ORIE 4741: Learning with Big Messy Data

Fairness

Professor Udell

Operations Research and Information Engineering
Cornell

November 23, 2021

Announcements 11/23/21

- no section this week
- no quiz this week
- rubrics for projects are posted on website
- ORIE 5741 project presentations due soon (Dec 3)
- my in-person OH are discontinued (too cold!)
- come to OH for project help
- happy Thanksgiving!

Poll

fairness is an important consideration in model selection

- yes
- no

Fairness in big data

Q: Why should algorithm designers think about fairness?

Fairness in big data

Q: Why should algorithm designers think about fairness?

- algorithms bias available information
  (search, recommendations, social media)
- algorithms can have big impacts
  (parole, credit)
- avoid unintended (and often unobservable) negative consequences
- legal requirements

Fairness in big data

Q: Why should algorithm designers think about fairness?

- algorithms bias available information (search, recommendations, social media)
- algorithms can have big impacts (parole, credit)
- avoid unintended (and often unobservable) negative consequences
- legal requirements

Important questions for algorithm designers:

- What is the harm of false positives? False negatives?
- How can errors change the data distribution in the future?

Examples

- Credit decisioning: HDMA `https://www.consumerfinance.gov/data-research/hmda/explore`, apple credit card
- Criminal justice: COMPAS `https://github.com/propublica/compas-analysis/`
- Advertising (eg, job openings)
- Hiring `https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-`
- Information feeds and recommender systems
- College admissions
- Medical diagnosis or treatment recommendation

Important questions for algorithm designers:

- What is the harm of false positives? False negatives?
- How can errors change the data distribution in the future?

Fairness: definitions

Fairness vs discrimination

Q: What does it mean for a classifier to be unfair?

- What groups or individuals should be protected from discrimination?
- How can we tell if the algorithm is unfair or discriminatory?
- What information is it permissible for the algorithm to use?

Fairness: legal definitions

Housing: The Equal Housing Opportunity Act states that someone seeking to rent a home has the right to expect that housing will be available to them without discrimination or other limitations based on race, sex, color, religion, sex, disability, familial status, or nationality.

Fairness: legal definitions

Housing: The Equal Housing Opportunity Act states that someone seeking to rent a home has the right to expect that housing will be available to them without discrimination or other limitations based on race, sex, color, religion, sex, disability, familial status, or nationality.

Credit and banking: The Equal Credit Opportunity Act makes discrimination unlawful in credit applications based on race, color, religion, national origin, sex, marital status, age, or because all or part of the applicant's income comes from a public assistance program.

Fairness: legal definitions

Housing: The Equal Housing Opportunity Act states that someone seeking to rent a home has the right to expect that housing will be available to them without discrimination or other limitations based on race, sex, color, religion, sex, disability, familial status, or nationality.

Credit and banking: The Equal Credit Opportunity Act makes discrimination unlawful in credit applications based on race, color, religion, national origin, sex, marital status, age, or because all or part of the applicant's income comes from a public assistance program.

Labor market: The Equal Employment Opportunity Act prohibits employment discrimination in its programs based on race, color, national origin, sex, religion, age, disability, political beliefs, and marital or familial status.

Legal vs algorithmic fairness

Given a supervised learning model (say, a project you're reviewing), could you assess whether it meets the legal requirement of fairness?

- yes
- no

how?

Definitions and notation

Definition

A **protected attribute** is a feature on which discrimination is prohibited
(eg, race, color, national origin, sex, religion, age, disability, political beliefs, and marital or familial status)

Notation: (assume classification problem)

- binary outcomes $y \in \mathcal{Y} = \{0, 1\}$
- covariates $x \in \mathbf{R}^d$
- binary protected attribute $a \in \{0, 1\}$
- prediction $\hat{y}$

Unawareness

Definition

A classifier is **unaware** of the protected attribute if the prediction is independent of the protected attribute given other covariates:

$$\hat{y} \perp a|x \quad \Longleftrightarrow \quad \mathbb{P}(\hat{y}|x, a = 0) = \mathbb{P}(\hat{y}|x, a = 1).$$

Unawareness

Definition

A classifier is **unaware** of the protected attribute if the prediction is independent of the protected attribute given other covariates:

$$\hat{y} \perp a|x \quad \iff \quad \mathbb{P}(\hat{y}|x, a = 0) = \mathbb{P}(\hat{y}|x, a = 1).$$

poll: is an unaware classifier always fair?

- yes
- no

Fair Lending laws

Law	FHA	ECOA
age		X
color	X	X
disability	X	
exercised rights under CCPA		X
familial status (household composition)	X	
gender identity	X	
marital status (single or married)		X
national origin	X	X
race	X	X
recipient of public assistance		X
religion	X	X
sex	X	X

Two doctrines of discrimination

- Disparate Treatment
  - *Treatment must not depend explicitly on group*
  - Example: **redlining** – refusing to finance mortgages in minority neighborhoods – violates disparate treatment

- Disparate Impact
  - *Outcomes must not differ (excessively) between groups*
  - Example: **four-fifths rule** (of thumb) says hiring rates of majority and minority should be within 80% of each other

Other quantifications of fairness: [Kleinberg Mullainathan Raghavan 2016], [Hardt Price Srebro 2016], …

Personal Information

FIRST NAME

MI

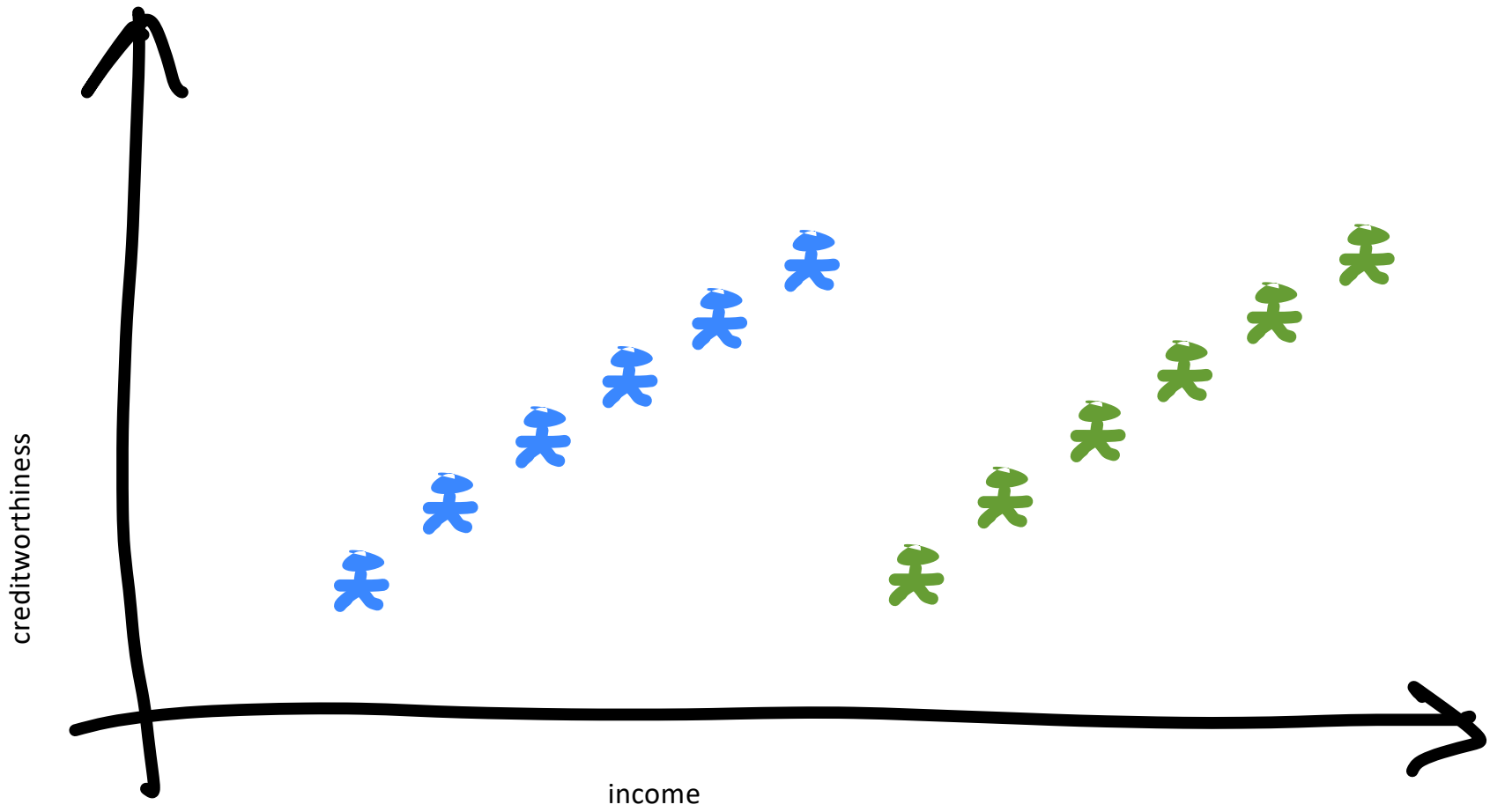
LAST NAME

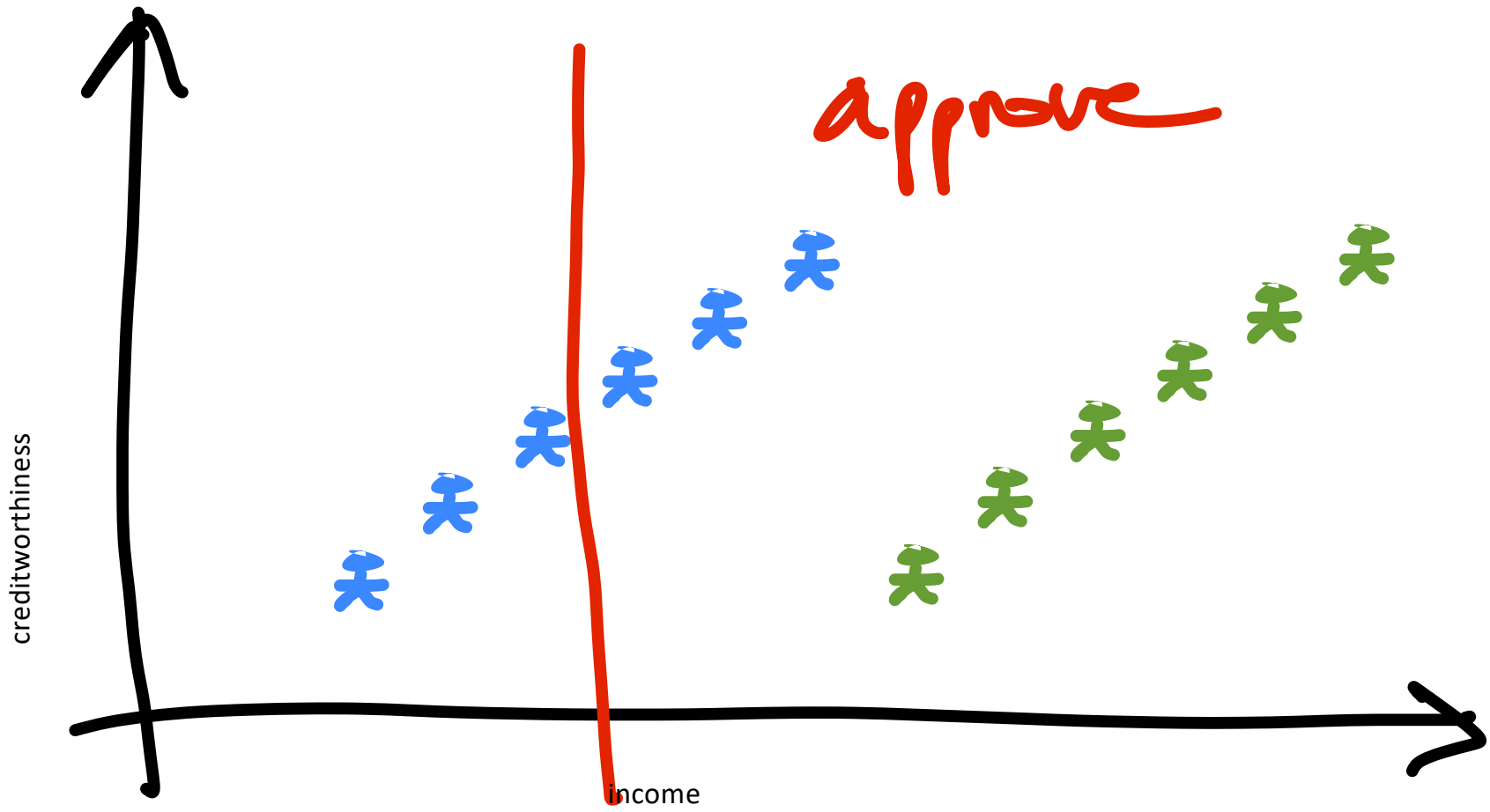
DATE OF BIRTH

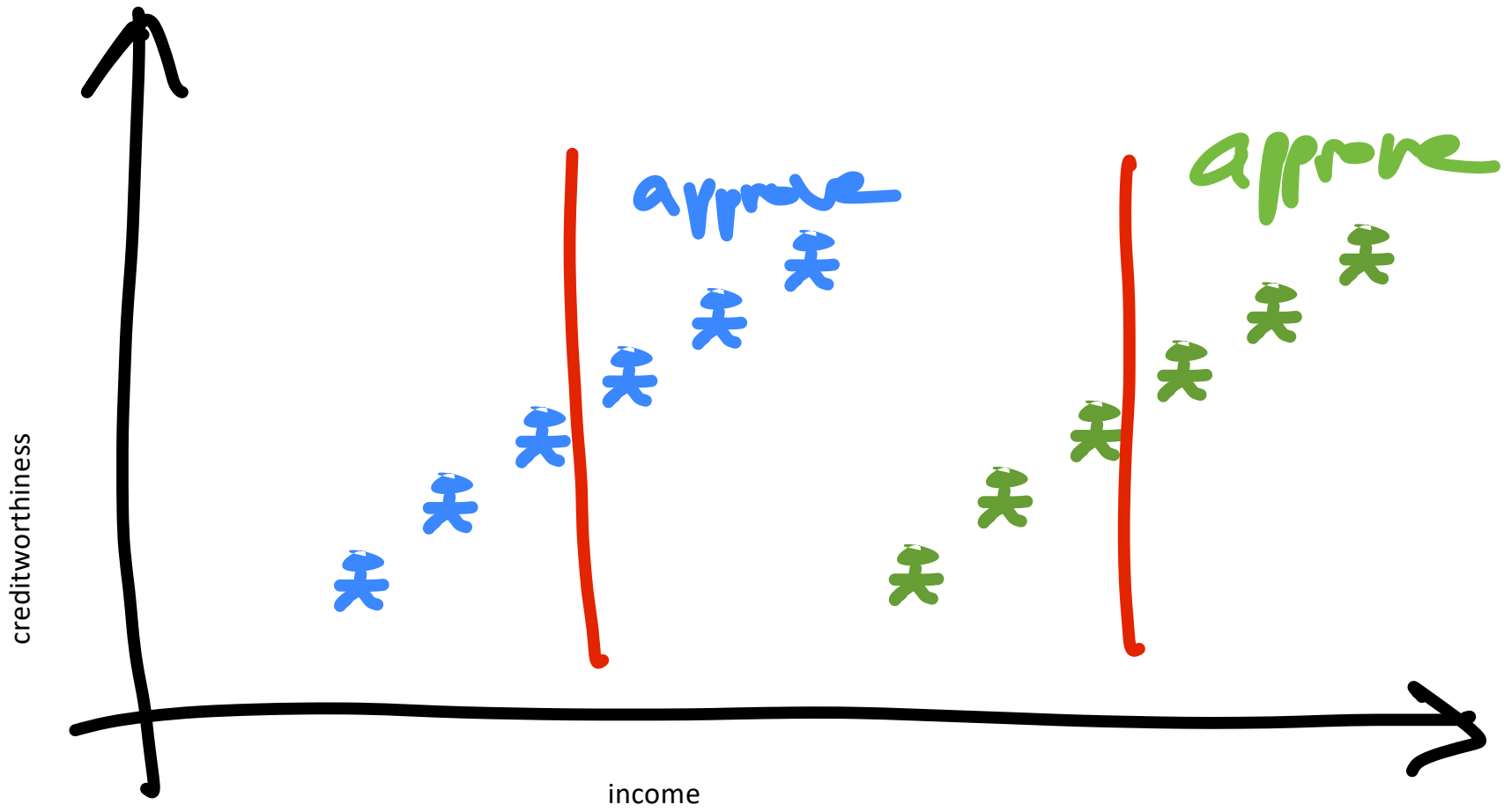
SOCIAL SECURITY NUMBER

ARE YOU A U.S. CITIZEN? YES NO

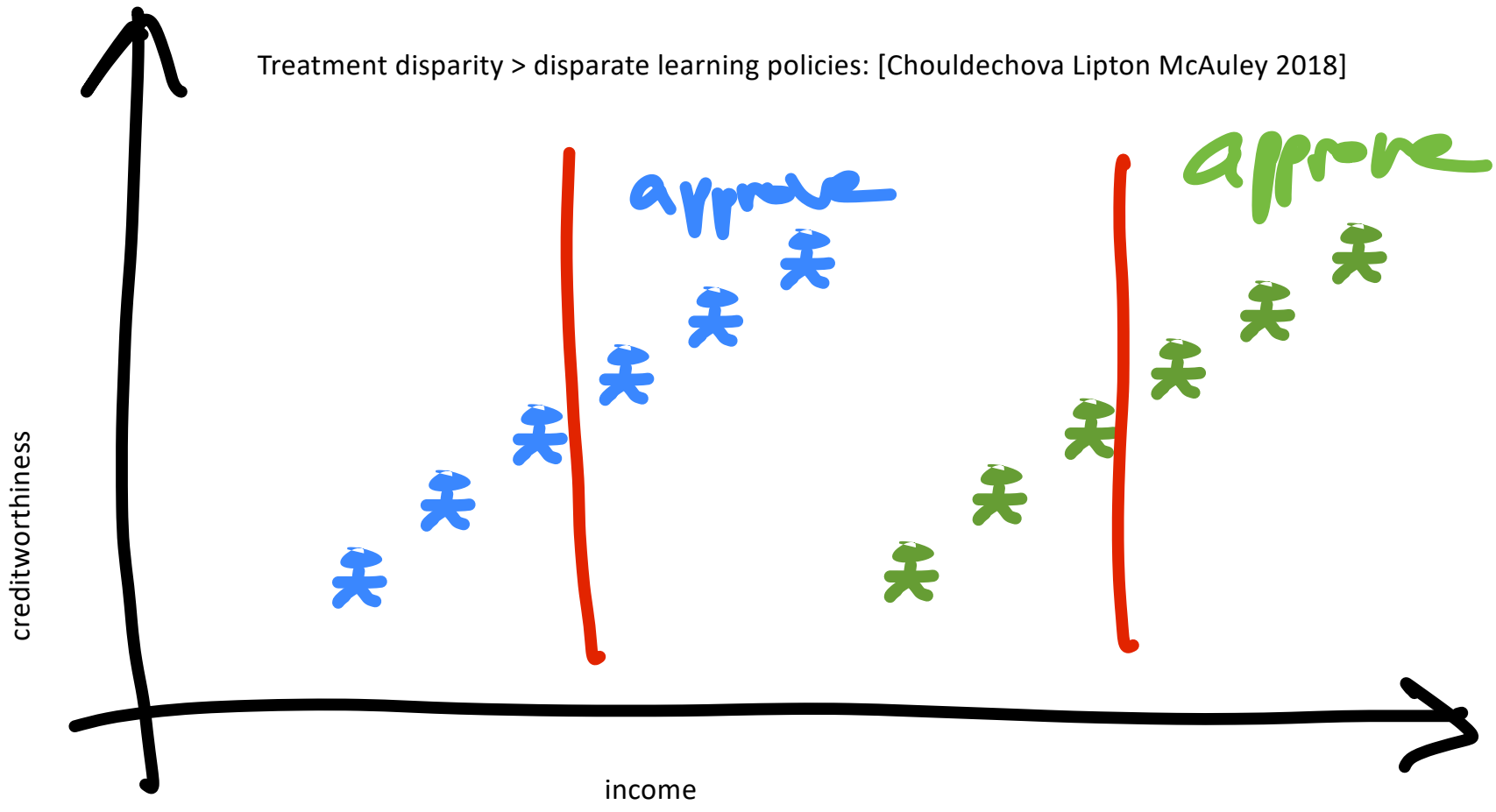
WHY ARE YOU ASKING ME THIS?


creditworthiness
income


approve
creditworthiness
income


creditworthiness
income
approve
approve

Treatment disparity > disparate learning policies: [Chouldechova Lipton McAuley 2018]



What's wrong with fairness through unawareness?

- **Reduced accuracy:** *e.g.*, gender and violent crime.

What's wrong with fairness through unawareness?

- **Reduced accuracy:** *e.g.*, gender and violent crime.
- **Proxies:** Other covariates (called **proxies**) may correlate with the protected attribute. How much correlation is too much? *e.g.*, zip code and race.

What's wrong with fairness through unawareness?

- **Reduced accuracy:** *e.g.*, gender and violent crime.
- **Proxies:** Other covariates (called **proxies**) may correlate with the protected attribute. How much correlation is too much? *e.g.*, zip code and race.
- **Allowable proxies:** *e.g.*, credit score and race; pinterest and gender

Fairness metrics

How can we define fairness? Active research area!

- Unawareness / anti-classification
- Demographic parity
- Equalized odds, Equality of opportunity
- Predictive Rate Parity
- Individual Fairness
- Counterfactual fairness

Demographic parity

Definition

The algorithmic classifier satisfies **demographic parity** if the prediction is independent of the protected attribute:

$$\mathbb{P}(\hat{y}|a = 1) = \mathbb{P}(\hat{y}|a = 0) = \mathbb{P}(\hat{y})$$

In other words, acceptance rates of applicants from both groups are the same.

Demographic parity

Definition

The algorithmic classifier satisfies **demographic parity** if the prediction is independent of the protected attribute:

$$\mathbb{P}(\hat{y}|a = 1) = \mathbb{P}(\hat{y}|a = 0) = \mathbb{P}(\hat{y})$$

In other words, acceptance rates of applicants from both groups are the same.

Q: Problems with demographic parity?

Demographic parity

Definition

The algorithmic classifier satisfies **demographic parity** if the prediction is independent of the protected attribute:

$$\mathbb{P}(\hat{y}|a = 1) = \mathbb{P}(\hat{y}|a = 0) = \mathbb{P}(\hat{y})$$

In other words, acceptance rates of applicants from both groups are the same.

Q: Problems with demographic parity?

- **Accuracy.** What if base rates are different? *e.g.*, gender and violent crime. Demographic parity rules out the perfect predictor when base rates are different.

Demographic parity

Definition

The algorithmic classifier satisfies **demographic parity** if the prediction is independent of the protected attribute:

$$\mathbb{P}(\hat{y}|a=1) = \mathbb{P}(\hat{y}|a=0) = \mathbb{P}(\hat{y})$$

In other words, acceptance rates of applicants from both groups are the same.

Q: Problems with demographic parity?

- **Accuracy.** What if base rates are different? *e.g.*, gender and violent crime. Demographic parity rules out the perfect predictor when base rates are different.
- **Unjust and misleading.** To achieve demographic parity, could admit qualified (eg, college) applicants from group $a = 0$ and random applicants from group $a = 1$. (And then complain: "Students from group 1 just aren't prepared!")

Equalized odds

Definition

The algorithmic classifier satisfies **equalized odds** if the prediction $\hat{y}$ is independent of the protected attribute a conditional on the outcome y.

$$\hat{y} \perp a|y \quad \Longleftrightarrow \quad \mathbb{P}(\hat{y}|y, a = 1) = \mathbb{P}(\hat{y}|y, a = 0)$$

As a consequence, the true positive, true negative, false positive, and false negative rates are the same for both groups.

Equalized odds

Definition

The algorithmic classifier satisfies **equalized odds** if the prediction $\hat{y}$ is independent of the protected attribute a conditional on the outcome y.

$$\hat{y} \perp a|y \quad \Longleftrightarrow \quad \mathbb{P}(\hat{y}|y, a = 1) = \mathbb{P}(\hat{y}|y, a = 0)$$

As a consequence, the true positive, true negative, false positive, and false negative rates are the same for both groups.

Variant: weaker condition, **equality of opportunity**, holds if $\mathbb{P}(\hat{y}|y = 1, a = 1) = \mathbb{P}(\hat{y}|y = 1, a = 0)$.

Equalized odds

Definition

The algorithmic classifier satisfies **equalized odds** if the prediction $\hat{y}$ is independent of the protected attribute a conditional on the outcome y.

$$\hat{y} \perp a|y \quad \iff \quad \mathbb{P}(\hat{y}|y, a = 1) = \mathbb{P}(\hat{y}|y, a = 0)$$

As a consequence, the true positive, true negative, false positive, and false negative rates are the same for both groups.

Variant: weaker condition, **equality of opportunity**, holds if $\mathbb{P}(\hat{y}|y = 1, a = 1) = \mathbb{P}(\hat{y}|y = 1, a = 0)$.

- $+$: optimality compatibility (perfect prediction) is allowed.
- $+$: provides an incentive to reduce errors uniformly in all groups.
- -: can require **artificially increasing misclassifications** in easy-to-classify group

Predictive rate parity

Definition

The algorithmic classifier satisfies **predictive rate parity** if the outcome y is independent of the protected attribute a conditional on the prediction $\hat{y}$.

$$y \perp a|\hat{y} \quad \Longleftrightarrow \quad \mathbb{P}(y|\hat{y}, a = 1) = \mathbb{P}(y|\hat{y}, a = 0)$$

In hiring, this would mean that the score returned from a prediction algorithm should reflect the candidate's real capability of doing this job. It is consistent with the employer's benefit.

Predictive rate parity

Definition

The algorithmic classifier satisfies **predictive rate parity** if the outcome y is independent of the protected attribute a conditional on the prediction $\hat{y}$.

$$y \perp a|\hat{y} \quad \Longleftrightarrow \quad \mathbb{P}(y|\hat{y}, a = 1) = \mathbb{P}(y|\hat{y}, a = 0)$$

In hiring, this would mean that the score returned from a prediction algorithm should reflect the candidate's real capability of doing this job. It is consistent with the employer's benefit.

- +: optimality compatibility (perfect prediction) is allowed.
- +: encourages (equal) error reduction in all groups.
- -: may not close the gap between different groups if true positive rates are quite different.

Impossibility theorem for fairness metrics

Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan.
"Inherent Trade-Offs in the Fair Determination of Risk Scores.":

> *Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously.*

- calibration within groups (strengthens predictive rate parity)
- equalized odds for $y = 1$ (equality of opportunity)
- equalized odds for $y = 0$

Individual fairness

Previous, statistical, notions of fairness judge fairness wrt **group**.

Individual fairness

Previous, statistical, notions of fairness judge fairness wrt **group**.

Q: Can we determine if a classifier is unfair to an **individual**?

Individual fairness

Previous, statistical, notions of fairness judge fairness wrt **group**.

Q: Can we determine if a classifier is unfair to an **individual**?

Definition

The algorithmic classifier satisfies **individual fairness** if similar individuals receive similar (distribution of) predictions.

Individual fairness

Previous, statistical, notions of fairness judge fairness wrt **group**.

Q: Can we determine if a classifier is unfair to an **individual**?

Definition

The algorithmic classifier satisfies **individual fairness** if similar individuals receive similar (distribution of) predictions.

- +: can assess fairness without defining groups
- -: how to define similarity?
- -: only makes sense for randomized predictions

Counterfactual fairness

Definition

The algorithmic classifier satisfies **counterfactual fairness** if flipping the group $a \to 1 - a$ doesn't change the prediction $\hat{y}$.

Counterfactual fairness is an indidual (not group) notion of fairness

Counterfactual fairness

Definition

The algorithmic classifier satisfies **counterfactual fairness** if flipping the group $a \to 1 - a$ doesn't change the prediction $\hat{y}$.

Counterfactual fairness is an indidual (not group) notion of fairness

- $+$: agrees with intuitive notion of fairness
- -: same problems as unawareness...
- -: how to assess given black-box classifier? need randomized experiments?

Best practice for assessing fairness?

- choose fairness metrics that make sense for your problem
- see how accuracy and fairness metrics change as you tweak your model
- report fairness (just like you'd report test set accuracy)

Exercise: fairness metrics

Pick an application (*e.g.*, parole, credit, admissions, hiring, your project). Which notion of fairness makes sense? Which have problems?

- Unawareness / anti-classification
- Demographic parity
- Equalized odds, Equality of opportunity
- Predictive Rate Parity
- Individual Fairness
- Counterfactual fairness

References

- Impossibility for fairness:
  `https://arxiv.org/pdf/1609.05807.pdf`
- Fairness under unawareness:
  `https://arxiv.org/abs/1811.11154`