Targeted Learning for Algorithmic Fairness

Alexander Asemota
University of California Berkeley
Berkeley, California, USA
alexander.asemota@berkeley.edu

Giles Hooker
University of Pennsylvania
Philadelphia, Pennsylvania, USA
ghooker@wharton.upenn.edu

ABSTRACT

With the rise of machine learning as a decision-making tool, there have been concerns regarding bias and discrimination in predictive models. Fairness metrics, such as equal opportunity and demographic parity, have been proposed as one mechanism to address potential unfairness in algorithms. Prior work has also developed methods for inference and hypothesis testing for fairness metrics. Here, we present a novel approach to inference in algorithmic fairness using targeted learning. Notably, targeted learning broadens the types of quantities we can perform inference on, and in doing so, broadens how we explore fairness in algorithms.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

Fairness, Targeted Learning, Inference, Statistics

ACM Reference Format:
Alexander Asemota and Giles Hooker. 2018. Targeted Learning for Algorithmic Fairness. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In the past few decades, machine learning algorithms have begun to augment and automate human decision-making processes. In theory, predictive models can remove bias that is ever-present in human discretion. In practice, however, predictive models can learn to perpetuate past discrimination, as witnessed by fields ranging from criminal justice to healthcare. Concerns around bias and discrimination in algorithmic decision-making have motivated research in algorithmic fairness. Work on algorithmic fairness is largely focused on detecting, measuring, and minimizing differences in decision between groups [1].

Fairness metrics form a significant portion of research in algorithmic fairness. Since data are random, some work has also developed methods for quantifying probabilistic uncertainty for fairness metrics. Current approaches to uncertainty quantification in fairness typically assume that the model is fixed so that the only uncertainty is associated with the data themselves. Here, we propose a novel approach to quantifying uncertainty in fairness metrics using targeted learning.

Though this project is still in the works, we seek to do the following:

(1) Introduce targeted learning to the fairness literature
(2) Derive efficient influence functions for several quantities of interest
(3) Demonstrate the usefulness of targeted learning for fairness
(4) Discuss the implications of targeted learning for fairness

2 INFERENCE WITH EFFICIENT INFLUENCE FUNCTIONS

Inference, estimating population quantities from data, is a central goal of statistics. Traditional approaches to inference tend to rely on strong (and often unrealistic) assumptions about the data-generating distribution. Moreover, traditional methods lead statisticians to use models that are overly simplistic and divorced from the quantity of interest. As discussed in [3], it is common to perform inference using a model that is 'close enough' to the quantity that we care about rather than performing inference on the exact quantity. Recent advancements in non-parametric inference have enabled statistical learning without strict distributional assumptions, in particularly leveraging so-called efficient influence functions (EIFs). We introduce EIFs and targeted learning below. For a fuller introduction, see [2].

Suppose we have an estimand $\Psi(\cdot) : \mathcal{P} \mapsto \mathbb{R}$, where $P \in \mathcal{P}$ is a distribution. For example, our estimand could be the mean of a random variable X, $\Psi_1(P) = \mathbb{E}_P[X]$. When we've defined an estimand, we can easily obtain a point estimate for our quantity of interest. However, in statistical inference we care not only about the point estimate, but the likely range of possible values for our quantity of interest. With general estimands, we cannot rely on the central limit theorem or delta method for uncertainty quantification. Instead, we can analyze how sensitive our estimand is to perturbations in P. The *efficient influence function* of an estimand is a measure of the aforementioned sensitivity to perturbation. For $\Psi_1(P)$, the EIF is $\phi(X, P) = X - \Psi_i(P) = X - \mathbb{E}_P X$. In the case of a point mass $\tilde{x}$, $\phi(\tilde{x}, P) = \tilde{x} - \mathbb{E}_P X$ implies that our estimand is shifted in the direction of $\tilde{x}$. We leave the details of how to calculate the EIF for a longer tutorial. In short, we can derive the EIF in several ways, all of which are centered on the effect of distributional perturbations on our estimand.

Given an estimand and its corresponding EIF, we can define an estimator. Since we do not have access to P, we plug-in an estimate of P into our estimand. In the case of the mean, a natural plug-in is the empirical distribution, P_n, where n is the sample size. However, the empirical distribution is not suitable for more complex estimands. Fortunately, we can plug-in a wide range of estimates for P, including machine learning models. With this flexibility, we need

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

a general approach for assessing uncertainty. Analysis of the plug-in estimator shows that the estimator's EIF provides information on bias and variance. Particularly, the average value of the EIF for held-out data is the bias of the estimator, and the variance of the EIF is the variance of the estimator.

This setup provides a toolkit to perform nonparametric inference for a quantity of interest. Rather than assuming a distribution and estimating a parameter of that distribution, we define an estimand and perform inference in a data-adaptive manner. That is, instead of telling the data how they should be distributed, we let the data speak for themselves.

3 TARGETED LEARNING FOR FAIRNESS

In the context of algorithmic fairness, there are a multitude of relationships and quantities we may want to make inferences about. Most current methods treat a model as fixed and seek to infer the value of a fairness metric for the model under a given data distribution. This approach is useful particularly when a fixed model is used to make decisions using data similar to the training data. Other methods seek to discover associations in data that might lead to biased models. Overall, existing methods tend to set up inference problems that can be solved with permutation tests, bootstrapping, or optimal transport.

Targeted learning offers a new approach to inference for fairness. In particular, targeted learning allows us to perform inference on population quantities we could not perform inference on before. Instead of using whatever existing approach is 'close enough' to our quantity of interest, we can perform inference on the exact quantity we care about. Consequently, targeted learning expands how we think about and investigate bias and fairness in algorithms.

3.1 Example: Demographic Parity

To show how targeted learning can be used in algorithmic fairness, we walk through an example using demographic parity. [1] defines demographic parity as

$$P(\hat{y} = 1|g_i) = P(\hat{y} = 1|g_j)$$

where $\hat{y}$ is the prediction from a machine learning model and g_i, g_j denote group membership. This definition itself inherently conditional on the estimated model: we treat the model as fixed. To test if our model violates demographic parity, we could perform a permutation test. If group membership does not impact predictions, we would expect that randomly shuffling group labels would not affect the conditional probability of a positive prediction. However, this test does not take into account any uncertainty in the model itself. This approach is often appropriate, for example, when we just want to perform a sanity check before deploying a model. Nonetheless, it does not capture the myriad ways in which we may want to do inference around demographic parity.

Suppose we want to investigate demographic (dis)parity in the Bayes optimal decision rule $f(x) = \mathbf{1}\{P(Y = 1|X = x) > c\}$. In this setup, we care not about the outputs of a specific model, but about the decisions that are made under the distribution of $Y|X$. This could be useful, for example, when deciding whether to use a dataset for a task. In that scenario, we may have a range of potential models in mind. Training and tuning multiple models may be a waste of resources, and we may want to know something about demographic parity before beginning the modeling process. Additionally, only looking at one model does not account for the inductive bias that choosing a model entails. Simply looking at the distribution of labels across groups is insufficient, since our models will be making predictions of the form $P(Y = 1|X)$. Even if there are differences across groups, if X is independent of G, our models may not demonstrate disparity. Therefore, we turn to targeted learning to perform inference about demographic parity.

Our estimand is

$$\Psi_{DP}(P) = P(f(X) = 1|G = g_i) - P(f(X) = 1|G = g_j)$$

that is, we are interested in the difference in the Bayes optimal rule between two groups. To simplify the EIF, we split our estimand into components by group, for example, $\Psi_{g_i}(P) = P(f(X) = 1|G = g_i)$. The corresponding EIF is

$$\phi(x, P) = (r_i(x) - r_j(x)) - (\Psi_{g_i}(P) - \Psi_{g_j}(P))$$

where $r_i(x)$ is 1 if $f(x) = 1$ and 0 otherwise. For brevity, we leave out the derivation of the EIF.

Finally, we need to define as estimator for our estimand $\Psi_{DP}(P)$. For a plug-in estimator, we need to estimate P. First, we note that $\Psi_{DP}(P)$ is defined by the distributions of X, $Y|X$, and $f(X)|G$, therefore, it is sufficient to only estimate these distributions. For X, we can use the empirical distribution from our data, since the distribution is marginal. For $f(X)|G$, we can also rely on the empirical distribution, since the distribution is a 2x2 table. However, $Y|X$ is more complex. We need to take into account the dependencies that Y has with X, and without any assumptions, it is difficult to say what the distribution of $Y|X$ is. Fortunately, we can estimate $Y|X$ with a flexible nonparametric approach, such as a machine learning model. [4] advocates for the use of a 'SuperLearner', a model that combines the predictions of several models. To perform inference, we plug in our estimate for P into our estimand using held-out data; this returns a point estimate. To analyze the bias and variance of our estimate, we evaluate the EIF on held-out data as well. The mean of the EIF is the bias, and the EIF's variance is roughly the variance of our estimate. [1]

4 NEXT STEPS

In this proposal, we have described the efficient influence function for demographic parity and discussed its use. Moving forward, we plan to derive EIFs for other fairness metrics, such as equal opportunity and conditional mutual information. We will walk through the derivation for a subset of metrics to demonstrate how to derive efficient influence functions in the context of fairness. After deriving EIFs, we will empirically analyze simulated and real data using targeted learning.

REFERENCES

[1] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7, Article 166 (apr 2024), 38 pages. https://doi.org/10.1145/3616865

[2] Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2022. Demystifying Statistical Learning Based on Efficient Influence Functions. *The American Statistician* 76, 3 (Feb. 2022), 292–304. https://doi.org/10.1080/00031305.2021.2021984

[1] This is the simplest approach to targeted learning, but it is not the most rigorous. We leave more rigorous approaches (e.g. one-step estimators, TMLE) for our full paper.

[3] Philip B. Stark and Andrea Saltelli. 2018. Cargo-cult Statistics and Scientific Crisis. *Significance* 15, 4 (July 2018), 40–43. https://doi.org/10.1111/j.1740-9713.2018.01174.x

[4] Mark J. van der Laan and Richard J. C. M. Starmans. 2014. Entering the Era of Data Science: Targeted Learning and the Integration of Statistics and Computational Data Analysis. *Advances in Statistics* 2014 (Sept. 2014), 1–19. https://doi.org/10.1155/2014/502678