Discrimination in Machine Learning:
From Principles to Measures and Mechanisms

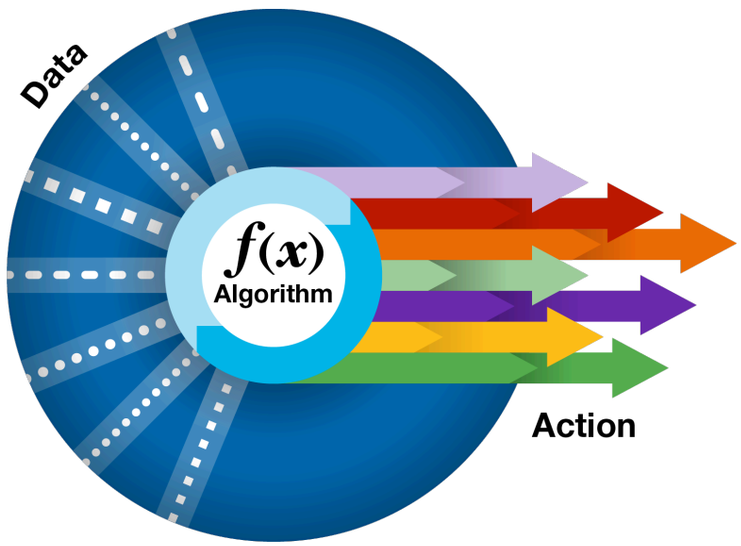
Muhammad Bilal Zafar

joint work with

Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

Algorithmic decision making in practice



Algorithms being used to assist or replace human-decision making

- Algorithmic decision making used in several domains
  - Banking: Loan approval
  - Recruiting: Filtering and ranking applicants
  - Judiciary: Bail decisions

- Learn from historical training data

Advantages of algorithmic decision making

Advantages of algorithmic decision making

An algorithm can predict human behavior better than humans

Advantages of algorithmic decision making

An algorithm can predict human behavior better than humans

The Upshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Advantages of algorithmic decision making

An algorithm can predict human behavior better than humans

TheUpshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Even Imperfect Algorithms Can Improve the Criminal Justice System

Discrimination in algorithmic decisions

Discrimination in algorithmic decisions

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Discrimination in algorithmic decisions

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Googling Politics

How the Google issue guide on candidates is biased.

Discrimination in algorithmic decisions

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Googling Politics

How the Google issue guide on candidates is biased.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

What constitutes discrimination?

[…] wrongful imposition of relative disadvantage on persons based on their membership in a salient social group e.g., gender, race

[Altman'16]

What constitutes discrimination?

[…] wrongful imposition of relative disadvantage on persons based on their membership in a salient social group e.g., gender, race

[Altman'16]

Challenge # 1: Measures

- How do we measure wrongful relative disadvantage?
  - Fuzzy notion, not readily measurable
  - Existing measures may be insufficient

What constitutes discrimination?

[…] wrongful imposition of relative disadvantage on persons based on their membership in a salient social group e.g., gender, race

[Altman'16]

Challenge # 2: Mechanisms

- How to incorporate nondiscrimination into algorithmic decision making?
- Algorithmic decision system training
  - Optimize prediction accuracy
  - Enabling efficient training (crucial for large training datasets)

Part 1

Computational Measures for Nondiscrimination

[WWW'17]

Recap: What constitutes discrimination?

[…] wrongful imposition of relative disadvantage on persons based on their membership in a salient social group e.g., gender, race

[Altman'16]

Toy example: University admission

Non sensitive features (**x**) | Sensitive (z) | Outcome (y) | Predictions (ŷ)

	F_1	F_2	...	F_m	**z**	Past decision
U_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,m}$	z_1	✔
U_2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,m}$	z_2	✖
...	...	...	...	...	...	...
U_n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,m}$	z_n	✔

→

Pred. Decision
✔
✖
...
✔

Gender, race, etc

Relative disadvantage measure 1: Disparate treatment

- Achieve parity (or equality) in treatment
  - Decisions should not change with change in sensitive feature

	z	x		ŷ
		School grade	**SAT score**	**Admit**
Bob				
Alice				

Relative disadvantage measure 1: Disparate treatment

- Achieve parity (or equality) in treatment
  - Decisions should not change with change in sensitive feature

	z	x		ŷ
		School grade	**SAT score**	**Admit**
Bob		90 / 100	700 / 800	
Alice		90 / 100	700 / 800	

Relative disadvantage measure 1: Disparate treatment

- Achieve parity (or equality) in treatment
  - Decisions should not change with change in sensitive feature

	z	x		ŷ
		School grade	**SAT score**	**Admit**
Bob		90 / 100	700 / 800	✔
Alice		90 / 100	700 / 800	✖

Disparate treatment

Disparate treatment: A measure of direct discrimination

$$P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$$

non-sensitive feature (education)

sensitive feature (gender)

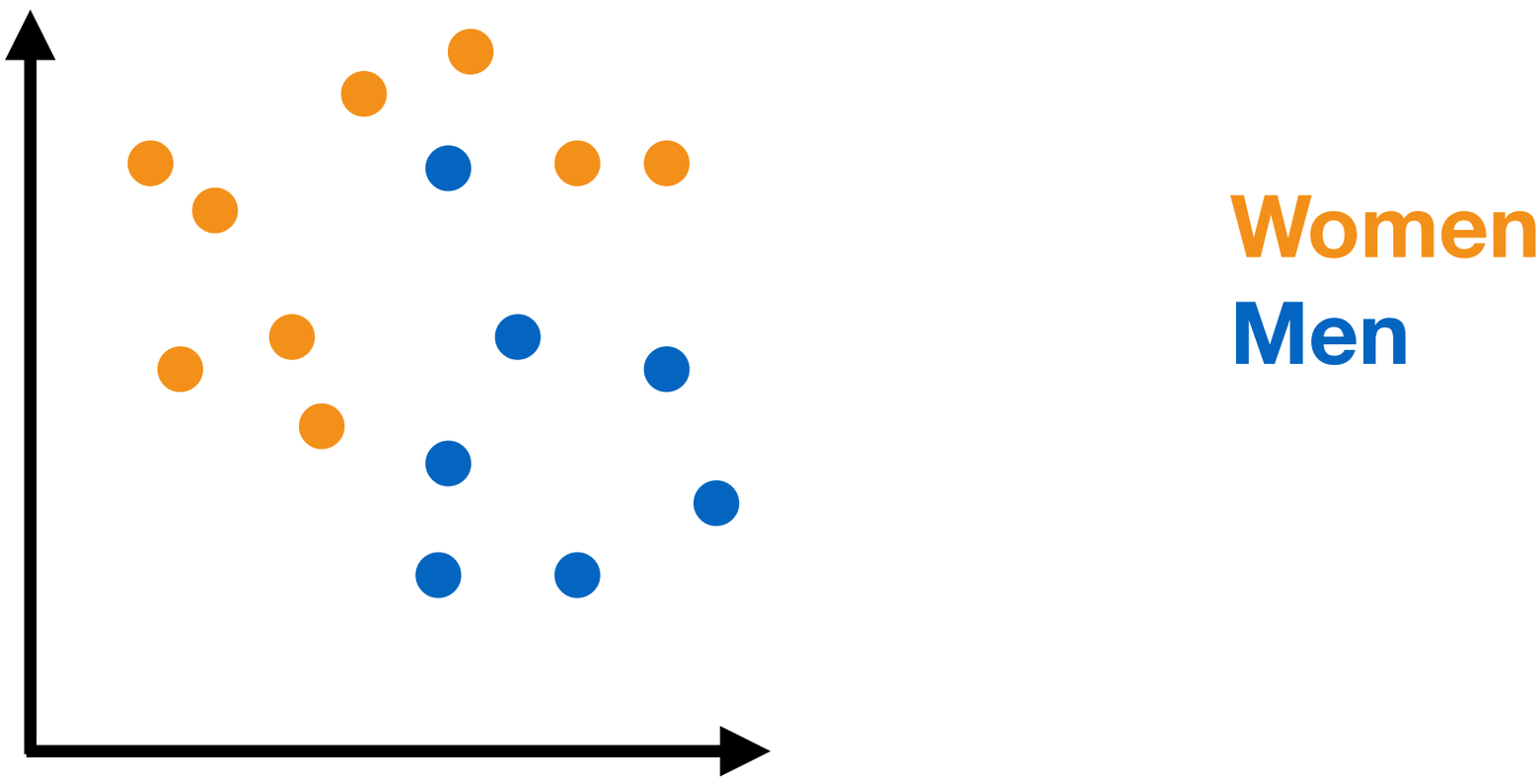
- Wrongful relative disadvantage: Basing decisions on gender, race, etc.
- Most intuitive notion of discrimination
- Focuses on direct or intentional discrimination

Relative disadvantage measure 2: Disparate impact

- Achieve parity (or equality) in impact
  - Positive outcome rates should be same for all groups

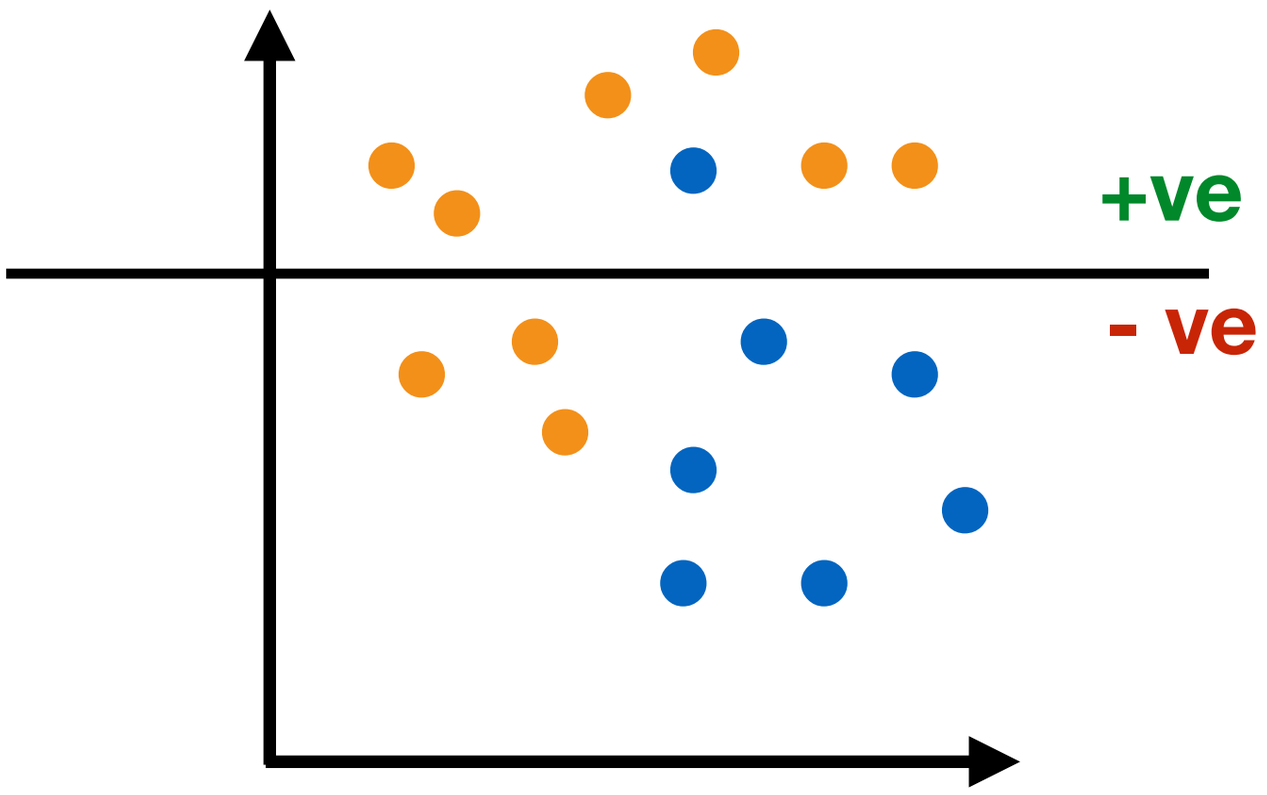
Relative disadvantage measure 2: Disparate impact

- Achieve parity (or equality) in impact
  - Positive outcome rates should be same for all groups



Relative disadvantage measure 2: Disparate impact

- Achieve parity (or equality) in impact
  - Positive outcome rates should be same for all groups



Women: 0.67

Men: 0.14

Disparate impact: A measure of indirect discrimination

$$P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$$

$\hat{y} = 1$ is the desired outcome

- Wrongful relative disadvantage: Disparity in positive outcome rates
- Historical biases in training data (perpetuated biases)

Toy example: University admission

Could be biased!

Non sensitive features (**x**) | Sensitive (z) | Outcome (y) | Predictions ($\hat{y}$)

	F_1	F_2	…	F_m	**z**
U_1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	z_1
U_2	$x_{2,1}$	$x_{2,2}$	…	$x_{2,m}$	z_2
…	…	…	…	…	…
U_n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$	z_n

Past decision
✔
✖
…
✔

→

Pred. Decision
✔
✖
…
✔

Disparate impact: A measure of indirect discrimination

$$P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$$

$\hat{y} = 1$ is the desired outcome

- Wrongful relative disadvantage: Disparity in positive outcome rates
- Historical biases in training data (perpetuated biases)
- Correlations between sensitive and non-sensitive features
  - Correlation between race and zip code
  - Helps where disparate treatment might fail

Until now: 2 existing measures of discrimination

- Disparate treatment: Targets direct discrimination
  - Requires: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$

- Disparate impact: Targets indirect discrimination, when biased historical labels
  - Requires: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$

Until now: 2 existing measures of discrimination

- Disparate treatment: Targets direct discrimination
  - Requires: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$

- Disparate impact: Targets indirect discrimination, when biased historical labels
  - Requires: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$

Unbiased historical labels (Ground truth)

Ground truth → No wrongful relative disadvantage?

Example: Credit risk assessment

Ground truth
(who returned the loan in the past)

Non sensitive features (**x**) | Sensitive (z) | Outcome (y) | Predictions (ŷ)

	F_1	F_2	…	F_m	z
U_1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	z_1
U_2	$x_{2,1}$	$x_{2,2}$	…	$x_{2,m}$	z_2
…	…	…	…	…	…
U_n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$	z_n

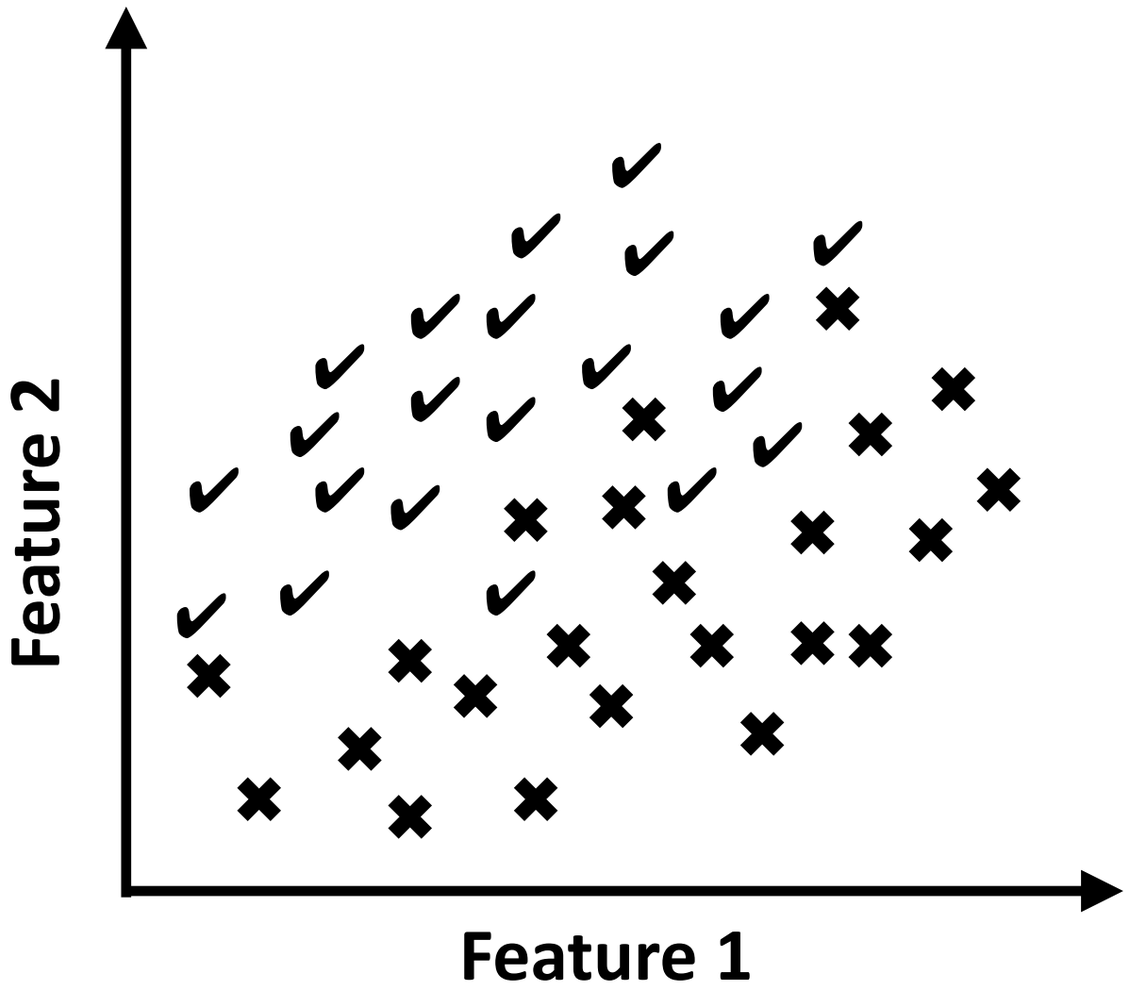
Returned loan
✔
✖
…
✔

→

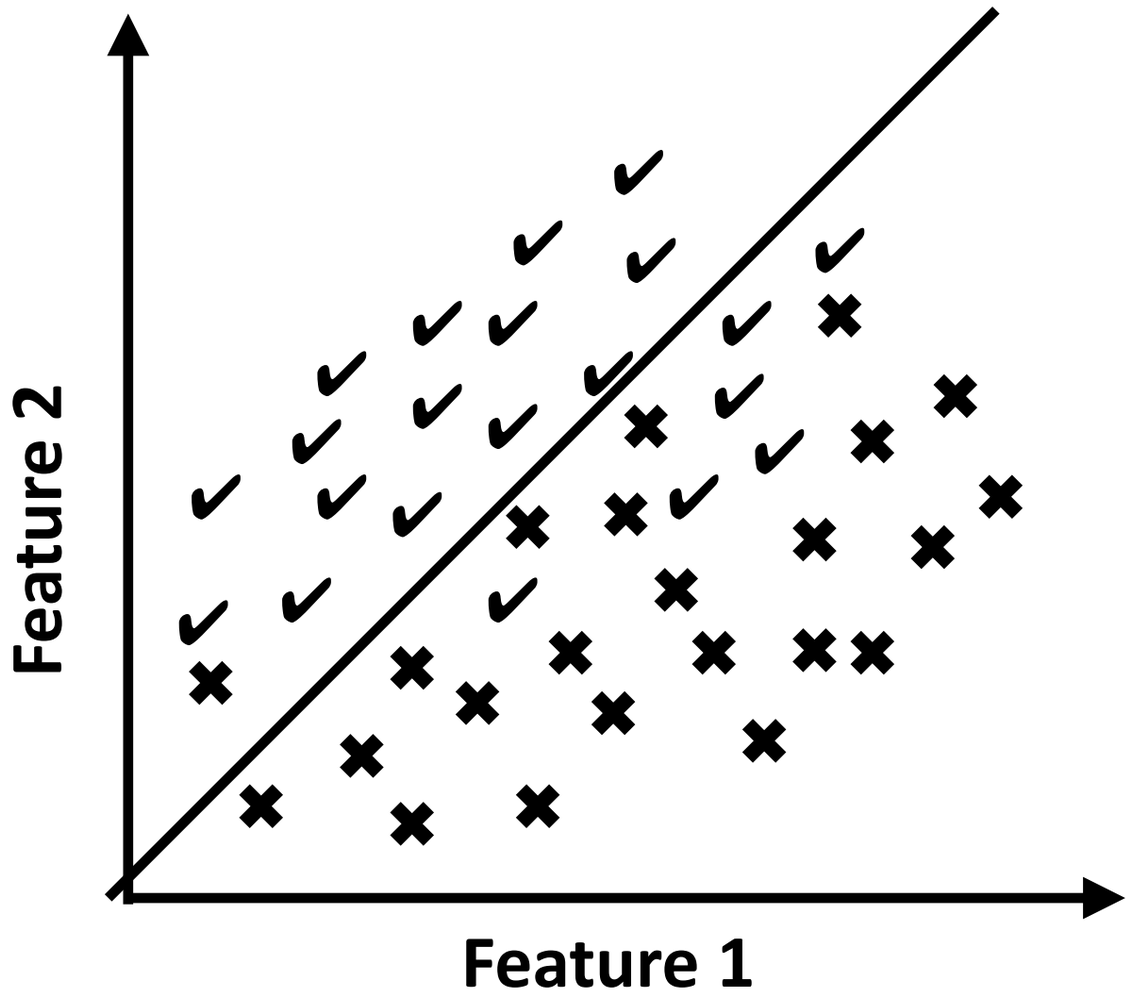
Pred. Decision
✔
✖
…
✔

No wrongful relative disadvantage?

A fictitious dataset: Predict who will return the loan



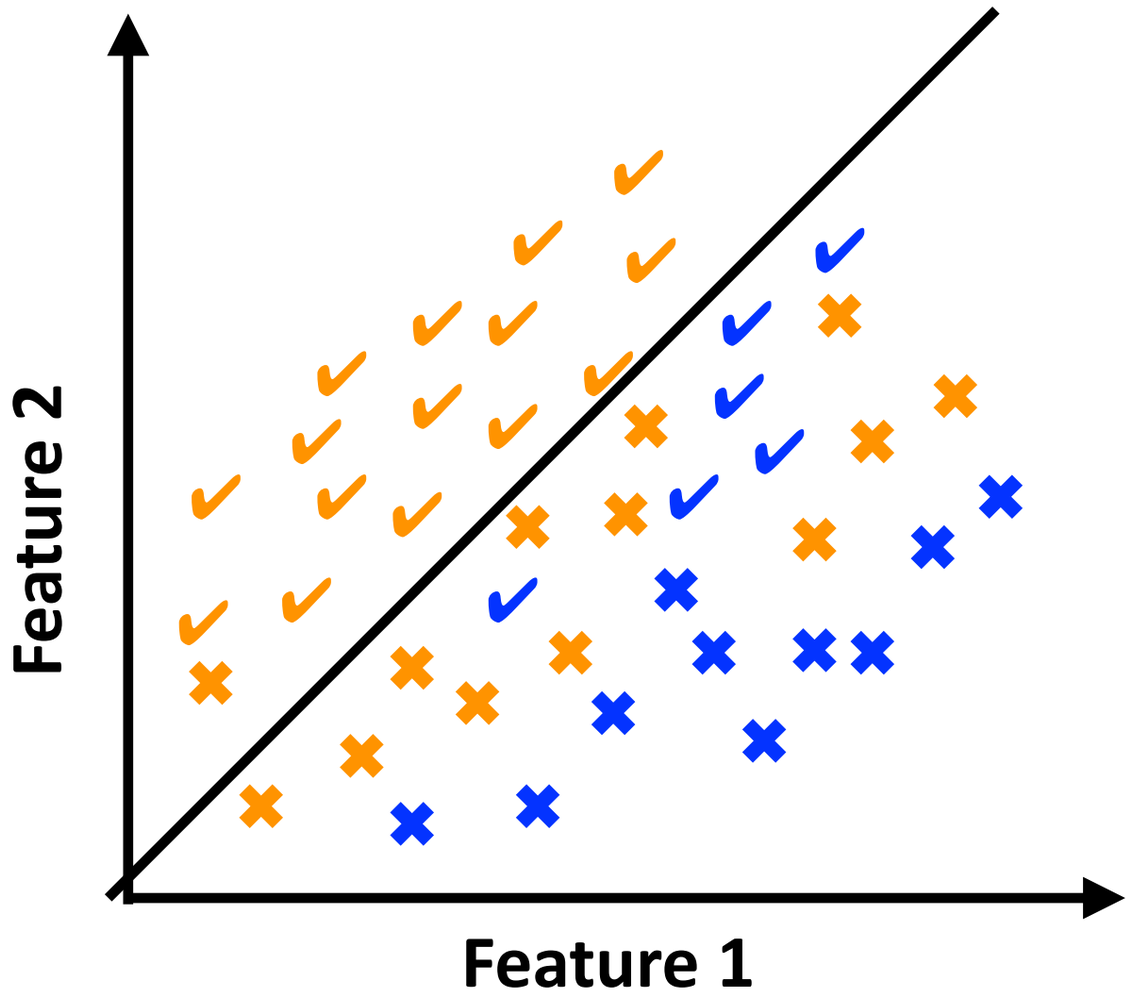
Learning the optimal boundary



$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

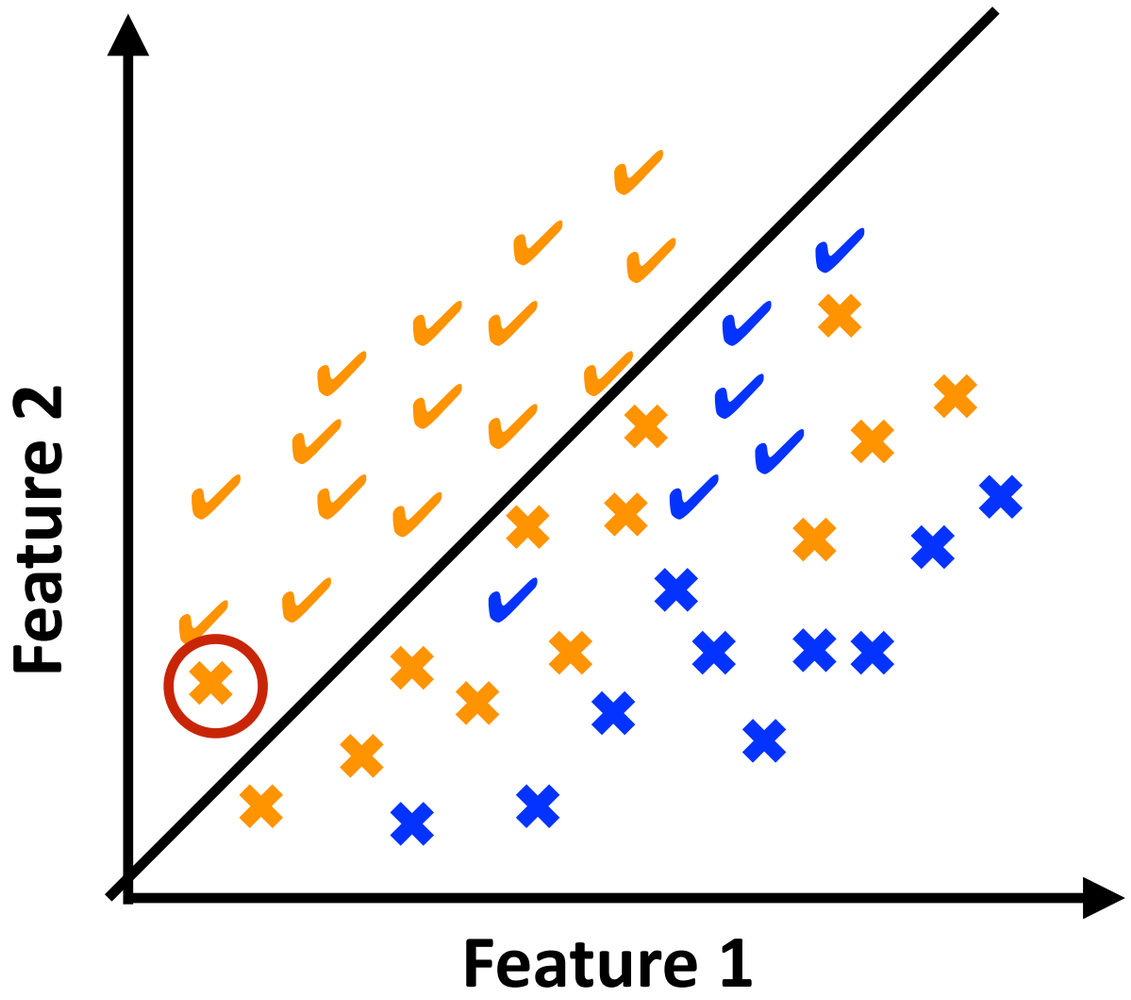
Optimal loss

Learning the optimal boundary



$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

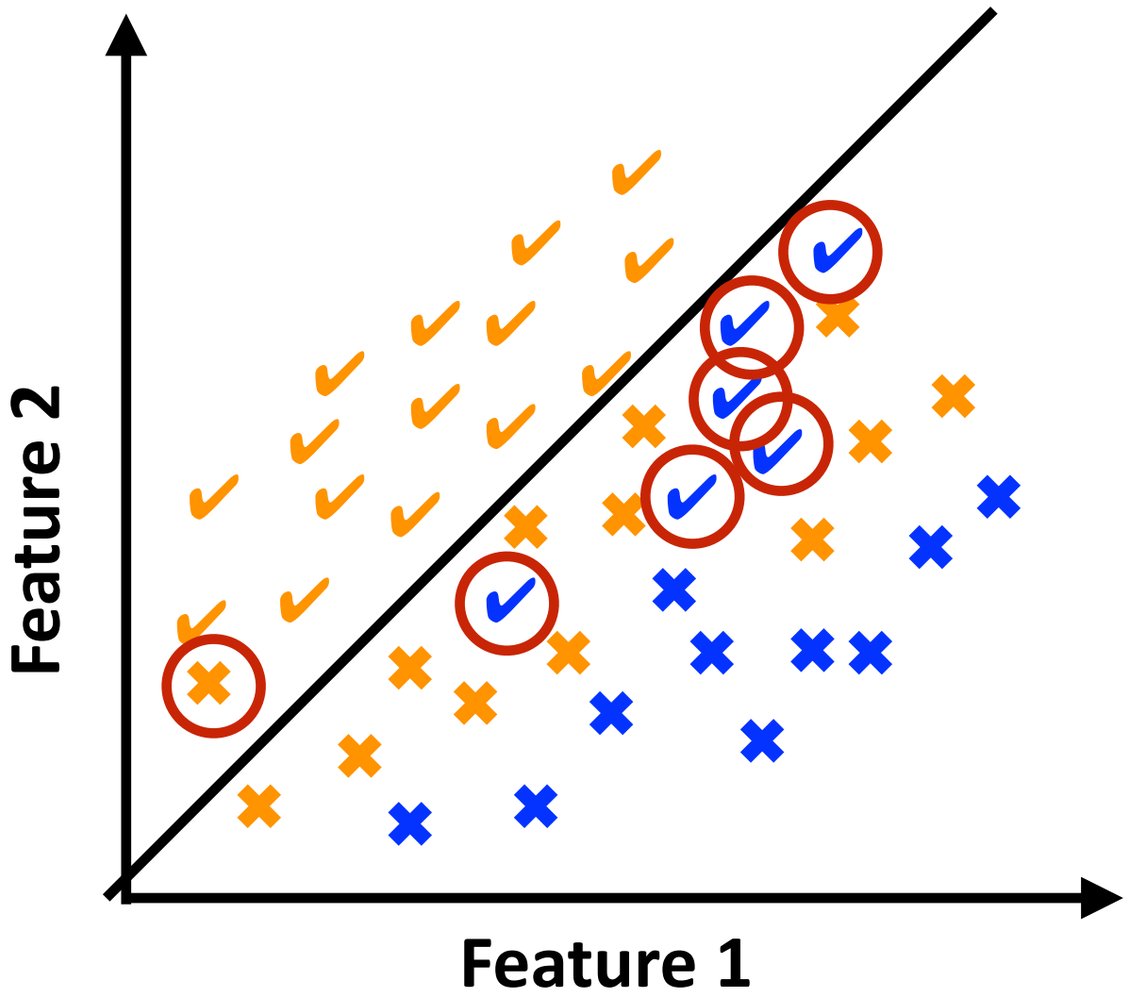
Learning the optimal boundary



$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

Women: Few errors

Learning the optimal boundary



$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

Women: Few errors
Men: Many errors
(unfair loan denial)

Disparate mistreatment: Different error rates

Relative disadvantage measure 3: Disparate mistreatment

- Achieve parity (or equality) in error rates
  - Error rates for all groups should be the same

- Wrongful relative disadvantage: Disparity in error rates

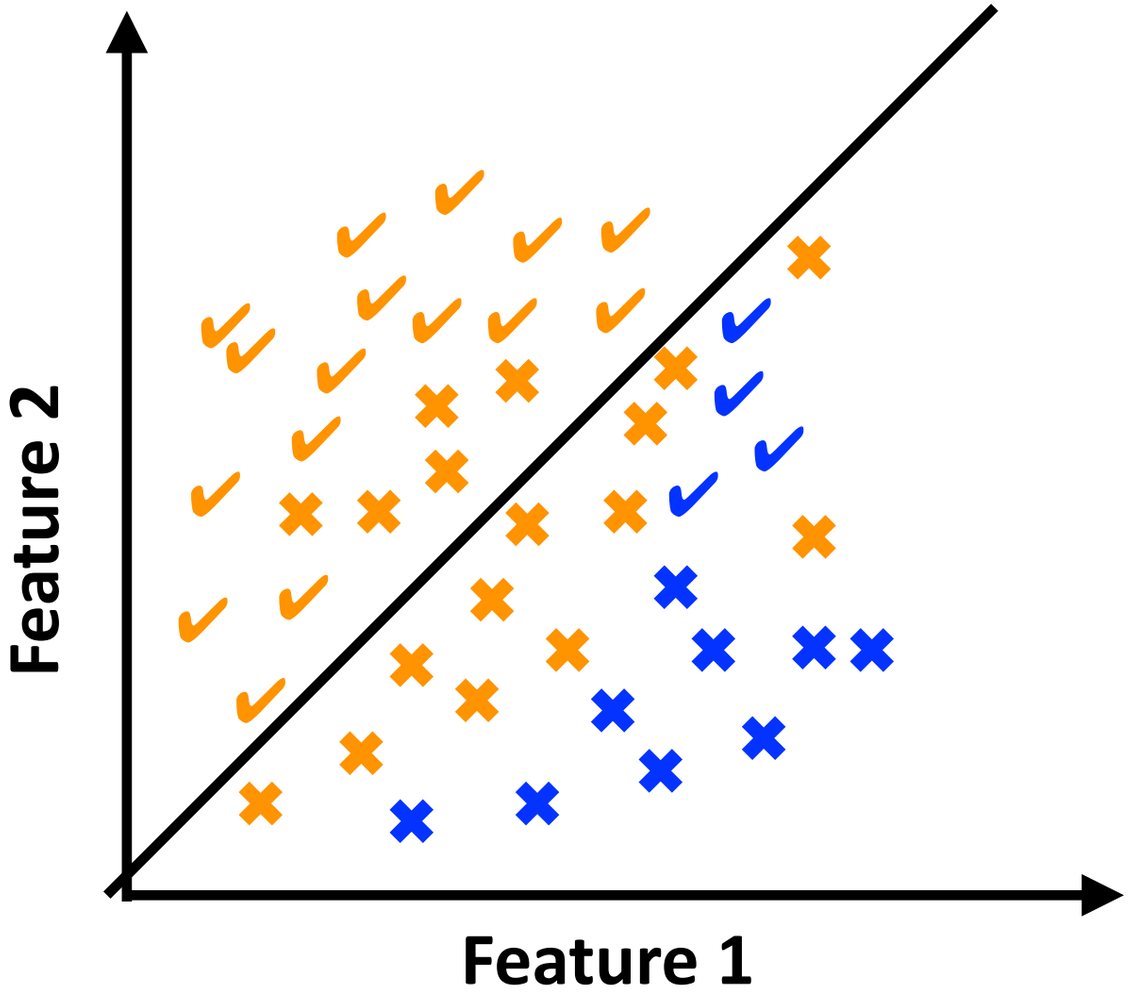
- Specially useful when ground truth labels are available

Formalizing disparate mistreatment

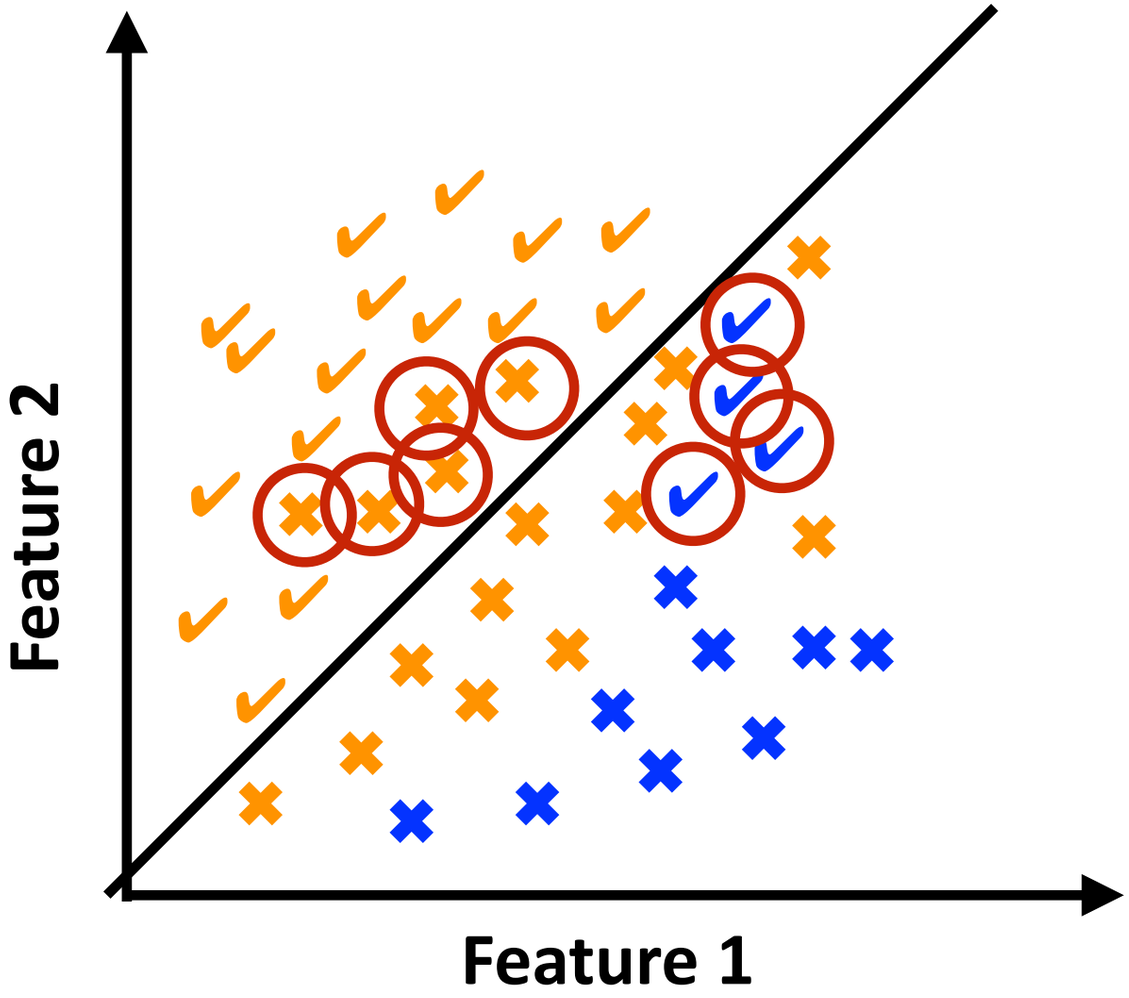
Equal error rates:

$$P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$$

Equality of overall error rates is not enough



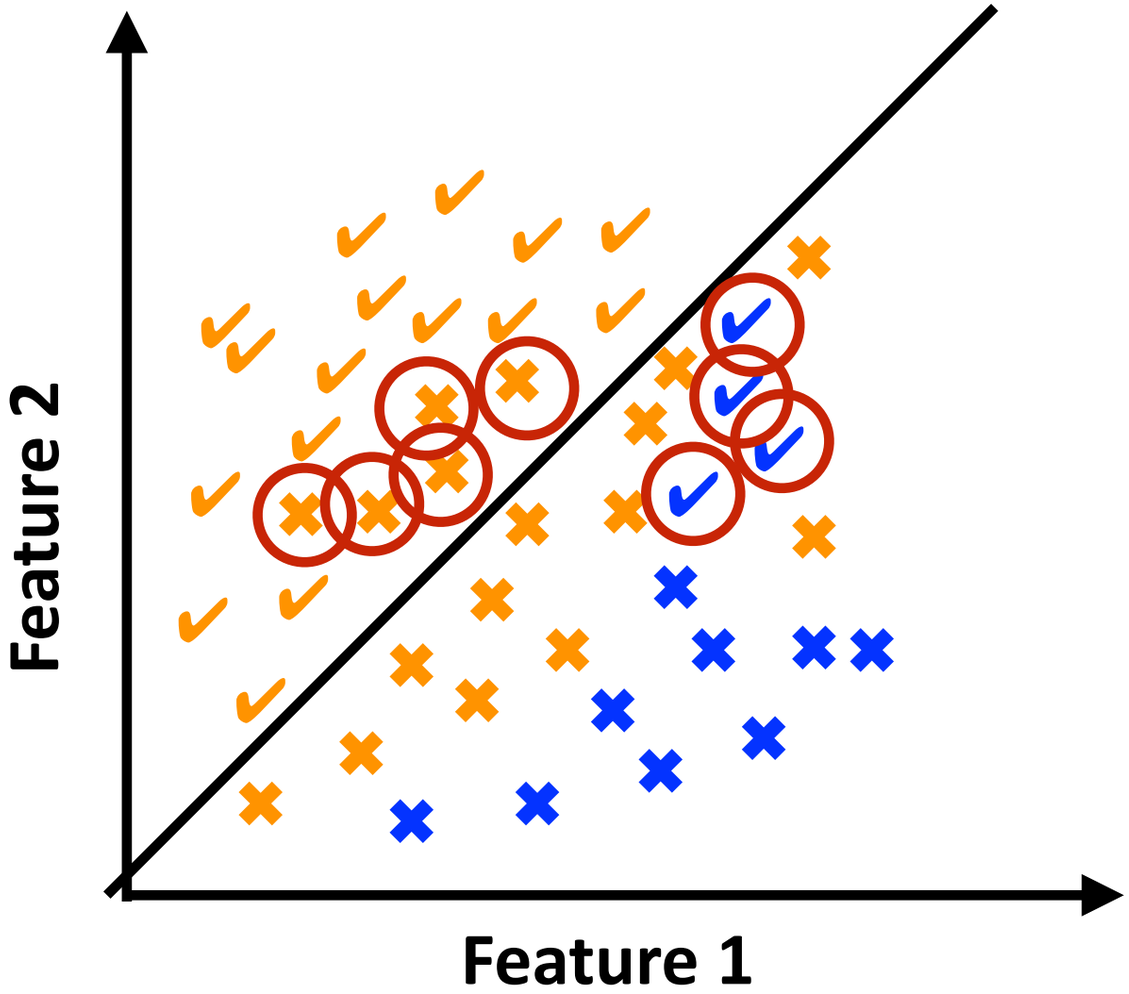
Equality of overall error rates is not enough



Similar error rate for men and women

No disparate mistreatment?

Equality of overall error rates is not enough



Similar error rate for men and women

Errors for women are false positives
(wrongly granted loan)

Errors for men are false negatives
(wrongly denied loan)

Disproportionate advantage to women

Error rates should be considered separately

Formalizing disparate mistreatment

Equal error rates:

$$P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$$

Equal false negative rates:

$$P(y \neq \hat{y} \mid y = +1, ♂) = P(y \neq \hat{y} \mid y = +1, ♀)$$

Equal false positive rates:

$$P(y \neq \hat{y} \mid y = -1, ♂) = P(y \neq \hat{y} \mid y = -1, ♀)$$

Formalizing disparate mistreatment

Equal error rates:

$$P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$$

Equal false negative rates:

$$P(y \neq \hat{y} \mid y = +1, ♂) = P(y \neq \hat{y} \mid y = +1, ♀)$$

Equal false positive rates:

$$P(y \neq \hat{y} \mid y = -1, ♂) = P(y \neq \hat{y} \mid y = -1, ♀)$$

Similar criteria for false omission and false discovery rates

Summary: 3 Measures of discrimination

- Disparate treatment: Targets direct discrimination
  - Requires: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$

- Disparate impact: Targets indirect discrimination, when biased historical labels
  - Requires: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$

- Disparate mistreatment: Targets indirect discrimination, when ground truth available
  - Requires: $P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$
  - Also for other misclassification rates

Part 2

Mechanisms for Nondiscriminatory Machine Learning

[AISTATS'17; WWW'17]

Mechanisms for nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$

- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$

- Disparate mistreatment: $P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$

Toy example: University admission

Non sensitive features (**x**) | Sensitive (z) | Outcome (y) | Predictions (ŷ)

	F_1	F_2	…	F_m	**z**	Past decision
U_1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	z_1	✔
U_2	$x_{2,1}$	$x_{2,2}$	…	$x_{2,m}$	z_2	✖
…	…	…	…	…	…	…
U_n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$	z_n	✔

→

Pred. Decision
✔
✖
…
✔

Learn a decision boundary (**w**) in the feature space, separating the two classes

Learning the optimal boundary

- Learning → Minimize loss on the historical data
- Convex boundary-based loss functions

Logistic loss	$\sum_{i=1}^{N} \log(1 + e^{-y_i d_{\mathbf{w}}(\mathbf{x}_i)})$
SVM loss	$\|\|\mathbf{w}\|\|^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i d_{\mathbf{w}}(\mathbf{x}_i))$

- Convex functions → Efficient learning

Classification free of disparate treatment

Non sensitive features (**x**) | Sensitive (z) | Outcome (y) | Predictions (ŷ)

	F_1	F_2	…	F_m	~~z~~	Past decision
U_1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	~~z_1~~	✔
U_2	$x_{2,1}$	$x_{2,2}$	…	$x_{2,m}$	~~z_2~~	✖
…	…	…	…	…	…	…
U_n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$	~~z_n~~	✔

→

Pred. Decision
✔
✖
…
✔

$$P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$$

Do not use the sensitive feature

Mechanisms for nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

Mechanisms for nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$

Classification free of disparate impact

Classification free of disparate impact

Key Idea: Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

Classification free of disparate impact

Key Idea: Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$$

z = 0 (Men)
z = 1 (Women)

- Minimizing loss → Optimizing accuracy
- Adding constraints → Nondiscrimination goals

Classification free of disparate impact

Key Idea: Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$$

z = 0 (Men)
z = 1 (Women)

Important insight

Tradeoff between accuracy & nondiscrimination

Classification free of disparate impact

Key Idea: Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$$

z = 0 (Men)
z = 1 (Women)

- Non-convex for many well-known classifiers (logistic regression, SVM)
- Hard to compute efficiently

Disparate Impact Constraints

Goal: $-\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$

Disparate Impact Constraints

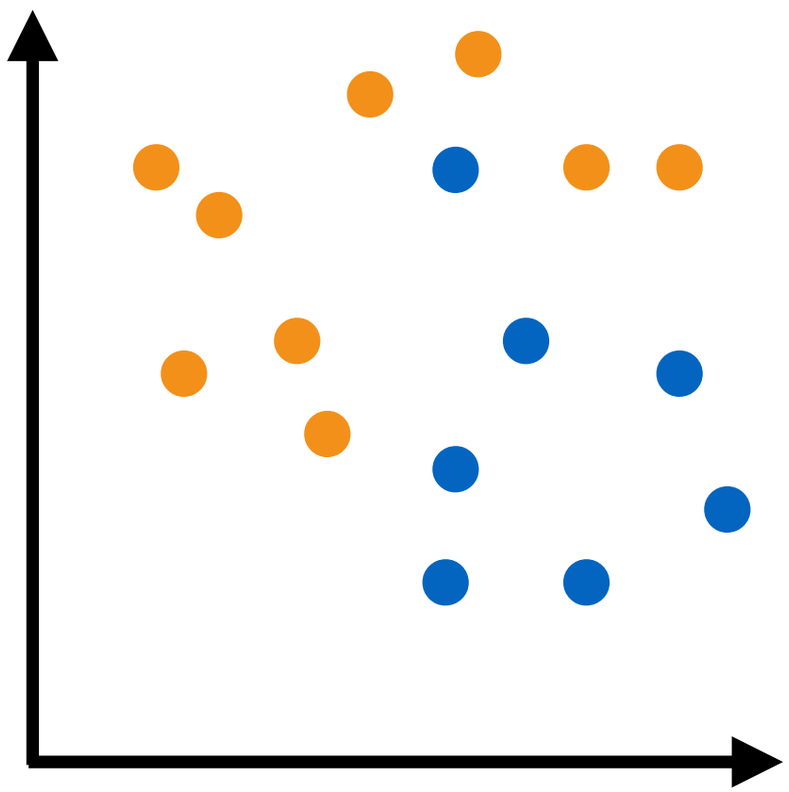
Goal: $-\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_{\mathbf{w}}(\mathbf{x}_i)$$

Key Idea: Limit the covariance between sensitive feature value and distance from decision boundary

Disparate Impact Constraints

Goal: $-\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$



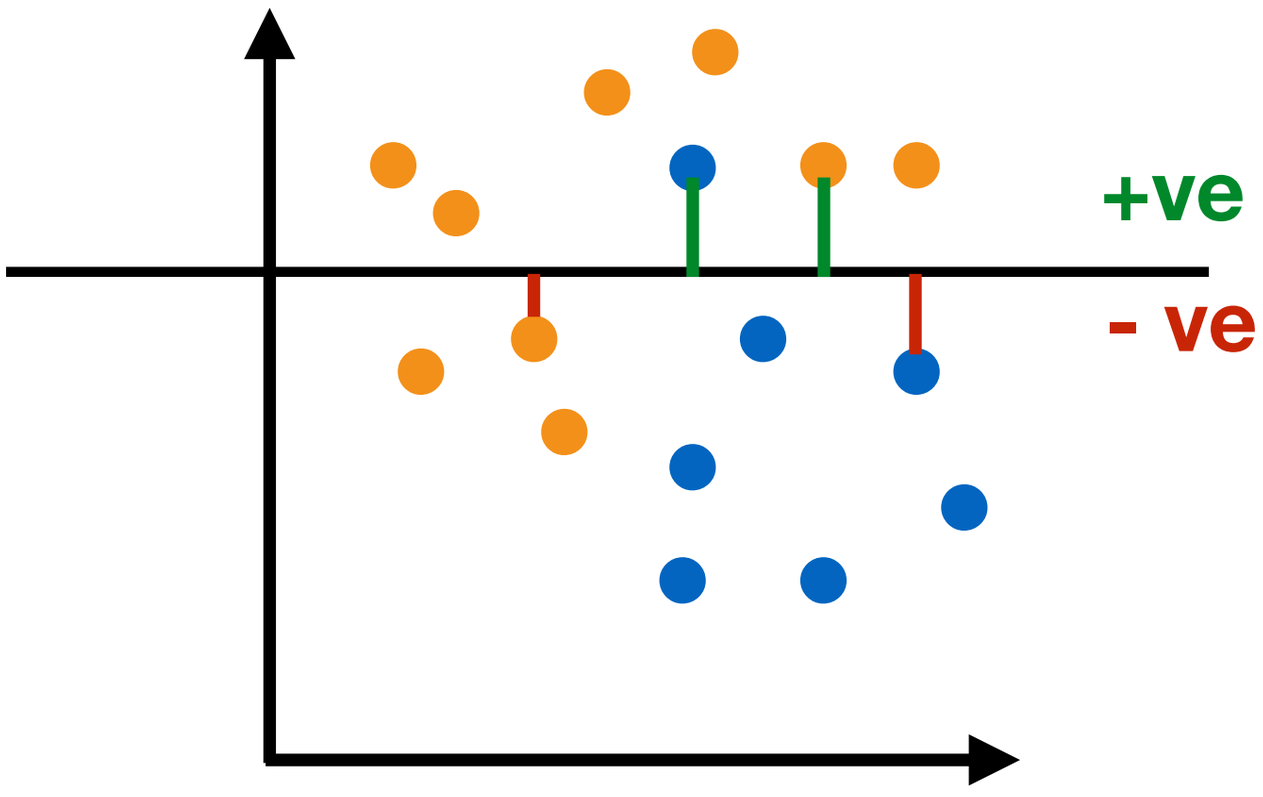
Women
Men

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_{\mathbf{w}}(\mathbf{x}_i)$$

Key Idea: Limit the covariance between sensitive feature value and distance from decision boundary

Disparate Impact Constraints

Goal: $-\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$



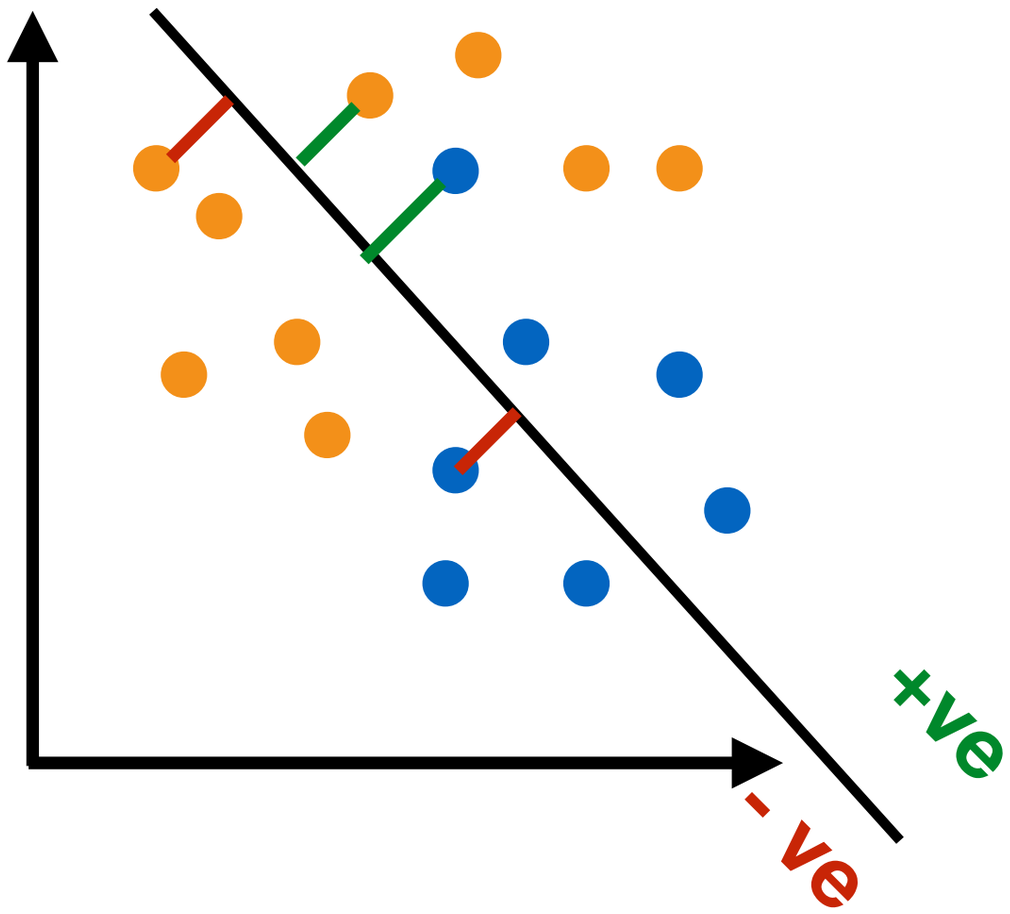
Women
Men

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_{\mathbf{w}}(\mathbf{x}_i)$$

Key Idea: Limit the covariance between sensitive feature value and distance from decision boundary

Disparate Impact Constraints

Goal: $-\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$



Women

Men

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_{\mathbf{w}}(\mathbf{x}_i)$$

In other words: Limit the difference in average strength of acceptance or rejection between sensitive feature groups

Classification free of disparate impact

Key Idea: Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\epsilon \leq \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_{\mathbf{w}}(\mathbf{x}_i) \leq \epsilon$$

$$d_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

Linear on **w**

Convex

$P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1)$

Minimizing disparate impact

(Optimal) Classifier without any constraints: $\mathbf{w}^* = \text{argmin} \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$

Minimizing disparate impact

(Optimal) Classifier without any constraints: $\mathbf{w}^* = \text{argmin} \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$

$$\min \quad \left| \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_{\mathbf{w}}(\mathbf{x}_i) \right|$$

Minimizing disparate impact

(Optimal) Classifier without any constraints: $\mathbf{w}^* = \text{argmin} \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$

$$\min \quad \left| \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_{\mathbf{w}}(\mathbf{x}_i) \right|$$

$$\text{s.t.} \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w}) \leq (1+\gamma) \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w}^*)$$

$\gamma = 0 \rightarrow$ No additional loss

$\gamma > 0 \rightarrow$ Allow additional loss

Allows for fine-grained control over loss in accuracy

Minimizing disparate impact

(Optimal) Classifier without any constraints: $\mathbf{w}^* = \text{argmin} \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$

$$\min \quad \left| \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_{\mathbf{w}}(\mathbf{x}_i) \right|$$

$$\text{s.t.} \quad L(\mathbf{x}_j, y_j, \mathbf{w}) \leq (1 + \gamma_j) L(\mathbf{x}_j, y_j, \mathbf{w}^*) \quad \forall j \in \{1, 2, \ldots, N\}$$

Set γ_j for individual points

$\gamma_j = 0$ → No additional loss

$\gamma_j > 0$ → Allow additional loss

Preserving accuracy for specific points

Extending constraints to non-linear models

$$\max \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C \quad \forall i$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$d_\alpha(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i)$$

Extending constraints to non-linear models

$$\max \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C \quad \forall i$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$-\epsilon \leq \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_\alpha(\mathbf{x}_i) \leq \epsilon$$

Convex

$$d_\alpha(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i)$$

Linear on α

Kernel trick on the constraints too

$P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1)$

Mechanisms for nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$
  - Convex constraints

Mechanisms for nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$
  - Convex constraints

- Disparate mistreatment: $P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$

Classification free of disparate mistreatment

Classification free of disparate mistreatment

Key Idea: Learn under constraints

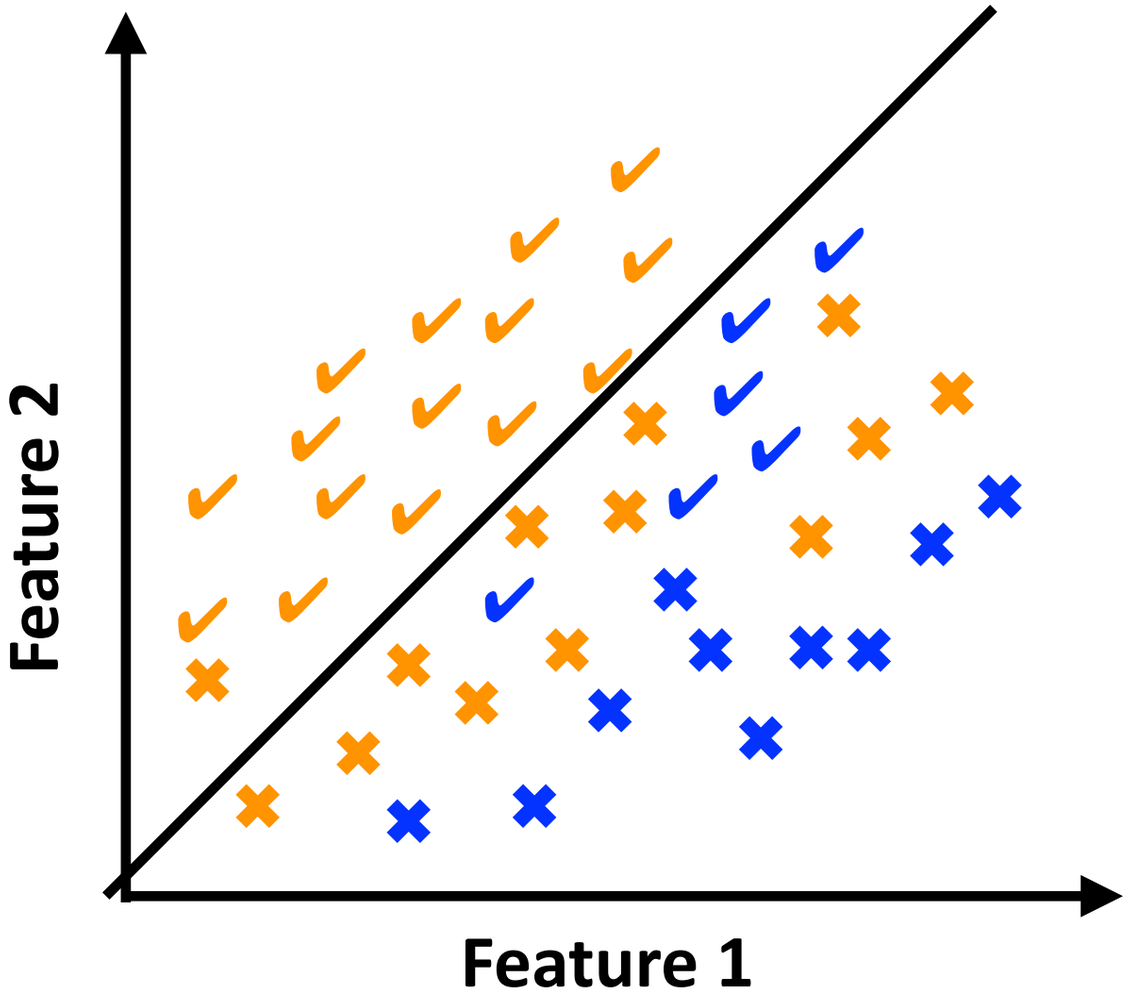
$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\varepsilon \leq P(y \neq \hat{y} \mid ♂) - P(y \neq \hat{y} \mid ♀) \leq \varepsilon$$

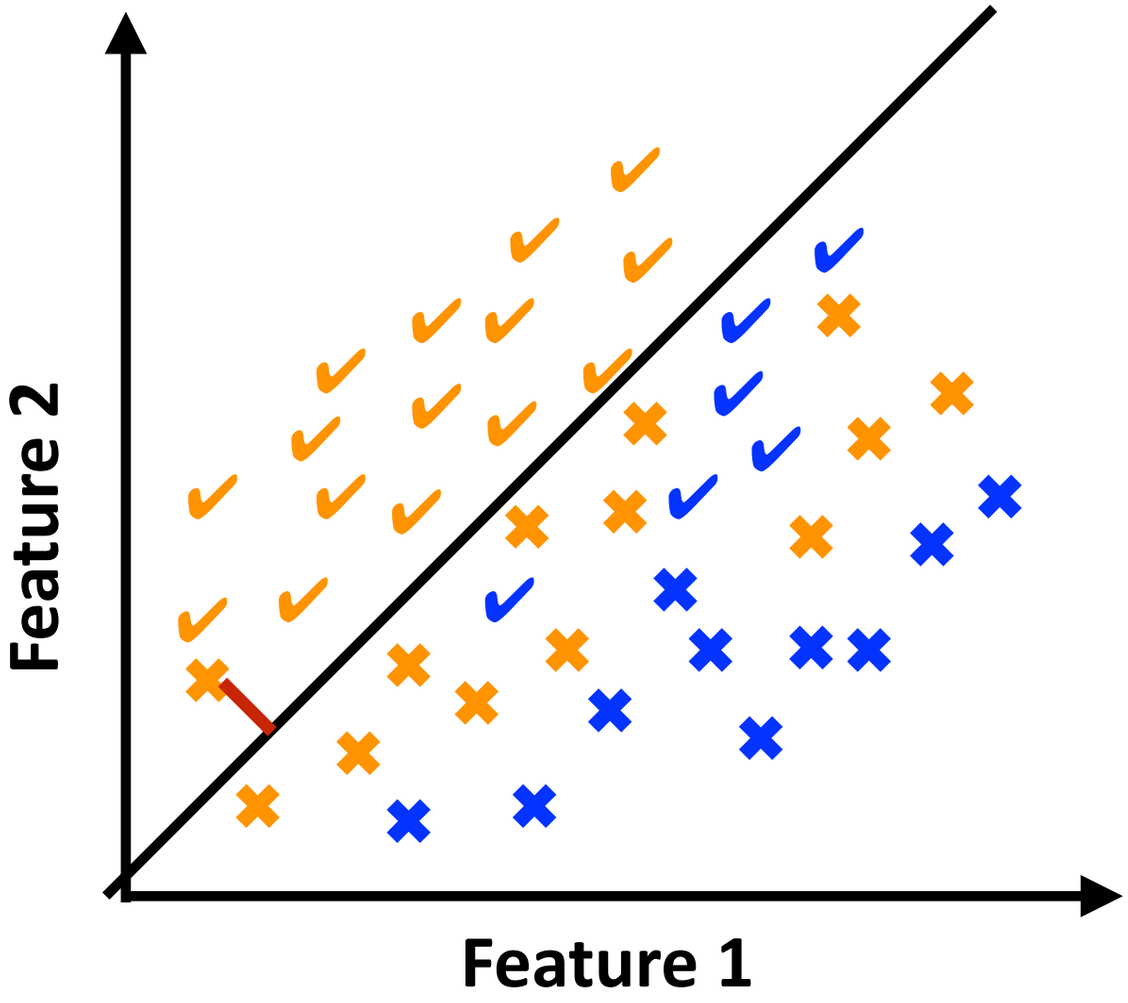
z = 0 (Men)
z = 1 (Women)

- Non-convex for many well-known classifiers (logistic regression, SVM)
- Hard to compute efficiently

Disparate mistreatment constraints

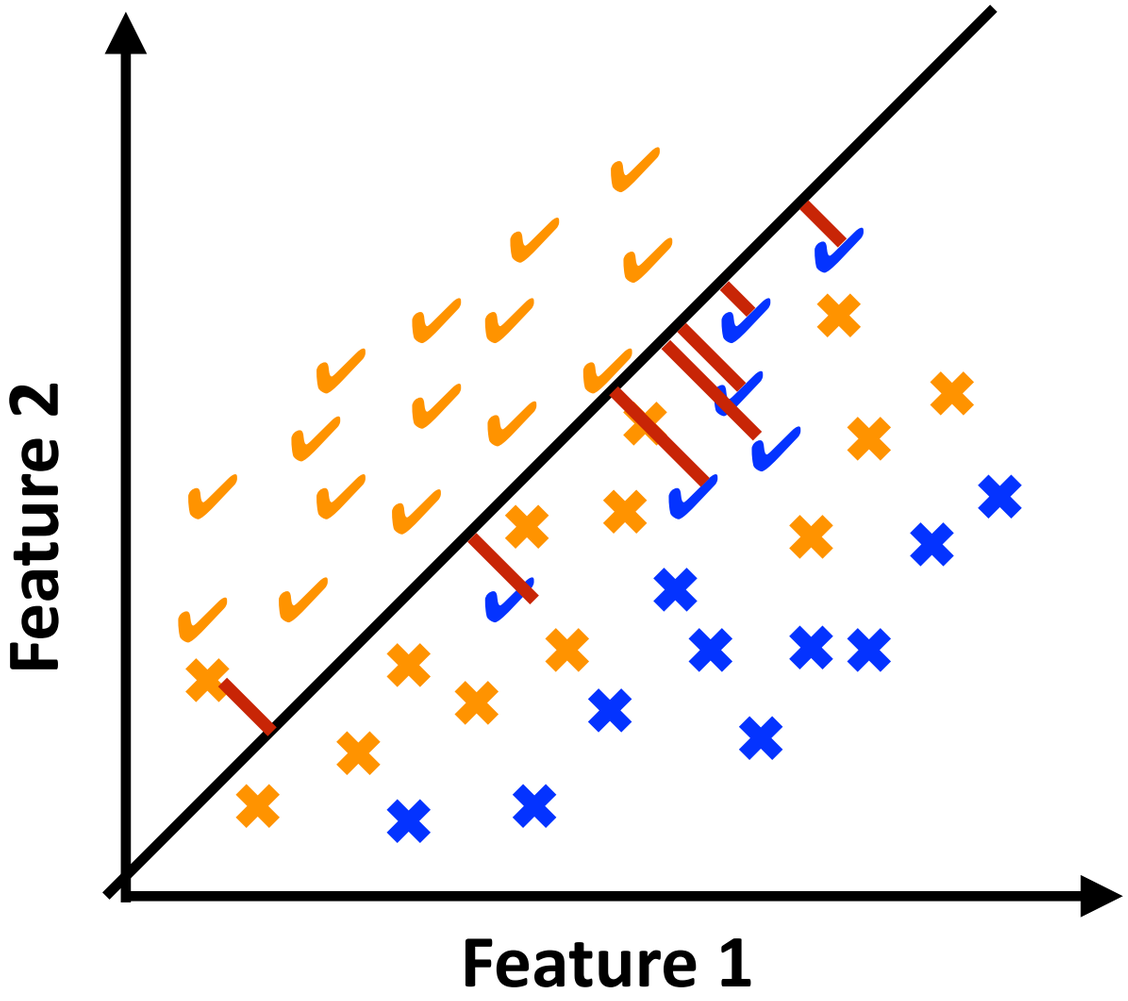


Disparate mistreatment constraints



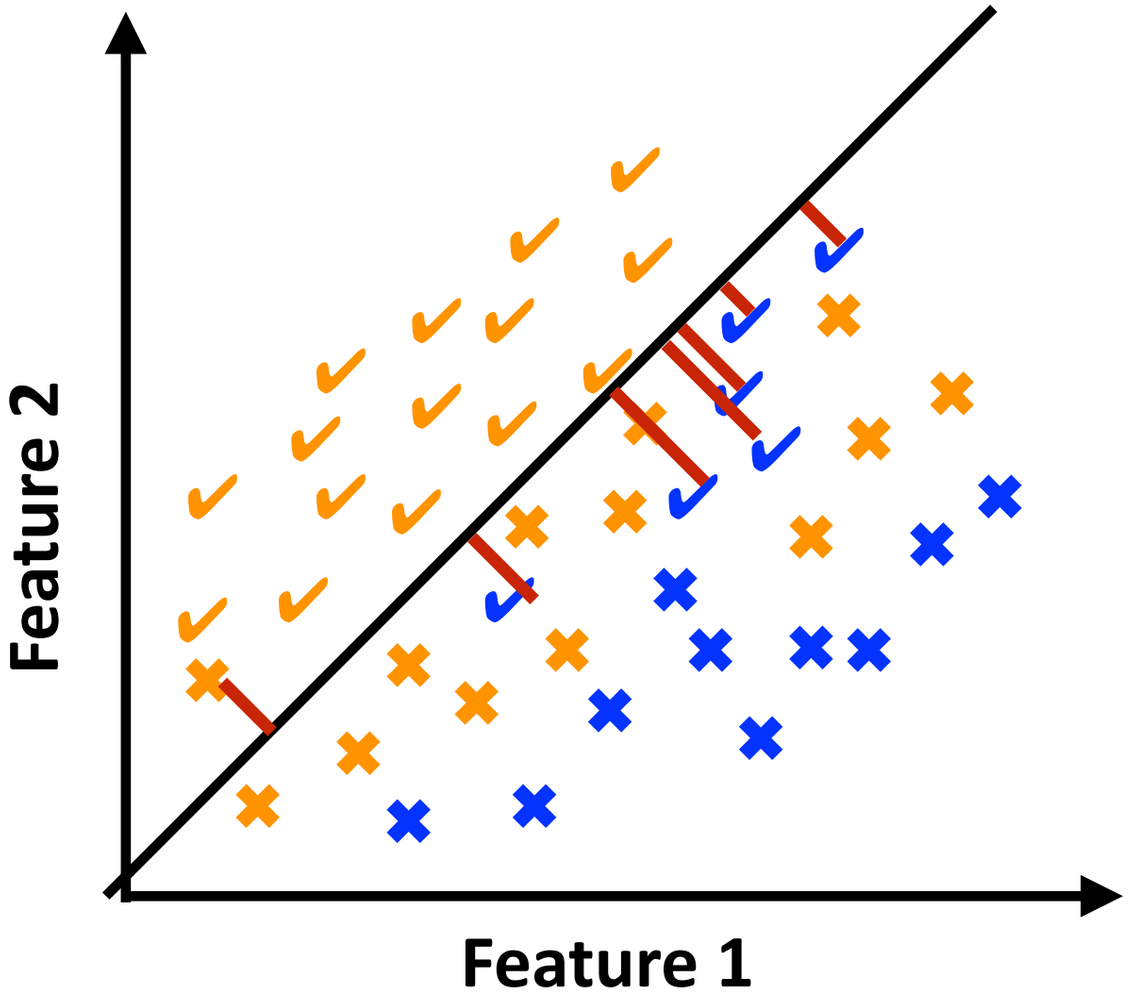
$$g_{\mathbf{w}}(\mathbf{x}, y) = \min(0, y d_{\mathbf{w}}(\mathbf{x}))$$

Disparate mistreatment constraints



$$g_{\mathbf{w}}(\mathbf{x}, y) = \min(0, y d_{\mathbf{w}}(\mathbf{x}))$$

Disparate mistreatment constraints



$$g_{\mathbf{w}}(\mathbf{x}, y) = \min(0, y d_{\mathbf{w}}(\mathbf{x}))$$

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z}) g_{\mathbf{w}}(\mathbf{x}_i, y_i)$$

Key Idea: Limit the covariance between sensitive feature value and misclassification distance from decision boundary

Disparate mistreatment constraints

Key Idea: Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\epsilon \leq \frac{1}{N} \sum (z_i - \bar{z}) g_{\mathbf{w}}(\mathbf{x}_i, y_i) \leq \epsilon$$

Linear on **w**

$$g_{\mathbf{w}}(\mathbf{x}, y) = \min(0, y d_{\mathbf{w}}(\mathbf{x}))$$

Concave on **w**

Convex-Concave

$$P(y \neq \hat{y} \mid ♂) - P(y \neq \hat{y} \mid ♀)$$

Disciplined Convex-Concave Program (DCCP)

(can be solved efficiently)

[Shen, Diamond, Gu, Boyd, 2016]

Mechanisms for nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$
  - Convex constraints

- Disparate mistreatment: $P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$
  - Convex-concave constraints

Mechanisms for nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$
  - Convex constraints

- Disparate mistreatment: $P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$
  - Convex-concave constraints

Case study: Crime risk prediction

Crime risk prediction: Background

- Person arrested on suspicion of a crime
  - Appear in front of a judge before the trial
  - Judge decides whether to give bail or not
  - Predict whether the person will recidivate
  - **Recidivism:** Commit a crime within two years

- COMPAS tool
  - Machine learning model to advise the judge
  - Predicts a recidivism probability

ProPublica COMPAS dataset

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

- Train a classifier on this dataset

Case study: Crime risk prediction

- Can traditional classifiers lead to disparate mistreatment?

- Can our approach help avoid disparate mistreatment?

Disparate mistreatment in risk prediction

Trained a logistic regression classifier to predict recidivism

Positive class: Will recidivate

Race	FPR	FNR
Black		
White		

Disparate mistreatment in risk prediction

Trained a logistic regression classifier to predict recidivism

Positive class: Will recidivate

Race	FPR	FNR
Black	34%	
White	15%	

- **False positive:** Non-recid. person wrongly classified as recid.

Disparate mistreatment in risk prediction

Trained a logistic regression classifier to predict recidivism

Positive class: Will recidivate

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

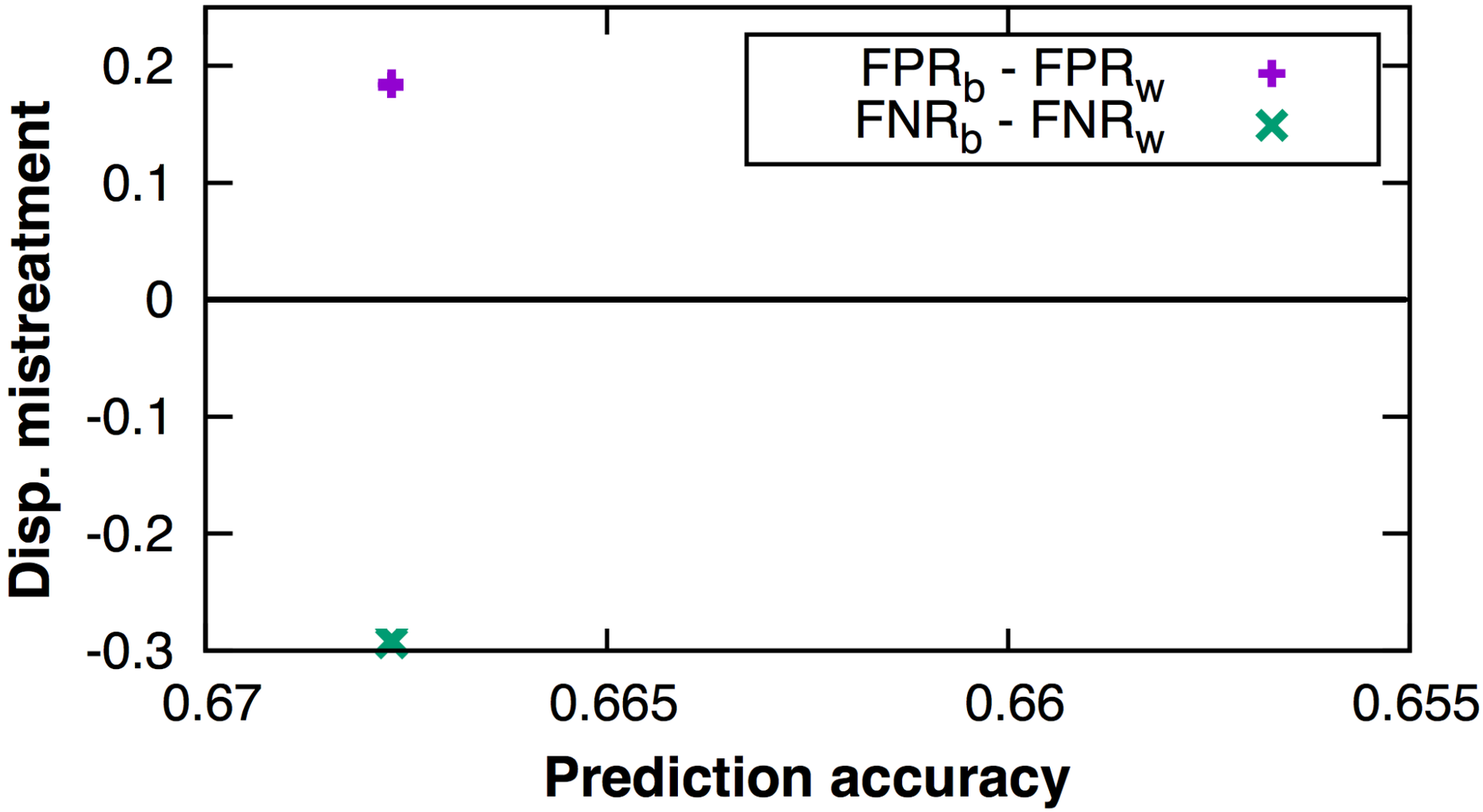
- **False positive:** Non-recid. person wrongly classified as recid.
- **False negative:** Recid. person wrongly classified as non-recid.

Case study: Crime risk prediction

- Can traditional classifiers lead to disparate mistreatment?
  - Disparity in both FPR and FNR!

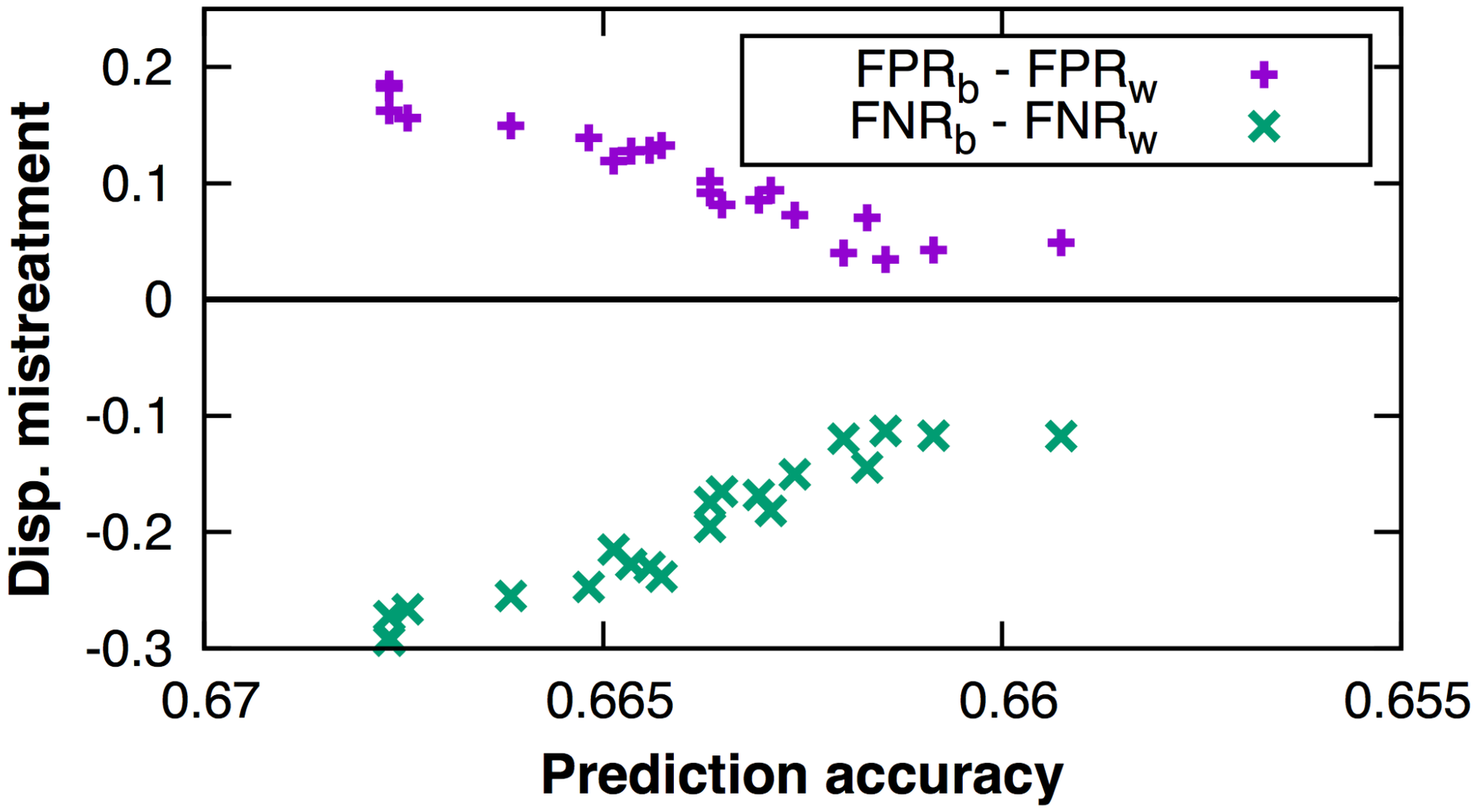
- Can our approach help avoid disparate mistreatment?

Removing disparate mistreatment



Removing disparate mistreatment

Introduce FPR and FNR constraints



Case study: Crime risk prediction

- Can traditional classifiers lead to disparate mistreatment?
  - Disparity in both FPR and FNR!

- Can our approach help avoid disparate mistreatment?
  - Yes!
  - Experiments with several synthetic datasets

Summary: Discrimination in machine leanring

- **Part 1:** Defined three measures of discrimination
  - Disparate treatment / impact / mistreatment

- **Part 2:** Designed mechanisms for each of them
  - Proposed tractable and efficient proxies

Summary: Discrimination in machine leanring

- **Part 1:** Defined three measures of discrimination
  - Disparate treatment / impact / mistreatment

- **Part 2:** Designed mechanisms for each of them
  - Proposed tractable and efficient proxies

- **Part 3:** Refined measures and designed mechanisms
  - Inspired by envy-freeness and Nash bargaining solution

Summary: Discrimination in machine leanring

- **Part 1:** Defined three measures of discrimination
  - Disparate treatment / impact / mistreatment

- **Part 2:** Designed mechanisms for each of them
  - Proposed tractable and efficient proxies

- **Part 3:** Refined measures and designed mechanisms
  - Inspired by envy-freeness and Nash bargaining solution

Code publicly available at
https://github.com/mbilalzafar/fair-classification