Certifying and Removing Disparate Impact

Sorelle Friedler Haverford College
Carlos Scheidegger University of Arizona
Suresh Venkatasubramanian University of Utah

Motivation

Data mining algorithms can replicate bias.

The algorithms are complex and opaque.

Legally, it's hard to challenge such algorithms.

Our goals:

- **Certification**: Help the law determine if such processes are fair.
- **Repair**: Make it possible to pass such tests.

Motivation

Disparate Treatment

- Blatant explicit discrimination [2]
- Direct discrimination [3]
- Reverse tokenism [2]

Disparate Impact

- Redlining [1,2,3]
- Discrimination based on redundant encoding [2]
- Indirect discrimination [3]
- Negative legacy [3]
- Underestimation [3]
- Self-fulfilling prophecy [2]

1: Calders and Verwer
2: Dwork, Hardt, Pitassi, Reingold, and Zemel
3: Kamishima, Akaho, Asoh, and Sakuma

Disparate Impact

Griggs v. Duke Power Co,

US Supreme Court, 1971

"tests or criteria for employment or promotion may not provide equality of opportunity merely in the sense of the fabled offer of milk to the stork and the fox"

EEOC 80% rule:

$$\frac{\text{percent of the minority group hired}}{\text{percent of the majority group hired}} \geq .8$$

Disparate Impact

EEOC 80% rule:

$$\frac{\text{percent of the minority group hired}}{\text{percent of the majority group hired}} \geq .8$$

$$\frac{(FP / (FP + TN)}{(TP / (TP + FN)} \geq .8$$

$$\frac{P(YES \mid minority)}{P(YES \mid majority)} \geq .8$$

	Classified Positive	**Classified Negative**
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Advantage: less sensitive to class imbalance

Our Trust Model

Given a data set D = (X,Y,C) with

- X: protected attributes
- Y: remaining attributes
- C: class outcome to predict

Goal: **Any** algorithm $g: Y \rightarrow C$ has no disparate impact

- we don't know the algorithm (complex or proprietary)
- we trust that the algorithm won't directly use X
- we are only given D

Certification of Disparate Impact

Main Theorem: **predictability if and only if disparate impact**

Predictability: There exists a function $f : Y \rightarrow X$ such that

$$BER(f(y), x) \leq \epsilon$$

Disparate Impact: For a classifier $g : Y \rightarrow C$

$$P(YES \mid \texttt{minority}) / P(YES \mid \texttt{majority}) \leq .8$$

Balanced Error Rate (BER):

$$\frac{\text{P(NO | majority) + P(YES | minority)}}{2}$$

Certification of Disparate Impact

Main Theorem: **predictability if and only if disparate impact**

Predictability: There exists a function $f: Y \to X$ such that

$$BER(f(y), x) \leq \epsilon$$

Disparate Impact: For a classifier $g: Y \to C$

$$P(YES \mid \texttt{minority}) / P(YES \mid \texttt{majority}) \leq .8$$

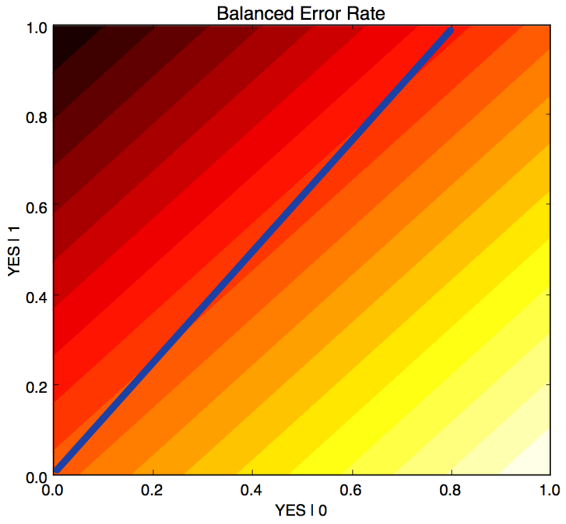
Main proof idea:

create purely biased classifier $\varphi: C \to X$

$$\varphi(YES) = \texttt{majority}$$

$$\varphi(NO) = \texttt{minority}$$

f has the same confusion matrix as g

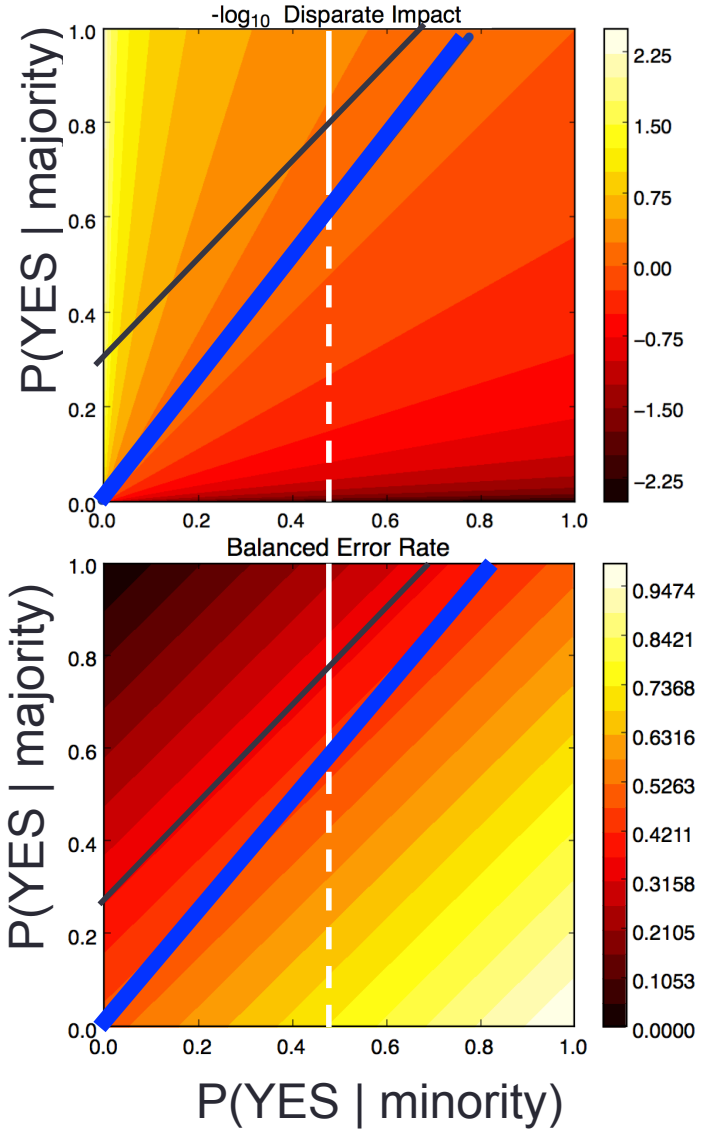


Certification of Disparate Impact

Goal: show that a data set *could* generate disparate impact.

Sketch:

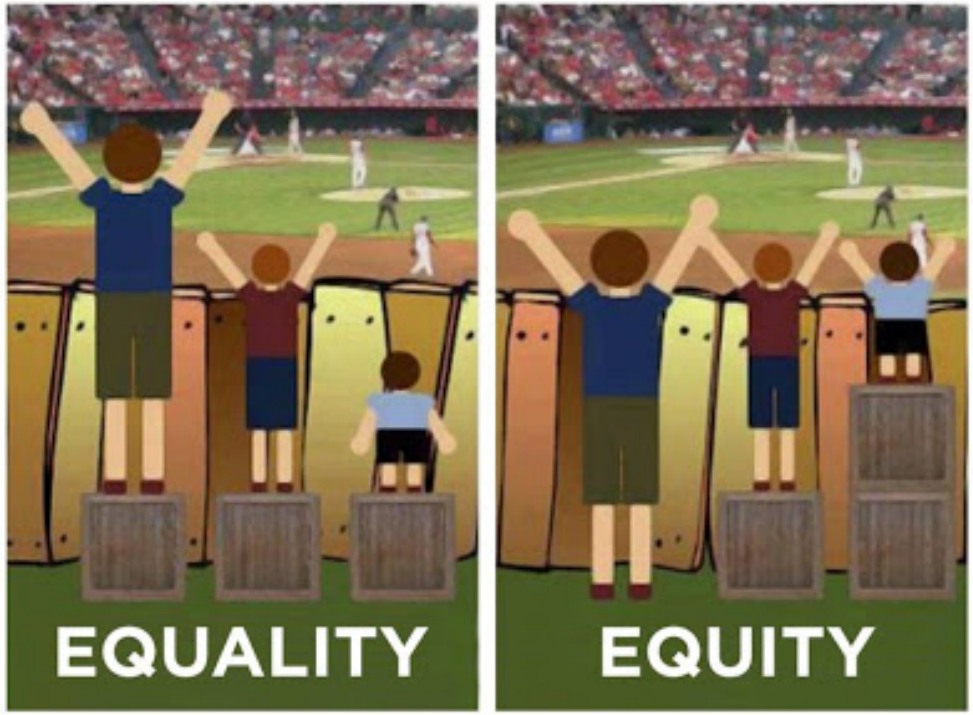
- Try to predict X from Y.
- Use a classifier that optimizes for BER.
- From the previous theorem, determine required BER for lack of disparate impact.
- If the BER is low enough, certify disparate impact.



Removal of Disparate Impact

Goals:

- preserve rank within each marginal distribution $P(Y| X = x)$
- generate a data set $D=(X,Y_{new},C)$ with no disparate impact

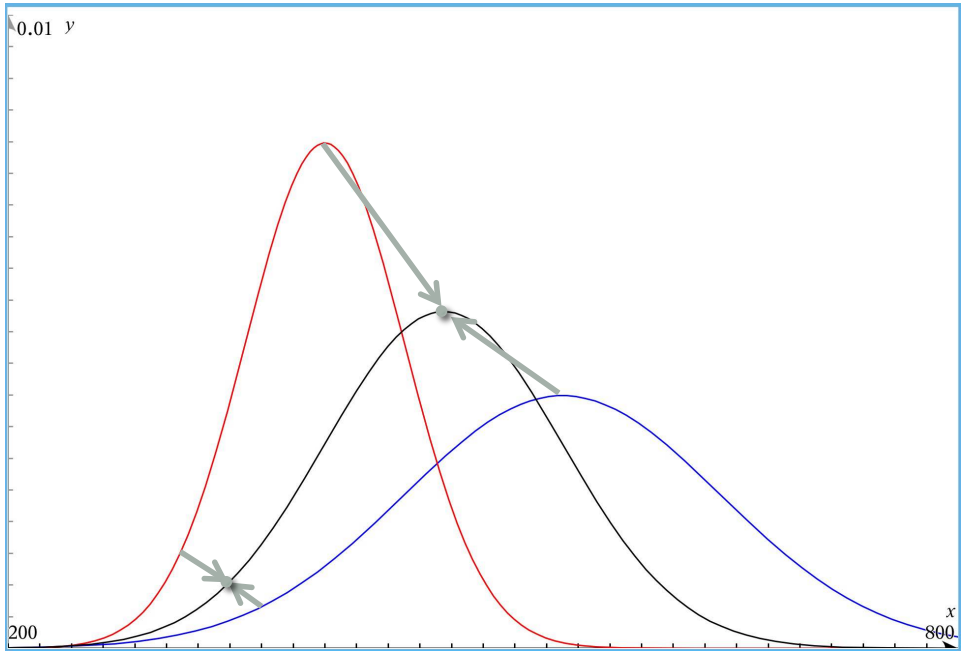


Note: This image was adapted by the City of Portland Office of Equity and Human Rights from the original graphic: http://indianfunnypicture.com/img/2013/01/Equality-Doesnt-Means-Justice-Facebook-Pics.jpg

Removal of Disparate Impact

Goals:

- preserve rank within each marginal distribution P(Y| X = x)
- generate a data set D=(X,Y_{new},C) with no disparate impact



This is like the Texas Top 10% Rule!

Further Discussion

CS Theory:

- Despite the law, do we believe disparate impact is the correct measure of fairness? When does it fall short?
- What is the tradeoff between fairness and utility when fairness is measured by disparate impact?
- Is our repair optimal given this tradeoff?

Legal:

- Could our certification algorithm be useful?
- Is our repair legal?
- How can we support liability and avoid business justifications for disparate impact?

THANKS!

full paper and repair implementation available at
fairness.haverford.edu