Melting together prediction and inference

Adel Daoud, Devdatt Dubhashi

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

Published by University of Pennsylvania Press

DOI: https://doi.org/10.1353/obs.2021.0035

For additional information about this article
https://muse.jhu.edu/article/807713

For content related to this article
https://muse.jhu.edu/related_content?type=article&id=807713
Melting together prediction and inference

Adel Daoud  adel.daoud@liu.se
Institute for Analytical Sociology & The Division for Data Science and Artificial Intelligence
Linköping University & Chalmers University of Technology
Norrköping, Sweden & Gothenburg, Sweden

Devdatt Dubhashi  dubhashi@chalmers.se
The Division for Data Science and Artificial Intelligence
Chalmers University of Technology
Gothenburg, Sweden

Abstract

In Leo Breiman’s influential article “Statistical modeling—the two cultures” he identified two cultures for statistical practices. The data modeling culture (DMC) denotes practices tailored for statistical inference targeting a quantity of interest, \( \hat{\beta} \). The algorithmic modeling culture (AMC) refers to practices defining an algorithm, or a machine-learning (ML) procedure, that generates accurate predictions about an outcome of interest, \( \hat{Y} \). As DMC was the dominant mode, Breiman argued that statisticians should give more attention to AMC. Twenty years later and energized by two revolutions—one in data-science and one in causal inference—a hybrid modeling culture (HMC) is rising. HMC fuses the inferential strength of DMC and the predictive power of AMC with the goal of analyzing cause and effect, and thus, HMC’s quantity of interest is causal effect, \( \hat{\tau} \). In combining inference and prediction, the result of HMC practices is that the distinction between prediction and inference, taken to its limit, melts away. While this hybrid culture does not occupy the default mode of scientific practices, we argue that it offers an intriguing novel path for applied sciences.

Keywords: causal inference, prediction, machine learning, data science, statistical cultures

1. Introduction

Breiman (2001a) identified two cultures for statistical modeling. The data modeling culture (DMC) refers roughly to practices aiming to conduct statistical inference on one or several quantities of interest. By unbiased statistical inference, we mean a procedure tailored to estimates a quantity \( \hat{\beta} \) such that the difference between the true quantity \( \beta \) is as small as possible: the procedure is unbiased when the difference \( \beta - \hat{\beta} \) is negligible in expectation. This true quantity \( \beta \) exists independently of the statistical model producing \( \hat{\beta} \). The algorithmic modeling culture (AMC) refers to practices defining a procedure, \( f \), that generates accurate predictions, \( \hat{Y} \), about an event (outcome), \( Y \). By accurate, we mean predictions that are as similar as possible to the true event that \( f \) has yet not encountered (Hastie et al., 2009). The smaller the difference \( Y - \hat{Y} \), the higher the similarity. A procedure is an algorithm, or a function, that takes some input \( X \), operates on this input \( f(X) \), and then, produces an output \( f(X) = \hat{Y} \). Often, this procedure are defined in terms of a
machine-learning (ML) algorithm (Hastie et al., 2009). Thus, \( \hat{Y} \)-prediction problems and \( \hat{\beta} \)-inference problems form distinct sets of practices in mobilizing data, algorithms, and results (Mullainathan and Spiess, 2017).

While Breiman argued that statisticians should devote more attention to \( \hat{Y} \)-problems than \( \hat{\beta} \)-problems, today a third culture is rising: the hybrid modeling culture (HMC). This third culture emerges from statistical practices where prediction and inference synthesize into new procedures (Daoud and Dubhashi, 2020; Kino et al., 2021). As the main concern of HMC is causality, its focus is on, what we denote, \( \hat{\tau} \)-problems. Using the Neyman-Rubin causal framework or Pearl’s do-calculus, we define causal effects as the difference between potential outcomes \( \tau = Y_1 - Y_0 \). The potential outcome \( Y_1 \) is the outcome when a cause \( W \) (e.g., a treatment, policy, or exposure) is active and \( Y_0 \) occurs when \( W \) is inactive. While the interest in identifying causal effects exist in DMC already, a key difference between DMC-powered models for causal inference and HMC-powered models for causal inference, is that the latter mobilizes the predictive power of ML. In other words, HMC uses tricks from AMC to achieve DMC goals. As these tricks rely on combining inference and prediction, the result of HMC is that the distinction between \( \hat{Y} \) and \( \hat{\beta} \)—taken to its limit—melts away. In this commentary, we delineate our “melting away” argument.

2. Machine learning for causal inference

Before describing how ML aids in inferring causality, we will refine our definition of what we mean by causal inference. We define a cause of interest as a binary variable, \( W \). Instead of merely recording each individual’s outcome as observed by the data, \( Y_i \), we assume that each individual \( i \) has two potential outcomes (Imbens and Rubin, 2015). One potential outcome records the outcome when the individual takes the treatment \( Y_{1i} \) and one where he or she does not take it \( Y_{0i} \). The causal effect for each individual \( i \) is then the difference between these two potential outcomes:

\[
\tau_i = Y_{1i} - Y_{0i}
\]

If we could observe both potential outcomes, we could then directly compute \( \tau_i \) and thus identify individual-level causal effects. However, the observed outcome—as supplied by the data—is a function of both the treatment and the two potential outcomes, \( Y_i = (W - 1) Y_{1i} + W Y_{0i} \). This function shows that the observed data reveals only one of these two potential outcomes, yet both are required to identify causal effects. Table 1 exemplifies an observed-data matrix of four individuals with fictitious variable values, and their missing potential outcomes. This impossibility of observing both potential outcomes is known as the fundamental problem of causal inference. Much of the causal-method development pertains to defining procedures for when the causal effect is identified from observational data (Hernan and Robins, 2020; Imbens and Rubin, 2015; Pearl, 2009; Peters et al., 2017). By identified, we mean a causal effect calculable from measured data.
Melting together prediction and inference

Table 1: A toy dataset illustrating the fundamental problem of causal inference

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>Y^1</th>
<th>Y^0</th>
<th>W</th>
<th>τ</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>20</td>
<td>20</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>10</td>
</tr>
<tr>
<td>John</td>
<td>30</td>
<td>30</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>11</td>
</tr>
<tr>
<td>Joe</td>
<td>25</td>
<td>?</td>
<td>25</td>
<td>0</td>
<td>?</td>
<td>10</td>
</tr>
<tr>
<td>Jan</td>
<td>22</td>
<td>?</td>
<td>22</td>
<td>0</td>
<td>?</td>
<td>11</td>
</tr>
</tbody>
</table>

Vibrant literature in the overlap between computer science, econometrics, and statistics combine ML and causal methodology to develop new estimators, used in various domains (Angrist and Pischke, 2014; Athey and Imbens, 2017; Athey et al., 2019; Chernozhukov et al., 2018; Daoud et al., 2020, 2019; Daoud and Johansson, 2020; Hedström and Manzo, 2015; Hernan and Robins, 2020; Hill, 2011; Hirshberg and Zubizarreta, 2017; Imai, 2018; Kraamwinkel et al., 2019; van der Laan and Rose, 2011; Morgan and Winship, 2014; Pearl and Mackenzie, 2018; Shiba et al., 2021; VanderWeele, 2015). A recurring theme in these methods is the many creative combinations where predictive AMC-type algorithms are used in DMC-type of causal inference. There are several ways in which ML algorithms help the scientific endeavour (for an overview see Daoud and Dubhashi, 2020), but the most important of them is the use of ML to predicting the missing-potential outcomes (Athey and Imbens, 2017).

As observed data only reveal one-half of the potential outcomes, the other half is regarded as missing data. One way of handling this fundamental problem is to cast it as a missing-data problem and proceed to identify conditions for imputing these data to populate all the $Y_i^1$ and $Y_i^0$ cells, based on covariates $X$ (for a critique of this missing-data definition see Pearl and Mackenzie, 2018). These imputation procedures rely on common identifiability assumptions. One such central assumption is conditional independence (also known as conditional ignorability and conditional exchangability), $W \perp Y_i^1, Y_i^0 | X$. This mathematical statement means that the treatment is as-if randomly assigned conditional on one or more covariates.

Because ML excels in prediction tasks compared to commonly used parametric models, HMC-influenced scholars have developed many different procedures to predict potential outcomes (Künzel et al., 2018). For example, the $T$-learner—“T” stands for two—procedure defines one ML-algorithm $f_{w=1}(x_i) = E[Y = y_i | W = 1, X = x_i]$ trained on the treated group and another algorithm $f_{w=0}(x_i) = E[Y = y_i | W = 0, X = x_i]$ trained on the control group. A Lasso, a random forest, or a collection of algorithms (an ensemble) are often used to define $f_{w=1}$ and $f_{w=0}$. The SuperLearner provides a well-tested framework to mobilize ensembles for causal inference (van der Laan and Rose 2011). After training, $f_{w=1}$ imputes potential outcomes for the control group and $f_{w=0}$ imputes these outcomes for the treated group. Based on the toy data of Table 1, $f_{w=1}$ trains on Jane and John, and imputes $Y_i^1$ of Joe and Jan; likewise, $f_{w=0}$ trains on Joe and Jan, and imputes $Y_i^0$ of Jane and John. This procedure culminates by calculating the difference $\tilde{\tau}_i = \tilde{Y}_i^1 - \tilde{Y}_i^0$ for the control group and $\hat{\tau}_i = Y_i^1 - \hat{Y}_i^0$ for the treated group, and then averaging over all groups $\hat{\tau} = E[E[\tilde{\tau}_i | W = w_i]]$ to calculate the average treatment effect.
The T-learner algorithm is one of several, but common to most of these ML algorithms is the procedure of imputing potential outcomes (Künzel et al., 2018) or imputing the treatment effect directly (Athey et al., 2019). Although much research is devoted to analyzing biases arising from ML regularization, these HMC algorithms demonstrate how $\hat{\tau}$-problems subsume $\hat{\beta}$-problems originating from DMC by mobilizing the algorithmic power of AMC used for $\hat{Y}$-problems. Thereby, the original distinction between $\beta$ and $\hat{Y}$ has dissipated—melted away.

3. Conclusions

It is perhaps a historical irony that one of the most popular HMC algorithms, the generalized random forest (Athey et al., 2019), uses a random-forest algorithm as a key ingredient for causal inference; the same algorithm that Breiman (2001a) used to exemplify what AMC-type of predictions had to offer the scientific endeavour. It is the same algorithm he devoted much research in developing (Breiman, 2001b).

Evidently, Breiman’s work has opened up new a new perspective not only for statistics but also for applied sciences. This perspective direct our attention towards the possibilities of our time—the era of data science. Twenty years later, based on the advances in machine learning and causal inference, scholars are enabled to move one step further. As identifying causal effects is one of the core goals of the scientific endeavor, we conclude that instead of retaining the dichotomy between AMC-prediction and DMC-inference, this endeavor gains more by embracing both synthetically. HMC provides a way to think about how this synthesis is possible in the era of data science (Meng, 2020) while still maintaining the scientific endeavor’s higher goal: explaining reality. Although this hybrid culture does not occupy the default mode of scientific practices, we argue that it offers an intriguing novel path forward for applied sciences (Daoud and Dubhashi, 2020).

Acknowledgments

We would like to thank Fredrik Johansson for constructive comments. The usual disclaimer applies.

References


