Focus 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

What is a computational perspective? Why is it 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
- Trained over big historical recidivism data across US
  - Excluding sensitive feature info like gender and race

COMPAS Goal: Criminal justice reform

- Many studies show racial biases in human judgments

- **Idea:** Nudge subjective human decision makers with objective algorithmic predictions
  - Algorithms have no pre-existing biases
  - They simply process information in a consistent manner

- Learn to make accurate predictions without race info.
  - Blacks & whites with same features get same outcomes
  - No disparate treatment & so non-discriminatory!

Is COMPAS fair to all groups?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

Is COMPAS fair to all groups?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Discovery Rate: 805 / (805 + 1369) = 0.37 349 / (349 + 505) = 0.40

Is COMPAS fair to all groups?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Discovery Rate: 805 / (805 + 1369) = 0.37 349 / (349 + 505) = 0.40

False Omission Rate: 532 / (532 + 990) = 0.35 461 / (461 + 1139) = 0.29

Is COMPAS fair to all groups?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Discovery Rate: 805 / (805 + 1369) = 0.37 349 / (349 + 505) = 0.40

False Omission Rate: 532 / (532 + 990) = 0.35 461 / (461 + 1139) = 0.29

- Northpointe: False discovery & omission rates for blacks & whites are comparable
- So **YES!**

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: 805 / (805 + 990) = 0.45

349 / (349 + 1139) = 0.23

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: 805 / (805 + 990) = 0.45   349 / (349 + 1139) = 0.23

False Negative Rate: 532 / (532 + 1369) = 0.29   461 / (461 + 505) = 0.48

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: 805 / (805 + 990) = 0.45 >> 349 / (349 + 1139) = 0.23

False Negative Rate: 532 / (532 + 1369) = 0.29 << 461 / (461 + 505) = 0.48

- ProPublica: False positive & negative rates are considerably worse for blacks than whites!
  - Constitutes discriminatory **disparate mistreatment**

Machine Bias

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

Why are error comparisons so different?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

Recidivism Ratio: (1369 + 532) : (805 + 990) = 1.06 : 1.00 (505 + 461) : (349 + 1139) = 0.65 : 1.00

- Impossibility result: **[Kleinberg '17, Chouldechova '17]**
  - When recidivism ratios for blacks & whites differ, no non-trivial solution can achieve equal FDR, FOR, FPR, FNR!
- Can equalize at most two out of the four error rates!

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

- Another example: Fairness of random judge selection
  - Suppose you have **N** fair / unfair judges
    - They have equal FPR / FNR / FOR / FDR for different racial groups
  - Does assigning cases to judges randomly affect fairness?

Computational Interpretations (measures) of Discrimination *[WWW '17]*

Defining discrimination

- A first approximate normative / moralized definition:

  wrongfully impose a **relative disadvantage** on persons **based on** their membership in some **salient social group** e.g., race or gender

- Challenge: How to operationalize the definition?
  - How to make it clearly distinguishable, measurable, & understandable in terms of empirical observations

Need to operationalize 4 fuzzy notions

1. What constitutes a relative disadvantage?
2. What constitutes a wrongful imposition?
3. What constitutes based on?
4. What constitutes a salient social group?

Need to operationalize 4 fuzzy notions

1. ~~What constitutes a relative disadvantage?~~

2. ~~What constitutes a wrongful imposition?~~

3. ~~What constitutes based on?~~

4. What constitutes a **salient social group?**
   - Defined by anti-discrimination laws: Race, Gender

Need to operationalize 4 fuzzy notions

1. ~~What constitutes a relative disadvantage?~~
2. ~~What constitutes a wrongful imposition?~~
3. What constitutes **based on?**
   - Do not use salient group information in training or deployment
   - Use during training, but not deployment
   - Use during both training and deployment
4. ~~What constitutes a salient social group?~~

Need to operationalize 4 fuzzy notions

1. What constitutes a **relative disadvantage?**

2. ~~What constitutes a wrongful imposition?~~

3. ~~What constitutes based on?~~

4. ~~What constitutes a salient social group?~~

Operationalizing discrimination

- Consider binary classification using user features

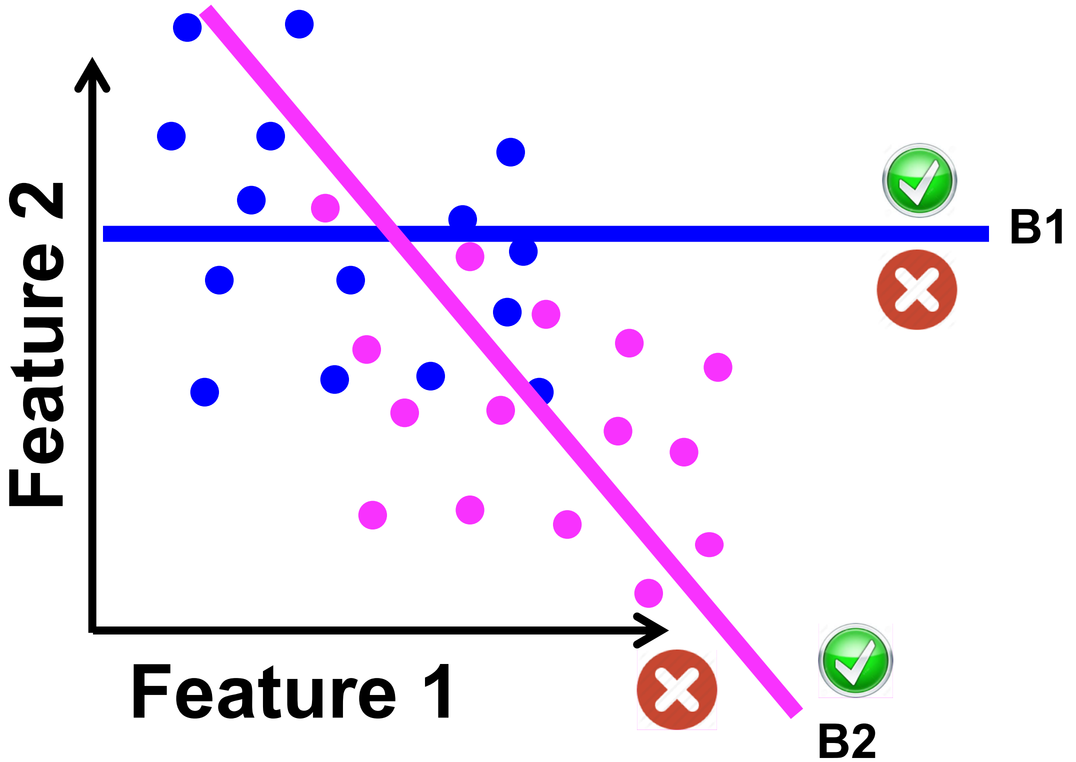
	F_1	F_2	…	F_m	Z
User$_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	Z_1
User$_2$	$x_{2,1}$			$x_{2,m}$	Z_2
User$_3$	$x_{3,1}$			$x_{3,m}$	Z_3
…	…			…	…
User$_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$	Z_n

Decision
Accept
Reject
Reject
…
Accept

Decision outcomes should not be **relatively disadvantageous** to social (sensitive feature) groups!

Relative disadvantage measure 1: Disparate treatment

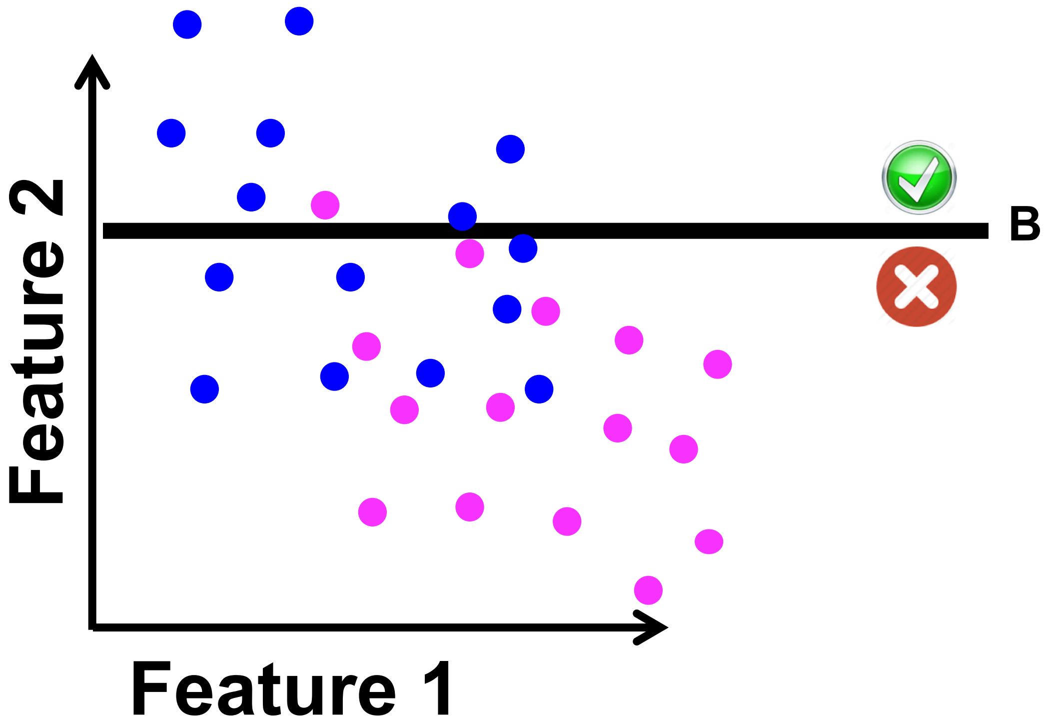
DT (B1, B2)
= 15 / 30 = 0.5



Measures the fraction of users whose outcomes change, when their sensitive features are changed

Relative disadvantage measure 1: Disparate treatment

DT (B)
= 0 / 30 = 0



Measures the fraction of users whose outcomes change, when their sensitive features are changed

Measures direct discrimination

- Based on counter-factual reasoning
  - Most intuitive measure of discrimination

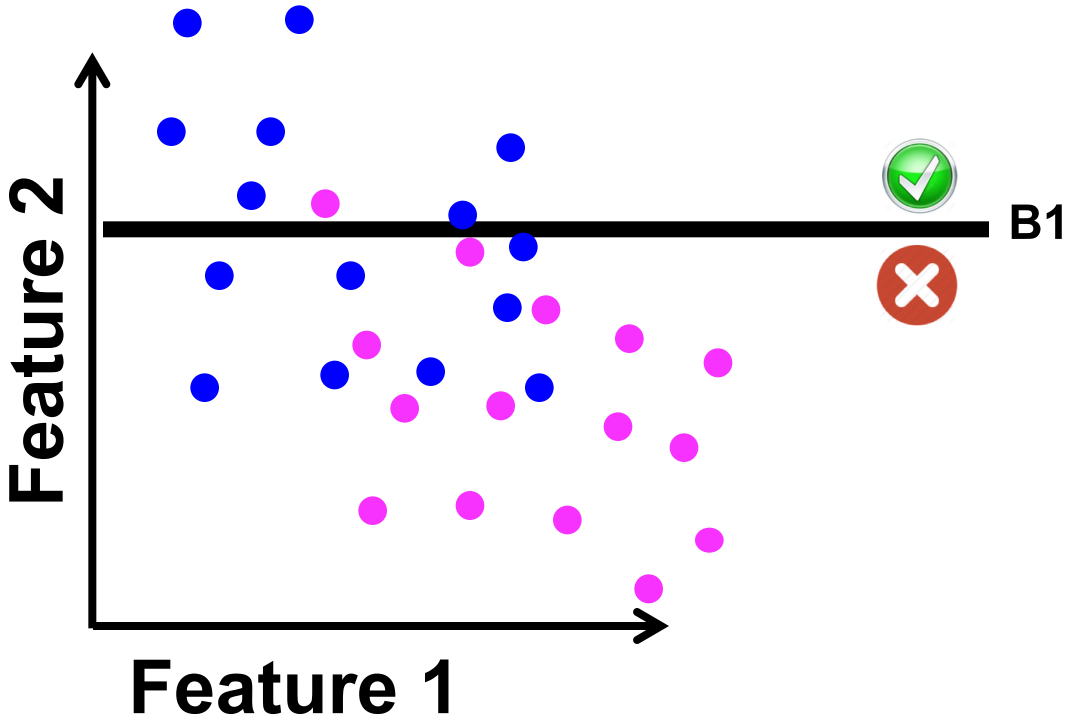
- To achieve parity treatment: Ignore sensitive features, when defining the decision boundary

- Situational testing for discrimination discovery checks for disparate treatment

- More formally: $P(\hat{y}|\mathbf{x}, z) = P(\hat{y}|\mathbf{x})$

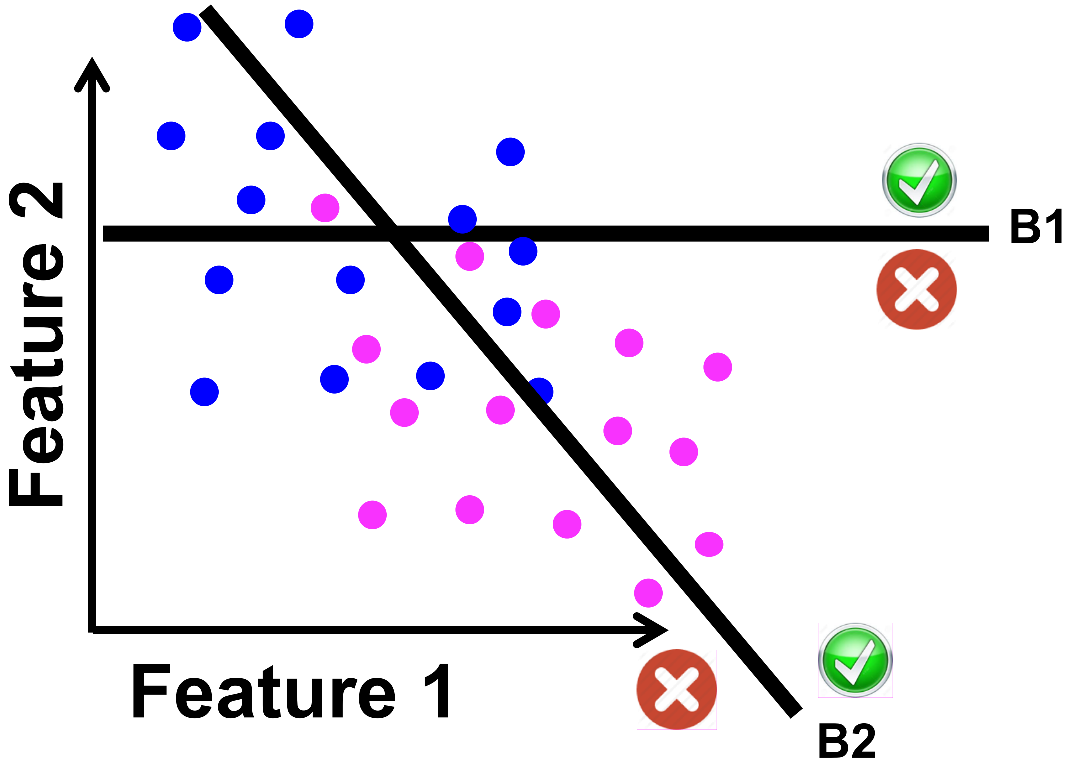
Relative disadvantage measure 2: Disparate impact

DI (B1)
= 7/15 - 1/15 = 0.4



Measures the difference in fraction of positive (negative) outcomes for different sensitive feature groups

Relative disadvantage measure 2: Disparate impact



DI (B1) = 0.4
DI (B2)
= 7/15 - 6/15 = .06

Measures the difference in fraction of positive (negative) outcomes for different sensitive feature groups

Measures indirect discrimination

- Observed in human decision making
- Indirectly discriminate against specific user groups using their correlated non-sensitive attributes
  - E.g., voter-id laws being passed in US states
- Doctrine of disparate impact
  - A US law applied in employment & housing practices
  - Proportionality tests over decision outcomes

A controversial measure

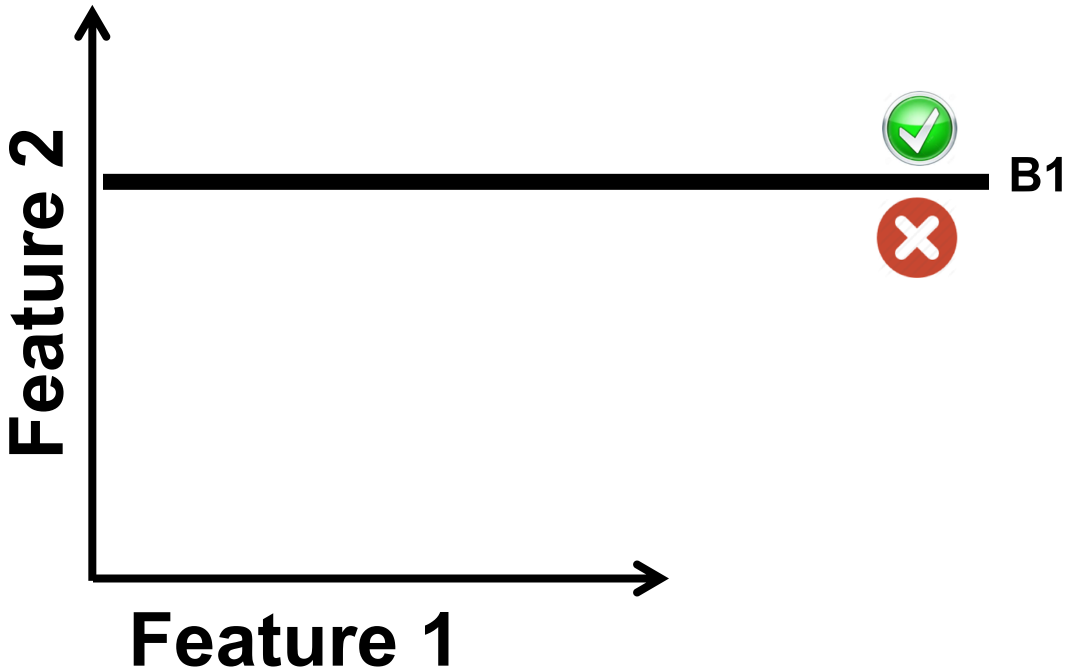
- To achieve parity impact: Select equal fractions of sensitive feature groups
  - More formally: $P(\hat{y} = 1 | z = 0) = P(\hat{y} = 1 | z = 1)$

- In Law:
  - Critics: There exist scenarios where disproportional outcomes are justifiable
  - Supporters: Provision for business necessity exists
    - Though the burden of proof is on employers

- In ML: Use, when labels in training data are biased

Relative disadvantage measure 3: Disparate mistreatment

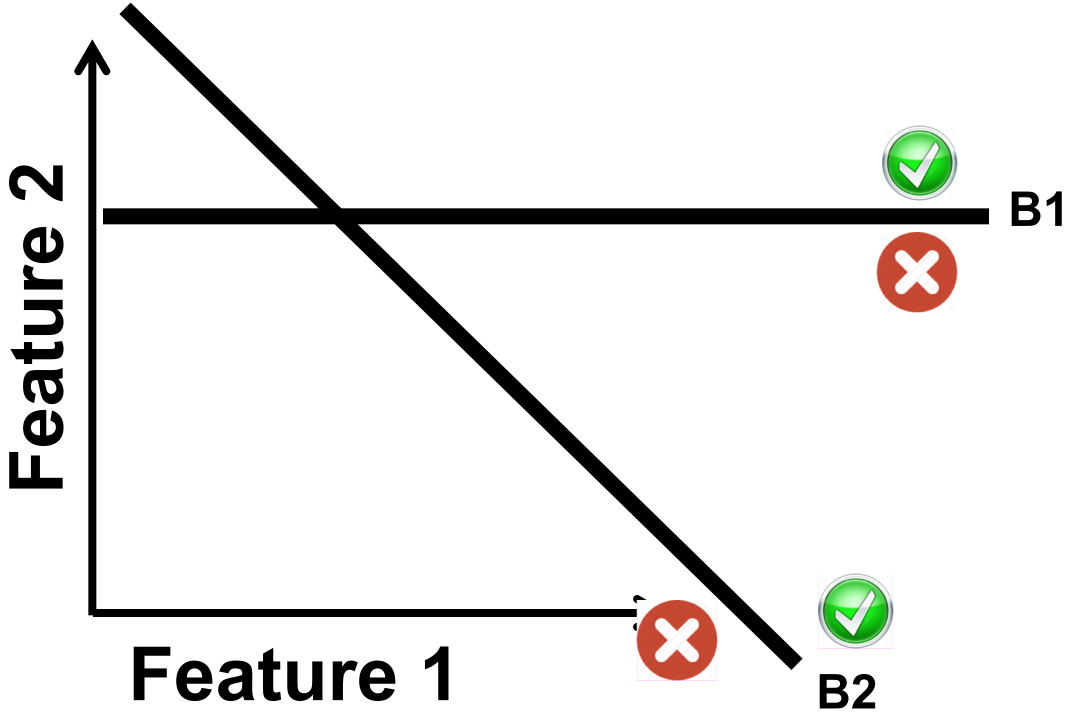
DM (B1)
= 13/15 - 9/15 = .26



Measures the difference in fraction of accurate outcomes for different sensitive feature groups

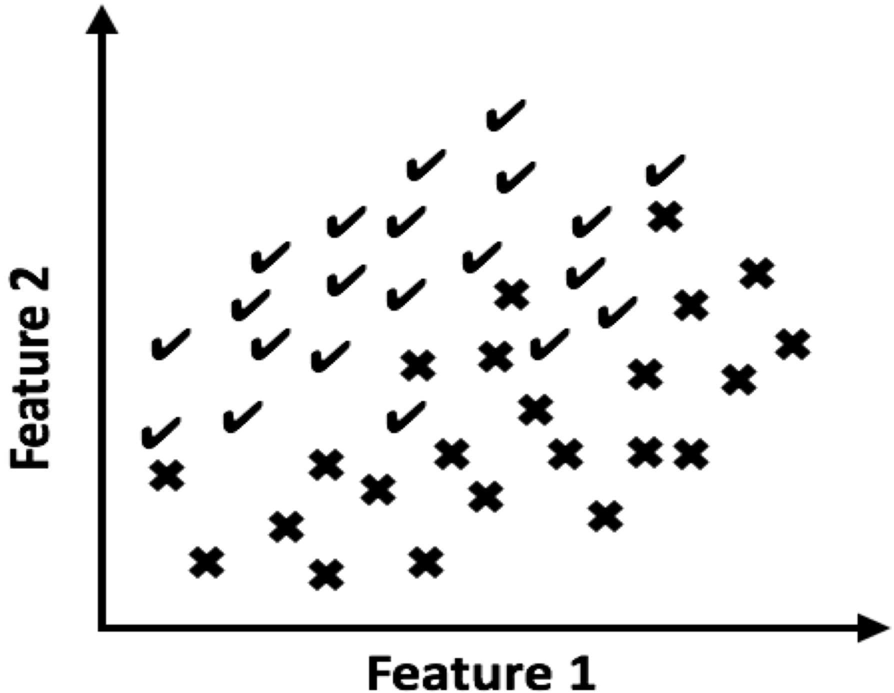
Relative disadvantage measure 3: Disparate mistreatment

DM (B1)
= 13/15 - 9/15 = .26
DM (B2)
= 10/15 – 10/15 = 0

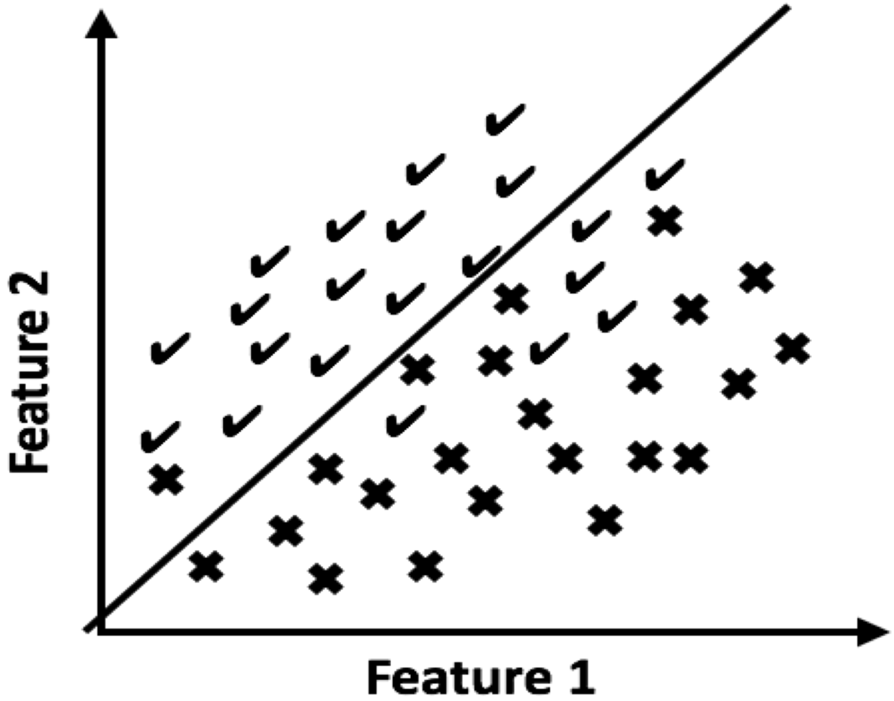


Measures the difference in fraction of accurate outcomes for different sensitive feature groups

Learning disparate mistreatment

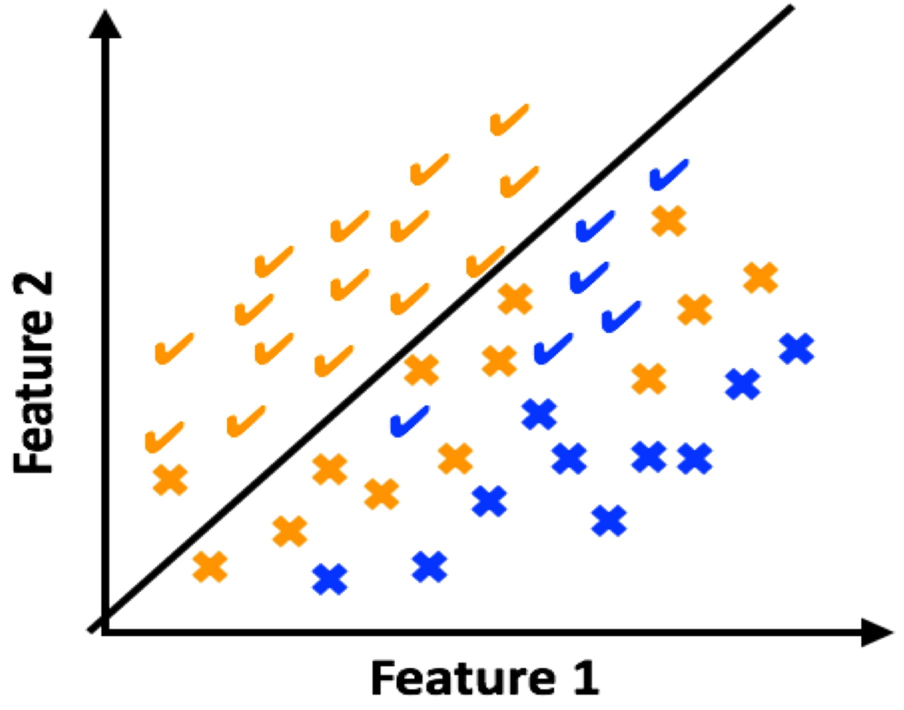


Learning disparate mistreatment



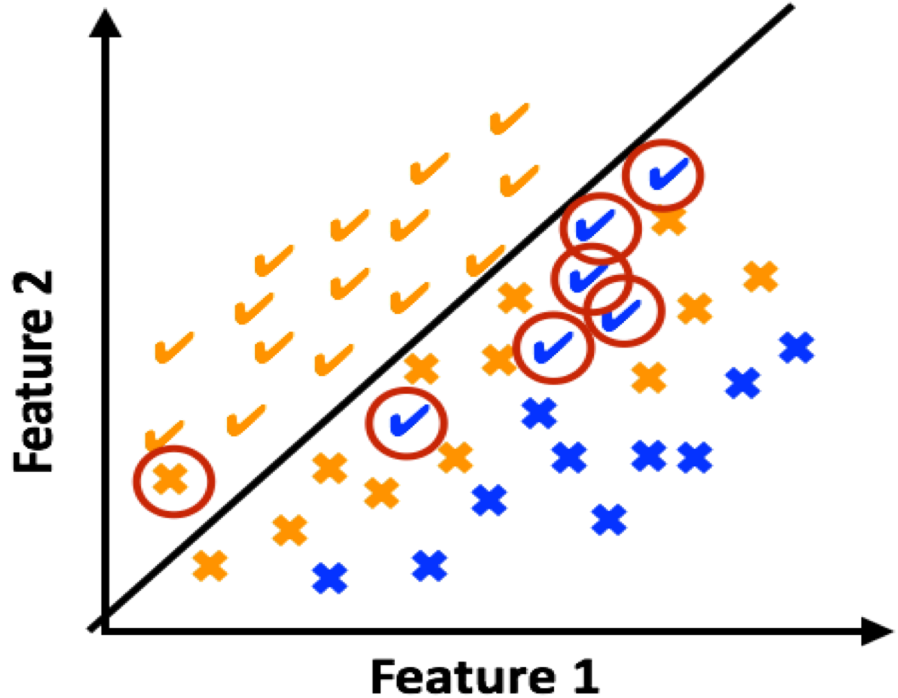
- Optimal (most accurate / least loss) linear boundary
- But, how do machines find (compute) it?
  - The boundary was computed using $\min \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$

Learning disparate mistreatment



- Optimal (most accurate / least loss) linear boundary

Learning disparate mistreatment



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

Measures indirect discrimination

- In decision making scenarios, where we have unbiased ground truth outcomes

- To achieve parity mistreatment: Provide accurate outcomes for equal fractions of sensitive feature groups

- More formally: $P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1)$

  - The above overall inaccuracy rate measure can be further broken down into its constituent FPR, FNR, FDR, and FOR

Summary: 3 discrimination measures

1. Disparate treatment: Intuitive direct discrimination
   - To avoid: $P(\hat{y}|\mathbf{x}, z) = P(\hat{y}|\mathbf{x})$

2. Disparate impact: Indirect discrimination, when ground-truth may be biased
   - To avoid: $P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$

1. Disparate mistreatment: Indirect discrimination, when ground-truth is unbiased
   - To avoid: $P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1)$

From Parity to Preference-based Discrimination Measures *[NIPS '17]*

Recap: Defining discrimination

- A first approximate normative / moralized definition:

wrongfully impose a **relative disadvantage** on persons **based on** their membership in some **salient social group** e.g., race or gender

Recap: Operationalize 4 fuzzy notions

1. What constitutes a relative disadvantage?

2. What constitutes a wrongful imposition?

3. What constitutes based on?

4. What constitutes a salient social group?

Need to operationalize 4 fuzzy notions

1. ~~What constitutes a relative disadvantage?~~

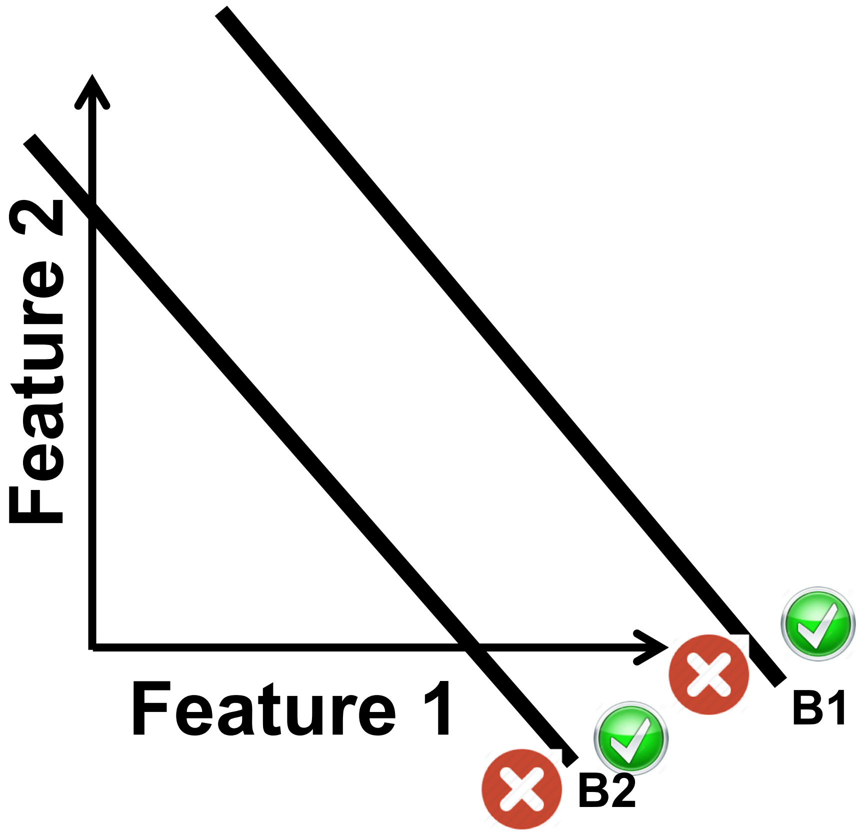
2. What constitutes a **wrongful imposition?**

3. ~~What constitutes based on?~~

4. ~~What constitutes a salient social group?~~

Is disparity in group error/acceptance rates wrong in all scenarios?

Parity error rates aren't pareto-optimal



Accuracy (B1) = 13/15 15/15

Accuracy (B2) = 09/15 09/15

Parity error rates: Picks non pareto-optimal **B2** over **B1**

Preferred error rates: Picks pareto-optimal **B1** over **B2**

Measures bargained discrimination

- Inspired by bargaining solutions in game-theory

- Disagreement (default) solution is parity!
  - Both groups try to avoid tragedy of parity

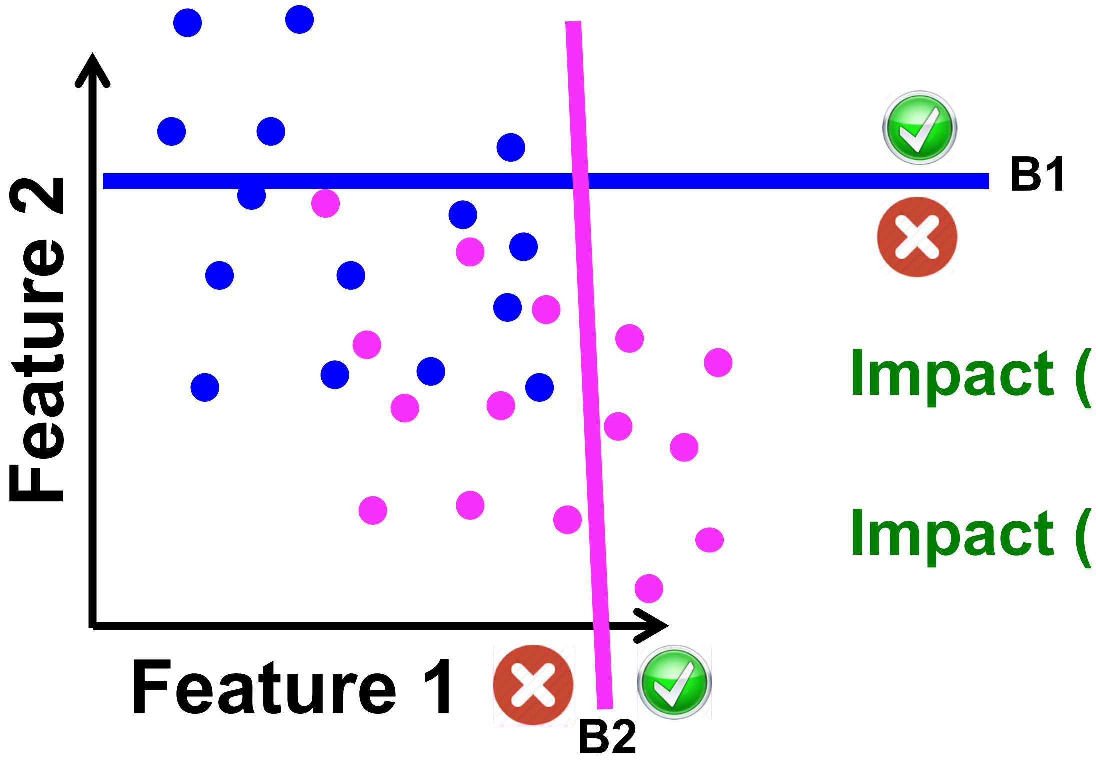
- Selects pareto-optimal boundaries over group accuracies

- More formally:

$$P(\hat{y} \neq y \mid X_{z=0}, W) \geq P(\hat{y} \neq y \mid X_{z=0}, W_{parity})$$
$$P(\hat{y} \neq y \mid X_{z=1}, W) \geq P(\hat{y} \neq y \mid X_{z=1}, W_{parity})$$

Are group-based decision boundaries discriminatory in all scenarios?

Group-based decisions can be envy-free



Impact (B1) = 5/15 0/15

Impact (B2) = 0/15 6/15

Parity treatment: Disallows group-based boundaries **B1, B2**

Preferred treatment: Allows envy-free boundaries **B1**, **B2**

Measures envy-free discrimination

- Preferred treatment allows group-conditional boundaries
- Yet, ensure they are envy-free
  - No lowering the bar to affirmatively select certain user groups
- Can be defined at individual or group-level
- More formally:

$$P(\hat{y} = 1 \mid X_{z=0}, W_{z=0}) \geq P(\hat{y} = 1 \mid X_{z=0}, W_{z=1})$$
$$P(\hat{y} = 1 \mid X_{z=1}, W_{z=1}) \geq P(\hat{y} = 1 \mid X_{z=1}, W_{z=0})$$

Summary: From parity to preference-based measures of discrimination

- Refined our three measures of discrimination
  - Disparate treatment / impact / mistreatment
  - Preferred treatment / impact / mistreatment

- The new measures allow group-conditional, envy-free, pareto-optimal boundaries
  - Can also be combined with one-another and parity measures

Operationalizing 4 fuzzy notions

- What constitutes a salient social group?
  1. Defined by anti-discrimination laws: Race, Gender
- What constitutes based on?
  1. Using salient group information in training or deployment
  2. Using salient group information in deployment, but not training
  3. Using salient group information in non envy-free boundaries
- What constitutes a relative disadvantage?
  1. Disparity in outcomes for similar users across groups
  2. Disparity in error rates across groups
  3. Disparity in acceptance rates across groups
- What constitutes a wrongful imposition?
  1. Any relative disadvantage for any group
  2. Non pareto-optimal or lower than parity advantage for any group