Towards a Non-Discriminatory Algorithm in Selected Data

David Arnold Will Dobbie Peter Hull

Algorithmic predictions increasingly guide many high-stakes decisions, contributing to growing concerns about algorithmic "fairness" or discrimination against legally protected groups. A common concern is that algorithmic predictions may systematically differ for individuals who are equally "qualified" for a treatment but are of (for example) different races—even if race is not used to train the algorithm. But addressing this concern is often complicated by a fundamental selection challenge: an individual's qualification is only revealed among those deemed appropriate for treatment by an existing algorithm or human decision-maker. This problem of "selective labels" [4, 5] can make it difficult to both measure discrimination and impose qualification-based fairness constraints when building algorithms. [3]

We develop new tools to overcome this selection challenge and both measure and reduce algorithmic discrimination in settings where decision-makers are as-good-as-randomly assigned. We show how such quasi-experimental assignment can be used to estimate a set of moments governing the relationship between the algorithmic inputs and individual qualification. We then show how these moments reveal algorithmic discrimination, and how the moment estimates can be used to build new less discriminatory algorithms which optimally predict qualification subject to a chosen fairness constraint.

We illustrate these new tools in the context of NYC pretrial decisions, leveraging the as-good-as-random assignment of bail judges to defendants. [1, 2] We show that building pretrial misconduct "risk scores" on selected data both reduces predictiveness and exacerbates algorithmic racial discrimination. We characterize the tradeoff between predictiveness and fairness after correcting for selective labels with our approach, and discuss extensions to alternative fairness constraints and other real-world considerations.

References

[1] D. Arnold, W. Dobbie, and P. Hull. Measuring racial discrimination in bail decisions. *NBER Working Paper No. 26999*, 2020.

[2] D. Arnold, W. Dobbie, and C. S. Yang. Racial bias in bail decisions. *Quarterly Journal of Economics*, 133(4):1885–1932, 2018.

[3] A. Coston, A. Mishler, E. H. Kennedy, and A. Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 2020.

[4] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *Quarterly Journal of Economics*, 133(1):237–293, 2018.

[5] H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.