Algorithmic Bias in Recidivism Prediction: A Causal Perspective

Aria Khademi[1] **and Vasant Honavar**[1]
[1]College of Information Sciences and Technology
The Pennsylvania State University
khademi@psu.edu

Abstract

ProPublica's analysis of recidivism predictions produced by Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software tool for the task, has shown that the predictions were racially biased against African American defendants. We analyze the COMPAS data using a causal reformulation of the underlying algorithmic fairness problem. Specifically, we assess whether COMPAS exhibits racial bias against African American defendants using FACT, a recently introduced causality grounded measure of algorithmic fairness. We use the Neyman-Rubin potential outcomes framework for causal inference from observational data to estimate FACT from COMPAS data. Our analysis offers strong evidence that COMPAS exhibits racial bias against African American defendants. We further show that the FACT estimates from COMPAS data are robust in the presence of unmeasured confounding.

Introduction

There is growing concern that AI technologies can perpetuate or amplify undesirable bias or discrimination based on race, gender, and other protected social attributes. An example is the COMPAS software used by the United States Judiciary to predict the likelihood of recidvism for defendants based on their characteristics and past criminal record. ProPublica's analysis of the COMPAS tool (Angwin et al. 2016) spurred extensive debate on whether the software was biased against African American defendants.

There have been many attempts to formalize various notions of algorithmic fairness (Barocas, Hardt, and Narayanan 2019). Of particular interest are notions of fairness that require that individuals do *not* experience differences in outcomes (e.g., recidivism score) *caused by* factors that are outside their control (e.g., race). Recent work has shown that tests of fairness expressed solely using the joint distribution (Hardt, Price, and Srebro 2016) of the observed variables are incapable of detecting unfairness. Hence, there is a growing interest in algorithmic fairness criteria that *causally* link protected attributes with the outputs (e.g., decisions, predictions) of the algorithm (Barocas, Hardt, and Narayanan 2019; Khademi et al. 2019). The key intuition behind such fairness criteria is that the question "Is the decision discriminatory with respect to a protected attribute?" can be reframed as: "Does the protected attribute have a causal effect on the decision?" Answering such a question is complicated by the fact that these factors can be meaningfully related to other characteristics that may be relevant in determining what is fair, and requires careful application of state-of-the-art tools for estimating causal effects from observational data.

We assess whether COMPAS exhibits racial bias against African American defendants using FACT, a recently introduced explicitly causal measure of algorithmic fairness (Khademi et al. 2019), using the Neyman-Rubin potential outcomes framework (Rubin 2005). Our analysis offers robust evidence that COMPAS exhibits racial bias against African American defendants.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Methods

Denote each individual i with $(\tilde{X}_i, A_i, Y_i)$ where $\tilde{X}$ is the vector of non-protected attributes, $A \in \{a, a'\}$ is race, and Y is the likelihood that COMPAS would predict recidivism ($Y = 1$) or non-recidivism ($Y = 0$). Let $Y_i^{(a)}$ be the *potential outcome* of individual i, if they had race a. For each individual, either $Y_i^{(a)}$ or $Y_i^{(a')}$ is observable. We use a causal notion of fairness, namely, fair in average causal effect on the treated (FACT) (Khademi et al. 2019): A decision function $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ is fair on average over individuals sharing a certain race if $\mathbb{E}[Y_i^{(a)} - Y_i^{(a')} \mid A_i = a] = 0$.

We estimate FACT using the state-of-the-art matching based methods for causal inference (Stuart 2010), i.e., for each African American defendant (we observe $Y_i^{(a)}$), we find their most similar "match" in terms of non-protected attributes among White defendants (and hence estimate $Y_i^{(a')}$). We use the following matching methods within the R package MatchIt (version 3.0.2) (Ho et al. 2011): **(i)** Nearest neighbor matching (NNM), **(ii)** Nearest neighbor matching with propensity caliper (NNMPC), **(iii)** Mahalanobis metric matching with propensity caliper (MMMPC), and **(iv)** Full matching (FM), all according to the parameters specified in (Khademi et al. 2019).

Table 1: Results of matching on the COMPAS data. Estimate of FACT is denoted by $\hat{\gamma}$. Statistical significance level is $\alpha = 0.05$.

COMPAS dataset						
Matching method	# of Treated Matches	# of Control Matches	$\overline{D}^m_{a,a'}$	$\hat{\gamma}$	Standard Error	P-value
NNM	1893	780	0.0002	0.734	0.258	0.004
NNMPC	1893	910	0.0123	0.251	0.222	0.257
MMMPC	1893	852	0.0073	0.331	0.292	0.253
FM	1893	1447	0.0002	0.624	0.223	0.005

To measure goodness-of-matches, we examined **(i)** absolute value of standardized difference in means of the treated (race a) and controlled (race a') in terms of the distance measure (propensity score), before ($\overline{D}_{a,a'}$) and after ($\overline{D}^m_{a,a'}$) matching, and **(ii)** jitter plots and histograms of the distribution of propensity scores after matching. For high quality matches, $\overline{D}^m_{a,a'}$ must be close to 0. As a result of the matching process, each individual is assigned a weight. Subsequently, we run the weighted regression $\mathbb{E}[Y^{(A)}] = \delta + \gamma A + \tilde{\theta}^\top \tilde{X}$ *on the matched data set* (having dropped the data points for which no match is found) and obtain $\hat{\gamma}$ as the estimated causal effect of A on Y measured by FACT.

In the absence of unmeasured confounding, estimates of FACT are doubly robust if either the matching model or the subsequent regression model are correct (Ho et al. 2011). To test for the effect of unmeasured confounding on our estimates of FACT, we run sensitivity analysis (SA) with the R package *rbounds* (version 2.1) (Keele 2010). We expose our estimates to a Γ factor of unmeasured confounding and measure the change in significance of estimates (see (Khademi et al. 2019; Rosenbaum 2005)).

Experiments

Data

The COMPAS data offer 2 years of data (2013-2014) from the COMPAS software tool. The question is whether COMPAS predicts different rates of recidivism for African Americans compared to Whites (all other things being equal). We designated African Americans as treated ($A = 1$) and Whites as control ($A = 0$). The binary outcome Y is the COMPAS prediction ($Y = 1$ indicating recidivism). We used the "Violent" data pre-processed using the procedure used by ProPublica yielding 3373 data points.[1]

Fairness Analysis Using FACT

We estimated the causal effect of race on COMPAS outcome using the techniques described in Section Methods. FM yielded the highest number of matched data points with the lowest $\overline{D}^m_{a,a'}$ (see Table 1) and hence highest quality of matches (histograms and jitter plots not shown).

The FACT estimates are summarized in Table 1. We were able to reject the null hypothesis $H_0 : \gamma = 0$ (in the case of NNM and FM) which suggests that the recidivism scores predicted by COMPAS exhibit racial bias against African Americans. We speculate that the propensity caliper in NNMPC and MMMPC disregards some data points that are important in rejecting H_0. In the case of FM, odds of the COMPAS software predicting that African American defendants would recidivate after release is $\exp(0.624) \approx 1.87$ times that of White defendants. This result is in agreement with previous work, e.g., (Angwin et al. 2016).

Robustness to Unmeasured Confounders

We ran SA with Γ ranging from 1 to 10. The larger Γ, the bigger the exposure to unmeasured confounders. Our estimates of NNM, NNMPC, MMMPC, and FM were robust to unmeasured confounding up to Γs of 9, 7.5, 8, and 5.5, respectively. We conclude that our FACT estimates are robust to unmeasured confounders.

[1] https://github.com/propublica/compas-analysis

References

Angwin, J.; Larson, J.; Surya, M.; and Kirchner, L. 2016. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)* 9.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

Ho, D. E.; Imai, K.; King, G.; and Stuart, E. A. 2011. Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* 42(8):1–28.

Keele, L. 2010. An overview of rbounds: An r package for rosenbaum bounds sensitivity analysis with matched data. *White Paper. Columbus, OH* 1–15.

Khademi, A.; Lee, S.; Foley, D.; and Honavar, V. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, 2907–2914. ACM.

Rosenbaum, P. R. 2005. Sensitivity analysis in observational studies. *Encyclopedia of Statistics in Behavioral Science* 4:1809–1814.

Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100(469):322–331.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science: a review journal of the Institute of Mathematical Statistics* 25(1):1.