Breiman's two Cultures from a Nonparametric Perspective

Seminar Paper for Non- and Semiparametric Modeling

- Name: Dingyi Lai
- Student-ID No: 615865
- Study Subject: M.Sc. Statistics

Abstract

Breiman’s two cultures once gave rise to a heated discussion in statistics. It mainly encouraged the algorithmic modeling culture instead of the traditional data modeling culture. With the development of data science, computer science and nonparametric statistics, debates around the boundary of the two cultures are going on; even a hybrid culture was proposed as a compromise. This paper argued that nonparametric statistics is still a strong backup to the data modeling culture, but it is an improved paradigm to interpret the blackbox, which could be used to derive another paradigm such as causal inference and some tenable theories that could give a better explanation to the models in the algorithmic modeling culture.

1. Breiman's Two Cultures and the Debates

Breiman came up with a partition in statistical thinking which is called "two cultures". In his paper (Breiman, 2001), he argued that statistics starts with data; therefore, he advised statisticians to come face to face with the problems and data, which requires them to find a reasonable solution by whatever methods that work. His strict criticism on the old fashion is under the context that the traditional statistician and students who studied statistics tended to start with a model, even these models could not work well in reality. Especially when the computer science has been revolutionized, a flush of new problems and idiosyncratic formats of data has been rising up. In many cases, the traditional statistical models could not meet the various demands nowadays.

Breiman defined a procedure about how outputs are generated by inputs in his paper. According to his proposing prototype, there is a blackbox lying between inputs and outputs. This blackbox has not been figured out by human, and we usually aim to predict and extract informations from the blackbox. The first culture in statistics, the data modeling culture, could be viewed as an attempt to figure out the blackbox. We assumed a stochastic data model to imitate the nature functions in blackbox, and defined a loss function (usually using goodness-of-fit tests and residual examination) to validate the model. Because it is difficult to model the blackbox once for all, some restricted assumptions are imposed in order to simplify the model. While the first culture is focusing on the mechanism of blackbox and it seems to go too far in the mathematical puzzle, the second culture, the algorithmic modeling one, decided to leave the mechanism alone. The adherents of the algorithmic modeling culture admitted that the blackbox is too complex and mysterious to be figured out, and they are not interested in its mechanism; rather, they just want to use the blackbox directly and apply some algorithms that could predict the response variables very well. In Breiman's view, the second culture focus on finding a good solution while facing with a practical problem, and it emphasizes the predictive accuracy, which is the ultimate goal in several fields. In a bid to prove the necessary of promoting the second culture, Breiman used three examples to demonstrate that the algorithms in machine learning could not only provide a lower prediction error, but also better rank important variables and reveal an intrinsic structure of data than the

traditional statistical models. Also, he threw out three lessons learned from the development of algorithms, which are Rashomon Effect, Occam's Razor and Bellman's Phrase. These lessons invited readers to think about the multiplicity of good models, the conflict between simplicity and accuracy, and whether dimensionality is a curse or blessing.(Breiman, 2001)

Breiman is not the only one who called for a reformation of academic statistics. John Tukey, John Chambers, Jeff Wu and Bill Cleveland all independently once urged for an enlargement of traditional statistics (Donoho, 2017). The direct result of this proposal is the birth of a new subject called "data science". This discipline is based on the idea of starting from data and focusing on data. Debatably, some statisticians argued that statistics has a larger scope whose subject of interest is learning from data, (e.g. Chambers, 1993; Dudewicz, 1987), while others insisted that statistics is only a small part of data science (e.g. Mosteller and Tukey, 1968). According to the debates that Donoho listed in his article, there was an overlap between data science and statistics originally. While lots of statisticians were occupied by studying the mathematical models, people from other subjects and industries got out ahead of the research in the big data and algorithms. In another article about the development of machine learning (Jordan and Mitchell, 2015), the authors viewed machine learning as an intersection between computer science and statistics, which is at the center of artificial intelligence and data science. After reading these debates on the similars concepts, we could see that the idea of Breiman's two cultures in fact leads to a key question: How should people decide between the first culture and the second culture? A step further: Does data science provide a tenable compromise between the two?

Admittedly, both data science and machine learning are inspired by the algorithmic modeling culture. It seems like that the second culture keeps one step ahead of competition. However, Breiman's two cultures set inference against prediction, which might fail to tell the whole story about analyzing the data. In these days, data and the problems we meet are in a variety and flexibility, but we not only long for accurate prediction but also desire a better inference and explanation. The emergence of explainable AI is a good example (Samek et al., 2019). Meanwhile, several concepts and techniques in nonparametric statistics, such as kernel, Bayesian, bootstrap etc., have been widely adopted in the modern algorithms. For example, k-

nearest neighbor (KNN) method, kernel-based methods and decision tree-based algorithms are parts of the statistical machine learning. Therefore, some scholars came up with a third culture, a hybrid one. Iain et al. (Iain et al., 2018) argued that the important way of thinking, scientific method, were taught in statistics courses, which was called "hypothetico-deductive" in another paper (Daoud and Dubhashi, 2020). An integrated statistical learning theory is proposed by gluing some underlying deeper principles from two cultures together (Mukhopadhyay and Wang, 2020). Another recent paper came up with a term called hybrid modeling culture (Daoud and Dubhashi, 2020). In this paper, the author took the food supply and famines as the inputs and outputs variables respectively, and concluded the difference of the data-modeling culture (DMC) and the algorithmic-modeling culture (AMC) in Table 1.

	Data-modeling culture (DMC)	**Algorithmic-modeling culture (AMC)**
Exemplifying question	What is the causal relationship between food supply and famines?	How well can famines be predicted from available data?
Goal	Estimating unbaised parameters for causal estimation, to populate the magnitudes of the edges of a DAG.	To develop and train an algorithm *f* for accurate prediction.
A key assumption	Assuming a DAG, a stipulated and interpretable statistical model such as $y_i = c_0 + \beta w_i + e_i$ produces unbiased estimates of the true causal quantity β.	The algorithm f can produce accurate predictions of Y from data source, D.
Limitation	Although the parametric model is interpretable, its statistical structure may be a poor representation of the causal system.	Although f produces accurate predictions, the model is a black-box restricting causal interpretations.
Quantity of interest	$\hat{\beta}$	$\hat{Y}$

Notes: a) in the equation $y_i = c_0 + \beta w_i + e_i$, the outcome is y_i and the treatment is w_i. The variable c_0 is the intercept and e_i is the residual.

TABLE 1: CENTRAL PRACTICES OF TWO STATISTICAL CULTURES

From Table 1 we can see, DMC and AMC pays attention to different aspects. DMC is a hypothetico-deductive scientific method (Hempel, 1965; Popper, 2002) and focus on explanation of causal relationships; AMC focus on accurate prediction that is also the key point in data analysis. Naturally, HMC aims to blend $\hat{Y}$-prediction problems and $\hat{\beta}$-inference problems to a mixing state (Molina and Garip, 2019; Mullainathan and Spiess, 2017; Yarkoni

and Westfall, 2017). In the paper of Daoud and Dubhashi, HMC emerged from the trend of data science and causal-inference revolution. Thus, data science is one of the contexts that HMC should be formulated and machine learning (ML) is viewed as a representation of AMC. The stylized characteristics of HMC is given in Table 2.

	ML for causal inference	**ML for data acquisition**	**ML for theory prediction**
Exemplifying question	What is the causal relationship between food supply and famines?	Can food availability be measured from satellite images?	How well does the Malthusian theory of famines predict new famines? How does it compare to a Senian theory?
Goal	Imputing potential outcomes for causal estimation, to populate the magnitudes of the edges of a DAG.	Producing new indicators from digital sources, D, to populate the nodes of a DAG.	Comparing the predictive power of two or more theories' DAGs, $\hat{Y}_{G_1}, \hat{Y}_{G_2}, \ldots, \hat{Y}_{G_k}$, for new realizations of an outcome, Y.
A key assumption	The algorithm f produces unbiased estimates of the true causal quantity τ, assuming a DAG.	The algorithm f can measure the true quantity of the variable of interest (X, W, Y) from a digital source, D.	The algorithm f is an appropriate representation of G_k to predict, Y.
Quantity of interest	$\hat{\tau} = \hat{Y}_i^1 - \hat{Y}_i^0$	$\hat{X}, \hat{W}, \hat{Y}$	$\epsilon_{\hat{Y}_{G_k}} \approx Y - \hat{Y}_{G_k}$ or $\epsilon_{\hat{\tau}_{G_k}} \approx \tau - \hat{\tau}_{G_k}$

TABLE 2: CENTRAL PRACTICE OF THE HYBRID-MODELING CULTURE (HMC)

In the ML for causal inference, four combinations are named in the paper: 1) Imputation of potential outcome, e.g. T-learner; 2) Initial estimation step with regard to prediction, e.g. improved two-stage instrumental variable methods; 3) Policy optimization in sequential decision-making by combing identification of causal effect and optimal choice based on sequential prediction; 4) Causal discovery. Furthermore, the collection of high-quality data is of importance. Since DMC mainly relies on an analog-measurement method controlled by humans and AMC extracts data from any kind of sources, HMC tends to synthesize analog and digital approaches taking data privacy and security into consideration. Theory prediction is about how to evaluate the predictive performance of their theories. AMC-predictions emphasize the contest among results, while DMC-predictions stress the competition among theories. HMC-predictions, ideally, are DMC-predictions submitting to the principles dictated by AMC. (Daoud and Dubhashi, 2020)

In the next part we will consider these theories from a nonparametric perspective.

2. Nonparametric perspective

The word “non-parametric” by J. Wolfowitz in 1942 originally referred to an opposite concept against “parametric” (Noether, 1967; Dudewicz, 1987). This term is sometimes used interchangeably with “distribution-free” (Dudewicz, 1987). Generally, it is acknowledged that the emergence and development of nonparametric statistics owed to the Wilcoxon rank sum test and the Wilcoxon signed rank test that were proposed by Wilcoxon in 1945 (Noether, 1992). Stone in 1982 defined nonparametric estimators belonging to a collection of functions in terms of a infinite number of unknown parameters (Stone, 1982). In 1985, the nonparametric statistics was defined to be a free-format and parameter-free statistical technique with general assumption (Gibbons, 1985). In the beginning of 21st century, the boundaries of nonparametric statistics were expanded further, such as the nonparametric smoothing concept (Härdle, 2004). From the history of nonparametric we can see that nonparametric statistics has been widely enriched.

Back to the theory of Breiman’s two cultures, nonparametric statistics has actually broaden the traditional data modeling culture (DMC), at least stripping off most of the strict assumptions in parametric procedure. But it is still concentrated in understanding the natural function. Therefore, if we think about Breiman’s two cultures from a nonparametric perspective, nonparametric relationship still captures the nature function but goes beyond the untenable assumptions of parametric model.

There are several applications of nonparametric statistics: 1) To improve the interpretability, it can combine with parametric model to build the semiparametric model, which might be useful in extracting explainable information from data; 2) To improve the precision, it embraces algorithmic modeling culture (AMC) and try to study the statistical properties of ML algorithms; 3) It uses ML algorithms to improve the original methods in statistics where a prediction process is needed. Also, in the third culture, the hybrid modeling culture (HMC), nonparametric statistics could be applied wherever a relationship needs to be described without committing to a specific functional form.

Take ML for causal relationship as an example. As we all know, “Correlation does not imply causation”. In reality, we conduct a scientific research to figure out the causal relationship between inputs and outputs. Causal inference can be used in this case, and it subsumes statistical inference. Although the causality is philosophically obscure, for example Hempel’s covering law model of explanation and Hume’s claim to that causality is a figment of our imagination (Okasha, 2002), there are already several models to depict the causality and widely used in reality. Table 3 is a toy dataset from Daoud and Dubhashi (Daoud and Dubhashi, 2020). The potential outcome framework proposed by Donald Rubin defined the causal effect as $\tau_i = Y_i^1 - Y_i^0$.

	Y	Y^1	Y^0	W	τ	X
Jane	20	20	?	1	?	10
John	30	30	?	1	?	11
Joe	25	?	25	0	?	10
Jan	22	?	22	0	?	11

TABLE 3: A TOY DATASET ILLUSTRATING THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

Because it is impossible to observe both potential outcomes, which is called the fundamental problem of causal inference, based on some assumptions ML algorithms can predict potential outcomes in this framework, e.g. T-learner (Daoud and Dubhashi, 2020). The more general causality framework that is popular in recent years, structural causal models (SCMs), replaces the parents-child relationship $P(X_i \mid \mathbf{PA_i})$ in directed acyclic graph (DAG) with its functional counterpart $X_i = f_i(\mathbf{PA}_i, U_i), i = 1,...,n$ which formalizes interventions (Schölkopf, 2019). According to the paper from Schölkopf in 2019, the natural function that lies in the blackbox is here the deterministic function f_i which depends on X_i’s parents in the graph (denoted by $\mathbf{PA}_i$) and a stochastic unexplained variable variable U_i; the causal mechanisms are the causal (or disentangled) factorization ($p(X_1, \ldots, X_n) = \prod_{i=i}^{n} p(X_i \mid \mathbf{PA}_i)$) and other entangled factorizations ($p(X_1, \ldots, X_n) = \prod_{i=i}^{n} p(X_i \mid X_{i+1}, \ldots, X_n)$). This causal learning expands the

statistical learning whose conceptual basis is a joint distribution, considering a richer class of assumptions and detecting the causality in the joint distribution (Schölkopf, 2019). In this framework, given a DAG, nonparametric methods could provide an alternative to capture the relationship. For example, there is an area called nonparametric structural equations (Pearl, 2009). Remarkably, Judea Pearl as one of the first to mathematize causal modeling wrote a comment very recently to Breiman's two cultures, in which the idea to separate data-fitting from interpretation should be encouraged, and Pearl stated that the task of interpretation can be done by causal analysis (Pearl, 2021).

Another example about the nonparametric view of Breiman's two cultures is the statistical study in ML, which is helpful for us to dispel the mystery about the natural function. Early in 2000, Breiman has written a paper to shed light on some infinity theory for predictor ensembles and bridged the gap between random forests and kernel (Breiman, 2000). Then in 2001, Breiman firstly provided the bound on the misclassification rate which is loose but suggestive. In 2007, Banerjee and McKeague derived that in a regression tree context with independent Gaussian noise, there is a limit law for the split location (Banerjee and McKeague, 2007; Peng, 2019). Ishwaran analyzed the CART-style splits further (Ishwaran, 2015; Peng, 2019) and Lin and Joen showed a relationship between potential nearest neighbors and tree-based ensembles (Lin and Joen, 2006; Peng, 2019). Other recent literature reviews are listed in the paper from Peng (Peng, 2019).

In conclusion, Breiman's two cultures from a nonparametric perspective are not mutually exclusive. The data modeling culture provides a hypothetico-deductive thinking way to understand the nature function, while the algorithmic modeling culture offers a more accurate simulation from the blackbox itself. The nonparametric study to the algorithm models with high accuracy is a relative ideal method to figure out the mechanism and better improve and explain the models. But still, nonparametric statistics is based on the traditional statistics that might be one of the paradigms to interpret the natural function. Causal inference might be an alternative paradigm.

3. Reference

Breiman, L. (2000). Some infinity theory for predictor ensembles. Technical Report 577, UC Berkeley.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), 199-215.

Carmichael, Iain, & Marron, J S. (2018). Data science vs. statistics: Two cultures? *Japanese Journal of Statistics and Data Science, 1*(1), 117-138.

Chambers, J. M. (1993), "Greater or Lesser Statistics: A Choice for Future Research," Statistics and Computing, 3, 182–184.

Charles J. Stone. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics, 10*(4), 1040-1053.

Daoud, A., Dubhashi, D. (2020). Statistical modeling: the three cultures. *ArXiv:2012.04570 [stat.ME]*.

Donoho, David. (2017). 50 Years of Data Science. Journal of Computational and Graphical Statistics, 26(4), 745-766.

Dudewicz, Edward J. "Nonparametric Methods: The History, the Reality, and the Future (with Special Reference to Statistical Selection Problems)." *Contributions to Stochastics*. Heidelberg: Physica-Verlag HD. 63-83. Web.

Gibbons, J. D. (1985). *Nonparametric Methods for Quantitative Analysis (Second Edition)*. American Sciences Press, Inc., Columbus, Ohio.

Härdle, W., Werwatz, A., Müller, M., Sperlich, S., & SpringerLink. (2004). *Nonparametric and Semiparametric Models* (Springer Series in Statistics). Berlin, Heidelberg.

Hemant Ishwaran. The effect of splitting on random forests. Mach. Learn., 99 (1):75–118, April 2015. ISSN 0885-6125. doi: 10.1007/s10994-014-5451-2. URL http://dx.doi.org/10.1007/s10994-014-5451-2. (peng, P. 34: 901)

Hempel, Carl G. 1965. Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. Place of publication not identified: Free Press.

Jordan M. I., & Mitchell T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science (American Association for the Advancement of Science), 349*(6245), 255-260.

Mosteller, F., and Tukey, J. W. (1968), "Data Analysis, Including Statistics," in Handbook of Social Psychology (Vol. 2), eds. G. Lindzey, and E. Aronson, Reading, MA: Addison-Wesley, pp. 80–203.

Molina, Mario, & Garip, Filiz. (2019). Machine Learning for Sociology. *Annual Review of Sociology, 45*(1), 27-45.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." Journal of Economic Perspectives 31(2):87–106. doi: 10.1257/jep.31.2.87.

Noether G.E. (1992) Introduction to Wilcoxon (1945) Individual Comparisons by Ranking Methods. In: Kotz S., Johnson N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4380-9_15

Okasha, Prof., Philosoph, Hochschullehrer, Oxford, Bristol, . . . Ca. 20./21. Jahrhundert. (2002). *Philosophy of science : A very short introduction* (1. publ. ed., Very short introductions BV013097034 67). Oxford [u.a.].

Pearl, J., & ProQuest. (2009). *Causality models, reasoning, and inference* (2nd ed.). Cambridge [England] ; New York: Cambridge University Press.

Pearl, Judea. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*(None), Statistics surveys, 2009-01-01, Vol.3 (none).

Pearl, Judea. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*(3), 54-60.

Pearl, Judea. (2021). Causally Colored Reflections on Leo Breiman's "Statistical Modeling: The Two Cultures" (2001).

Peng, Wei, Coleman, Tim, & Mentch, Lucas. (2019). Asymptotic Distributions and Rates of Convergence for Random Forests via Generalized U-statistics. *ArXiv:2005.13596 [stat.ML]*
Popper, Karl. 2002. The Logic of Scientific Discovery. London: Routledge.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Muller, K., & SpringerLink. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (1st ed. 2019. ed., Lecture Notes in Artificial Intelligence ; 11700). Cham.

Schölkopf, Bernhard. (2019). Causality for Machine Learning. *ArXiv:1911.10500 [cs.LG]*.

Subhadeep, M. & Wang, K. (2020). Breiman's "Two Cultures" Revisited and Reconciled. *ArXiv:2005.13596 [stat.ML]*

Yarkoni, Tal, and Jacob Westfall. 2017. "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning." Perspectives on Psychological Science 12(6):1100–1122. doi: 10.1177/1745691617693393.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association, 101(474):578–590, 2006.*