Introducing ‘Algorithimic decision making and the cost of fairness’ and Accompanying Normative Considerations

Jacob Thoma

2024-03-19

Introduction

If birds needed a seasonal calendar to know when to migrate, they would be dead. Instead, migratory bids use changes in temperature, winds, precipitation, and day length as proxies for seasonal change. As such, migratory birds have developed a basic algorithm that informs migration decisions crucial for survival. For birds and humans alike, seeing the whole picture is not always possible at the time a decision is required. Advances in machine learning have facilitated construction of powerful algorithms designed to discern differences across data points with limited information and considerable accuracy. But with so many applications of algorithmic decision making and so much at stake, fairness in how these algorithms make conclusions must be considered alongside accuracy. The research paper “Algorithmic decision making and the cost of fairness”, authored by Corbett-Davies, Pierson, Feller, Goel, and Huq, makes use of formal proofs to explore an accuracy vs safety debate surrounding optimization algorithms (such as COMPAS), and is linked here. This introductory writing aims to summarize the methodology and identify normative considerations of the above paper to create a solid foundation for further mathematical and philosophical analysis.

Approach to Formal Proofs (Methodology Part 1)

The authors approach the topic of fairness vs. accuracy by mathematically defining fairness with three metrics: *statistical parity*, *conditional statistical parity* and *predictive equality*. *Statistical parity* is defined with the following notation: $E[d(X) \mid g(X)] = E[d(X)]$ for a decision rule $d(X)$ and indicator variable $g(X)$ of an individual’s attributes X. As race is not directly considered within the COMPAS algorithm, $g(X)$ is inferred from proxies of race in the COMPAS data set, such as zip code [1]. In the context of this research, statistical parity would be achieved if the expected probability of the COMPAS algorithm classifying someone as a repeat offender does not change based on knowing their inferred race. *Conditional statistical parity*, notated $E[d(X) \mid l(x), g(X)] = E[d(X)]$, is an extension of statistical parity that also controls for a set of “legitimate” risk factors $l(x)$ (such as prior offenses) that could influence $d(x)$ [1]. *Predictive equality* is measured by false positive rate (FPR) and notated: $E[d(X) \mid Y = 0, g(X)] = E[d(X) \mid Y = 0]$, where Y is the true class of the individual with attributes x. As such, predictive equality in the COMPAS use case involves comparable false positive rates of COMPAS classification with and without the racial inference $g(x)$.

With definitions for fairness intact, the researchers developed mathematical framework for constrained and unconstrained optimization problems. The goal for both optimization problems was to maximize *immediate utility* u(d,c), expressed $E[d(X)(p_{Y|X} - c)]$ where c is the cost of detaining an individual in terms of crimes prevented between 0 and 1 and $p_{Y|X}$ is the probability of recidivism for a given X. Under this definition, u(d,c) can be thought of as a balance between the expected number of crimes prevented and the expected number of people detained [1]. To maximize u(d,c) in this case is to find an optimal decision rule, d*(x) that allows for optimal classification of both repeat offenders and non-repeat offenders. The constrained optimization of u(d,c) is based on the three restrictions of statistical fairness discussed earlier, while the unconstrained optimization lacks these restrictions.

Formal Proof Results and Applications (Methodology Part 2)

In unconstrained optimization, the optimal decision d* is to develop optimal *decision thresholds* for d(x) on any information in the data using all observable attributes of X, including inferred race. Two different mathematical proofs are used to evaluate constrained optimization, finding that the optimal decision d* can guarantee statistical fairness with varying decision thresholds across different inferred values for race [1]. The first proof verifies that both types of statistical parity (conditional and general) are satisfied under the constrained d* by choosing decision thresholds that ensure equal proportions p* of detainment across groups. As such, every group g(x) would have the p* proportion riskiest defendants detained. The second proof argues that predictive equality of an optimal constrained algorithm is only possible with varying decision thresholds across l(x) and g(x). This is done by considering that for an optimal rule d with no decision thresholds, there exists a more optimal rule d' that simultaneously uses thresholds to maintain the same predictive equality between d and d' but also improve u(d,c) by detaining additional offenders through better true positive rate [1]. It then logically follows that if FPR is the same between d and d', that d' does a better job of correctly classifying people that should be detained since $u(c,d) < u(c,d')$. In essence, the authors conclude that the optimal constrained and unconstrained algorithms require decisions thresholds across inferred race.

This optimization theory was applied directly to a subset of the COMPAS data from Broward County, Florida. Four optimization frameworks were constructed, one that included each fairness definition as a constraint one that does not consider any of the fairness conditions. The researchers found that each decision algorithm had different optimal decision thresholds with varying implications. Compared to the unconstrained algorithm, the constrained algorithms detained more low-risk offenders in order to satisfy statistical fairness conditions, leading to estimated increases in violent crime. Ensuring statistical parity, predictive equality, or conditional statistical parity alone entailed estimated increases in violent crime of 9%, 7%, and 4% respectively when compared to the optimal unconstrained algorithm [1]. In addition, the percentage of detainees in the unconstrained algorithms that are low risk is 17%, 14%, and 10% respectively [1].

Normative Considerations

The formal proofs in “Algorithmic decision making and the cost of fairness” bring up several normative considerations, the most consequential being the idea that different decision thresholds are necessary to optimize any of the algorithms mentioned above. It is expected that an algorithm unconstrained on fairness has decision thresholds based on any variable in the data. However, the optimal constrained algorithms guarantee “statistical fairness” by deliberately using different decision thresholds across inferred race. If these fairness constraints are only satisfied in the optimal algorithms by treating different races differently, then the idea of an optimal algorithm with “fairness constraints” should be questioned from the angles discussed below.

As stated in the name “Algorithmic decision making and the cost of fairness”, the authors approach the disturbing results of their formal proofs by discussing trade offs between favoring statistical fairness or classification accuracy in an optimized decision algorithm. The abstractness of comparing optimal decision algorithms should not take away from the clear-cut consequences of mis-classification. In the case of predicting recidivism, false positive is a leper, shunned away from a better chance at life for crimes they will not commit. A false negative is a threat to society from the second they are released to the second they are caught for their recidivism. A 4% increase in violent crime to justify conditional statistical parity is just a number to some, but to the people affected it means everything. With all of these different viewpoints, the importance of correct classification cannot be overstated. However, should this be prioritized at the expense of violating statistical fairness?

An even broader clash between the idea of fairness and statistical fairness follows from the proofs conducted. The idea that statistical fairness can be achieved in algorithms through considering factors out of one’s control is as alarming as it is confusing. When an optimized decision algorithm is only possible with varying decision thresholds for different races, statistical fairness is found to be on the wrong side of fairness. Statisticians can tinker away with their optimization algorithms and decision thresholds all they like, but peoples lives change when these decisions are made. Even the most brilliant statistician could not look a white offender in the eye and say, “Our algorithm thinks you and [black cellmate] have the same probability of recidivism, but we are keeping you in jail instead of him so our algorithm can be statistically fair” without questioning the idea of fairness. How do we balance statistical fairness in relation to the foundational idea of fairness?

With various mathematical and philosophical components, the unanswered questions left from “Algorithmic Decision Making and the Cost of Fairness” are intimidating. But when the implications of recidivism risk algorithms impact human lives, hard questions have to be asked.

References

[1] Corbett-Davies, Sam, et al. “Algorithmic Decision Making and the Cost of Fairness.” In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM, 2017. https://doi.org/10.1145/3097983.3098095.