Building a Robot Judge: Data Science for Decision-Making

11. Algorithms and Decisions II

Weekly Q&A

https://padlet.com/eash44/fl7y3ims5r5pzkli

Week 10 End-of-Class Activity

`https://padlet.com/eash44/u4rfjxc8mbgjd587`

Under what conditions are predictions sufficient for...
... optimal allocation of inspectors:
(1) Benefits of fixing problems are mostly homogeneous.
(2) Establishments do not change behavior in response to the algorithm.
(3) Inspectors respond predictably to the algorithm.

Week 10 End-of-Class Activity

https://padlet.com/eash44/u4rfjxc8mbgjd587

Under what conditions are predictions sufficient for...

... optimal allocation of inspectors:

(1) Benefits of fixing problems are mostly homogeneous.

(2) Establishments do not change behavior in response to the algorithm.

(3) Inspectors respond predictably to the algorithm.

... making pre-trial release decisions:

When are recividism risk scores sufficient for bail/release decisions?

(1) Benefits of detention are mostly homogeneous.

- why not?

When are recividism risk scores sufficient for bail/release decisions?

(1) Benefits of detention are mostly homogeneous.
- why not?

(2) Potential offenders do not change behavior in response to the algorithm.
- why not?

When are recividism risk scores sufficient for bail/release decisions?

(1) Benefits of detention are mostly homogeneous.
- why not?

(2) Potential offenders do not change behavior in response to the algorithm.
- why not?

(3) Judge respond predictably to the algorithm.
- why not?

Alternative: Doctor's testing decision

Mullainathan and Obermeyer (2019)

- Consider the problem of a doctor deciding whether to order a test for a heart blockage.
  - if blockage is detected, useful treatment can be given
  - if no blockage, then test was wasted (test is costly to administer)

Alternative: Doctor's testing decision

Mullainathan and Obermeyer (2019)

- Consider the problem of a doctor deciding whether to order a test for a heart blockage.
  - if blockage is detected, useful treatment can be given
  - if no blockage, then test was wasted (test is costly to administer)
- Optimal testing strategy:
  - form predicted prior probability of a positive test $\hat{Y}(X_i)$
  - test all i with predicted prior probability above some threshold $\bar{Y}$.

When are test result priors sufficient for testing decisions?

(1) Benefits of _____ are mostly homogeneous.

- why not?

When are test result priors sufficient for testing decisions?

(1) Benefits of _____ are mostly homogeneous.

- why not?

(2) _____ do not change behavior in response to the algorithm.

- why not?

When are test result priors sufficient for testing decisions?

(1) Benefits of _____ are mostly homogeneous.

- why not?

(2) _____ do not change behavior in response to the algorithm.

- why not?

(3) _____ responds predictably to the algorithm.

- why not?

When are test result priors sufficient for testing decisions?

(1) Benefits of _____ are mostly homogeneous.

- why not?

(2) _____ do not change behavior in response to the algorithm.

- why not?

(3) _____ responds predictably to the algorithm.

- why not?

Note: Given (1) through (3), the doctor testing decision is a **prediction problem**.

Generalizing these points

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?
  - $D^* = \max_D u(D, Y, X)$

Generalizing these points

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?
  - $D^* = \max_D u(D, Y, X)$

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
   - $u(D, Y, X) = u(D, Y)$, and hence $D^*(Y, X) = D^*(Y)$

Generalizing these points

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?
  - $D^* = \max_D u(D, Y, X)$

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
   - $u(D, Y, X) = u(D, Y)$, and hence $D^*(Y, X) = D^*(Y)$
2. Environment factors (i.e. decision subjects) do not respond to the algorithm.
   - X is not a function of $D^*(\cdot)$

Generalizing these points

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?
  - $D^* = \max_D u(D, Y, X)$

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
   - $u(D, Y, X) = u(D, Y)$, and hence $D^*(Y, X) = D^*(Y)$
2. Environment factors (i.e. decision subjects) do not respond to the algorithm.
   - X is not a function of $D^*(\cdot)$
3. Each decision-maker j responds predictably to the algorithm.
   - $D(X, \hat{Y}, j) = D^*(\hat{Y})$

Outline

Prediction / Causation / Decisions

Kleinberg, Ludwig, Mullainathan and Obermeyer (AER P&P 2015)

- Rain Dance:
  - farmer 1 is facing a drought, should she pay a shaman to give a rain dance to bring rain?

Prediction / Causation / Decisions

Kleinberg, Ludwig, Mullainathan and Obermeyer (AER P&P 2015)

- Rain Dance:
  - farmer 1 is facing a drought, should she pay a shaman to give a rain dance to bring rain?
- Umbrella:
  - farmer 2 is walking to work, should she bring an umbrella?

Prediction / Causation / Decisions

Kleinberg, Ludwig, Mullainathan and Obermeyer (AER P&P 2015)

- Rain Dance:
  - farmer 1 is facing a drought, should she pay a shaman to give a rain dance to bring rain?
- Umbrella:
  - farmer 2 is walking to work, should she bring an umbrella?
- Common elements:
  - both are decisions with payoffs Π.
  - both rely on the same dataset: $Y =$ rain, $X =$ variables correlated with rain.
  - both want to estimate a function $Y = f(X)$

Prediction / Causation / Decisions

Kleinberg, Ludwig, Mullainathan and Obermeyer (AER P&P 2015)

- Rain Dance:
  - farmer 1 is facing a drought, should she pay a shaman to give a rain dance to bring rain?
- Umbrella:
  - farmer 2 is walking to work, should she bring an umbrella?
- Common elements:
  - both are decisions with payoffs Π.
  - both rely on the same dataset: $Y =$ rain, $X =$ variables correlated with rain.
  - both want to estimate a function $Y = f(X)$
- One is a prediction problem, and one is a causation problem.
  - which is which?

Prediction / Causation / Decisions

Kleinberg, Ludwig, Mullainathan and Obermeyer (AER P&P 2015)

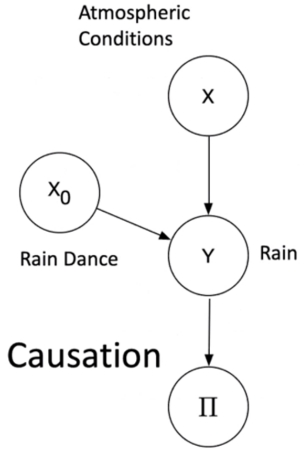
- Rain Dance:
  - farmer 1 is facing a drought, should she pay a shaman to give a rain dance to bring rain?
  - $\rightarrow$ **causation problem: Will the rain dance affect the rain?**

Prediction / Causation / Decisions

Kleinberg, Ludwig, Mullainathan and Obermeyer (AER P&P 2015)

- Rain Dance:
  - farmer 1 is facing a drought, should she pay a shaman to give a rain dance to bring rain?
  - $\rightarrow$ **causation problem: Will the rain dance affect the rain?**
- Umbrella:
  - farmer 2 is walking to work, should she bring an umbrella?
  - $\rightarrow$ **prediction problem (like doctor testing): Will it rain?**

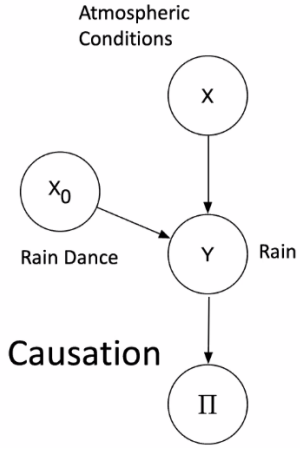
Rain Dances and Umbrellas

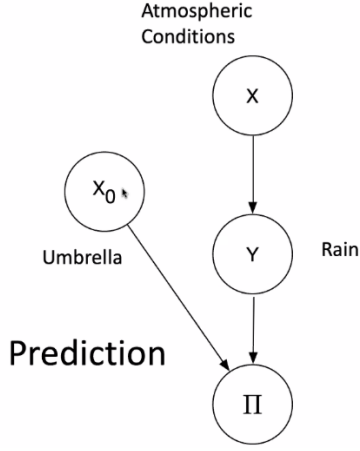


- X_0 is the decision, payoff is Π
  - **Causation**: we care about $X_0 \to Y$, potentially conditional on X

Source: Sendhil Mullainathan Slides.

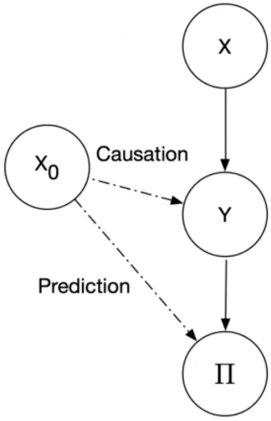
Rain Dances and Umbrellas

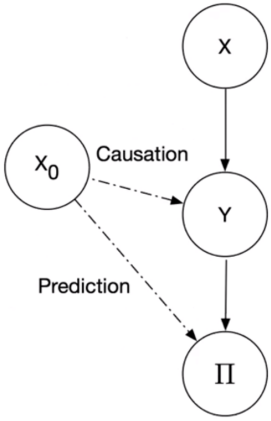




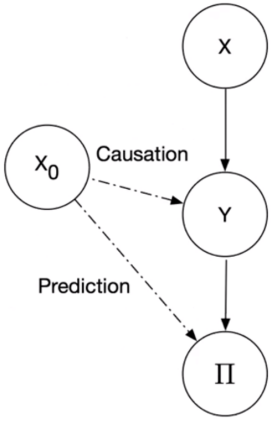
- X_0 is the decision, payoff is Π
  - **Causation**: we care about $X_0 \to Y$, potentially conditional on X
  - **Prediction**: we care about $X_0 \to \Pi$, potentially conditional on Y

Source: Sendhil Mullainathan Slides.

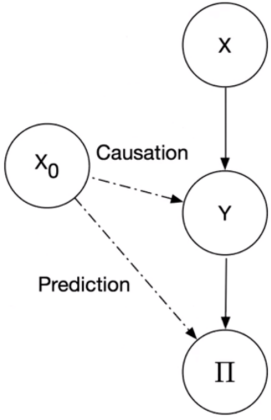

X
X_0
Causation
Y
Prediction
Π



$$\frac{d\Pi}{dX_0} =$$

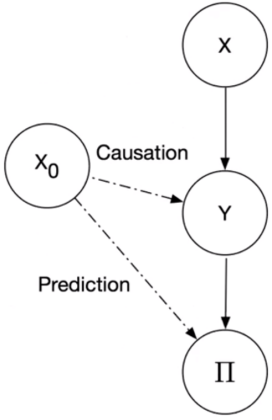


$$\frac{d\Pi}{dX_0} = \frac{\partial \Pi}{\partial X_0}(Y) + \frac{\partial \Pi}{\partial Y}\frac{\partial Y}{\partial X_0}$$



$$\frac{d\Pi}{dX_0} = \frac{\partial \Pi}{\partial X_0}(\hat{Y}) + \frac{\partial \Pi}{\partial Y}\frac{\partial Y}{\partial X_0}$$

- $\frac{\partial \Pi}{\partial X_0}(\hat{Y})$ = the direct change in payoff from the decision, conditional on a **prediction** about Y
- $\frac{\partial Y}{\partial X_0} = \hat{\rho}$ = the **causal effect** of the decision on Y



$$\frac{d\Pi}{dX_0} = \frac{\partial \Pi}{\partial X_0}(\hat{Y}) + \frac{\partial \Pi}{\partial Y}\frac{\partial Y}{\partial X_0}$$

- $\frac{\partial \Pi}{\partial X_0}(\hat{Y})$ = the direct change in payoff from the decision, conditional on a **prediction** about Y
- $\frac{\partial Y}{\partial X_0} = \hat{\rho}$ = the **causal effect** of the decision on Y

- Rules for good decision-making:
  - good predictions require machine learning
  - consistent causal effect estimates require causal inference and experiments.

Outline

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
2. **Environment factors (i.e. decision subjects) do not respond to the algorithm.**
3. **Decision-makers respond predictably to the algorithm.**

Outline

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
2. **Environment factors (i.e. decision subjects) do not respond to the algorithm.**
3. Decision-makers respond respond predictably to the algorithm.

Incentive Responses to Decision Systems

A policy implemented today $D_t(\cdot)$ could change features tomorrow X_{t+1}.

Incentive Responses to Decision Systems

A policy implemented today $D_t(\cdot)$ could change features tomorrow X_{t+1}.

- Take the case of ML-based credit scoring:
  - Some strategic responses are benign/helpful – e.g., pay back existing debts to improve scores

Incentive Responses to Decision Systems

A policy implemented today $D_t(\cdot)$ could change features tomorrow X_{t+1}.

- Take the case of ML-based credit scoring:
  - Some strategic responses are benign/helpful – e.g., pay back existing debts to improve scores
  - Other responses could be costly manipulation – e.g., open more credit accounts to increase credit score, which increase default risk.

Incentive Responses to Decision Systems

A policy implemented today $D_t(\cdot)$ could change features tomorrow X_{t+1}.

- Take the case of ML-based credit scoring:
  - Some strategic responses are benign/helpful – e.g., pay back existing debts to improve scores
  - Other responses could be costly manipulation – e.g., open more credit accounts to increase credit score, which increase default risk.
- More generally:
  - ML subjects can pay some cost and manipulate their features to improve their predicted label.

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)
- Subject gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$:

$$u(X'; X) = \hat{y}(X') - c(X, X')$$

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)
- Subject gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$:

$$u(X'; X) = \hat{y}(X') - c(X, X')$$

- The subject reports

$$\Delta(X) = \arg\max_{X'} u(X'; X)$$

- and therefore the institution's objective is

$$\max_{\hat{y}(\cdot)} \Pr(\hat{y}(\Delta(X)) = Y).$$

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)
- Subject gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$:

$$u(X'; X) = \hat{y}(X') - c(X, X')$$

- The subject reports

$$\Delta(X) = \arg\max_{X'} u(X'; X)$$

- and therefore the institution's objective is

$$\max_{\hat{y}(\cdot)} \Pr(\hat{y}(\Delta(X)) = Y).$$

- Equilibrium:
  - features x_j that are costly to change (high $\frac{\partial c}{\partial x_j}$) will be used by the designer. features that are less costly to change will not be used.
  - in strategic context, designer chooses overall more conservative decision threshold.

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)
- Subject gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$:

$$u(X'; X) = \hat{y}(X') - c(X, X')$$

- The subject reports

$$\Delta(X) = \arg\max_{X'} u(X'; X)$$

- and therefore the institution's objective is

$$\max_{\hat{y}(\cdot)} \Pr(\hat{y}(\Delta(X)) = Y).$$

- Equilibrium:
  - features x_j that are costly to change (high $\frac{\partial c}{\partial x_j}$) will be used by the designer. features that are less costly to change will not be used.
  - in strategic context, designer chooses overall more conservative decision threshold.
- The costs $c(\cdot)$ are socially wasteful, but responses to manipulation increase them.
  - $c(\cdot)$ could be different across groups, causing inequity

Bjorkegren et al (2021): Manipulation-Proof Machine Learning

- Assume a cost function for individual i

$$c(X, X') = \frac{1}{2}(X - X')^{\mathsf{T}} \Gamma_i (X - X')$$

  - with matrix of cost parameters Γ_i.

Bjorkegren et al (2021): Manipulation-Proof Machine Learning

- Assume a cost function for individual i

$$c(X, X') = \frac{1}{2}(X - X')^{\mathsf{T}} \Gamma_i (X - X')$$

  - with matrix of cost parameters Γ_i.
- Assume a linear decision rule

$$\hat{y}(x_i) = \beta^{\mathsf{T}} X$$

Bjorkegren et al (2021): Manipulation-Proof Machine Learning

- Assume a cost function for individual i

$$c(X, X') = \frac{1}{2}(X - X')^{\mathsf{T}} \Gamma_i (X - X')$$

  - with matrix of cost parameters Γ_i.
- Assume a linear decision rule

$$\hat{y}(x_i) = \beta^{\mathsf{T}} X$$

- Individual's optimal report is

$$X_i^*(\beta) = X + \Gamma_i^{-1} \beta$$

Bjorkegren et al (2021): Manipulation-Proof Machine Learning

- Assume a cost function for individual i

$$c(X, X') = \frac{1}{2}(X - X')^{\mathsf{T}} \Gamma_i (X - X')$$

  - with matrix of cost parameters Γ_i.
- Assume a linear decision rule

$$\hat{y}(x_i) = \beta^{\mathsf{T}} X$$

- Individual's optimal report is

$$X_i^*(\beta) = X + \Gamma_i^{-1} \beta$$

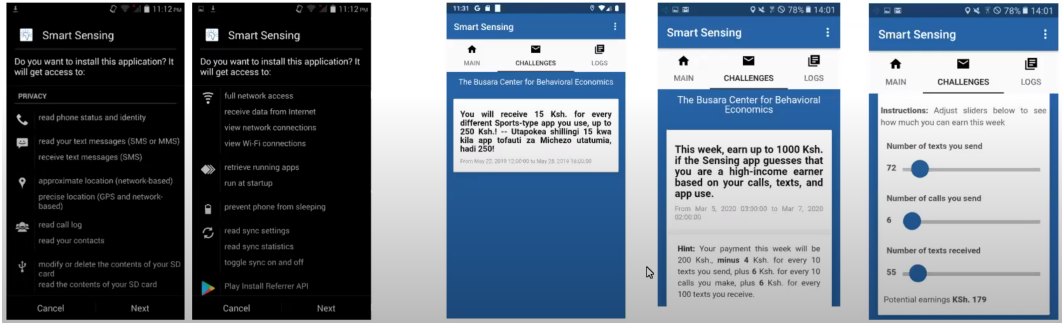
- Principal's "manipulation-proof" decision rule is

$$\begin{aligned} \beta &= \arg\min_{\beta} \sum_i [y_i - \beta^{\mathsf{T}} X_i^*(\beta)]^2 \\ &= \arg\min_{\beta} \sum_i [y_i - \beta^T \underbrace{(X + \Gamma_i^{-1} \beta)}_{\text{strategic response}}]^2 \end{aligned}$$

Bjorkegren et al (2021): Experiment

We built a new smartphone app, with Busara Center (Nairobi)

1. App collects **behavioral data** (same as with a digital credit app)
2. App provides **rewards** based on user behavior
3. We experimentally vary **transparency** of the algorithm [timeline]



Bjorkegren et al (2021): Learning cost parameters Γ_i

		Change in observed behavior					
		Missed Calls	People Called (Workday)	Battery Charges	Missed Calls (outgoing)	Non-Contact Calls (Weekend)	Text messages sent
Behavior incentivized	Missed Calls	**0.709 (0.05)****	**0.152 (0.044)***	0.026 (0.865)	0.825 (0.124)	-0.002 (0.994)	**4.16 (0.035)****
	People Called (Workday)	0.395 (0.165)	**0.227 (0.0)*****	-0.06 (0.609)	0.121 (0.773)	0.068 (0.7)	-1.537 (0.321)
	Battery Charges	-0.053 (0.913)	-0.03 (0.766)	-0.038 (0.85)	-0.616 (0.391)	-0.015 (0.96)	0.687 (0.795)
	Missed Calls (outgoing)	0.324 (0.491)	**0.197 (0.045)***	0.313 (0.11)	**1.187 (0.089)**	**0.502 (0.089)***	-0.206 (0.936)
	Non-Contact Calls (on Weekend)	-0.056 (0.906)	-0.054 (0.585)	-0.138 (0.481)	**1.234 (0.078)***	**1.233 (0.0)*****	-2.022 (0.433)
	Text messages sent	-0.052 (0.921)	-0.014 (0.898)	0.005 (0.981)	-0.836 (0.286)	-0.022 (0.948)	**24.508 (0.0)*****
	Week & Individual FE's	✓	✓	✓	✓	✓	✓
	N (person-weeks)	7976	7976	7976	7976	7976	7976

Bjorkegren et al (2021): Decision Rule

Robust model selects features that are harder to manipulate

Response variable: MONTHLY INCOME	**Naive β (3 coefficients)** ¢/action	**Robust β (3 coefficients)** ¢/action	**Cost α_{jj} of manipulation** ¢/action^2
# Outgoing Calls	0.625	0.542	0.591
# Outgoing SMS	-0.395	-0.107	0.035
# Incoming SMS	0.065	0	0.038
# Evening SMS	0	-0.121	0.058
N (unique individuals)	1377	1377	

Bjorkegren et al (2021): Results

In experiment, robust rule performs better

Response variable: MONTHLY INCOME	Naive β (3 coefficients)	Robust β (3 coefficients)	Cost α_{jj} of manipulation
# Outgoing Calls	0.625	0.542	0.59087919
# Outgoing SMS	-0.395	-0.107	0.03454505
# Incoming SMS	0.065	0	0.03834774
# Evening SMS	0	-0.121	0.05849779
RMSE (w/o Manipulation)	3.553	3.554	
RMSE (with Manipulation)	3.867	3.655	
Δ (%)	8.837%	2.841%	

The naïve model does better when people don't have incentives to game

The robust model does better when people are gaming

The robust model is less impacted by strategic behavior

Outline

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
2. Environment factors (i.e. decision subjects) do not respond to the algorithm.
3. **Decision-makers respond predictably to the algorithm.**

Decision-makers are usually separate from the algorithm

- So far we have treated the decision D as a deterministic function of $\hat{Y}$: $D = 1$ if $\hat{Y} > \bar{Y}$, $D = 0$ otherwise.
  - means that $\frac{\partial D}{\partial x_j} = 0, \forall j$: decisions are not sensitive to case characteristics, after conditioning on $\hat{Y}$.

Decision-makers are usually separate from the algorithm

- So far we have treated the decision D as a deterministic function of $\hat{Y}$: $D = 1$ if $\hat{Y} > \bar{Y}$, $D = 0$ otherwise.
  - means that $\frac{\partial D}{\partial x_j} = 0, \forall j$: decisions are not sensitive to case characteristics, after conditioning on $\hat{Y}$.
- But there could be many reasons that this assumption does not hold, e.g.:
  - judges caring about whether a defendant has children or not.
  - tax/fraud auditors not wanting to audit their friends / family members
  - doctor wanting to save people with more years of life left / not terminally ill

Decision-makers are usually separate from the algorithm

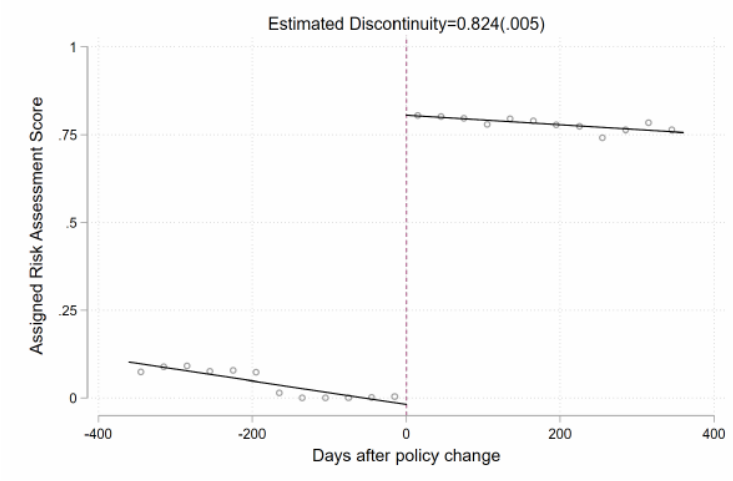
- So far we have treated the decision D as a deterministic function of $\hat{Y}$: $D = 1$ if $\hat{Y} > \bar{Y}$, $D = 0$ otherwise.
  - means that $\frac{\partial D}{\partial x_j} = 0, \forall j$: decisions are not sensitive to case characteristics, after conditioning on $\hat{Y}$.
- But there could be many reasons that this assumption does not hold, e.g.:
  - judges caring about whether a defendant has children or not.
  - tax/fraud auditors not wanting to audit their friends / family members
  - doctor wanting to save people with more years of life left / not terminally ill

$\rightarrow$ empirical evidence is needed on how decision-makers respond to algorithms.

First Stage: Discrete Reform introducing risk scoring

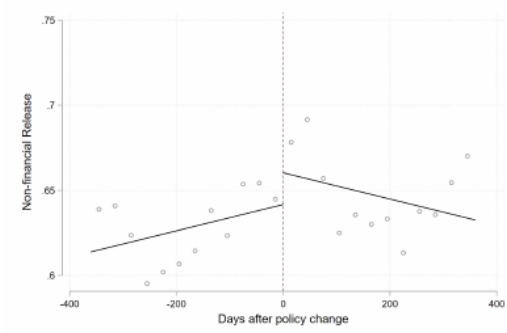
Sloan et al 2018

Figure 4: Regression Discontinuity Results for the Probability of Receiving a Risk Assessment Score

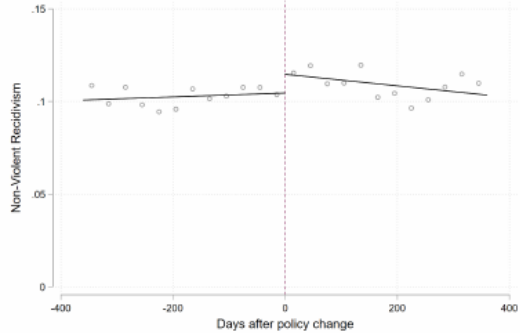


Risk scoring increases release rates and recidivism

Sloan et al 2018



(a) Non-financial Bond



(a) Probability of Non-Violent Recidivism

- In response to risk scoring, judges release more poor defendants.

Stevenson and Doleac: Method

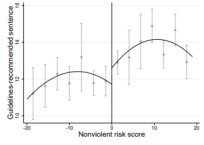
- RD using a continuous risk score – above a discrete cutoff, defendant is labeled "risky".

Stevenson and Doleac: Method

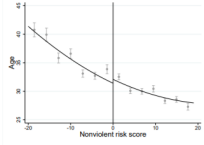
- RD using a continuous risk score – above a discrete cutoff, defendant is labeled "risky".
- Identification check: Other predetermined characteristics are flat around the cutoff (covariate balance):

Figure 2: Covariate balance across risk score cutoffs

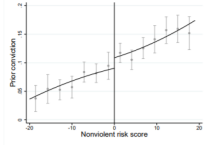
(a) Nonviolent risk score and the guidelines-recommended sentence



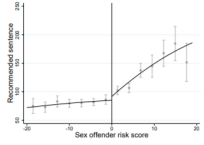
(b) Nonviolent risk score and age



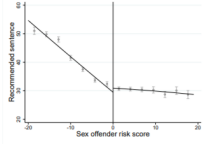
(c) Nonviolent risk score and prior convictions



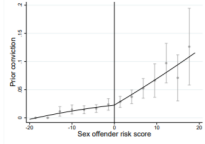
(d) Sex offender risk score and the guidelines-recommended sentence



(e) Sex offender risk score and age



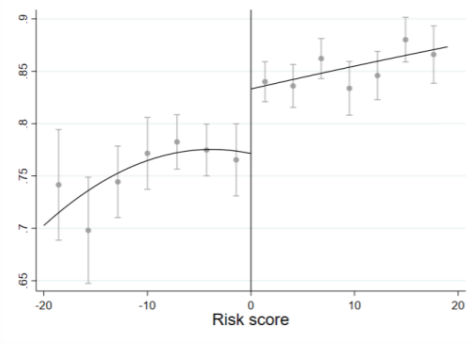
(f) Sex offender risk score and prior convictions



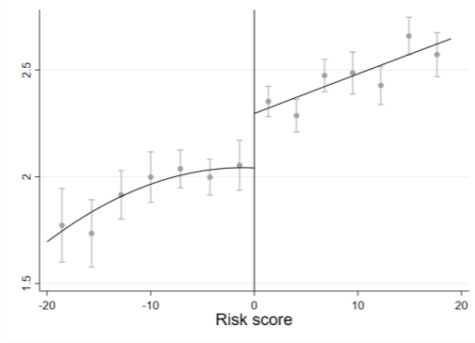
Stevenson and Doleac: Result (RDD)

Figure 3: Does the risk classification affect defendants' sentences at the margin?

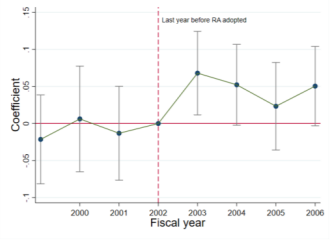
(a) Probability of incarceration



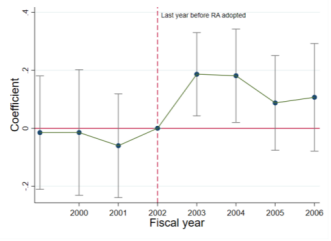
(b) The sentence length



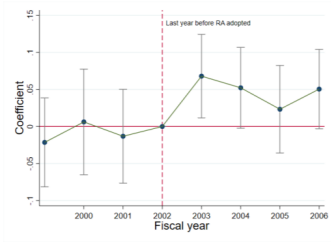
(c) Predicted risk score event study (outcome = pr(incarceration))



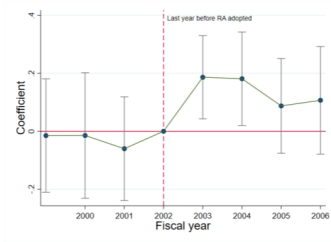
(d) Predicted risk score event-study (outcome = sentence length)



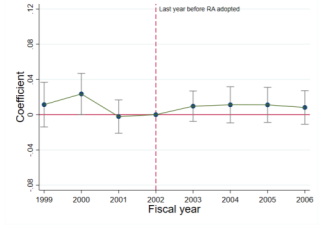
(c) Predicted risk score event study (outcome = pr(incarceration))



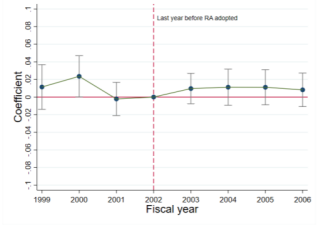
(d) Predicted risk score event-study (outcome = sentence length)



(a) Risk assessment's impact on pr(incarceration)



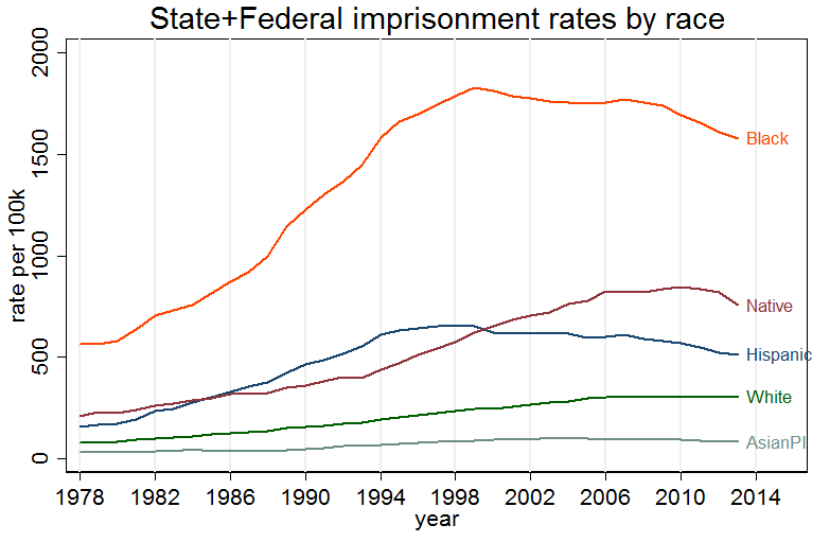
(b) Risk assessment's impact on sentence length (arcsinh)



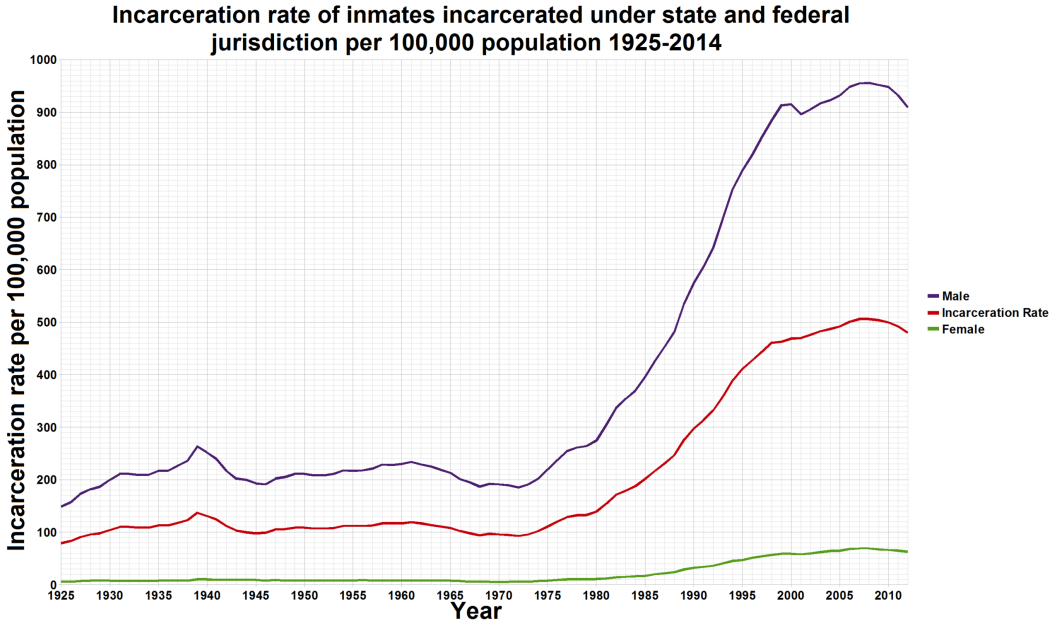
"...despite explicit instructions that risk assessment was supposed to lower prison populations, there was no net reduction in incarceration. Nor do we detect any public safety benefits from its use..."

Outline

Incarceration Rates by Race in U.S.A, 1978-2014



Incarceration Rates by Gender in U.S.A, 1925-2014



Homicide Offending Rates, by Race and Gender

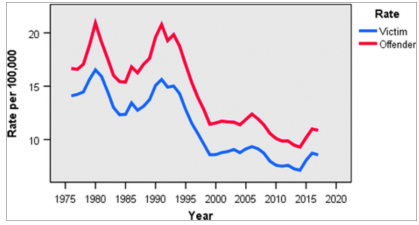
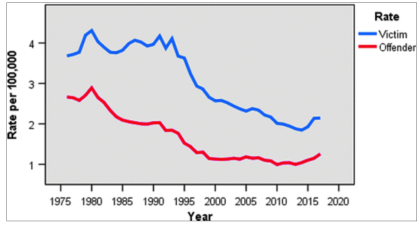


FIG. 1. Offending and victimization rates for men, 1976–2017.



Can group differences in preferences/ability explain variation in crime and incarceration?

- Is answer different for race and gender?

Can group differences in preferences/ability explain variation in crime and incarceration?

- Is answer different for race and gender?
- Preferences/ability could be the result of past discrimination/disadvantage.
  - disparities in health/education
  - prejudice leading to demotivation
  - etc.

Taste-Based Discrimination (Prejudice)

- "Taste for discrimination" (associated with Becker)
  - Firms willing to pay to associate with some persons, not others
  - E.g. act as if racial minorities are more expensive to hire than they are

Taste-Based Discrimination (Prejudice)

- "Taste for discrimination" (associated with Becker)
  - Firms willing to pay to associate with some persons, not others
  - E.g. act as if racial minorities are more expensive to hire than they are
- Prejudice will reduce profits $\rightarrow$ in a competitive market, discriminating firms will be competed out.

Taste-Based Discrimination (Prejudice)

- "Taste for discrimination" (associated with Becker)
  - Firms willing to pay to associate with some persons, not others
  - E.g. act as if racial minorities are more expensive to hire than they are
- Prejudice will reduce profits $\rightarrow$ in a competitive market, discriminating firms will be competed out.
  - could remain with other labor market frictions, e.g. imperfect competition

Taste-Based Discrimination (Prejudice)

- "Taste for discrimination" (associated with Becker)
  - Firms willing to pay to associate with some persons, not others
  - E.g. act as if racial minorities are more expensive to hire than they are
- Prejudice will reduce profits $\rightarrow$ in a competitive market, discriminating firms will be competed out.
  - could remain with other labor market frictions, e.g. imperfect competition
  - could remain in public sector (e.g. judicial decisions)

Jury Race in Criminal Trials

Anwar, Bayer, and Hjalmarsson (2012)

- Examine jury racial composition and trial outcomes in Florida, 2000-2010
- Exogenous treatment: day-to-day variation in composition of jury pool

Jury Race in Criminal Trials

Anwar, Bayer, and Hjalmarsson (2012)

- Examine jury racial composition and trial outcomes in Florida, 2000-2010
- Exogenous treatment: day-to-day variation in composition of jury pool
  - Identification check: composition of jury pool uncorrelated with characteristics of the defendant and case.

Results

Anwar, Bayer, and Hjalmarsson (2012)

TABLE IV

REDUCED-FORM BENCHMARK REGRESSIONS

	(1)	(2)	(3)	(4)
Dependent variable	Any guilty conviction		Proportion guilty convictions	
Black defendant	0.150***	0.164***	0.156***	0.160***
	[0.056]	[0.058]	[0.055]	[0.057]
Any black in pool	0.069	0.105**	0.063	0.090*
	[0.048]	[0.051]	[0.047]	[0.050]
Black defendant * any black in pool	−0.168**	−0.166**	−0.174**	−0.155**
	[0.070]	[0.074]	[0.069]	[0.072]
Constant	0.656***	0.627***	0.600***	0.576***
	[0.039]	[0.041]	[0.038]	[0.040]
Includes controls for:				
Gender/age of pool	No	Yes	No	Yes
County dummy	No	Yes	No	Yes
Year of filing dummies	No	Yes	No	Yes
Observations	712	712	712	712
R-squared	0.01	0.07	0.01	0.08

Statistical discrimination (stereotypes)

- Employers/judges have different information/beliefs about identity groups.
  - race/gender could in fact be correlated with productivity/criminality

Statistical discrimination (stereotypes)

- Employers/judges have different information/beliefs about identity groups.
  - race/gender could in fact be correlated with productivity/criminality
- Different priors (stereotypes) about productivity/criminality.

Statistical discrimination (stereotypes)

- Employers/judges have different information/beliefs about identity groups.
  - race/gender could in fact be correlated with productivity/criminality
- Different priors (stereotypes) about productivity/criminality.
  - could be self-confirming: employer/judge doesn't give the stereotyped group a chance to prove themselves.

Statistical discrimination (stereotypes)

- Employers/judges have different information/beliefs about identity groups.
  - race/gender could in fact be correlated with productivity/criminality
- Different priors (stereotypes) about productivity/criminality.
  - could be self-confirming: employer/judge doesn't give the stereotyped group a chance to prove themselves.
  - another channel for self-confirmation: minority workers expect to be discriminated against, and therefore don't invest in education/skills.

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs
  - Taste-based discrimination: police are prejudiced

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs
  - Taste-based discrimination: police are prejudiced
- Can formally test in this context:
  - statistical discrimination $\rightarrow$ contraband discovery (successful search) rates will be the same for both groups.

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs
  - Taste-based discrimination: police are prejudiced
- Can formally test in this context:
  - statistical discrimination $\rightarrow$ contraband discovery (successful search) rates will be the same for both groups.
  - prejudice $\rightarrow$ contraband discovery rates will be lower for black drivers, as threshold for search is lower.

Racial bias in vehicle searches

Knowles, Persico, and Todd (2001)

- Motivation: Black drivers are searched more often by police for drugs.
  - is this prejudice?
- Two reasons blacks get searched more:
  - Statistical discrimination: blacks are more likely to have drugs
  - Taste-based discrimination: police are prejudiced
- Can formally test in this context:
  - statistical discrimination $\rightarrow$ contraband discovery (successful search) rates will be the same for both groups.
  - prejudice $\rightarrow$ contraband discovery rates will be lower for black drivers, as threshold for search is lower.
- Empirical test:
  - data on 1500 traffic searches in Maryland, 1995-1999
  - contraband discovery rates are the same across races, consistent with statistical discrimination, but not taste-based discrimination

Outline

“Fair ML” / “AI Fairness”

- There is growing concern about social harms and disparities produced by AI decisions.

“Fair ML” / “AI Fairness”

- There is growing concern about social harms and disparities produced by AI decisions.
- “ML” or “AI” refer to statistical algorithms
  - can learning algorithms be fair or not?

"Fair ML" / "AI Fairness"

- There is growing concern about social harms and disparities produced by AI decisions.
- "ML" or "AI" refer to statistical algorithms
  - can learning algorithms be fair or not?
- Rather: *fairness* is a property of *decisions*.
  - so *"AI Fairness"* should be understood as "*fairness of AI-supported decision-making*".

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.
- Firms using ML to screen job applicants might wish to incorporate diversity objectives.

Examples

- Lending laws (e.g. in the United States) prohibit practices that discriminate on the basis of race.
- Firms using ML to screen job applicants might wish to incorporate diversity objectives.
- Judges might want to reduce biases in legal decisions.

List of Protected Attributes Specified in US Fair Lending Laws

- Fair Housing Acts (FHA)
- Equal Credit Opportunity ACts (ECOA)

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

- Machine learning researchers take these as given.

Data can be biased

- e.g. Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).

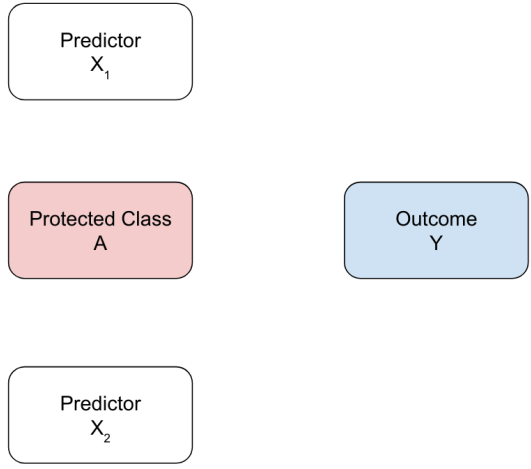
Data can be biased

- e.g. Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.

Data can be biased

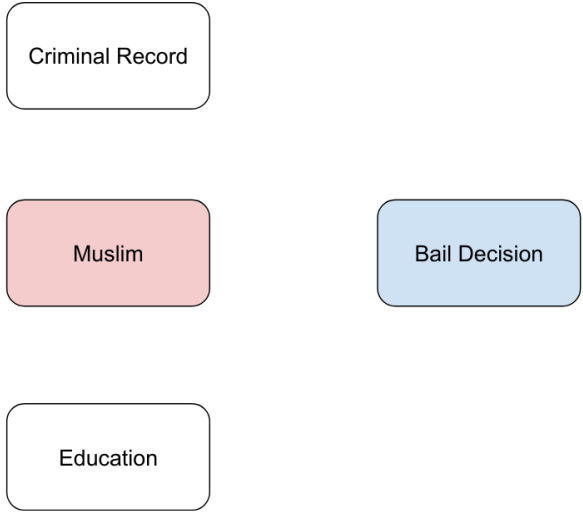
- e.g. Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
- Today we will ignore the sources of bias in machine learning models
  - in week 13, we will look at how machine learning can be used to help detect bias

Overview: Fairness in Decision-Making

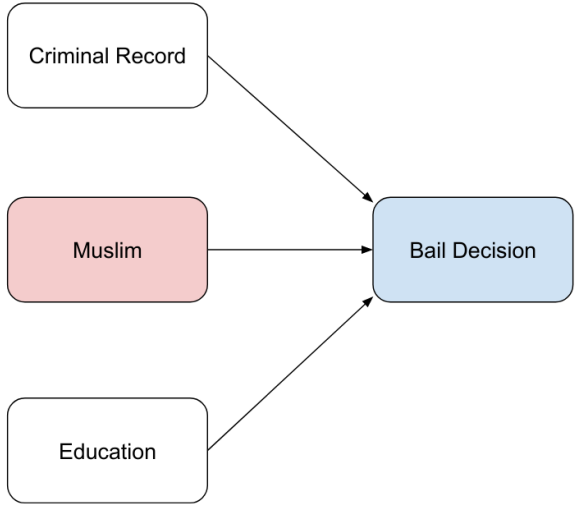


- $A \in \{0,1\}$= protected class, X= other predictors, Y= outcome.
- let $\hat{Y}(X,A)$ be our model predictions.

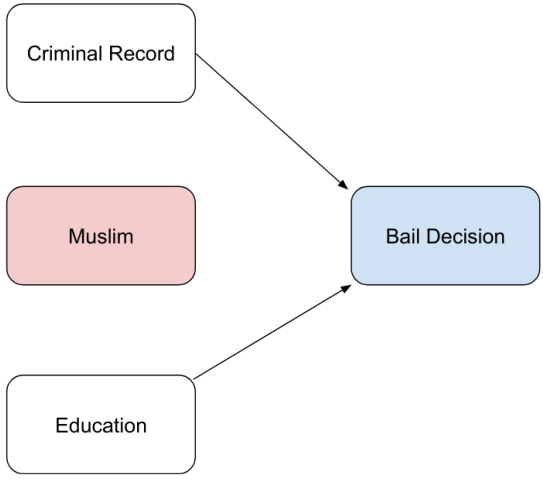
For example:



Standard Approach: Use All Data

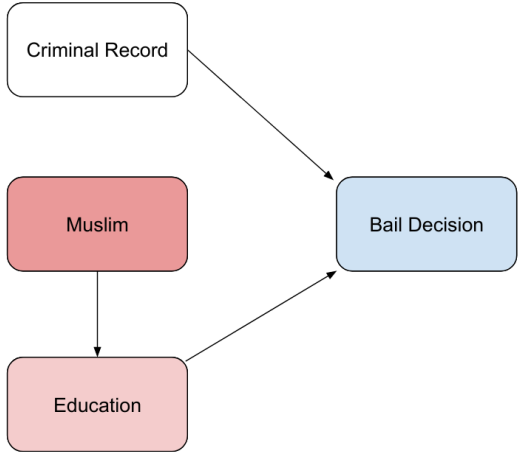


Fairness through Unawareness



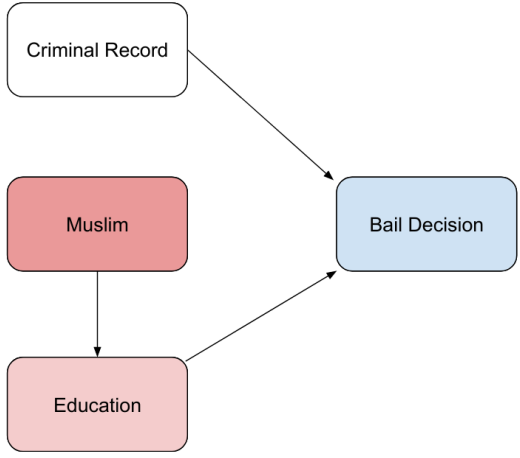
- **Fairness through unawareness:** protected attributes are not explicitly used in the prediction process.
  - that is, $\hat{Y}(X,0) = \hat{Y}(X,1)$, $\forall X$.

Problem: Indirect Discrimination



- sensitive factors are implicitly being used by the model, to the extent that they are correlated with included predictors.
  - e.g., muslims have lower education than rest of population.

Problem: Indirect Discrimination



- sensitive factors are implicitly being used by the model, to the extent that they are correlated with included predictors.
  - e.g., muslims have lower education than rest of population.
- in most datasets, if you drop the sensitive attribute and train a new classifier, the resulting predictions will be the same or very close.

Outline

Review: Classification Metrics

	Predicted Positive	Predicted negative
Actual Positive	$TP = \#$ true positives	$FN = \#$ false negatives
Actual Negative	$FP = \#$ false positives	$TN = \#$ true negatives

- Identify the correct sequence of labels for the following four metrics, separated by commas.
  1. $\frac{TP+TN}{TP+TN+FP+FN}$
  2. $\frac{TP}{TP+FP}$
  3. $\frac{TP}{TP+FN}$
  4. $\frac{FP}{FP+TN}$

Classification Metrics

Event	Condition	Associated metric: Pr{event | condition}	Formula: $\frac{\#\text{ event}}{\#\text{ condition}}$
$\hat{Y}=1$	$Y=1$	True positive rate [Recall for positive class]	$\frac{TP}{TP+FN}$
$\hat{Y}=0$	$Y=0$	True negative rate [Recall for negative class]	$\frac{TN}{TN+FP}$
$Y=1$	$\hat{Y}=1$	Positive predictive value [Precision for positive class]	$\frac{TP}{TP+FP}$
$Y=0$	$\hat{Y}=0$	Negative predictive value [Precision for negative class]	$\frac{TN}{TN+FN}$
$\hat{Y}=1$	$Y=0$	False positive rate	$\frac{FP}{TN+TN}$
$\hat{Y}=0$	$Y=1$	False negative rate	$\frac{FN}{TP+FN}$
$Y=1$	$\hat{Y}=0$	?	$\frac{TP}{TN+FN}$
$Y=0$	$\hat{Y}=1$	?	$\frac{TN}{TP+FP}$

- $Y \in \{0,1\}$= outcome label, e.g. reoffends or not; $\hat{Y} \in \{0,1\}$ = classifier output label
- TP = # true positives, FN = # false negatives, FP = # false positives, TN = # true negatives

Classifier Setup

- $Y \in \{0,1\}$ = outcome label, e.g. reoffends or not
- X = predictors, e.g. criminal history
- $A \in \{0,1\}$ = protected class, e.g. gender

Classifier Setup

- $Y \in \{0,1\}$ = outcome label, e.g. reoffends or not
- X = predictors, e.g. criminal history
- $A \in \{0,1\}$ = protected class, e.g. gender

Classifier output:

- $\hat{y}(X,A) \in [0,1]$ = the **score**, usually interpreted as a predicted probability

Classifier Setup

- $Y \in \{0,1\}$ = outcome label, e.g. reoffends or not
- X = predictors, e.g. criminal history
- $A \in \{0,1\}$ = protected class, e.g. gender

Classifier output:

- $\hat{y}(X,A) \in [0,1]$ = the **score**, usually interpreted as a predicted probability
- $\hat{Y}(X,A) \in \{0,1\}$ = the assigned class label
  - usually assigned by a threshold rule: $\hat{Y} = 1$ if $\hat{y} \geq \bar{y}$, $\hat{Y} = 0$ if $\hat{y} < \bar{y}$, for some $\bar{y} \in (0,1)$.
  - if $\hat{y}(\cdot)$ is well-calibrated, would typically set $\bar{y} = 0.5$.

Statistical Fairness Criteria

Based on Berk et al (2017) and Barocas et al (2021):

1. Equalizing outcomes across groups (statistical parity / independence)
2. Equalizing recall across groups (separation)
3. Equalizing precision across groups (calibration / sufficiency)

1. Equalizing Outcomes Across Groups

Statistical Parity

Average predicted outcome ($\frac{\#\text{ predicted positive}}{\text{sample size}}$) should be the same across groups.

$$\Pr(\hat{Y} = 1 | A = a) = \Pr(\hat{Y} = 1 | A = b)$$

- also called "demographic parity" or "disparate impact". This is probably the most commonly used fairness metric.

1. Equalizing Outcomes Across Groups

Statistical Parity

Average predicted outcome ($\frac{\#\text{ predicted positive}}{\text{sample size}}$) should be the same across groups.

$$\Pr(\hat{Y} = 1|A = a) = \Pr(\hat{Y} = 1|A = b)$$

- also called "demographic parity" or "disparate impact". This is probably the most commonly used fairness metric.
- Pros:
  - simple and intuitive
  - sometimes legally required (e.g. EEOC's four-fifths rule)
- Cons:
  - enforcing statistical parity tends to reduce accuracy, especially when the true label varies across groups (different base rates).
  - e.g.: if decision to grant bail is based on $\hat{Y}$, can lead to undesirable outcomes, such as imprisoning a lot more women who are not risky.

- In practice, achieving equal outcomes could be too restrictive.
- Instead, could impose a slack condition:

$$|\Pr(\hat{Y} = 1|A = a) - \Pr(\hat{Y} = 1|A = b)| \leq \epsilon$$

  - where, e.g. ϵ could be set to satisfy the "four-fifths rule" from disparate impact law.

1. Equalizing Outcomes Across Groups

Relaxed Statistical Parity and Independence

- In practice, achieving equal outcomes could be too restrictive.
- Instead, could impose a slack condition:

$$|\Pr(\hat{Y} = 1|A = a) - \Pr(\hat{Y} = 1|A = b)| \leq \epsilon$$

  - where, e.g. ϵ could be set to satisfy the "four-fifths rule" from disparate impact law.
- "Independence" (Barocas et al 2021)
  - a stronger criterion that implies statistical parity
  - requires independence of the **score** and the protected attribute: $\hat{y} \perp A$.
  - If $I(z,x)$ is mutual information between z and x, equivalent to requiring $I(A;\hat{y}) = 0$ or $I(A;\hat{y}) \leq \epsilon$.

2. Equalizing Recall Across Groups

- A disadvantage of statistical parity is that it will penalize one of the groups if there are different base rates.

2. Equalizing Recall Across Groups

- A disadvantage of statistical parity is that it will penalize one of the groups if there are different base rates.
- Another set of fairness criteria penalize divergence in error metrics across groups, conditional on the true label.
  - e.g. accuracy, recall, FPR, FNR
  - allows for different treatment of groups if justified by variation in base rates

2. Equalizing Recall Across Groups

- A disadvantage of statistical parity is that it will penalize one of the groups if there are different base rates.
- Another set of fairness criteria penalize divergence in error metrics across groups, conditional on the true label.
  - e.g. accuracy, recall, FPR, FNR
  - allows for different treatment of groups if justified by variation in base rates
  - e.g. men and women should have same model accuracy

2. Equalizing Recall Across Groups

- A disadvantage of statistical parity is that it will penalize one of the groups if there are different base rates.
- Another set of fairness criteria penalize divergence in error metrics across groups, conditional on the true label.
  - e.g. accuracy, recall, FPR, FNR
  - allows for different treatment of groups if justified by variation in base rates
  - e.g. men and women should have same model accuracy
- Can also combine multiple criteria:
  - e.g., the ratio of false positives to false negatives should be the same for men and women.

2. Equalizing Recall Across Groups

Separation

Barocas et al (2021) discuss the more general criteria, "**separation**":

- requires $\hat{y} \perp A|Y$: that is, the score is independent of the sensitive attribute, conditional on the true label.
- In the binary case, equivalent to equalizing **both** true positive rates (recall for positive class) **and** false positive rates across groups.

3. Equalizing Precision Across Groups

Definition

- A third set of metrics requires equalizing precision across groups
  - precision for both positive and negative outcomes
  - i.e. positive/negative predictive value
  - also called “predictive parity”

3. Equalizing Precision Across Groups

Definition

- A third set of metrics requires equalizing precision across groups
  - precision for both positive and negative outcomes
  - i.e. positive/negative predictive value
  - also called "predictive parity"
- Barocas et al (2021) call this "**sufficiency**" and formalize it as

$$\Pr(Y = 1 | \hat{Y}, A = a) = \Pr(Y = 1 | \hat{Y}, A = b)$$

  - that is, conditioning on the score, both groups get the same label.

3. Equalizing Precision Across Groups

Calibration

- An intuitive way to achieve sufficiency (equalizing precision across groups) is to require that the classifier is **well-calibrated** for each group.
- that is,

$$\hat{y}(X, A) = \Pr(Y = 1), \forall A$$

the scores provide the probability that the true label equals one, for all groups.

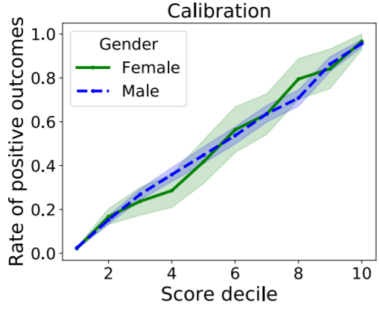
3. Equalizing Precision Across Groups

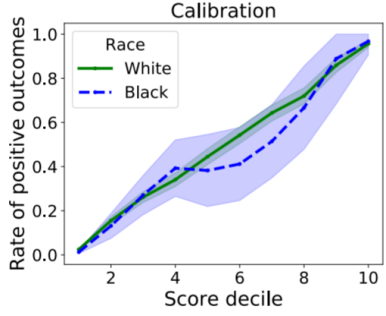
Calibration

- An intuitive way to achieve sufficiency (equalizing precision across groups) is to require that the classifier is **well-calibrated** for each group.
- that is,

$$\hat{y}(X, A) = \Pr(Y = 1), \forall A$$

the scores provide the probability that the true label equals one, for all groups.





What notions of fairness does this classifier satisfy?

Group A			
	$\hat{Y}=1$	$\hat{Y}=0$	
$Y=1$	30	20	TPR=.6
$Y=0$	20	20	TNR=.5
	PPV = .6	NPV = .5	
avg $\hat{Y}$ = .55, acc. = .55, FP/FN = 1			

Group B			
	$\hat{Y}=1$	$\hat{Y}=0$	
$Y=1$	60	40	TPR=.6
$Y=0$	60	60	TNR=.5
	PPV = .5	NPV = .4	
avg $\hat{Y}$ = .55, acc. = .55, FP/FN = 1.5			

1. Equality of outcomes (statistical parity / independece)
2. Equality of recall (separation)
3. Equality of precision (sufficiency)

Outline

1. Equality of outcomes (statistical parity / independece)
2. Equality of recall (separation)
3. Equality of precision (sufficiency)

Except in highly artificial datasets, Criteria (1), (2), and (3) are all mutually incompatible with each other!

(1) and (2), (1) and (3): Statistical Parity vs. Equal Recall/Precision

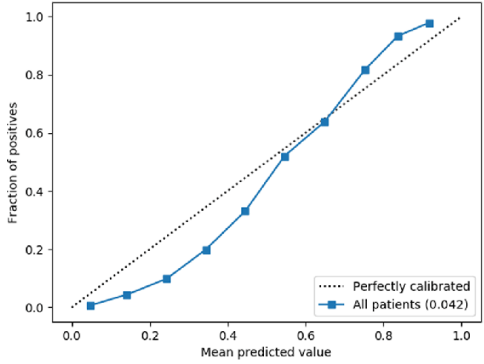
- If the outcome Y varies by group status A, a classifier achieving statistical parity means that average $\hat{Y}$ does not equal average Y for at least one of the groups.
  - hence, there will be differences in both recall (error rates conditional on true label) and precision (error rates conditional on predicted label) across groups.
- Hence, satisfying (1) precludes satisfying (2) or (3) except in the unrealistic case of identical base rates across groups.

(2) and (3): Equal Error Rate vs Equal Precision

- If base rates differ by group, these requirements cannot hold simultaneously:
  - error rate balance (equality of FPR/FNR across groups)
  - predictive parity (equality of PPV/NPV across groups)
- this is often called the precision-recall tradeoff.

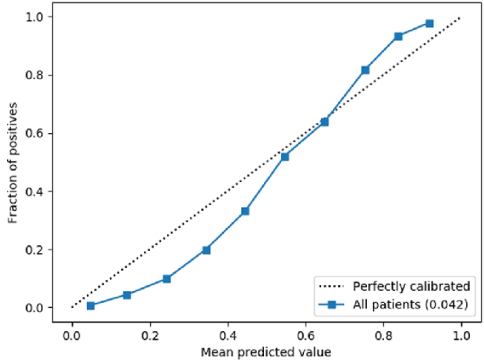
(2) and (3): Equal Error Rate vs. Calibration

- recall that in a well-calibrated model, we can bin observations by their predicted outcome probabilities, and the outcome rates should roughly match in those bins.
- good calibration requires equalizing false positive and false negative rates.



(2) and (3): Equal Error Rate vs. Calibration

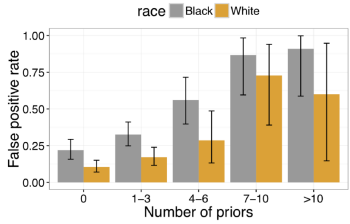
- recall that in a well-calibrated model, we can bin observations by their predicted outcome probabilities, and the outcome rates should roughly match in those bins.
- good calibration requires equalizing false positive and false negative rates.



Trade-off: If base rates differ by group, error rate balance (equality of FPR/FNR across groups) precludes calibration.

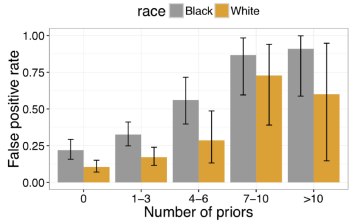
Example: COMPAS

FPR is higher for black defendants! (Chouldechova'17):

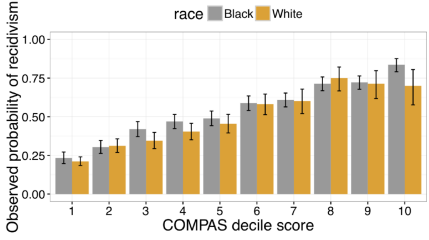


Example: COMPAS

FPR is higher for black defendants! (Chouldechova'17):



But the scores are well-calibrated (or PPV similar across all groups)! (Chouldechova'17):



COMPAS: Dressel and Farid (2018)

COMPAS has higher false positive rate and lower false negative rate for black defendants.

- errors disfavor black defendants.

Dressel and Farid (2018):

- also asked human annotators to produce recidivism predictions, and race info was not provided.

COMPAS: Dressel and Farid (2018)

COMPAS has higher false positive rate and lower false negative rate for black defendants.

- errors disfavor black defendants.

Dressel and Farid (2018):

- also asked human annotators to produce recidivism predictions, and race info was not provided.
- humans were almost identically biased.

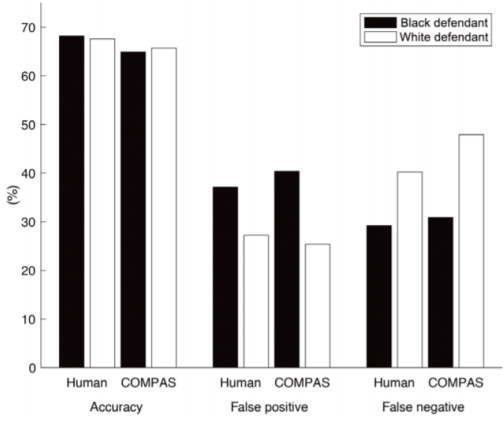


Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

- giving the human annotators information on the race of the defendant made no difference.

Outline

How to make ML-Based Decisions fair?

- So far, our metrics can be used to assess the fairness of classifiers and the resulting decisions.
- What if our decisions is biased? What do we do?

How to make ML-Based Decisions fair?

- So far, our metrics can be used to assess the fairness of classifiers and the resulting decisions.
- What if our decisions is biased? What do we do?
- There are three groups of approaches:
  - **Pre-processing**: Adjust the feature space to be uncorrelated with the sensitive attribute.
  - **At training time**: Work the constraint into the optimization process that constructs a classifier from training data.
  - **Post-processing**: Adjust a learned classifier so as to be uncorrelated with the sensitive attribute.

Outline

Post-Processing with the Score

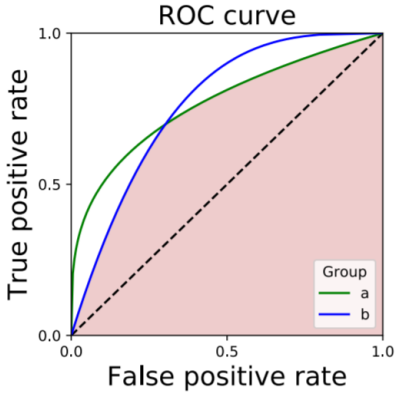
- Given a score function $\hat{y}(\cdot)$ and a cost for false negatives and false positives, find the derived classifier that minimizes the expected cost of false positive and false negatives subject to the fairness constraint at hand.
  - can depend on the sensitive attribute
  - can add randomness
- Advantages:
  - simple and transparent
  - works for any black-box classifier regardless of its inner workings.
  - no need for re-training models
- Disadvantage:
  - requires and uses the protected attribute.

Achieving Fairness with Post-Processing

- Statistical Parity: Just set the thresholds for each group k such that the average $\hat{Y}$ is the same.

Achieving Fairness with Post-Processing

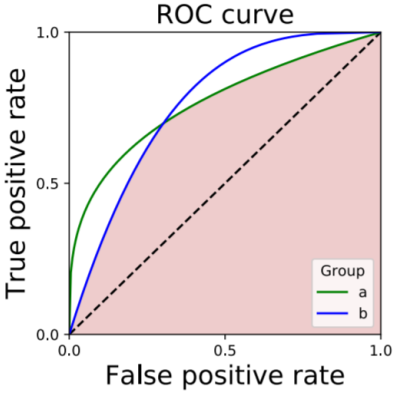
- Statistical Parity: Just set the thresholds for each group k such that the average $\hat{Y}$ is the same.
- Separation (equality of true positive rates and false positive rates):



- In the binary case, a classifier satisfying separation is limited to the region in red.
- Set separate group thresholds and randomize across multiple classifiers to equalize the rates.

Achieving Fairness with Post-Processing

- Statistical Parity: Just set the thresholds for each group k such that the average $\hat{Y}$ is the same.
- Separation (equality of true positive rates and false positive rates):



- In the binary case, a classifier satisfying separation is limited to the region in red.
- Set separate group thresholds and randomize across multiple classifiers to equalize the rates.

- Calibration:
  - just calibrate the classifier separately by group.

"An Economic Approach to Regulating Algorithms"

Kleinberg et al (2018); Rambachan et al (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.

"An Economic Approach to Regulating Algorithms"

Kleinberg et al (2018); Rambachan et al (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

"An Economic Approach to Regulating Algorithms"

Kleinberg et al (2018); Rambachan et al (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.

"An Economic Approach to Regulating Algorithms"

Kleinberg et al (2018); Rambachan et al (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.
- i.e., there should be no limit on the use of sensitive attributes.

Outline

- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.

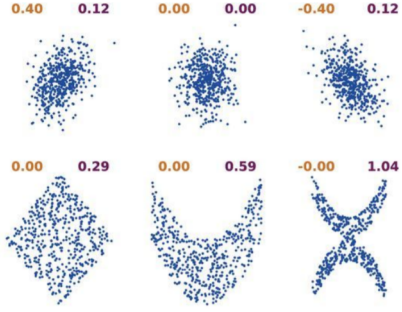
- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.
- simple approach: shuffle the values of the protected attribute during training.
  - will accomplish calibration, but not statistical parity or error rate balance

- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.
- simple approach: shuffle the values of the protected attribute during training.
  - will accomplish calibration, but not statistical parity or error rate balance
- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.

- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.
- simple approach: shuffle the values of the protected attribute during training.
  - will accomplish calibration, but not statistical parity or error rate balance
- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.
  - Then $\text{corr}(\tilde{X}_j, A) = 0$ by construction.

- Disadvantage of post-processing is it requires knowledge of the protected attribute.
  - pre-processing adjusts the dataset so that the attribute is no longer in the inputs.
- simple approach: shuffle the values of the protected attribute during training.
  - will accomplish calibration, but not statistical parity or error rate balance
- E.g., take education, residualize it on muslim:
  - for each predictor j, regress X_j on A, produce $\tilde{X}_j = X_j - \hat{X}_j$, then use $\tilde{X}_j$ in the ML model.
  - Then $\text{corr}(\tilde{X}_j, A) = 0$ by construction.

Problem: Uncorrelated $\neq$ Independent (e.g. Ince et al 2016)



- relations could be non-linear
- could be interactions between predictors, $X_j X_k$, $j \neq k$, correlated with A.
- X_j and A could have an interaction effect on Y.

correlation $\neq$ mutual information

Purging information on the protected class

Goal: remove any dependence between X and A while preserving information in X that is predictive for Y.

Purging information on the protected class

Goal: remove any dependence between X and A while preserving information in X that is predictive for Y.

- Zemel et al (2013), "Learning fair representations" and follow-up papers for sophisticated approach to this problem.

Purging information on the protected class

Goal: remove any dependence between X and A while preserving information in X that is predictive for Y.

- Zemel et al (2013), "Learning fair representations" and follow-up papers for sophisticated approach to this problem.
- Double ML methods would seem to also work, but I have not seen that (potential project idea).
  - Seemingly unrecognized problem: unobserved confounders relating A to X and Y.

Wang et al (adversarial de-biasing approach using gender and images)

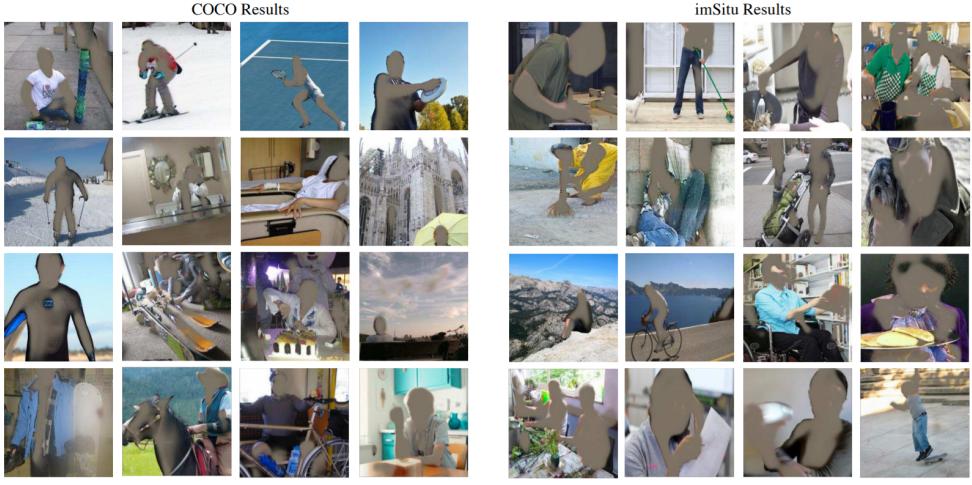


Figure 6. Images after adversarial removal of gender in image space by using a U-Net based autoencoder as inputs to the recognition model. While people are clearly being obscured from the image, the model selectively chooses to obscure only parts that would reveal gender such as faces but tries to keep information that is useful to recognize objects or verbs. 1st row: WWWM MMWW; 2nd row: MWWW WMWW; 3rd row: MMMW MMWM; 4th row: MMMW WWMM. W: woman; M: man.

Outline

The "Reductions" approach (Agarwal 2018)

Minimize model loss, subject to the expected outcome being similar across groups (beneath some threshold ϵ):

$$\min_{\theta} L(\theta)$$
$$\text{s.t.} |\mathbb{E}(\hat{Y}|\text{white}) - \mathbb{E}(\hat{Y}|\text{black})| \leq \epsilon$$

The "Reductions" approach (Agarwal 2018)

Minimize model loss, subject to the expected outcome being similar across groups (beneath some threshold ϵ):

$$\min_{\theta} L(\theta)$$
$$\text{s.t.} |\mathbb{E}(\hat{Y}|\text{white}) - \mathbb{E}(\hat{Y}|\text{black})| \leq \epsilon$$

- Reductions Approach: solve a series of cost-sensitive classification problems using off-the-shelf methods.
  - also works for error rate balance (but not predictive parity)
  - see paper for details.

The "Reductions" approach (Agarwal 2018)

Minimize model loss, subject to the expected outcome being similar across groups (beneath some threshold ϵ):

$$\min_{\theta} L(\theta)$$
$$\text{s.t.} |\mathbb{E}(\hat{Y}|\text{white}) - \mathbb{E}(\hat{Y}|\text{black})| \leq \epsilon$$

- Reductions Approach: solve a series of cost-sensitive classification problems using off-the-shelf methods.
  - also works for error rate balance (but not predictive parity)
  - see paper for details.
- in general, there appear to be many approaches and no consensus.

Constrained Optimization with TensorFlow Keras

- The TFCO package in TensorFlow integrates constrained optimization into the training process.
- not that easy to use yet – check out the notebooks linked the syllabus.

Lightning Essay

For last minutes of class:

`https://bit.ly/BRJ-W9-Essay`