is now. The project was a failure because the false positive rate was too high. Breiman looks back at this issue wishing he had access to the algorithmic methods available today.

The Chlorine Project was research project in which the EPA wanted to determine if a compound’s mass spectra could be used to determine the presence of chlorine reliably. This was motivated by the fact that the EPA samples thousands of compounds in a year trying to determine their potential toxicity and determining the chemical structure required an intensive and expensive analysis by a trained chemist. However, getting the mass spectra was fast and cheap. Breiman had access to a large dataset of 30,000 compound’s mass spectra. This data was split into 25,000 for a training set and 5,000 for the test set. Linear discriminant analysis and quadratic discriminant analysis was tried but due to the varying dimensionality of the mass spectrum predictor vector the only successful approach was using a design tree algorithm largely influenced by domain knowledge. The model predicted both chlorines and nonchlorines with 95% accuracy.

Data models are not entirely useless, but at its core the idea that a statistician can look at data and then invent a “reasonably good parametric class of models for a complex mechanism devised by nature” is flawed. Based on this common thinking illustrated by the all too common “Assume that the data are generated by the following model”. Furthermore, conclusions that are drawn from the devised model become gospel instead of being taken with a grain of salt. The paper cites a famous example. “Assume the data is generated by independent draws from the model:

Given the data being generated in this way there are tests of hypotheses, confidence intervals and distributions of residual sum-of-squares as well as asymptotics that can be derived. However, it is often true that the data could not have plausibly been constructed with a linear model, yet thousands of articles have been published claiming proof of something because the coefficient is significant at the 5% ignoring the evidence against linearity. Mosteller and Tukey sums up this issue succinctly “The whole area of guided regression is fraught with intellectual, statistical, computational, and subject matter difficulties.”

Rashomon means the multiplicity of good models. Occam means the conflict between simplicity and accuracy. Bellman means dimensionality. These three words; Rashomon, Occam and Bellman in large part explain the upside to the algorithmic approach.

Rashomon (named after a Japanese movie in which four people from different vantage points witness an event and all testify in court all stating what they saw but have vastly different ideas of what happened) means that there are different equations in a class of functions that give the same minimum error rate. This issue unfortunately is present in both algorithmic and data model driven approaches. When data is sensitive to differing models contingent on a small change in the data it is considered to be unstable. One potential solution proposed by Breiman is to aggregate a large set of competing models, thus the term “bagging” was coined.

Traditionally Occam’s Razor is interpreted as if there are competing explanations that have similar evidence the simpler one is better. In this paper however, Breiman thinks that there is a fundamental tradeoff between accuracy and complexity and more complex models are usually more accuracy. Although linear regression may provide an easily interpretable picture of the covariance between x and y it is usually not as accurate as a neural net. “Accuracy generally requires more complex prediction methods”. For academia this may seem like a big issue in regard to interpretation, but to an end focused application of statistics it doesn’t seem like there is much of an issue at all. Additionally, statisticians need to think about whether the tradeoff between accuracy and interpretability is really worth it. Often academics may think that they are losing interpretability by using a more complex model, but if the conclusions drawn are for a model that doesn’t actually model nature correctly how valuable are these interpretations?

Bellman refers to dimensionality and its potential to be dangerous with traditional statistical methods. In fact in 1993 Willian Cleveland considered to be one of the fathers of residual analysis stated that residual analysis cannot uncover lack of fit in more than four to five dimensions. However, with machine learning dimensionality can be a blessing by providing more information to improve the predictive accuracy of the algorithm (Shape Recognition Forest) and (Support vector Machines).

Overall the argument made is not one suggesting that the algorithmic approach is any better than the data modeling approach. The argument is one instead that we should keep our options open when we are trying to solve a problem. Going forward we should let the data tell us what to do. We should not decide how we are going to analyze a problem without understanding the context in which it is present. The world we live in today has increasingly accepted the algorithmic culture. Likely partially because of Breiman himself.