Algorithmic Fairness in Criminal Justice

Marcello Di Bello

Predictive algorithms

They assign risk scores to individuals in order to predict a behavior or condition, such as recidivism or being a crime victim

They rely on machine learning methods to identify correlations between risk factors, such as prior convictions, and crime

Examples: Chicago SSL, COMPAS, PSA[1]

The good: They can end the bail system that disproportionally targets the poor (see Criminal Justice Reform in New Jersey)

The bad: They may exacerbate existing racial inequities in society[2]

Northpointe/ProPublica Debate

ProPublica's 2016 analysis[3] of COMPAS showed that

> *False positives (FP):* 23.5% of whites who didn't reoffend were misclassified as 'high risk' (score ≥ 5) versus 44.9% of blacks.
> *False negatives (FN)*: 47.7% of whites who reoffended were misclassified as 'low risk' (score < 5) versus 28% of blacks.[4]

ProPublica singled out the group of non-reoffenders and compared the percentage of whites in that group misclassified as high risk (FP) to the percentage of blacks in the same group also misclassified as high risk (also FP). It also singled out the group of reoffenders and compared the percentage of whites in that group misclassified as low risk (FN) to the percentage of blacks in the same group also misclassified as low risk (also FN).

Equality along this dimension is called classification fairness

Northpointe, the company that designed COMPAS, responded

> *Wrong positive prediction*: Among those labeled 'high risk,' 41% of whites and 37% of blacks did not reoffend
> *Wrong negative prediction*: Among those labeled 'low risk,' 29% of whites and 35% of blacks reoffended

Northpointe singled out the group of those labeled 'high risk' by COMPAS and compared the percentage of whites in this group who are non-reoffenders to the percentage of blacks in the same group who are non-reoffenders. It also singled out the group of those labeled 'low risk' by COMPAS and compared the percentage of whites in this groups who are reoffenders to the percentage of blacks in the same group who are reoffenders.

Equality along this dimension is called prediction fairness

[1] *www.psapretrial.org*

[2] Data about US criminal justice system: *www.prisonpolicy.org/reports/pie2019.html*

[3] *www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*

[4] Rearrest was used as a proxy for actual recidivism

Is COMPAS fair or not towards blacks v. whites?

▷ Racial disparities in *classification errors* (FPs and FNs) are huge, but racial disparities in *prediction errors* are not significant

♠ COMPAS satisfies prediction fairness, not classification fairness

♣ No algorithms can satisfies both conception of fairness under realistic conditions (see Chouldechova's *impossibility theorem*)

QUESTION: Which of the two conceptions of fairness should we pick?

Mayson v Hellman v Huq

Mayson and Hellman believe that fairness requires to treat *similarly situated* individuals the same, but they disagree on what this means

MAYSON: Against classification fairness[5]

> The question of what makes two people (or groups) relevantly "alike" for purposes of a particular action is really a question about the permissible grounds for that action. To judge that two people with equivalent skill and experience are relevantly "alike" for purposes of a hiring decision is to judge that skill and experience are good grounds on which to make such a decision (p. 2273).
>
> [Ultimate outcomes] cannot be the basis for risk assessment because at the time of assessment they are unknown. This is why we resort to risk assessment in the first place (p. 2275).
>
> The demand for equal algorithmic treatment for same-outcome groups amounts to a judgment that outcomes are the appropriate basis for prediction. And that judgment is nonsensical (p. 2275).

HELLMAN: In favor of classification fairness[6]

> *Fair testing analogy*: Two students are similarly situated when they are equally prepared. A fair test should treat equally prepared students the same. Likewise, a fair algorithm should treat reoffenders the same and should treat non-reoffenders the same.

HUQ: Neither prediction fairness nor classification fairness[7]

> The key question for racial equity is whether the costs that an algorithmically driven policy imposes upon a minority group outweigh the benefits accruing to that group (p. 1111).
>
> The spillover costs of coercion of minority individuals for the minority group will be greater on a per capita basis than the costs of coercing majority group members (p. 1113).
>
> There is no particular reason to believe that any of these spillover costs are less if the person subject to the coercion is in fact a true rather than false positive (pp. 1127)
>
> Accounting for both the immediate and spillover costs of crime control . . . conduces to a bifurcated risk threshold—one rule for the majority, and one for minority (p. 1131).

Huq is thinking about maximizing expected utility

[5] Sandra Mayson, 'Bias In, Bias Out,' *Yale Law Journal*, 2019, 128: 2218-2300,

[6] Deborah Hellman, 'Measuring Algorithmic Fairness,' *Virginia Law Review*, forthcoming

[7] Aziz Huq, 'Racial Equity in Algorithmic Criminal Justice,' *Duke Law Journal*, 2019, 68: 1043-1134.