Machine Learning and Causal Inference for Policy Evaluation

Susan Athey
Stanford Graduate School of Business
655 Knight Way
Stanford, CA 94305
1-650-725-1813
athey@stanford.edu

ABSTRACT

A large literature on causal inference in statistics, econometrics, biostatistics, and epidemiology (see, e.g., Imbens and Rubin [2015] for a recent survey) has focused on methods for statistical estimation and inference in a setting where the researcher wishes to answer a question about the (counterfactual) impact of a change in a policy, or "treatment" in the terminology of the literature. The policy change has not necessarily been observed before, or may have been observed only for a subset of the population; examples include a change in minimum wage law or a change in a firm's price. The goal is then to estimate the impact of small set of "treatments" using data from randomized experiments or, more commonly, "observational" studies (that is, non-experimental data). The literature identifies a variety of assumptions that, when satisfied, allow the researcher to draw the same types of conclusions that would be available from a randomized experiment. To estimate causal effects given non-random assignment of individuals to alternative policies in observational studies, popular techniques include propensity score weighting, matching, and regression analysis; all of these methods adjust for differences in observed attributes of individuals. Another strand of literature in econometrics, referred to as "structural modeling," fully specifies the preferences of actors as well as a behavioral model, and estimates those parameters from data (for applications to auction-based electronic commerce, see Athey and Haile [2007] and Athey and Nekipelov [2012]). In both cases, parameter estimates are interpreted as "causal," and they are used to make predictions about the effect of policy changes.

In contrast, the supervised machine learning literature has traditionally focused on prediction, providing data-driven approaches to building rich models and relying on cross-validation as a powerful tool for model selection. These methods have been highly successful in practice. This talk will review several recent papers that attempt to bring the tools of supervised machine learning to bear on the problem of policy evaluation, where the papers are connected by three themes.

The first theme is that it important for both estimation and inference to distinguish between parts of the model that relate to the causal question of interest, and "attributes," that is, features or variables that describe attributes of individual units that are held fixed when policies change. Specifically, we propose to divide the features of a model into causal features, whose values may be manipulated in a counterfactual policy environment, and attributes. A second theme is that relative to conventional tools from the policy evaluation literature, tools from supervised machine learning can be particularly effective at modeling the association of outcomes with attributes, as well as in modeling how causal effects vary with attributes. A final theme is that modifications of existing methods may be required to deal with the "fundamental problem of causal inference," namely, that no unit is observed in multiple counterfactual worlds at the same time: we do not see a patient at the same time with and without medication, and we do not see a consumer at the same moment exposed to two different prices. This creates a substantial challenge for cross-validation, as the ground truth for the causal effect is not observed for any individual.

The talk reviews several lines of research that incorporate these themes. The first, exemplified by Athey and Imbens [2015a], focuses on estimating heterogeneity in treatment effects, identifying (based on unit attributes) subpopulations of units that have larger or smaller than average treatment effects. The method enables valid inference: confidence intervals for the size of the treatment effect in each subpopulation are derived. Thus, large-scale randomized experiments for drugs or A/B tests in online settings can be evaluated systematically, with the method discovering the magnitude of treatment effect heterogeneity. The challenge in this setting is to find a method that is optimized for the problem of predicting causal effects, rather than for predicting outcomes. The approach can also be applied to observational studies under some additional conditions. Our approach addresses the problem of cross-validation by constructing an unbiased (but noisy) estimate of each unit's treatment effect. More generally, we pose the question of how to best modify supervised machine learning methods to use estimated parameters rather than observed data in cross-validation. In ongoing research, we explore this question in a variety of settings.

A second line of research analyzes robustness of causal estimates. In applied social science studies of the impact of policy changes, it is common for researchers to present a handful of alternative models to assess the robustness of the causal estimates. Although the importance of model robustness has been highlighted by many researchers (e.g. Leamer [1983]), to date no metric for the robustness of a model has gained widespread adoption in the policy evaluation literature. Athey and Imbens [2015b] propose a measure of robustness of parameter estimates. A starting point is to define the causal estimand of interest as well as the attributes of

KDD '15, August 10-13, 2015, Sydney, NSW, Australia.
ACM 978-1-4503-3664-2/15/08..
DOI: http://dx.doi.org/10.1145/2783258.2785466

individuals in the dataset (features that may affect the robustness of the causal estimate). The method for constructing the robustness measure is inspired by the machine learning technique of regression trees. The sample is split according to each attribute in turn, and the original model re-estimated on the two subsamples. The split point is determined as the one that leads to the greatest improvement in model fit. An alternative estimate of the causal effect is constructed by taking a weighted average of the estimates in the two subsamples. The robustness measure is then defined as the standard deviation of the estimates, taken over all of the alternative estimates (one for each attribute). This measure has some attractive properties: there is no need to define an estimation approach other than the one used in the baseline model, and the measure is robust to monotone transformations of the individual attributes. The measure lacks other desirable properties, however: it can be reduced by adding irrelevant attributes to the model, for example. An ongoing research agenda addresses this and other issues.

Finally, Abadie et al. [2014] consider the problem of inference in environments where the researcher may observe a large part of a population, or an entire population. It is typical in social science to treat causal features and attributes symmetrically when conducting inference about parameter estimates, and to justify inference by appealing to the idea that the data are a random sample from a larger population. We argue that this convention is not appropriate, and that the source of uncertainty for causal estimands is not purely sampling variation, but rather uncertainty arises because we do not observe all of the potential outcomes for any unit. The distinction is especially clear if we observe the entire population of interest: we may observe average income for all fifty states or all countries in the world, or we may observe all advertisers or sellers or consumers on an electronic commerce platform. When the population is observed, there is no uncertainty about the answers to questions such as, what is the average difference in income or average online purchases between coastal and interior states. On the other hand, if we attempt to estimate the effect of changing minimum wage policy or prices, we have residual uncertainty about the effect of making such a change even if we observe a randomized experiment comparing the two policies, as we do not observe any given unit under multiple policies at the same time. We propose an alternative approach to conducting inference in regression models that takes these factors into account, showing that in general conventional standard errors are conservative. More broadly, this paper highlights the theme that the theory of inference is different for causal estimates than it is for parameter estimates associated with fixed attributes of individuals.

ACM Classification

• Computing methodologies~Supervised learning by regression • Computing methodologies~Classification and regression trees • Computing methodologies~Cross-validation

Keywords

Supervised machine learning, cross-validation, causal inference, model robustness, policy evaluation, counterfactual prediction, randomized experiments, A/B tests, treatment effects.

Short Biography

Susan Athey is The Economics of Technology Professor at Stanford Graduate School of Business. She received her bachelor's degree from Duke University and her Ph.D. from Stanford, and she holds an honorary doctorate from Duke University. She previously taught at the economics departments at MIT, Stanford and Harvard. In 2007, Professor Athey received the John Bates Clark Medal, awarded by the American Economic Association to "that American economist under the age of forty who is adjudged to have made the most significant contribution to economic thought and knowledge." She was elected to the National Academy of Science in 2012 and to the American Academy of Arts and Sciences in 2008. Professor Athey's research focuses on the economics of the internet, online advertising, the news media, marketplace design, virtual currencies and the intersection of computer science, machine learning and economics. She advises governments and businesses on marketplace design and platform economics, notably serving since 2007 as a long-term consultant to Microsoft Corporation in a variety of roles, including consulting chief economist.

REFERENCES

[1] Guido Imbens and Donald Rubin. 2015. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press: Cambridge, United Kingdom.

[2] Susan Athey and Philip Haile. 2007. Nonparametric approaches to auctions. In James J. Heckman and Edward E. Leamer, eds. *Handbook of Econometrics Volume 6*, Elsevier, 3847-3965.

[3] Susan Athey and Denis Nekipelov. 2012. A Structural Model of Sponsored Search Advertising Auctions. Working paper, Stanford University. Retrieved May 30, 2015 from http://faculty-gsb.stanford.edu/athey/documents/Structural_Sponsored_Search.pdf.

[4] Susan Athey and Guido Imbens. 2015a. Machine learning methods for estimating heterogeneous causal effects. ArXiv e-print number 1504.01132. Retrieved May 30, 2015 from http://arxiv.org/abs/1504.01132.

[5] Edward Leamer. 1983. Let's take the con out of econometrics. *American Economic Review* 73, 1 (Mar. 1983), 725-736.

[6] Susan Athey and Guido Imbens. 2015b. A measure of robustness to misspecification. *American Economic Review*. 105, 5 (May 2015), 476-80. DOI=10.1257/aer.p20151020.

[7] Alberto Abadie, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2014. Finite population standard errors. NBER Working Paper Number 20325. Retrieved May 30, 2015 from http://www.nber.org/papers/w20325.