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

Raise concerns about their unfairness

- Implicit biases in search and recommender systems

How Google Shapes the News You See About the Candidates

Who would Google vote for? An analysis of political bias in internet search engine results

Donald Trump Accuses Google of Bias in Search Engine Results

How Google's search algorithm spreads false information with a rightwing bias

Raise concerns about their unfairness

- Discrimination in predictive risk analytics

Artificial Intelligence's White Guy Problem - The New York Times
https://www.nytimes.com/2016/06/26/.../artificial-intelligences-white-guy-problem.html
Jun 25, 2016 - Sexism, racism and other forms of **discrimination** are being built into the machine-learning **algorithms** that underlie the technology behind many ...

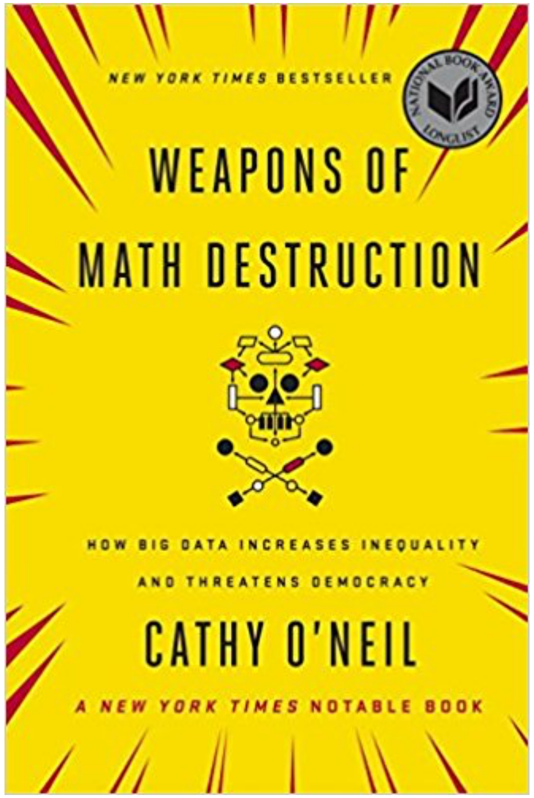
Racism is Poisoning Online Ad Delivery, Says Harvard Professor - MIT ...
https://www.technologyreview.com/.../racism-is-poisoning-online-ad-delivery-says-ha... ▾
Feb 4, 2013 - So begins Latanya Sweeney at Harvard University in a compelling paper arguing that racial **discrimination** plagues **online ad delivery**.

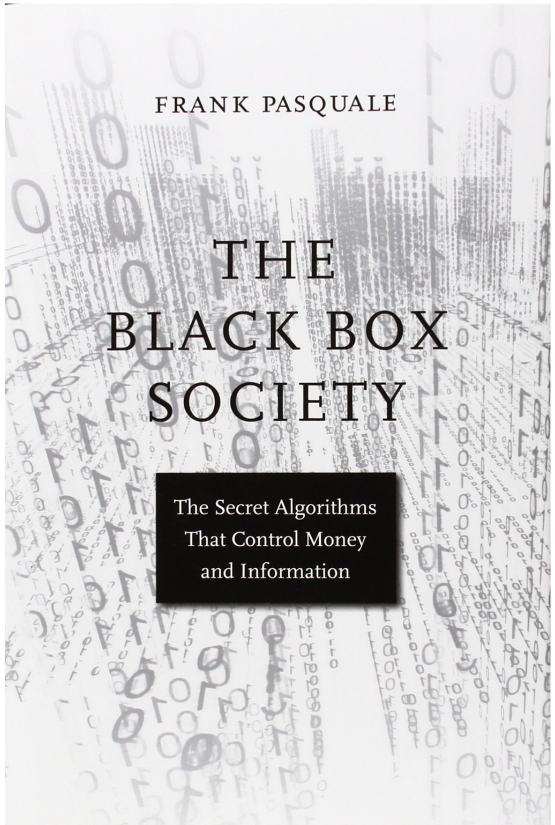
Machine Bias

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

Raise concerns about their unfairness

- Opacity of algorithmic (data-driven) decision making



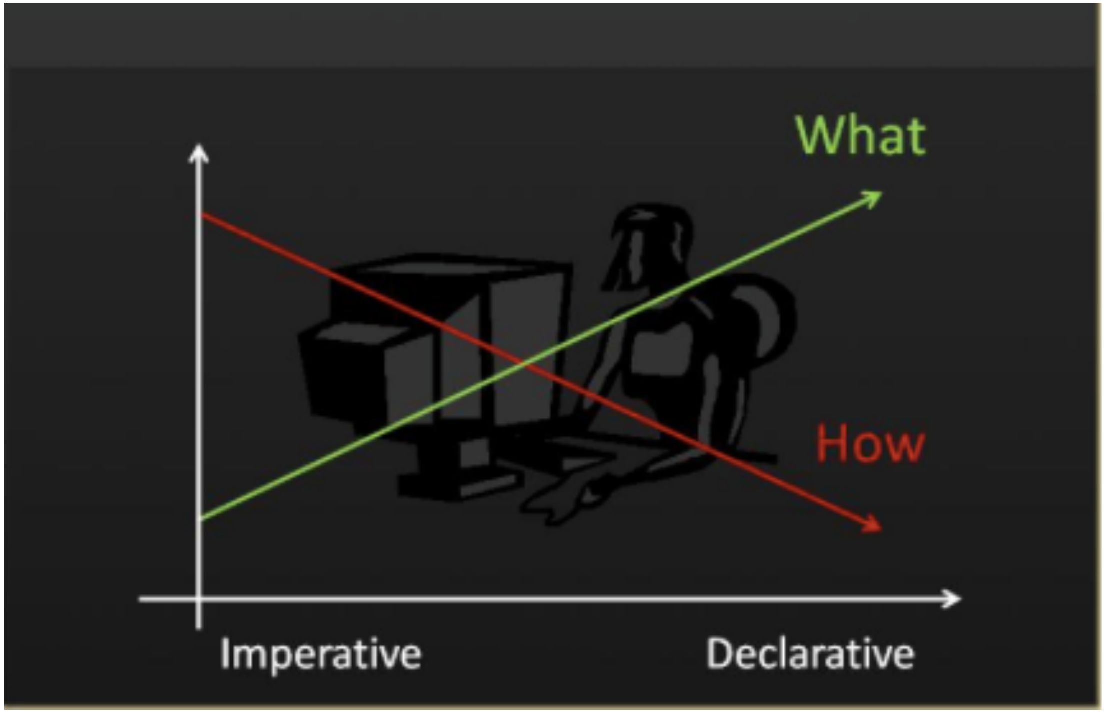


Are the unfairness concerns justified?

How we engineer machine decisions

- Imperative programming:
  - You describe the procedure for making decisions
    - Not what you want from the decisions
- Declarative programming:
  - You declare the outcome goals of your decision making
    - Not how you want to make decisions
  - Leveraging machines to find optimal decision procedure

Imperative vs. Declarative Engineering



- Is one programming style better than other?

The excitement about AI/ML

- Can get away with lazy declarative engineering
  - Get some training data – examples of past decisions
  - Declare a default goal – decision prediction accuracy

- Miraculously, lazy engineering appears to work!
  - But, does it really work?

The achilles heels of lazy AI/ML

Even assuming no training data biases, AI/ML decisions

1. Optimize for a single decision outcome goal, ignoring
   - Fairness: Equal prediction accuracy for all salient social groups
   - Worst-cases: Lower bound worst-case prediction accuracy
   - Norms: Should use or not use data in a specific manner

2. Optimal for a static NOT an evolving society, because
   - Training data becomes unrepresentative
   - Feedback loops are not accounted for in the first place
   - Decision outcome goals change over time!

Can we guard the achilles heels?

- Can we account for fairness & other norms in ML decision making?
  - Maybe! Even with declarative engineering
    - Declare multiple decision outcome objectives when training

- Can we design ML decision making for an evolving society?
  - Not sure! Need more imperative / procedural engineering

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

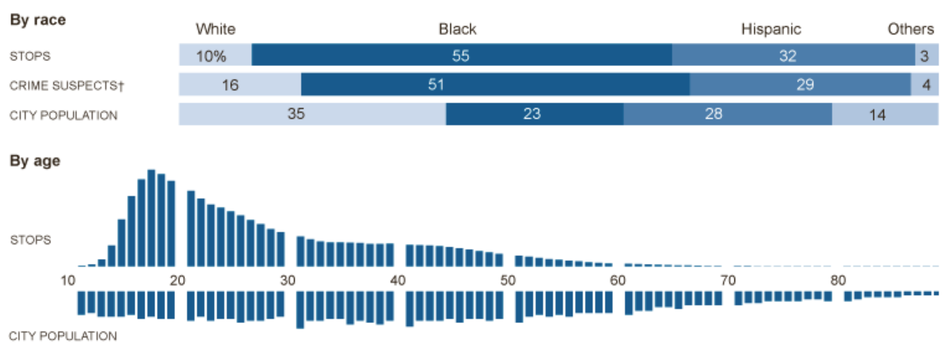
Part 1:

Why is a computational perspective on discrimination needed?

Why, a computational perspective?

Datamining/ML is increasingly being used to detect discrimination in human/machine decision making

- Examples: NYPD stop and frisk, Airbnb rentals



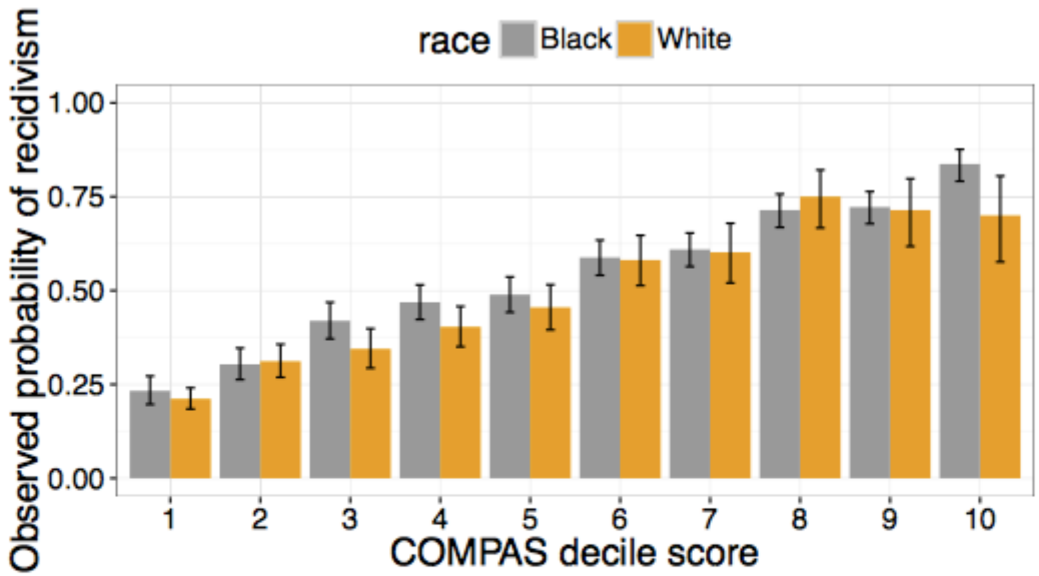
Case study: Recidivism risk prediction

- 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 non-intuitive 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?

Part 2:

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?~~

Operationalizing discrimination

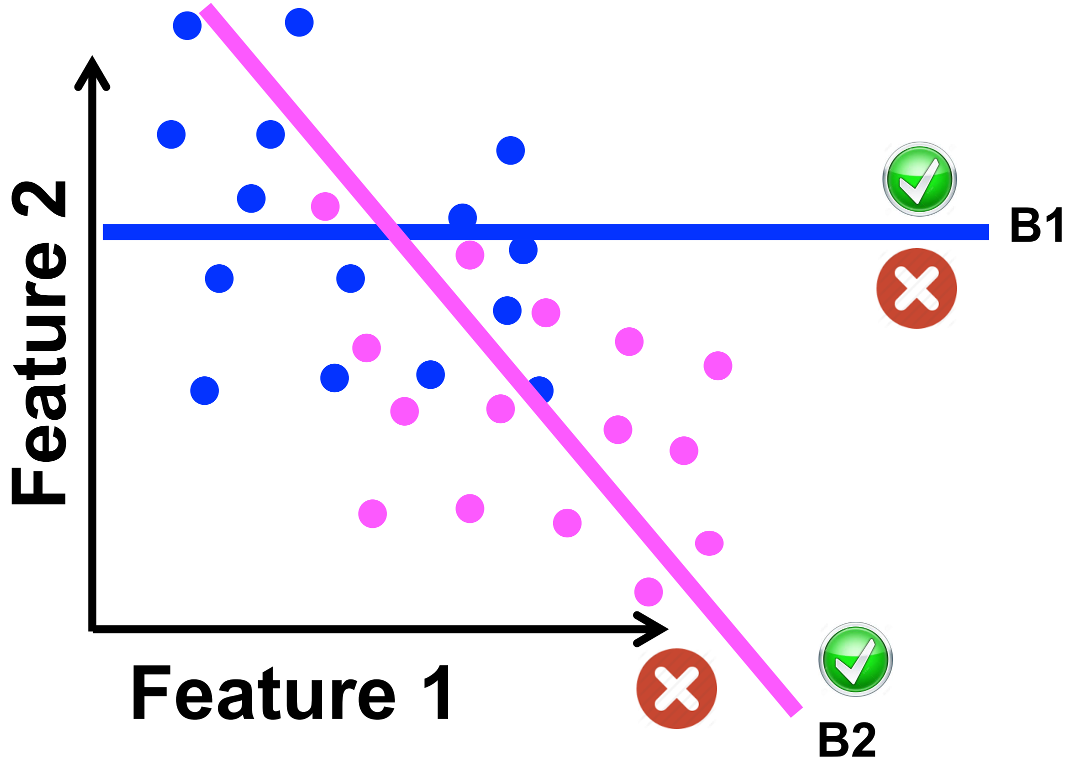
- Consider binary classification using user features

	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

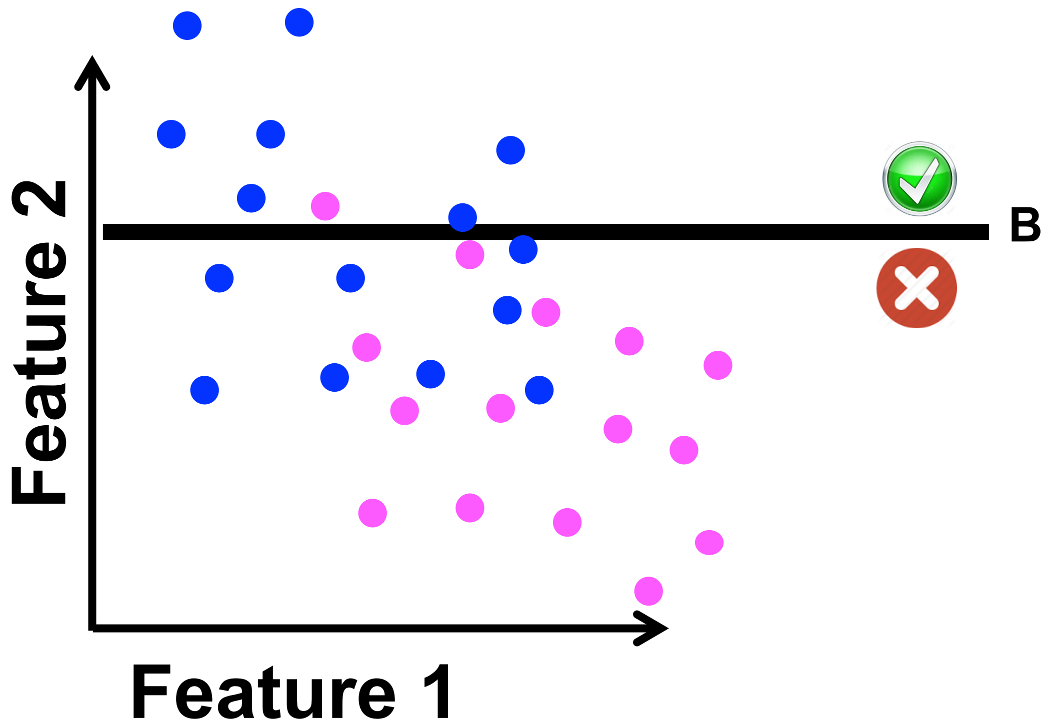
Decision outcomes should not be **relatively disadvantageous** to social (sensitive feature) groups!

Relative disadvantage measure 1: Disparate treatment



Measures the difference in outcomes for users, when their sensitive features are changed

Relative disadvantage measure 1: Disparate treatment



Measures the difference in outcomes for users, when their sensitive features are changed

Measures direct discrimination

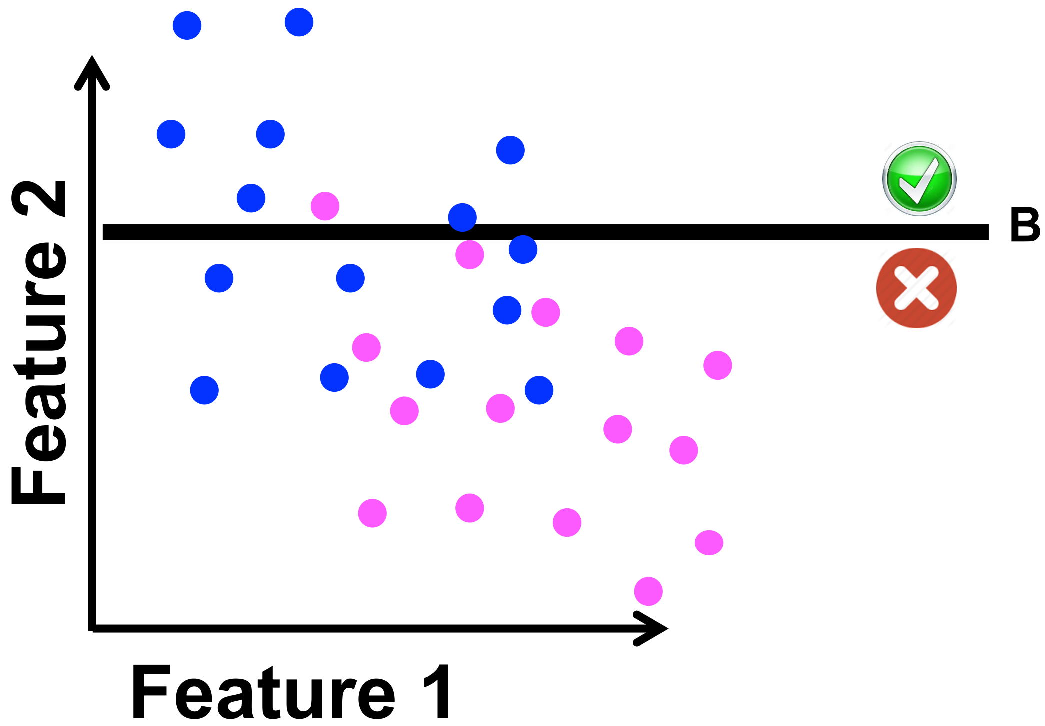
- Based on counter-factual reasoning
  - Most intuitive measure of discrimination

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

- Situational testing for discrimination discovery checks for disparate treatment

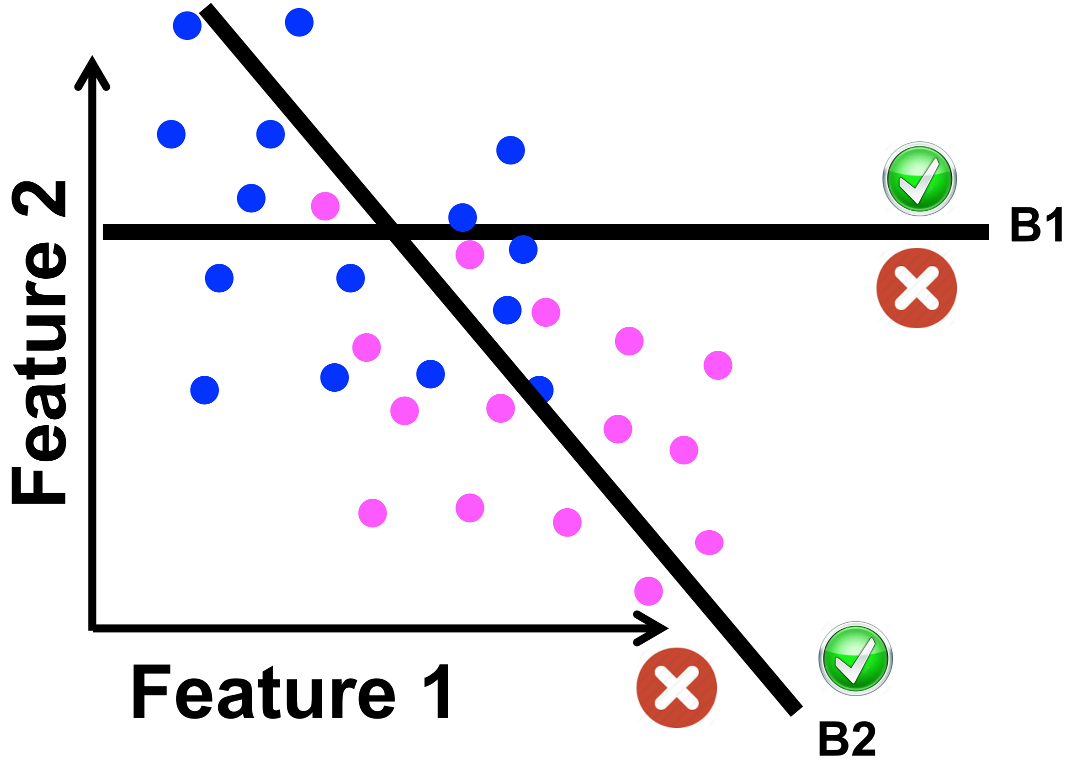
- More formally: $P(\hat{y}|\mathbf{x}, z) = P(\hat{y}|\mathbf{x})$

Relative disadvantage measure 2: Disparate impact



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

Relative disadvantage measure 2: Disparate impact



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

- Notoriously hard to detect indirect discrimination
  - In decision making scenarios where ground truth on intent is unknown or ground truth on outcomes may be biased

Detecting indirect discrimination

- Doctrine of disparate impact
  - A US law applied in employment & housing practices
- Proportionality tests over decision outcomes
  - E.g., in 70’s and 80’s, some US courts applied the 80% rule for employment practices
    - If 50% (P1%) of male applicants get selected at least 40% (P2%) of female applicants must be selected
  - UK uses P1 – P2; EU uses (1-P1) / (1-P2)
  - Fair proportion thresholds may vary across different domains

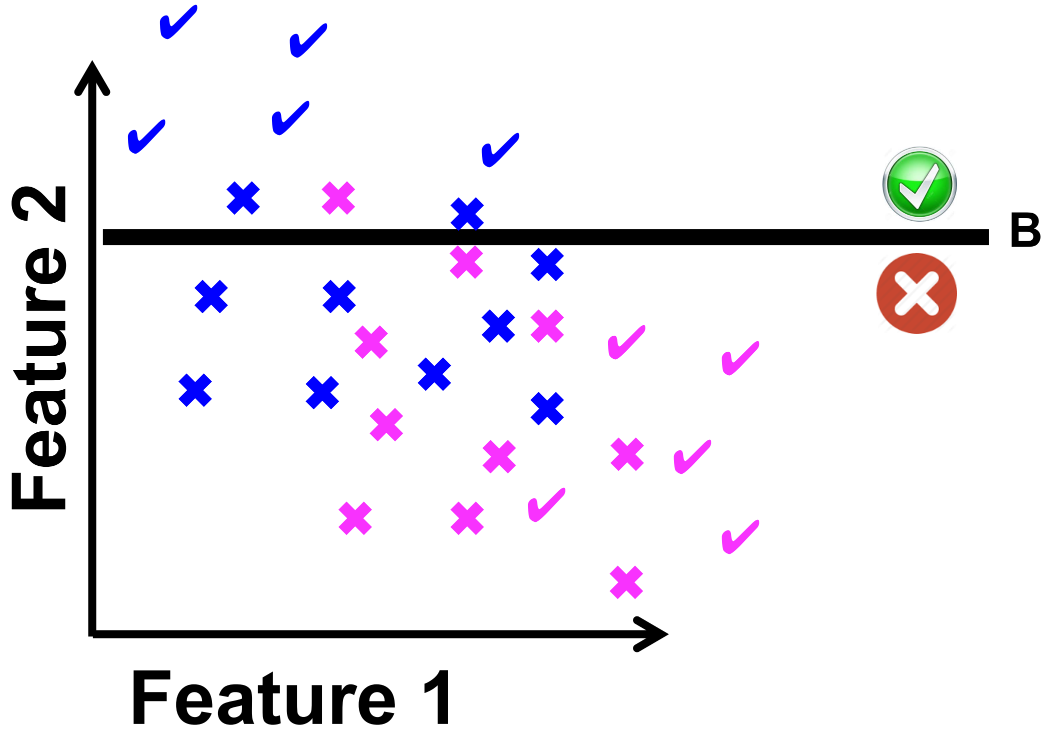
A controversial measure

- To achieve parity impact: Select equal fractions of sensitive feature groups

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

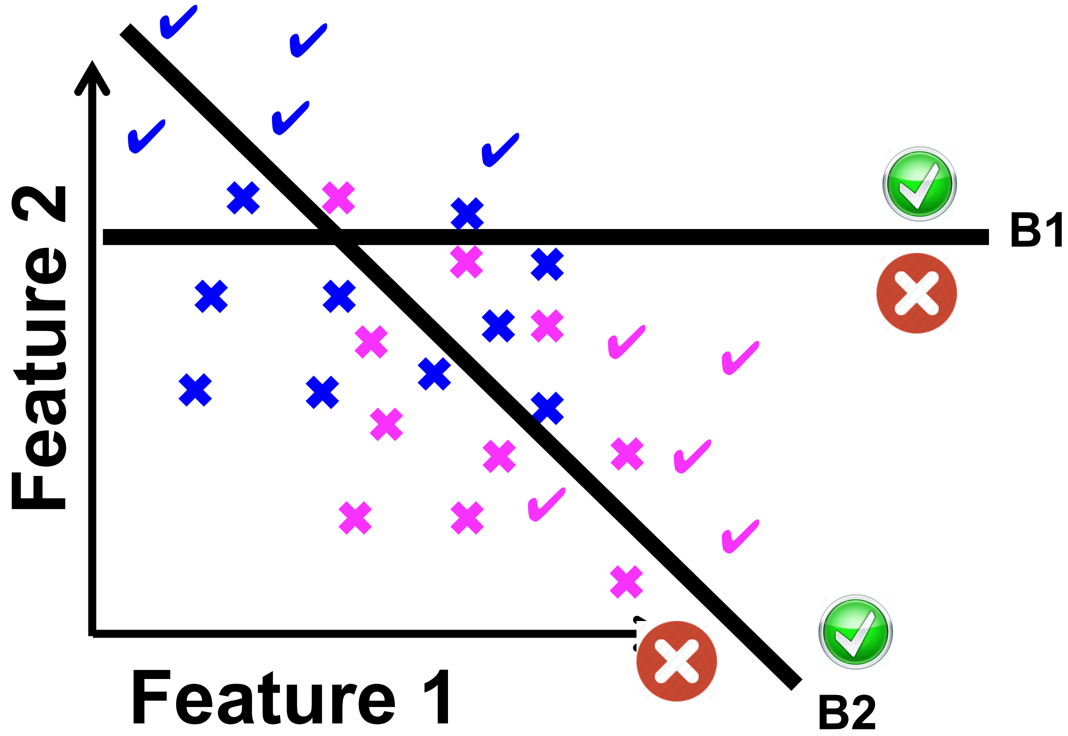
  - Critics: There exist scenarios where disproportional outcomes are justifiable
  - Supporters: Provision for business necessity exists
    - Though the burden of proof is on employers
    - Law is necessary to detect indirect discrimination!

Relative disadvantage measure 3: Disparate mistreatment



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

Relative disadvantage measure 3: Disparate mistreatment



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

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: $P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$

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

Part 3:

Mechanisms for Non-discriminatory Machine Learning *[AISTATS '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

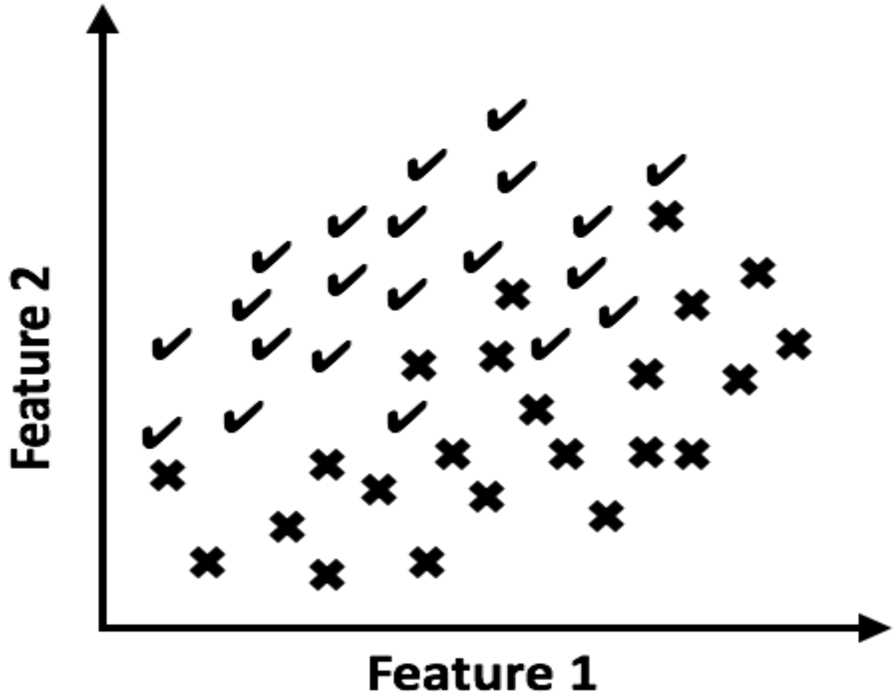
- By training over historical data
- Example task: Predict who will return loan

	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

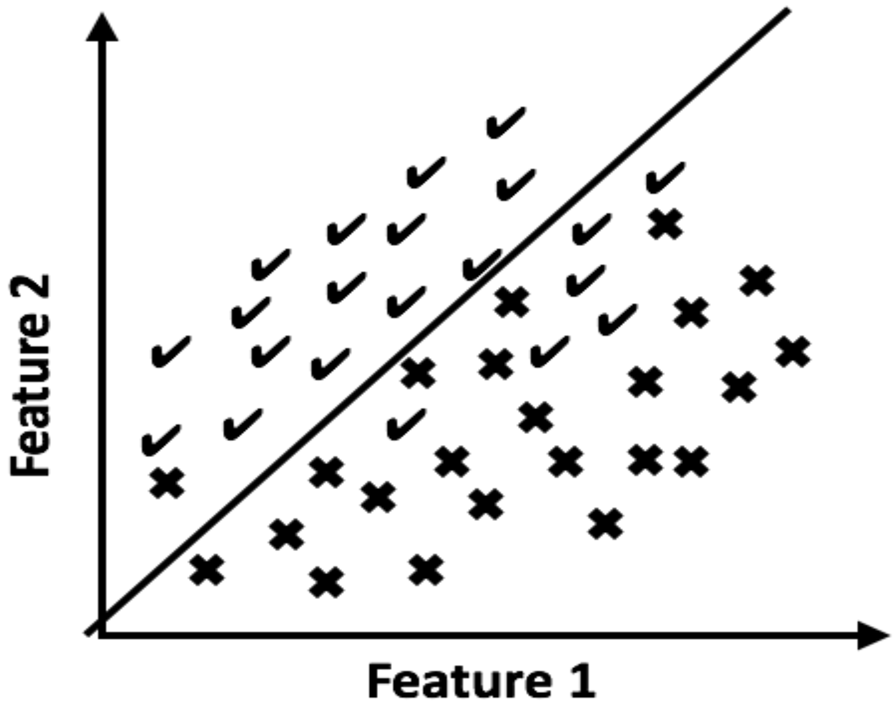
Past Outcomes
Returned
Defaulted
Defaulted
…
Returned

- 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

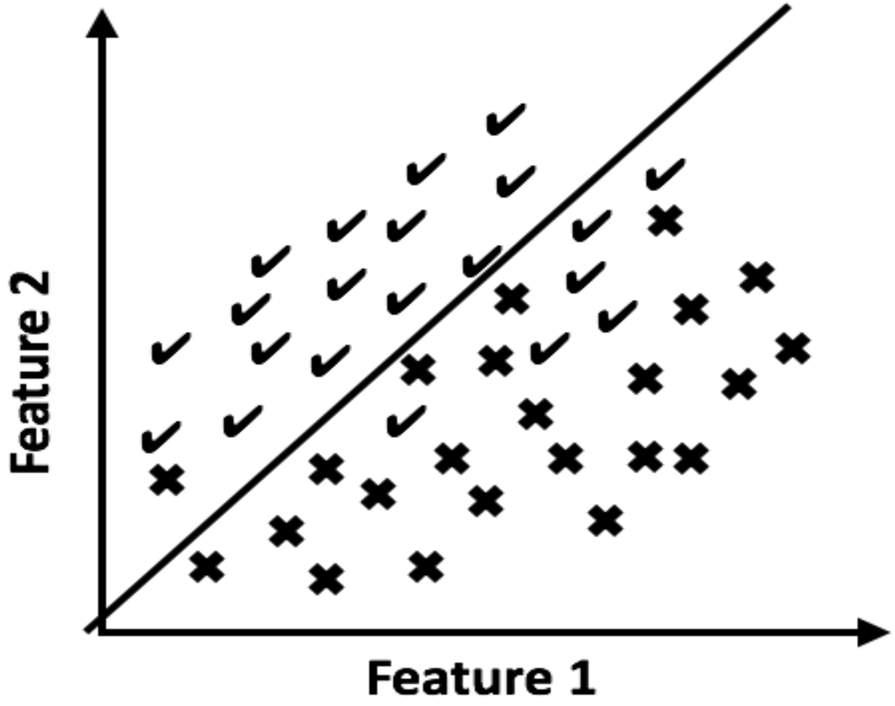
$$L(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \qquad L(\mathbf{w}) = \sum_{i=1}^{N} -\log p(y_i|\mathbf{x}_i, \mathbf{w})$$

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

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

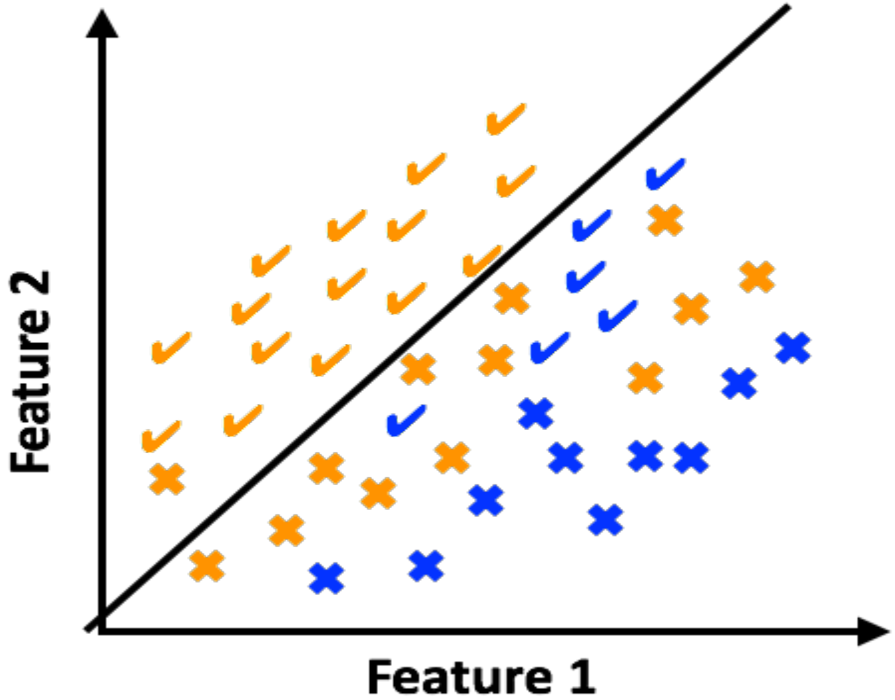
- Central challenge in machine learning
  - Finding loss function that capture prediction loss, yet be efficiently optimized
  - Many loss functions used in learning are convex

Predict who will return loans



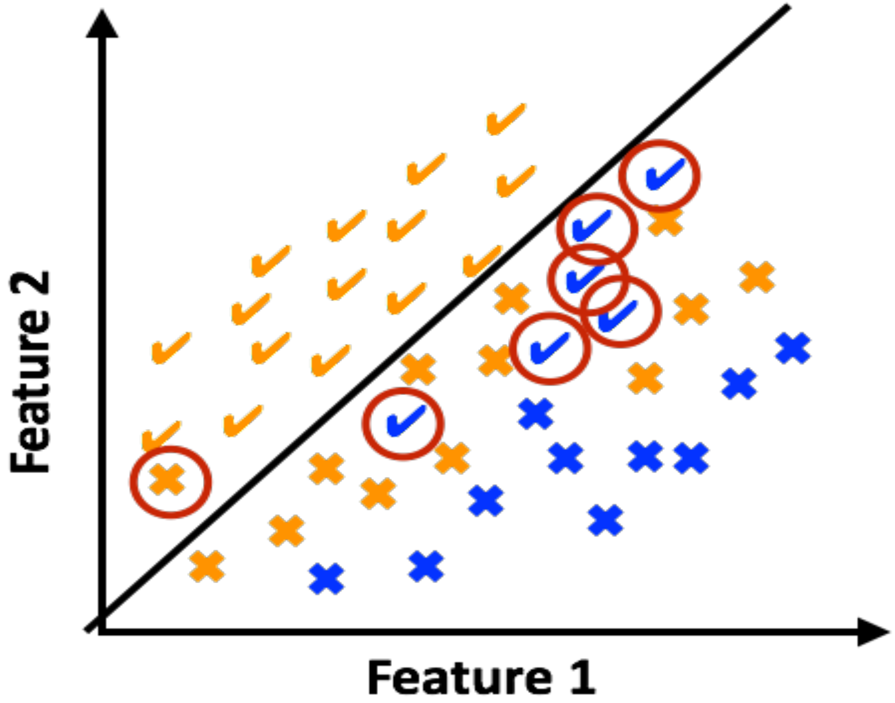
- Optimal (most accurate / least loss) linear boundary
- But, how do machines find (compute) it?
  - The boundary was computed using $\min \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$

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

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

$$\textit{subject to} \quad P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1)$$

- The constraints embed ethics & values when learning
- No free lunch: Additional constraints lower accuracy
  - Tradeoff between performance & ethics (avoid discrimination)

A few observations

- Any discrimination measure could be a constraint

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

- Might not need all constraints at the same time
  - E.g., drop disp. impact constraint when no bias in data
  - When avoiding disp. impact / mistreatment, we could achieve higher accuracy without disp. treatment

Key technical challenge

- How to learn efficiently under these constraints?

$$\mathit{minimize}\ L(\mathbf{w})$$
$$\mathit{subject\ to}\quad P(\hat{y}=1|z=0)=P(\hat{y}=1|z=1)$$

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

- 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

Specifying disparate impact constraints

- Instead of requiring: $P(\hat{y} = 1 | z = 0) = P(\hat{y} = 1 | z = 1)$

- Bound covariance between items' sensitive feature values and their signed distance from classifier's decision boundary to less than a threshold

$$\boxed{\left| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}}) \, \mathbf{w}^{\mathbf{T}} \mathbf{x}_i \right| \leq \mathbf{c}}$$

Learning classifiers w/o disparate impact

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

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

- New formulation: Convex, easy-to-learn

$$\begin{array}{ll} \textit{minimize} & L(\mathbf{w}) \\ \textit{subject to} & \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}}) \, \mathbf{w}^T \mathbf{x}_i \leq \mathbf{c} \\ & \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}}) \, \mathbf{w}^T \mathbf{x}_i \geq -\mathbf{c} \end{array}$$

A few observations

- Our formulation can be applied to any convex-margin (loss functions) based classifiers
  - hinge-loss, logistic loss, linear and non-linear SVM
- Can easily change our formulation to optimize for fairness under accuracy constraints
  - Useful in practice, when you want to be fair but have business necessity to meet a certain accuracy threshold

Learning classifiers w/o disparate mistreatment

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

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

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

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

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

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

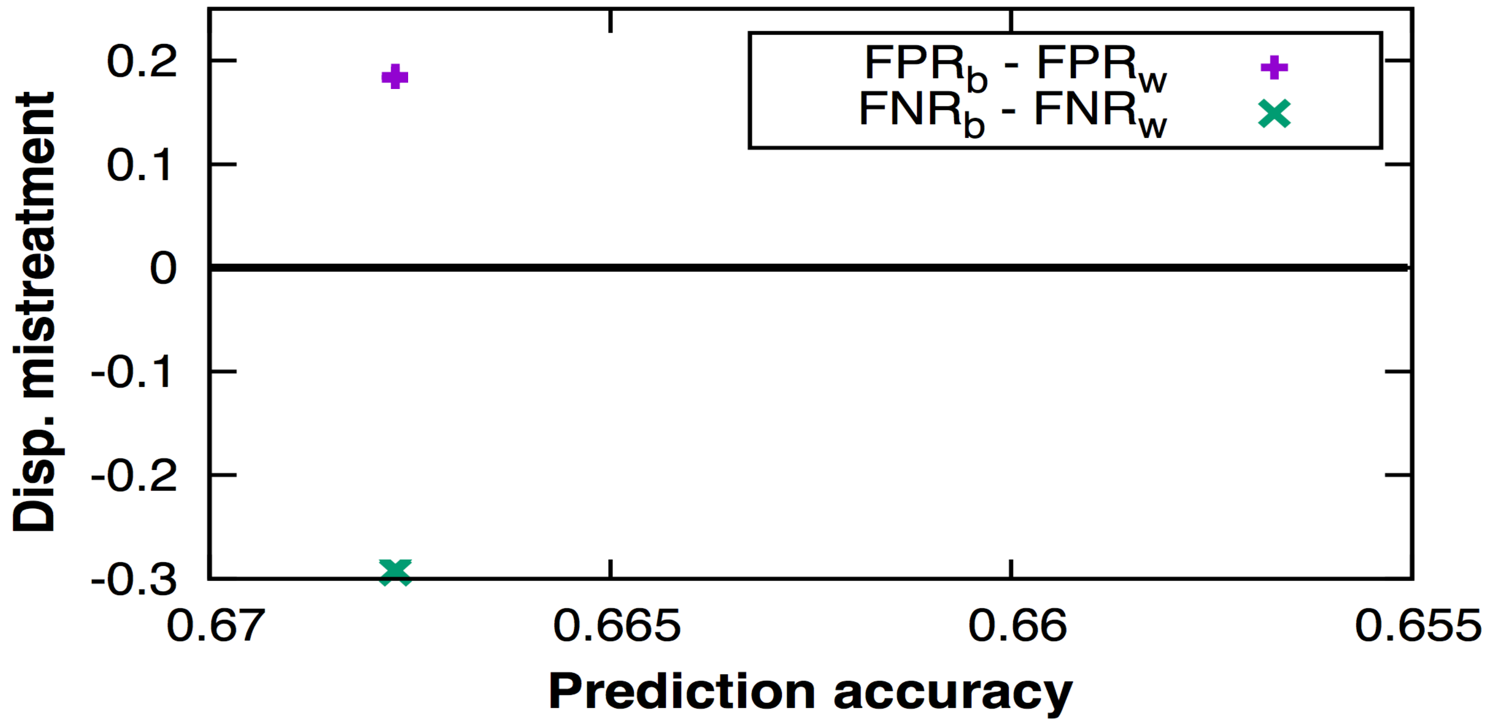
Key evaluation questions

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

- Can our approach help avoid disparate mistreatment?

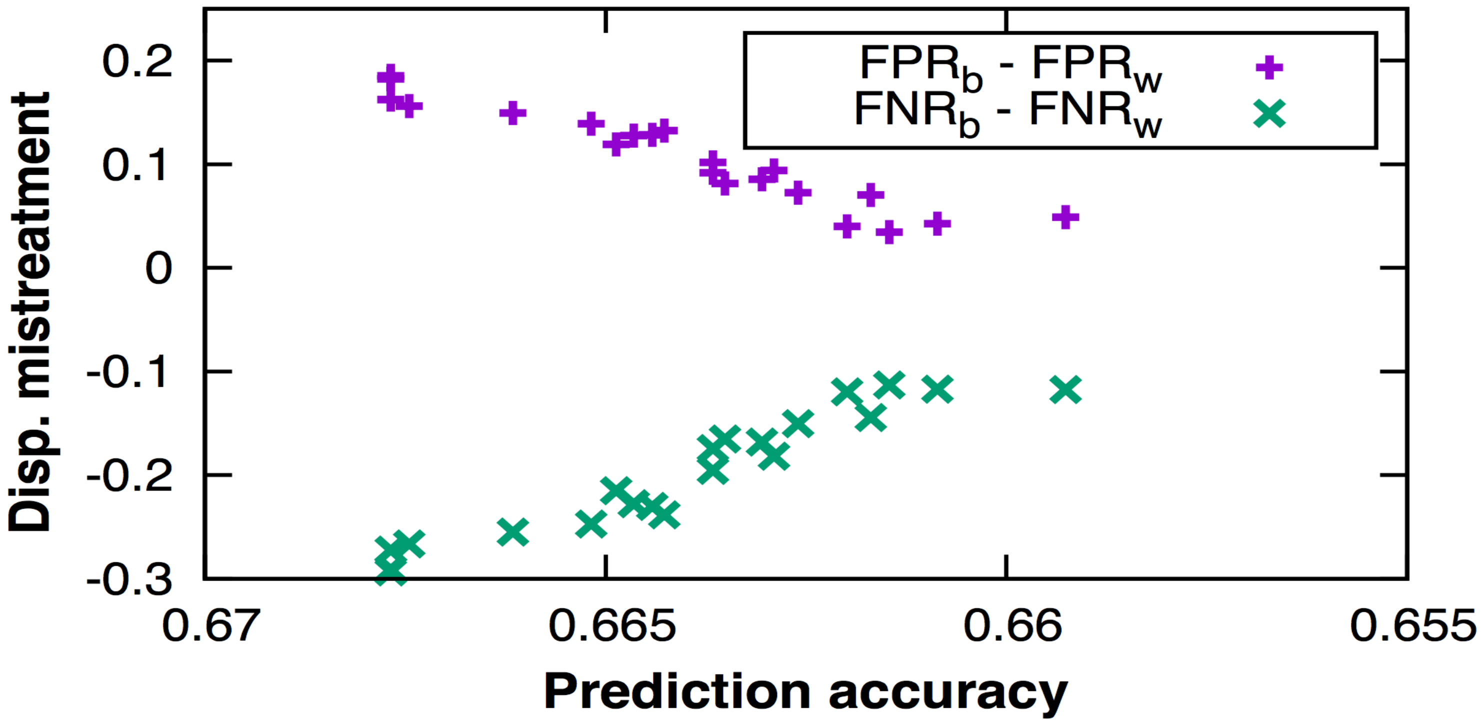
Removing disparate mistreatment

- 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 three measures of discrimination
  - disparate treatment / impact / mistreatment
  - 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

Part 4:

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?~~

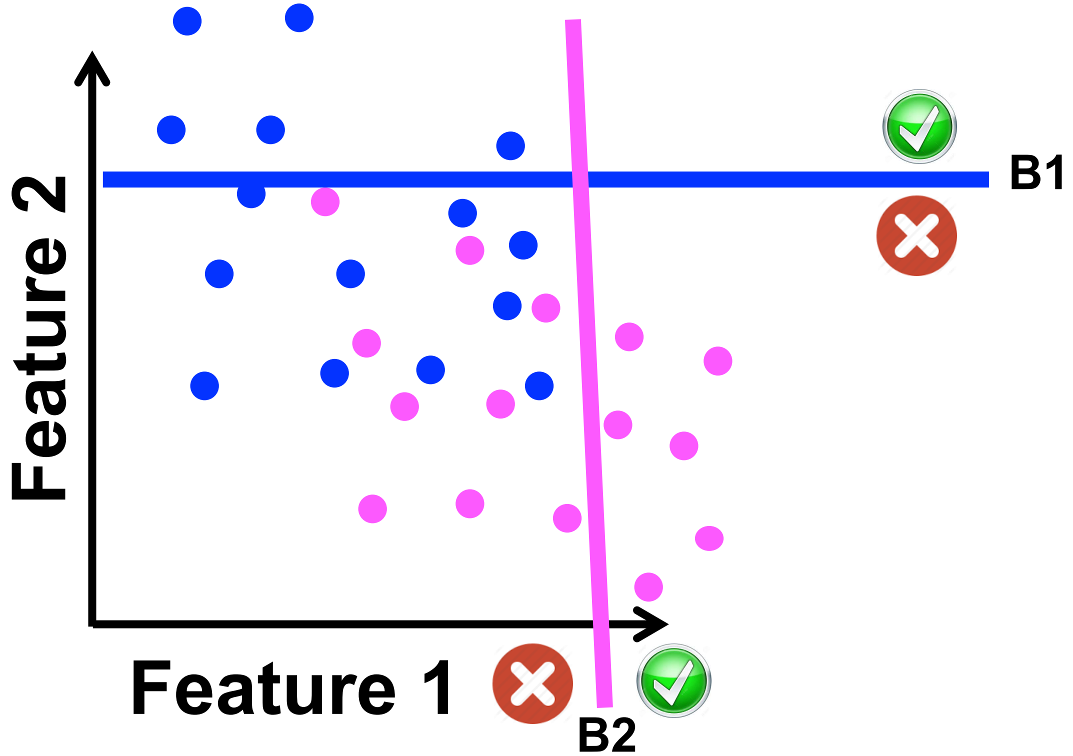
Revisit relative disadvantage measure 1: Disparate treatment

- Parity treatment: Changing sensitive feature should not change outcomes

- Equivalent to having same boundary for all groups

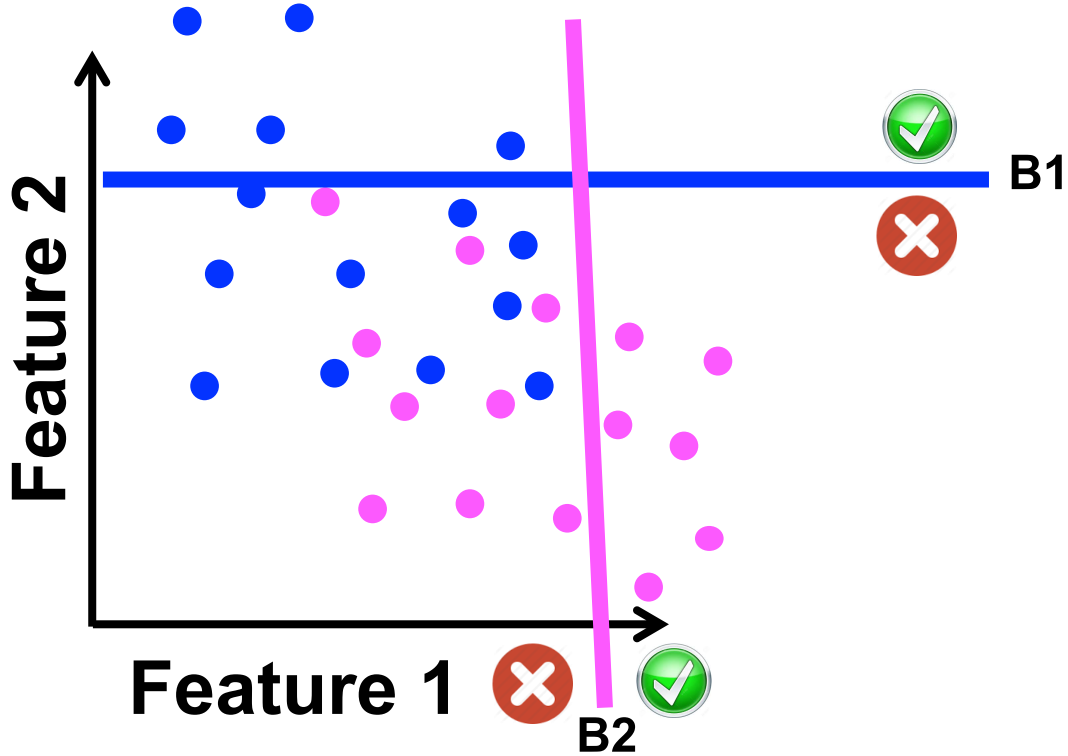
- Do there exist scenarios where group-conditional boundaries are not wrong?

Relative disadvantage measure 4:
From Disparate treatment to Preferred treatment



Disparate treatment: Measures the difference in outcomes for users, when their sensitive features are changed

Relative disadvantage measure 4: From Disparate treatment to Preferred treatment



Preferred treatment: Measures the **increase** in positive outcomes for users, when their sensitive features are changed

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:

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

Learning preferred treatment classifiers

$$\text{Minimize } L_{z=0}(W_{z=0}) + L_{z=1}(W_{z=1})$$

Subject to

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

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

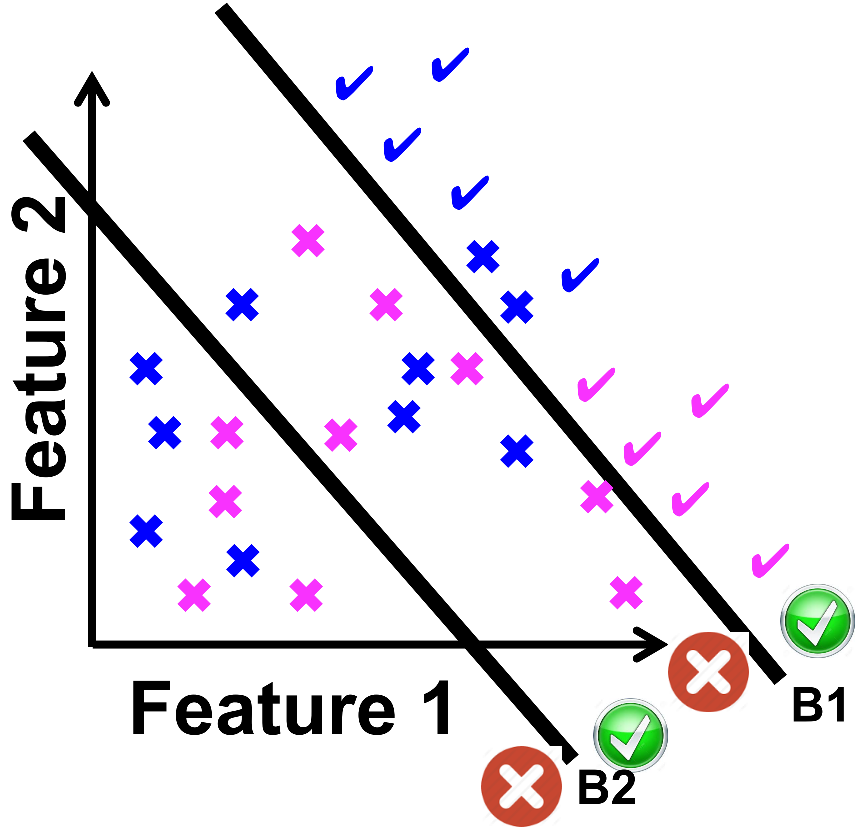
- Preferred treatment subsumes parity treatment
  - Every parity treatment classifier offers preferred treatment
- Preferred treatment constraint is weaker than parity
  - Suffers lower cost of fairness

Revisit relative disadvantage measure 3: Disparate mistreatment

- Parity mistreatment: Provide accurate outcomes for equal fractions of sensitive feature groups

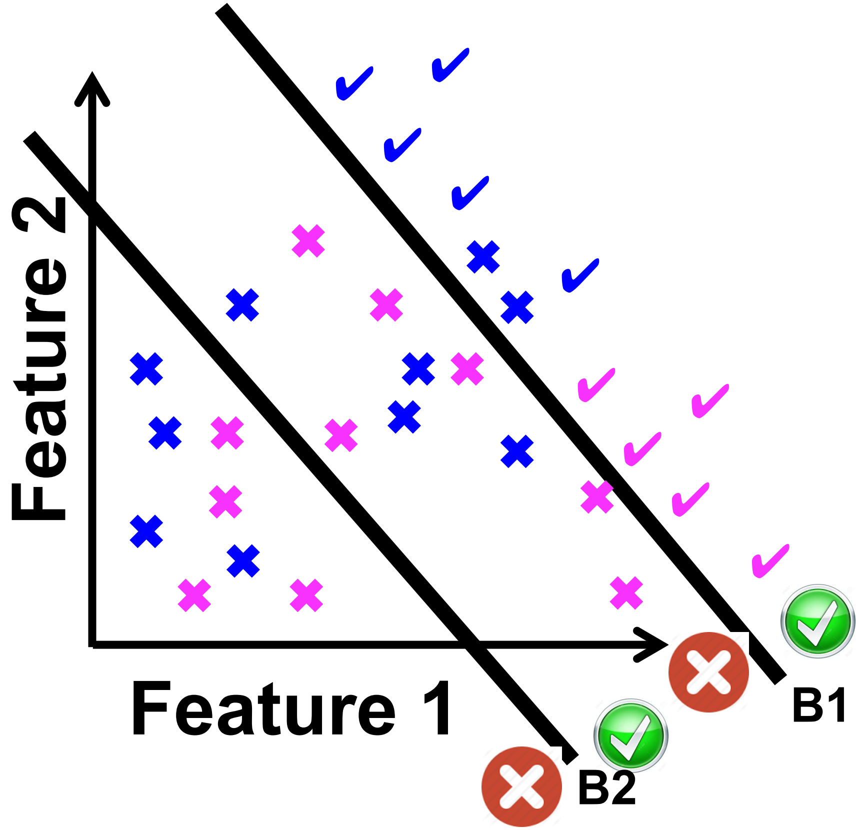
- Do there exist scenarios where differences in outcome accuracies for groups are not wrong?

Relative disadvantage measure 5: From Disparate to Preferred mistreatment



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

Relative disadvantage measure 5:
From Disparate to Preferred mistreatment



Preferred mistreatment: Measures the difference in fraction of accurate outcomes **relative to parity** for different sensitive feature groups

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:

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

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
- Proposed mechanisms for mitigating each of them
  - Formulated the measures as constraints that can be learned

Part 4:

Open Challenges Towards Non-Discriminatory Decision Making

Beyond binary classification

How to learn

- Fair **regression**
  - Applicable principle: **Non-Discrimination**
- Fair **multi-class classification**
  - Applicable principle: **De-Segregation**
- Fair **set selection**
  - Applicable principle: **Fair Representation**
- Fair **ranking**
  - Applicable principle: **Fair Scheduling**

From distributive to procedural fairness

- Current fairness notions based on **outcomes**

- Ignores fairness of the **process** of making decisions
  - Today's recidivism risk prediction tools use features like
    - Juvenile crime history, family criminality, work/social environment
  - Raise concerns about their usage because of
    - Privacy norms, their non-volitional nature, reliability of assessment, relevance to decision, vicious causal cycle

- How can we account for these factors in decisions?

Foundations for Fair Machine Decision Making

- Distributive fairness: Fairness of outcomes
  - Non-discriminatory, de-segregation, fair representation, fair sharing
- Procedural fairness: Fairness of process
  - Privacy of inputs, diversity of decision processes, evolution of decision processes
- Informational fairness: Transparency of outcomes and process
  - Understandability for designers, controllability for end users, and verifiability for regulators

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.

Fair classifier implementation at:

fate-computing.mpi-sws.org

Related References

- Dino Pedreshi, Salvatore Ruggieri and Franco Turini. *Discrimination-aware Data Mining*. In Proc. KDD, 2008.
- Faisal Kamiran and Toon Calders. *Classifying Without Discriminating*. In Proc. IC4, 2009.
- Faisal Kamiran and Toon Calders. *Classification with No Discrimination by Preferential Sampling*. In Proc. BENELEARN, 2010.
- Toon Calders and Sicco Verwer. *Three Naive Bayes Approaches for Discrimination-Free Classification*. In Data Mining and Knowledge Discovery, 2010.
- Indrė Žliobaitė, Faisal Kamiran and Toon Calders. *Handling Conditional Discrimination*. In Proc. ICDM, 2011.
- Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh and Jun Sakuma. *Fairness-aware Classifier with Prejudice Remover Regularizer*. In PADM, 2011.
- Binh Thanh Luong, Salvatore Ruggieri and Franco Turini. *k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention*. In Proc. KDD, 2011.

Related References

- Faisal Kamiran, Asim Karim and Xiangliang Zhang. *Decision Theory for Discrimination-aware Classification*. In Proc. ICDM, 2012.
- Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold and Rich Zemel. *Fairness Through Awareness*. In Proc. ITCS, 2012.
- Sara Hajian and Josep Domingo-Ferrer. *A Methodology for Direct and Indirect Discrimination Prevention in Data Mining*. In TKDE, 2012.
- Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, Cynthia Dwork. *Learning Fair Representations*. In ICML, 2013.
- Andrea Romei, Salvatore Ruggieri. *A Multidisciplinary Survey on Discrimination Analysis*. In KER, 2014.
- Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. *Certifying and Removing Disparate Impact*. In Proc. KDD, 2015.
- Moritz Hardt, Eric Price, Nathan Srebro. *Equality of Opportunity in Supervised Learning*. In Proc. NIPS, 2016.
- Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan. *Inherent Trade-Offs in the Fair Determination of Risk Scores*. In FATML, 2016.