Discrimination in Machine Decision Making

Krishna P. Gummadi

Max Planck Institute for Software Systems

Machine decision making

- Refers to data-driven algorithmic decision making
  - By learning over data about past decisions
- To assist or replace human decision making
- Increasingly being used in several domains
  - Recruiting: Screening job applications
  - Banking: Credit ratings / loan approvals
  - Judiciary: Recidivism risk assessments
  - Welfare: Welfare benefit eligibility
  - Journalism: News recommender systems

The talk: Focuses on discrimination

- Discrimination is a specific type of unfairness
- Well-studied in social sciences
  - Political science
  - Moral philosophy
  - Economics
  - Law
    - Majority of countries have anti-discrimination laws
    - Discrimination recognized in several international human rights laws
- But, less-studied from a computational perspective

Part 1:

Why is a computational perspective on discrimination needed?

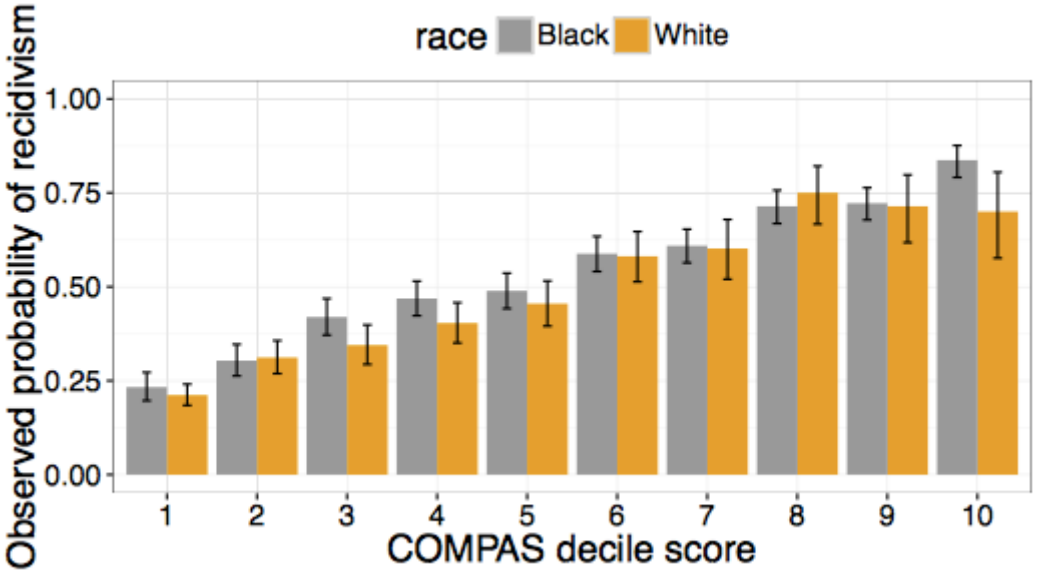
Case study: Recidivism risk prediction

- COMPAS recidivism prediction tool
  - Built by a commercial company, Northpointe, Inc.

- Estimates likelihood of criminals re-offending in future
  - Inputs: Based on a long questionnaire
  - Outputs: Used across US by judges and parole officers

- Are COMPAS' estimates fair to salient social groups?

Is COMPAS fair to all groups?



- Northpointe: In each estimated risk level, false discovery rates for blacks & whites are similar
- So **YES!**

Is COMPAS fair to all groups?

Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369
FP rate: 44.85		
FN rate: 27.99		

White Defendants	Low	High
Survived	1139	349
Recidivated	461	505
FP rate: 23.45		
FN rate: 47.72		

- ProPublica: False positive & false negative rates are considerably worse for blacks than whites
- So **NO!**

Who is right about COMPAS?

- **Both!** Depends on how you measure fairness!
- How many fairness measures can one define?
  - How many different error rate measures can one define?

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

But, aren't the measures similar?

- **NO!** They present inherent trade-offs!
- When base recidivism rates for blacks & whites differ, no non-trivial solution to achieve similar FPR, FNR, FDR, FOR!
- No non-trivial solution can be simultaneously fair according to both ProPublica & Northpointe analyses!

Why, a computational perspective?

- Formal interpretations of discrimination can help us understand the notions better

- Reveals the inherent trade-offs between multiple measures of discrimination and their utility

Part 2:

Mechanisms for Non-discriminatory Machine Learning *[WWW '17]*

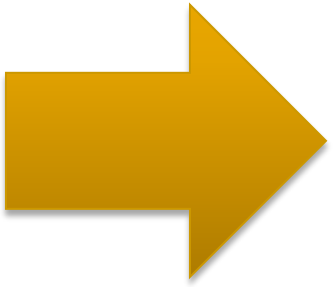
Can machines even discriminate?

- Aren't machine decisions inherently objective?
  - Don't algorithms simply process information?
  - Don't people with same features get the same treatment?
- In contrast to subjective human decisions
- Doesn't that make them fair & non-discriminatory?
- Objective decisions can be objectively unfair & discriminatory!

How machines learn

- By training over historical data
- Example task: Predict who will return loan

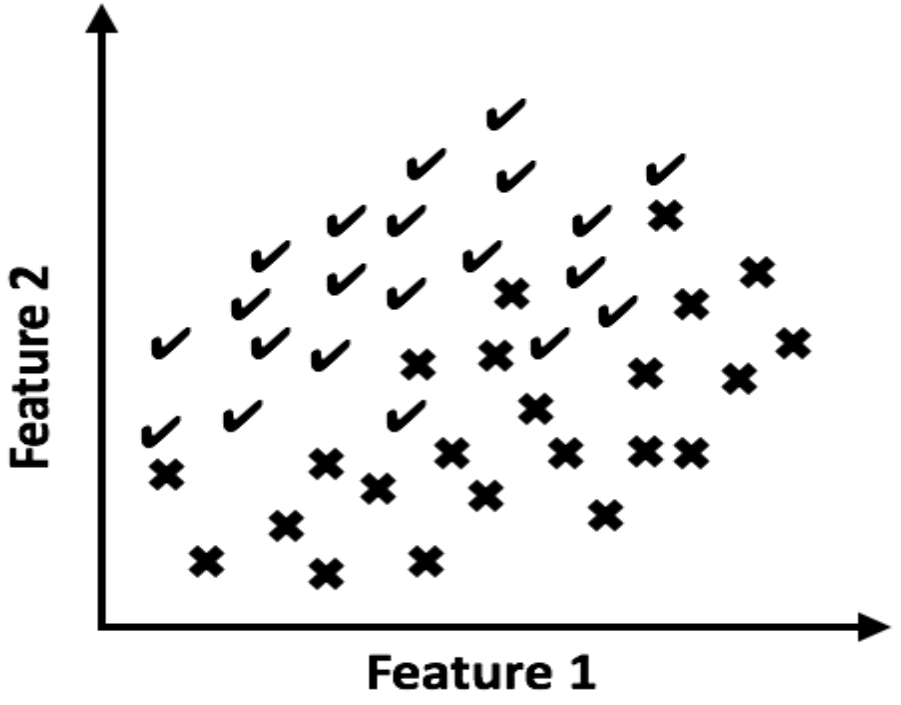
	F_1	F_2	…	F_m	Z
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	Z_1
$User_2$	$x_{2,1}$			$x_{2,m}$	Z_2
$User_3$	$x_{3,1}$			$x_{3,m}$	Z_3
…	…			…	…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$	Z_n



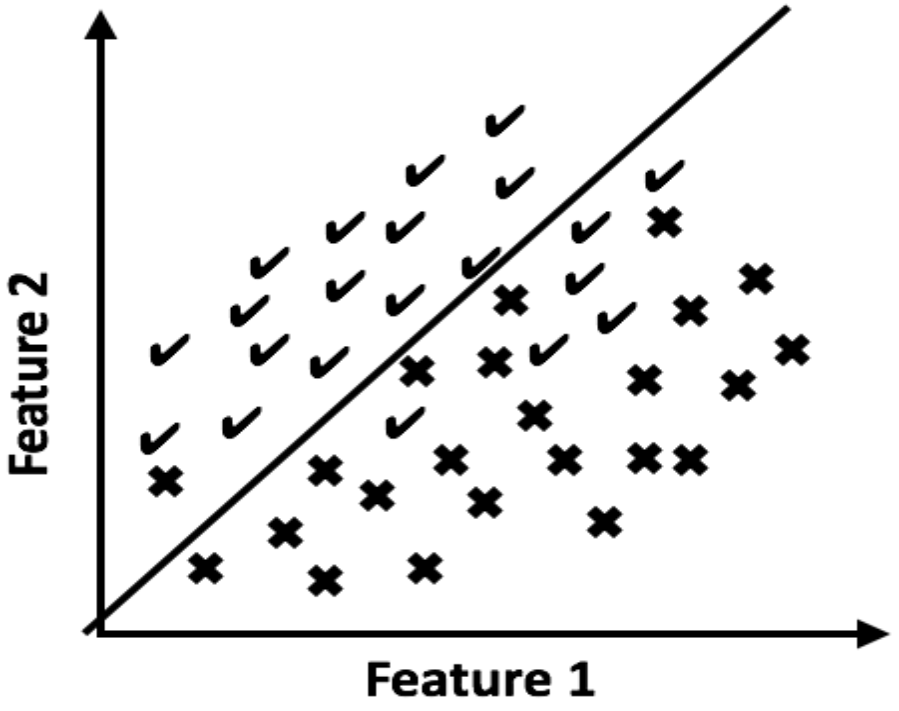
Past Outcomes
Returned
Defaulted
Defaulted
…
Returned

- Learning challenge: Learn a decision boundary (**W**) in the feature space separating the two classes

Predict who will return loans



Predict who will return loans



- Optimal (most accurate / least loss) linear boundary
- But, how do machines find (compute) it?

Learning (computing) the optimal boundary

- Define & optimize a loss (accuracy) function
  - The loss function captures inaccuracy in prediction

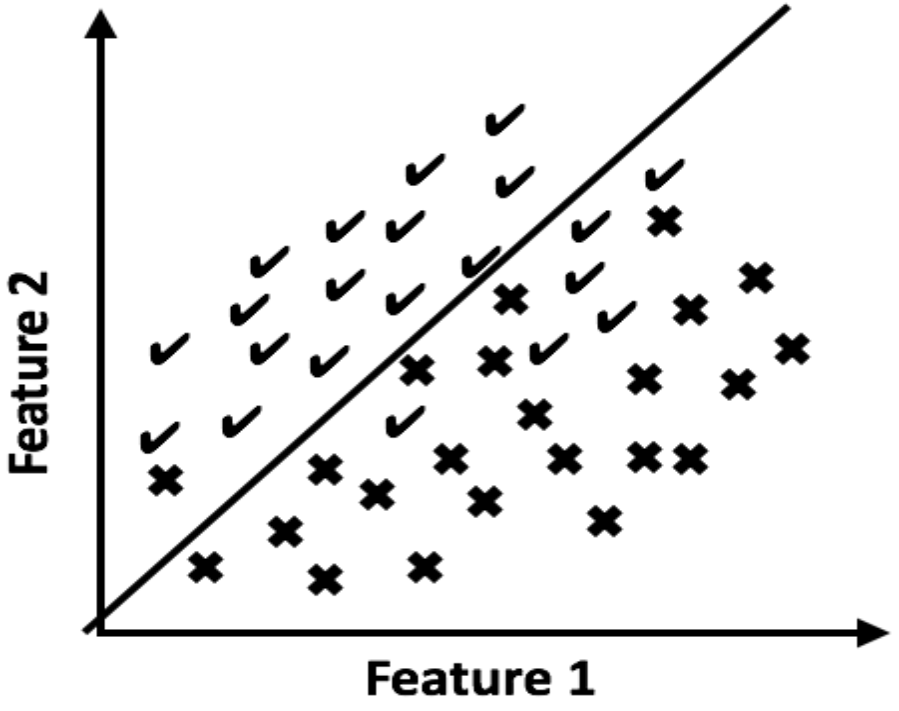
$$L(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \qquad L(\mathbf{w}) = \sum_{i=1}^{N} -\log p(y_i | \mathbf{x}_i, \mathbf{w})$$

  - Minimize (optimize) it over all examples in training data

$$minimize\ L(\mathbf{w})$$

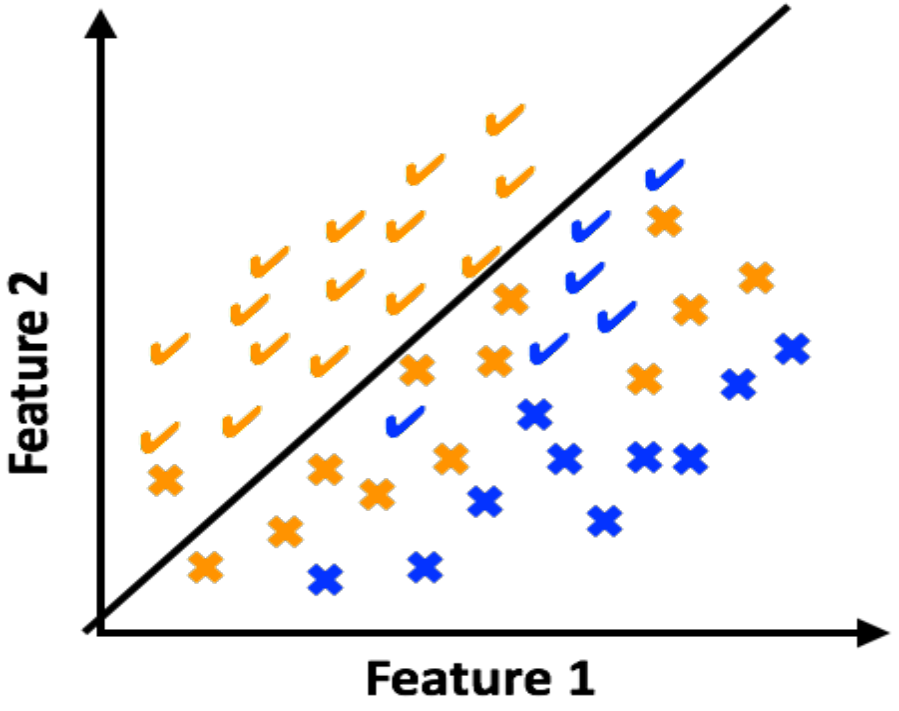
- Central challenge in machine learning
  - Finding loss function that capture prediction loss, yet be efficiently optimized
  - Many loss functions used in learning are convex

Predict who will return loans



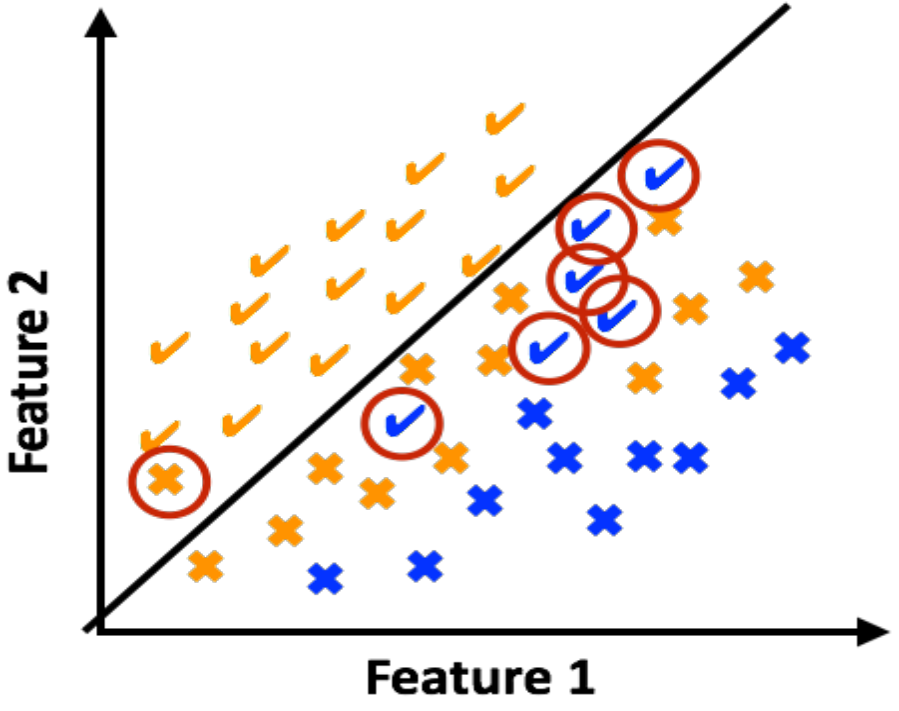
- Optimal (most accurate / least loss) linear boundary
- But, how do machines find (compute) it?
  - The boundary was computed using $\min \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$

How machines learn to discriminate



- Optimal (most accurate / least loss) linear boundary

How machines learn to discriminate



- Optimal (most accurate / least loss) linear boundary
- Makes few errors for yellow, lots of errors for blue!
  - Commits disparate mistreatment: $P(\hat{y} \neq y | z = 0) \neq P(\hat{y} \neq y | z = 1)$

How to learn to avoid discrimination

- Specify discrimination measures as constraints on learning
- Optimize for accuracy under those constraints

$$\begin{aligned} &minimize\ L(\mathbf{w}) \\ &subject\ to \quad P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1) \end{aligned}$$

- The constraints embed ethics & values when learning

- No free lunch: Additional constraints lower accuracy
  - Tradeoff between performance & ethics (avoid discrimination)

Key technical challenge

- How to learn efficiently under these constraints?

$$\begin{aligned} &\textit{minimize } L(\mathbf{w}) \\ &\textit{subject to } \quad P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1) \end{aligned}$$

- Problem: The above formulations are not convex!
  - Can't learn them efficiently

- Need to find a better way to specify the constraints
  - So that loss function under constraints remains convex

Learning classifiers w/o disparate mistreatment

- Previous formulation: Non-convex, hard-to-learn

$$\begin{aligned} &minimize\ L(\mathbf{w}) \\ &subject\ to\ \ P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1) \end{aligned}$$

Learning classifiers w/o disparate mistreatment

- New formulation: Convex-concave, can learn efficiently using convex-concave programming

$$
\begin{array}{l|l}
minimize & L(\mathbf{w}) \\
subject\ to & \frac{-N_1}{N}\sum_{i=1}^{N_0} g_{\mathbf{w}}(y_i,\mathbf{x}_i) + \frac{N_0}{N}\sum_{i=1}^{N_1} g_{\mathbf{w}}(y_i,\mathbf{x}_i) \leq \mathbf{c} \\
 & \frac{-N_1}{N}\sum_{i=1}^{N_0} g_{\mathbf{w}}(y_i,\mathbf{x}_i) + \frac{N_0}{N}\sum_{i=1}^{N_1} g_{\mathbf{w}}(y_i,\mathbf{x}_i) \geq -\mathbf{c}
\end{array}
$$

All misclassifications $g_{\mathbf{w}}(y,\mathbf{x}) = min(0, yd_{\mathbf{w}}(\mathbf{x})),$

False positives $g_{\mathbf{w}}(y,\mathbf{x}) = min\left(0, \frac{1+y}{2} yd_{\mathbf{w}}(\mathbf{x})\right),$ or

False negatives $g_{\mathbf{w}}(y,\mathbf{x}) = min\left(0, \frac{1-y}{2} yd_{\mathbf{w}}(\mathbf{x})\right),$

Evaluation: Recidivism risk estimates

- **Recidivism:** To re-offend within a certain time

- COMPAS risk assessment tool
  - Assign recidivism risk score to a criminal defendant
  - Score used to advise judges' decision

- ProPublica gathered COMPAS assessments
  - Broward Country, FL for 2013-14
  - Features: arrest charge, #prior offenses, age,…
  - Class label: 2-year recidivism

Key evaluation questions

- Do traditional classifiers suffer disparate mistreatment?
- Can our approach help avoid disparate mistreatment?

Disparity in mistreatment

- Trained logistic regression for recidivism prediction

Race	FPR	FNR
Black	34%	32%
White	15%	55%

- **False positive:** Non-recidivating person wrongly classified as recidivating

- **False negative:** Recidivating person wrongly classified as non-recidivating

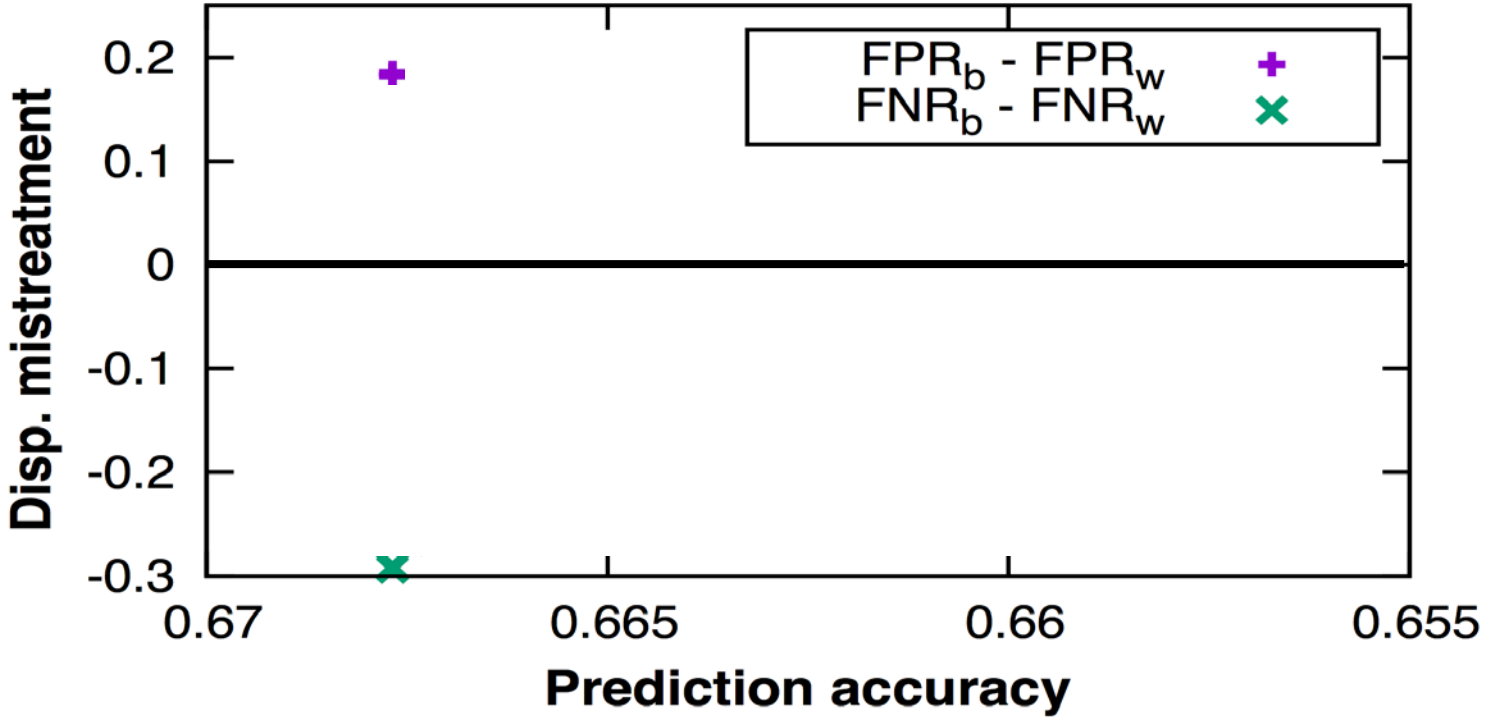
Key evaluation questions

- Do traditional classifiers suffer disparate mistreatment?
  - Yes! Considerable disparity in both FPR and FNR

- Can our approach help avoid disparate mistreatment?

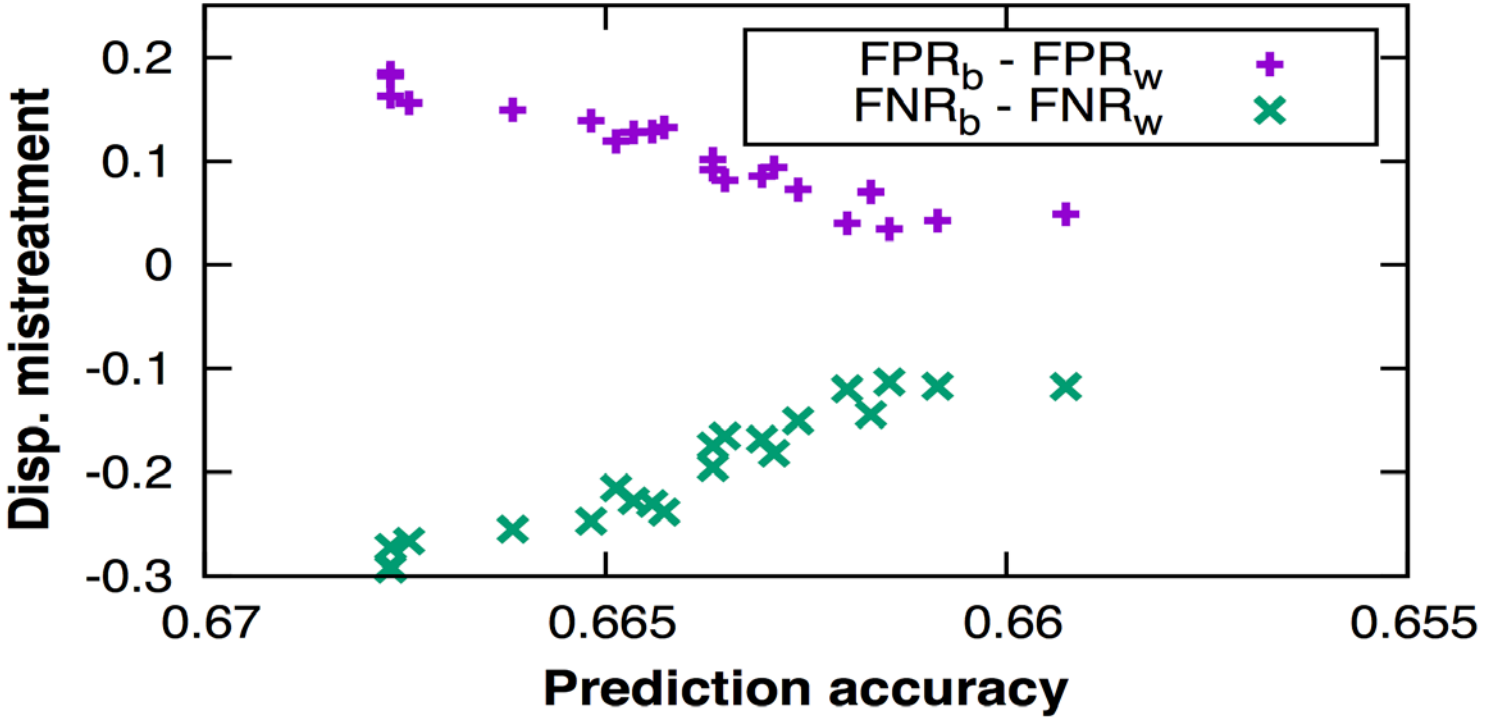
Removing disparate mistreatment

- Traditional classifiers without constraints



Removing disparate mistreatment

- Introducing our FPR and FNR Constraints



Key evaluation questions

- Do traditional classifiers suffer disparate mistreatment?
  - Yes! Considerable disparity in both FPR and FNR
- Can our approach help avoid disparate mistreatment?
  - Yes! For a small loss in accuracy

Summary: Discrimination through computational lens

- Defined a measure of discrimination
  - Disparate mistreatment
  - There exist other measures: Disparate treatment / impact
    - They are applicable in different contexts
- Proposed mechanisms for mitigating each of them
  - Formulate the measures as constraints on learning
  - Proposed proxy functions that can be efficiently learned

Our works

- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez and Krishna P. Gummadi. *Fairness Constraints: A Mechanism for Fair Classification*. In FAT-ML 2015, AISTATS 2017
- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez and Krishna P. Gummadi. *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment*. In FAT-ML 2016, WWW 2017
- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. *From Parity to Preference-based Notions of Fairness for Classification*. In FAT-ML 2017, NIPS 2017
- Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi and Adrian Weller. *The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making*. In NIPS Symposium on ML and the Law, 2016, AAAI 2017.

Fair classifier implementation at:

fate-computing.mpi-sws.org