Principal Fairness for Human and Algorithmic Decision-Making

Kosuke Imai

Harvard University

DeMo Meeting
Research Section, Royal Statistical Society
February 8, 2022

Joint work with Zhichao Jiang (Univeristy of Massachusetts, Amherst)

Fair Decision-Making

- What is a fair decision?
- How should we assess the fairness of decision?
- How should we improve the fairness of decision-making from data?
- Examples: courts, medicine, admissions, lending, insurance, hiring, ...
- Fair decision-making in public policies
- Literature on algorithmic fairness

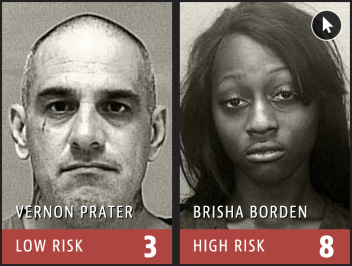
- Imai, K. and Jiang, Z. (2020). "Principal fairness for human and algorithmic decision-making." arXiv preprint, `https://arxiv.org/pdf/2005.10400.pdf`

Statistical Fairness Criteria

- Developed for assessing the fairness of prediction algorithms
- But also used for assessing the fairness of algorithmic/human decision
- Setup:
  - outcome: Y
  - prediction or decision: D
  - protected attribute (e.g., race, gender): A
- 3 Statistical fairness criteria:
  1. Equal decision: $D \perp\!\!\!\perp A$
     $\Pr(D = 1 \mid A = a) = \Pr(D = 1 \mid A = a')$
  2. Equal accuracy: $D \perp\!\!\!\perp A \mid Y$
     $\Pr(D = 1 \mid Y = 1, A = a) = \Pr(D = 1 \mid Y = 1, A = a')$
     $\Pr(D = 0 \mid Y = 0, A = a) = \Pr(D = 0 \mid Y = 0, A = a')$
  3. Equal calibration: $Y \perp\!\!\!\perp A \mid D$
     $\Pr(Y = 1 \mid D = 1, A = a) = \Pr(Y = 1 \mid D = 1, A = a')$
     $\Pr(Y = 0 \mid D = 0, A = a) = \Pr(Y = 0 \mid D = 0, A = a')$

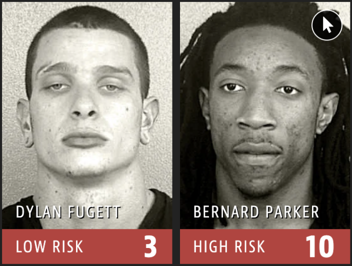
The COMPAS Debate (Correctional Offender Management Profiling for Alternative Sanctions)

Two Petty Theft Arrests



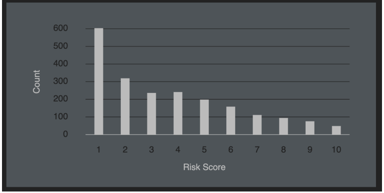
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests

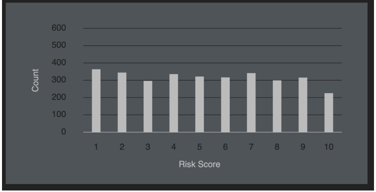


Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

White Defendants' Risk Scores



Black Defendants' Risk Scores



Impossibility Results

- Propublica: false positive rate is higher for blacks
  $\Pr(\text{risky} \mid \text{not rearrested}, \text{black}) \gg \Pr(\text{risky} \mid \text{not rearrested}, \text{white})$
- Northpointe: calibration is equal
  $\Pr(\text{rearrested} \mid \text{risky}, \text{black}) \approx \Pr(\text{rearrested} \mid \text{risky}, \text{white})$

- It is impossible to satisfy both criteria unless:
  - recidivism rate and score distribution are identical across racial groups
  - or, some racial groups never experience recidivism

- In general, we cannot satisfy all three statistical fairness criteria
  - If equal decision ($D \perp\!\!\!\perp A$) and equal accuracy ($D \perp\!\!\!\perp A \mid Y$) hold, then either the base rate is equal ($Y \perp\!\!\!\perp A$) or the decision is useless ($D \perp\!\!\!\perp Y$)
  - If equal decision ($D \perp\!\!\!\perp A$) and equal calibration ($Y \perp\!\!\!\perp A \mid D$) hold, then the base rate has to be equal ($Y \perp\!\!\!\perp A$)

Principal Fairness: Taking Causality into Account

- The statistical fairness criteria ignore the fact that the decision may affect the outcome
  1. observed data are contaminated (related to selective labels problem)
  2. fairness should address how individuals are affected by the decision

- Causality framework:
  - potential outcomes: $Y(1)$ and $Y(0)$
  - causal effect: $Y(1) - Y(0)$
  - fundamental problem of causal inference
  - different from the observed outcome: $Y = Y(D)$
  - potential outcomes are pre-treatment characteristics
  - principal strata: $R = (Y(1), Y(0)) = (y_1, y_0)$

- Principal fairness: individuals who are similarly affected by the decision should be treated similarly

$$D \perp\!\!\!\perp A \mid R$$

An Illustrative Example

Group A		$Y(0) = 1$	$Y(0) = 0$
		Dangerous	Backlash
$Y(1) = 1$	Detained $(D = 1)$	120	30
	Released $(D = 0)$	30	30
		Preventable	Safe
$Y(1) = 0$	Detained $(D = 1)$	70	30
	Released $(D = 0)$	70	120
Group B		$Y(0) = 1$	$Y(0) = 0$
		Dangerous	Backlash
$Y(1) = 1$	Detained $(D = 1)$	80	20
	Released $(D = 0)$	20	20
		Preventable	Safe
$Y(1) = 0$	Detained $(D = 1)$	80	40
	Released $(D = 0)$	80	160

- Detention rate within each principal strata is identical for Groups A&B
  - "Dangerous" group $(y_0 = 1, y_1 = 1)$: 80%
  - "Safe" group $(y_0 = 0, y_1 = 0)$: 20%
  - "Preventable" group $(y_0 = 1, y_1 = 0)$: 50%
  - "Backlash" group $(y_0 = 0, y_1 = 1)$: 50%

The Example Does Not Satisfy Statistical Fairness

	Group A		Group B	
	Detained	Released	Detained	Released
$Y = 1$	150	100	100	100
$Y = 0$	100	150	120	180

- Equal decision
  - Group A: 50%
  - Group B: 44%
- Equal accuracy
  - Group A: 60% ($Y = 1$), 60% ($Y = 0$)
  - Group B: 50% ($Y = 1$), 40% ($Y = 0$)
- Equal calibration
  - Group A: 60% ($D = 1$), 60% ($D = 0$)
  - Group B: 45% ($D = 1$), 64% ($D = 0$)

Relations between Principal Fairness and Statistical Fairness

Theorem 1

If $A \perp\!\!\!\perp R$ holds, principal fairness implies all three statistical fairness criteria

Assumption 1 (Monotonicity)

$Y(1) \leq Y(0)$

Theorem 2

If $A \perp\!\!\!\perp R$ and monotonicity hold, principal fairness is equivalent to the three statistical fairness criteria

- $A \perp\!\!\!\perp R$ is the equal base rate condition with potential outcomes
- The results hold conditional on covariates
- Monotonicity assumption eliminates the "Backlash" group in our example

Other Causality-based Fairness Criteria

1. Counterfactual equalized odds criterion
   - condition on a "natural baseline":
     $\Pr(D \mid Y(0), A = a) = \Pr(D \mid Y(0), A = a')$
   - does not account for the impact of decision
   - if $Y(1) \perp\!\!\!\perp A \mid Y(0)$, principal fairness implies this criterion
   - a special case: $Y(1)$ is constant across groups
   - principal fairness as a generalization

2. Counterfactual fairness
   - protected attribute as a causal variable: $D(a)$, $D = D(A)$
   - fairness criteria: $\Pr(D(a) = 1) = \Pr(D(a') = 1)$
   - no causation without manipulation
   - decision rule that does not depend on the protected attribute satisfies counterfactual fairness but can fail to meet principal fairness

Empirical Evaluation and Policy Learning

- Difficulty: principal strata are unobserved
- The approach taken in our JRSSA discussion paper

Assumption 2 (Unconfoundedness)

$Y(d) \perp\!\!\!\perp D \mid \mathrm{X}$ *for any* d *where* X *is the decision variables*

- Plausible if the decision variables are known (e.g., algorithmic decision)
- Under monotonicity and unconfoundedness, we can
  - identify principal score: $e_r(\mathrm{X}, A) = \Pr(R = r \mid \mathrm{X}, A)$
  - evaluate principal fairness by computing $\Pr(D = 1 \mid R, A)$
- Policy learning:
  - decision rule: $D = \delta(\mathrm{X})$
  - $\Pr(\delta(\mathrm{X}) = 1 \mid R = r, A) \; = \; \mathbb{E}\Big[\underbrace{\frac{e_r(\mathrm{X}, A)}{\mathbb{E}\{e_r(\mathrm{X}, A) \mid A\}}}_{\text{weight}} \delta(\mathrm{X}) \mid A\Big]$
  - optimal policy subject to the fairness constraint

Concluding Remarks

- Fairness of human and algorithmic decision-making needs to be placed in the causal inference framework
- We must consider how the decision affects individuals
- Important extensions:
  - algorithm-assisted human decision making (our JRSSA paper)
  - evaluation and policy learning in real world applications
  - selection biases, and dynamic systems