Published in final edited form as:

Eur J Epidemiol. 2018 May ; 33(5): 437–440. doi:10.1007/s10654-018-0405-9.

You are smarter than you think: (super) machine learning in context

Alexander P. Keil and Jessie K. Edwards

Department of Epidemiology, University of North Carolina

Abstract

We discuss an article on super learning by Naimi and Balzer in the current issue of this journal in the context of machine learning. We give a brief example that emphasizes the need for human intelligence in the rapidly evolving field of machine learning.

In 1992, David Wolpert introduced what he called “stacked generalization” as a method for prediction that had the nice property of producing predictions that were accurate across samples of data drawn from similar populations(1). “Stacking,” as it has become known, is a member of a class of methods known as ensemble-learners, or algorithms that combine information from multiple models to (ideally) generate more accurate results. In the current issue of this journal, Naimi and Balzar revisit stacking under the name “super learning” (or super learner) (2). As those authors demonstrate, super learning is a useful tool for prediction that allows researchers to worry less about how to specify prediction models and focus more on the predictions themselves. One of the primary anxieties of graduate students in epidemiology is how to build and specify models. Remarkably, however, stacking and ensemble learning in general, the very treatments for this anxiety, do not often show up in the toolbox of epidemiologists.

One of the barriers to implementation of stacking has been the lack of available software, a problem that has been overcome by van der Laan and colleagues who developed free, open source software for super learner. A second barrier is unfamiliarity with the software itself, which is available in R and (perhaps necessarily) has a syntax and environment that provides a strong layer of abstraction between the programmer and the relationship between model inputs and outputs; Naimi and Balzer address this second barrier through an admirably simple demonstration of super learning for prediction. Beyond the technical barriers, however, is the basic question of “why.” Why ought we to care about marginal or even substantial gains in black-box predictive models when we call so little of what we do “prediction?” This is not to demean the current use of predictive models in epidemiology. Rather, this is to suggest that much of our work is prediction in disguise: we would like to predict what would happen if we could, somehow, manipulate our exposures. That is, causal inference is an exercise in prediction, and we would do well to pay attention to the advances

in machine learning not only for “pure” prediction, but as a way to improve our quest for uncovering relationships between exposures and health.

Super learner is a machine learning method, but so is regression

Murphy defines machine learning as a “set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data...” (3) [pg 1]. In the special case described by Naimi and Balzer, we have some set of predictors $\mathbf{x}$ that we wish to use make predictions of some outcome y (given by $\hat{y}(\mathbf{x})$), via a model given by $f(\mathbf{x}; \boldsymbol{\beta})$. When we have y measured in some individuals, then we can “train” the parameters of the model $\boldsymbol{\beta}$, which characterize the “patterns” in the data. This “trained” model is then used to generate predictions $\hat{y}(\mathbf{x})$ for new individuals in whom $\mathbf{x}$ is measured. Super learner is a clever approach to defining the model such that it produces accurate predictions in target population data when it is trained in data that are generated under the same probability distribution as the data from the target population (but which may vary randomly or according to different sampling schemes). In this sense, Super learner is just another machine learning method, albeit one with good prediction capabilities.

The favorite tools of epidemiologists, parametric regression models (e.g. linear regression, logistic regression) are also models of the form $f(\mathbf{x}; \boldsymbol{\beta})$, which are “trained” via maximum likelihood or sampling methods. In this sense, most of us already use the tools of machine learning, though interest often lies in $\boldsymbol{\beta}$ rather than the model predictions. Thus, machine learning should not feel foreign to us. Parametric regression models are often default choices to include within a super learner library.

Prediction versus inference

When epidemiologists learn about exposure-disease relationships, a traditional form of this learning is through the parameters of the model, rather than the predictions from that model. For example, the model may be a logistic model, in which case the parameters $\boldsymbol{\beta}$ are log-odds ratios. In contrast, machine learning algorithms treat $\boldsymbol{\beta}$ as a nuisance parameter and so very rarely do they estimate parameters that can be interpreted as simply as a log-odds ratio.

The utility of machine learning, however, becomes much more obvious when we divorce our particular model form from the effect measure. Classical examples in epidemiology include algorithms to estimate absolute risk (and hence, risk differences or ratios) from a logistic model (4–6). More generally, the field of causal inference has given rise to a particular type of prediction as the object of inference itself: potential outcomes. A “pure” prediction $\hat{y}(\mathbf{x})$ is the output of the model $f(\mathbf{x}; \boldsymbol{\beta})$ for some arbitrary choice of $\mathbf{x}$. In contrast, a predicted potential outcome ($\hat{y}_{\mathbf{x}}$) is an output of the model $f(\mathbf{x}; \boldsymbol{\beta})$ in which $\mathbf{x}$ (and the model) is chosen carefully so as to yield predicted outcomes we would observe, had exposure been set to some specific value (e.g. unexposed). In concert with machine learning methods to estimate propensity scores (7), this type of prediction is used in service of modern epidemiologic methods for inference, such as targeted maximum likelihood (minimum loss) estimation (8).

As Naimi and Balzer point out, there are still theoretical details to work out when using machine learning for causal inference, but this is an area of active research with some promising avenues (9, 10) and practical epidemiologic examples (11, 12). Describing super learner as simply a way to make predictions vastly undersells its potential value for epidemiology.

Disregarding whether inference is based on β or $\hat{y}$, the practical difference between regression modeling and super learning is (in principle) the way that a model is selected. A typical workflow for selecting a particular model for a study involves selecting from a set of candidate models based on some particular utility function. Our utility function may dictate that we prefer parsimony over complexity, that we select a model that maximizes some objective measure of model fit, or that we use models consistent with prior literature. In contrast, with super learner we could simply include each candidate model as a member of the super learner library, and allow the algorithm to select the best performing model or some weighted average of models. Super learner does this in a principled way that follows well defined rules to maximize predictive accuracy without over-fitting. In contrast the human learner may follow practices that are informal and, at worst, biased towards findings of a particular direction or strength (even if no nefarious intentions are present (13)).

Are we there yet?

The rules that allow super learner to avoid some of the pitfalls of more informal model selection procedure can be a double-edged sword. The startling advances in machine learning for making inference under unknown generative models make it easy to neglect the importance of the analyst. Whereas machine learning algorithms see only the data at hand, the human analyst carries a wealth of prior knowledge that can inform model choice. The things that we “just know” can play a more important role than is currently appreciated in the theoretical world of algorithmic prediction.

We illustrate the importance of the analyst with a simple simulation. In contrast to the simulation of Naimi and Balzer, our data represent realistic epidemiologic data in that the noise to signal ratio is high such that the relationship between variables may not be obvious via graphical analysis. We simulated data for 50 individuals on two variables X and Y , where X is a standard normal variable and $Y_i = x_i - 0.2x_i^2 + \varepsilon_i$, where ε_i follow a normal distribution with a standard deviation of 2. These data are shown in Figure 1.

We also used super learner (using R defaults) to make predictions with these data, using a library that included multivariate adaptive regression splines, a linear model with a quadratic polynomial for X (the ‘true’ model), generalized additive models, Loess, and the sample mean. The predictions from this algorithm given by the dashed line in Figure 1 roughly pass the “sniff test” – that is this model meets our (human) expectation of a reasonable dose-response curve. Based on prior knowledge, we might expect that such curves are smooth or have features such as plateaus that reflect physical laws or biologic processes such as the conservation of momentum or receptor saturation.

We then added the Random Forest algorithm to the super learner library, and generated new predictions. These predictions, shown in Figure 2, do not pass the sniff test. The predicted dose response is very jagged and violates our principal that nature has some degree of smoothness. Most of us would scoff at the idea that any biologic process subject to homeostatic regulation could result in such apparently unregulated “jumps.” The usual advice for characterizing prior information within super learner is to include a set of plausible model forms (14); notably, our analysis includes the true model, suggesting this prior knowledge can be easily overruled by the algorithm. Indeed, knowing the truth is the ultimate form of background knowledge, and even that cannot save us here.

This is not to say we could not be incorrect in real examples that disagree with prior knowledge; this knowledge may be flawed. This is also not to devalue any particular algorithm – we use them rather incautiously in this example. This example is merely to point out that we run the risk of devaluing our own knowledge about the world if we neglect the important role of the researcher in evaluating the plausibility of the output of automated algorithms. This message is not new to epidemiology, but it bears repeating (15).

Conclusions

The promise of machine learning for improving epidemiologic inference is vast, especially in large datasets that prevent close scrutiny of the ways that variables relate to each other. Super learner has the potential to be an important tool to help epidemiologists keep pace with the data in a principled way. We should be cautious, however, that the billions of years of evolution and experience leading up to current levels of human intelligence is not ignored in the context of advances to computing in the last 30 years.

References

1. Wolpert DH. Stacked generalization. *Neural networks*. 1992; 5(2):241–59.
2. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018
3. Murphy, K. *Machine learning: a probabilistic perspective*. 1. Cambridge, MA: The MIT press; 2012.
4. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*. 2004; 160(4):301–5. DOI: 10.1093/aje/kwh221 [PubMed: 15286014]
5. Greenland S. Multivariate estimation of exposure-specific incidence from case-control studies. *J Chronic Dis*. 1981; 34(9–10):445–53. [PubMed: 7276134]
6. Muller CJ, MacLehose RF. Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. *Int J Epidemiol*. 2014; 43(3):962–70. DOI: 10.1093/ije/dyu029 [PubMed: 24603316]
7. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*. 2010; 63(8):826–33. [PubMed: 20630332]
8. Van der Laan, MJ, Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer; 2011.
9. van der Laan MJ, Starmans RJ. Entering the era of data science: Targeted learning and the integration of statistics and computational data analysis. *Advances in Statistics*. 2014; 2014
10. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*. 2017

11. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in medicine*. 2010; 29(3):337–46. [PubMed: 19960510]
12. Watkins S, Jonsson-Funk M, Brookhart MA, Rosenberg SA, O'Shea TM, Daniels J. An empirical comparison of tree-based methods for propensity score estimation. *Health Serv Res*. 2013; 48(5): 1798–817. DOI: 10.1111/1475-6773.12068 [PubMed: 23701015]
13. Gelman, A, Loken, E. Department of Statistics. Columbia University; 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time.
14. Polley, EC, van der Laan, MJ. Super learner in prediction: Division of Biostatistics. University of California; Berkeley: 2010. Contract No.: 222
15. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology*. 2001; 12(3):313–20. [PubMed: 11338312]

Figure 1.

Super learner predictions for a small (N=50) simulation example. The super learner library included multivariate adaptive regression splines, a linear model with a quadratic polynomial for X (the 'true' model), generalized additive models, Loess, and the sample mean.

Figure 2.
Super learner predictions for the simulation example after adding Random Forest to the algorithm described in Figure 1.