Fairness Interventions as (Dis)Incentives for Strategic Manipulation

Xueru Zhang[1], Mohammad Mahdi Khalili[2], Kun Jin[3], Parinaz Naghizadeh[1], Mingyan Liu[3]

1 The Ohio State University 2 yahoo! 3 University of Michigan

Background

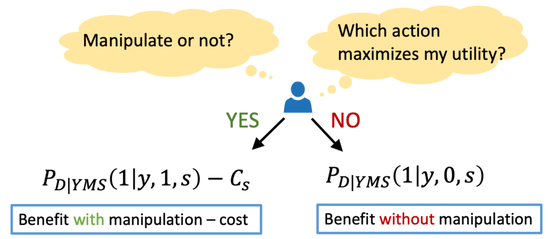
- ML has been increasingly used to help make decisions about people
  - lending, hiring, college admission, $\cdots$
- Two challenges:
  - ML is vulnerable to strategic manipulation
  - ML can be biased against certain social groups

Model: Strategic Interaction

Two demographic groups $\mathcal{G}_a$, $\mathcal{G}_b$

- Sensitive attribute $S \in \{a, b\}$
- Features $X \in \mathbb{R}^d$
  - ***feature generation*** $P_{X|YS}(x|y,s)$
- Qualification state $Y \in \{0,1\}$
  - ***qualification rate*** $\alpha_s = P_{Y|S}(1|s)$
- Decision $D \in \{0,1\}$
  - ***policy*** $\pi_s(x) = P_{D|XS}(1|x,s)$
- Manipulation action $M \in \{0,1\}$
  - Manipulation doesn't affect Y but results in better feature distribution
  - Manipulation cost $C_s \geq 0$

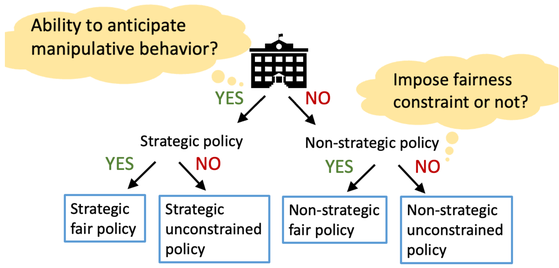
Individual best response



- Individual chooses to manipulate at cost if manipulation brings the higher utility
- Manipulation probability:

$$\Pr(C_s \leq P_{D|YMS}(1|y,1,s) - P_{D|YMS}(1|y,0,s))$$

Decision-maker's optimal (fair) policies



$$\max_{\pi_a, \pi_b} \quad \mathbb{E}[R(D,Y)]$$

s.t. fairness constraint

Existing Work

Fair machine learning

$$\min \quad Loss \qquad s.t., \quad \phi(\text{👤}) \approx \phi(\text{👤})$$

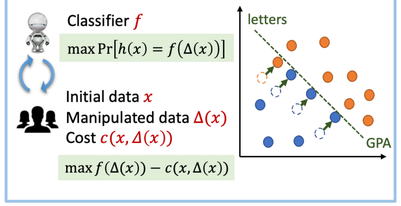
Fairness constraint

- **Demographic parity**: equal positive rate
- **Equal Opportunity**: equal true positive rate

Strategic classification

- Stackelberg game formulation

Hardt et al., 2016a; Dong et al. 2018; Milli et al., 2019; Hu et al., 2019; Braverman & Garg, 2020



- Most works studied these two problems separately
- Existing Stackelberg game formulation assumes:
  - Manipulation outcome is deterministic & known
  - Manipulation cost is a deterministic function of features before & after manipulation

Theoretical Results

- Characterize the equilibrium strategies of individuals & decision-maker (four types of policies)
- Impact of decision-maker's anticipation of strategic manipulation
  - Strategic policy over(under) accepts majority-qualified(majority-unqualified) group
  - Anticipation of manipulation can **worsen** the fairness of a strategic policy when one group is majority-qualified while the other is majority-unqualified
  - When both groups are majority-unqualified, a strategic policy may **mitigate** unfairness and even **flip the disadvantaged group**
- Impact of fairness interventions on policies and individuals' manipulation
  - Conditions under which non-strategic decision maker may **benefit from fairness constraints**
  - Conditions under which fairness constraints serve as **(dis)incentives** for strategic manipulation

This Work

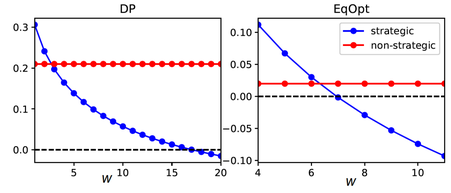
Motivating Example: students cheat on exams to get admitted to college

- A new Stackelberg game formulation:
  - Uncertain manipulation outcomes
  - Manipulation cost is determined before observing manipulation outcomes
- Understand the impacts strategic manipulation and fairness intervention have on each other

Strategic Manipulation ⇄ **Fairness Constraint**

Experiments: FICO Credit Score

- ***Hispanic & Black***: strategic policy mitigates unfairness



- ***White & Asian***: non-strategic fair policy has higher utilities

δ_u	C_a	$U_a(\widehat{\theta}_a^{\text{UN}})$	$U_a(\widehat{\theta}_a^{\mathcal{C}})$	$U_b(\widehat{\theta}_b^{\text{UN}})$	$U_b(\widehat{\theta}_b^{\mathcal{C}})$
0.8	Beta(10, 10)	-0.190	-0.189	0.024	0.034
0.756	Beta(10, 1)	0.396	0.397	0.181	0.201

- ***Black & $\mathcal{G}_a$***: strategic policy worsens unfairness

	$\mathcal{G}_a$	strategic $C_a = C_b$	strategic $C_a \neq C_b$	non-strategic
EqOpt	Caucasian	0.355	0.556	0.136
	Hispanic	0.292	0.493	0.034
	Asian	0.333	0.533	0.123
DP	Caucasian	0.611	0.680	0.449
	Hispanic	0.421	0.490	0.242
	Asian	0.634	0.703	0.522

- ***More results***: