Discrimination in Decision Making: Humans vs. Machines

Muhammad Bilal Zafar, Isabel Valera,

Manuel Gomez-Rodriguez, 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
  - Journalism: News recommender systems

The concept of discrimination

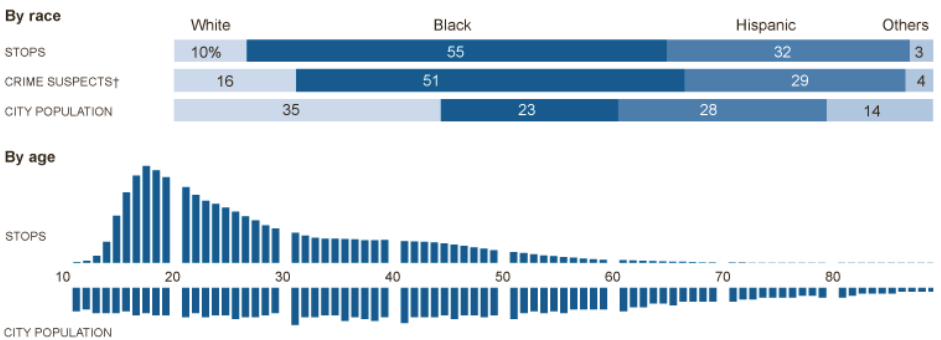
- Discrimination is a special type of unfairness
- Well-studied in social sciences
  - Political science
  - Moral philosophy
  - Economics
  - Law
    - Majority of countries have anti-discrimination laws
    - Discrimination recognized in several international human rights laws

- But, less-studied from a computational perspective

Why, a computational perspective?

1. Datamining is increasingly being used to detect discrimination in human decision making

- Examples: NYPD stop and frisk, Airbnb rentals



Why, a computational perspective?

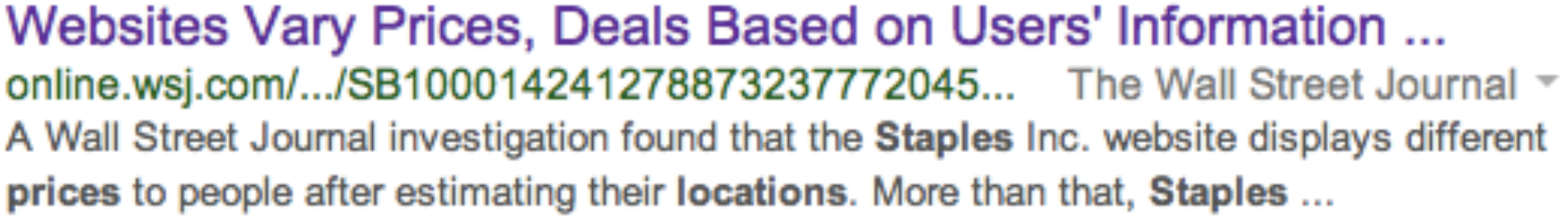
2. Learning to avoid discrimination in data-driven (algorithmic) decision making

- Aren't algorithmic decisions inherently objective?
  - In contrast to subjective human decisions
- Doesn't that make them fair & non-discriminatory?

- Objective decisions can be unfair & discriminatory!

Why, a computational perspective?

- Learning to avoid discrimination in data-driven (algorithmic) decision making
  - *A priori* discrimination in biased training data
    - Algorithms will objectively learn the biases
  - Learning objectives target decision accuracy over all users
    - Ignoring outcome disparity for different sub-groups of users



Our agenda: Two high-level questions

1. How to detect discrimination in decision making?
   - Independently of who makes the decisions
     - Humans or machines

2. How to avoid discrimination when learning?
   - Can we make algorithmic decisions more fair?
   - If so, algorithms could eliminate biases in human decisions
     - Controlling algorithms may be easier than retraining people

This talk

1. ~~How to detect discrimination in decision making?~~
   - ~~Independently of who makes the decisions~~
     - ~~Humans or machines~~

2. How to avoid discrimination when learning?
   - Can we make algorithmic decisions more fair?
   - If so, algorithms could eliminate biases in human decisions
     - Controlling algorithms may be easier than retraining people

The concept of 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

The concept of 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

The devil is in the details

- What constitutes a salient social group?
  - A question for political and social scientists
- What constitutes relative disadvantage?
  - A question for economists and lawyers
- What constitutes a wrongful decision?
  - A question for moral-philosophers
- What constitutes **based on?**
  - A question for **computer scientists**

Discrimination: A computational perspective

- Consider binary classification using user attributes

	A_1	A_2	…	A_m
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
…	…			…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

Discrimination: A computational perspective

- Consider binary classification using user attributes

	SA_1	NSA_2	…	NSA_m
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
…	…			…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

- Some attributes are sensitive, others non-sensitive

Discrimination: A computational perspective

- Consider binary classification using user attributes

	SA_1	NSA_2	...	NSA_m
$User_1$	$x_{1,1}$	$x_{1,2}$	...	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
...	...			...
$User_n$	$x_{n,1}$	$x_{n,2}$	...	$x_{n,m}$

Decision
Accept
Reject
Reject
...
Accept

- Some attributes are sensitive, others non-sensitive

Decisions should **not be based on** sensitive attributes!

What constitutes "not based on"?

- Most intuitive notion: Ignore sensitive attributes
  - Fairness through blindness or veil of ignorance
- When learning, strip sensitive attributes from inputs
- Avoids disparate treatment
  - Same treatment for users with same non-sensitive attributes
    - Irrespective of their sensitive attribute values

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

    - Situational testing for discrimination discovery checks for this condition

Two problems with the intuitive notion

When users of different sensitive attribute groups have different non-sensitive feature distributions, we risk

1. Disparate Mistreatment
   - Even when training data is unbiased, sensitive attribute groups might have different misclassification rates

2. Disparate Impact
   - When labels in training data are biased, sensitive attribute groups might see different beneficial outcomes to different extents
     - Training data bias due to past discrimination

Background: Two points about learning

1. To learn, we define & optimize a risk (loss) function
   - Over all examples in training data

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

   - Risk function captures inaccuracy in prediction
   - So learning is cast as an optimization problem

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

2. For efficient learning (optimization)
   - We define loss functions so that they are convex

Origins of disparate mistreatment

	SA_1	NSA_2	…	NSA_m
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
…	…			…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

Origins of disparate mistreatment

	SA_1	NSA_2	…	NSA_m
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
…	…			…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

- Suppose users are of two types: blue and pink

Origins of disparate mistreatment

	SA_1	NSA_2	…	NSA_m
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
…	…			…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

- Minimizing ***L(W)***, does not guarantee ***L(W)*** and ***L (W)*** are equally minimized
  - Blue users might have a different risk / loss than red users!

Origins of disparate mistreatment

	~~SA_1~~	NSA_2	…	NSA_m
$User_1$	~~$x_{1,1}$~~	$x_{1,2}$	…	$x_{1,m}$
$User_2$	~~$x_{2,1}$~~			$x_{2,m}$
$User_3$	~~$x_{3,1}$~~			$x_{3,m}$
…	~~…~~			…
$User_n$	~~$x_{n,1}$~~	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

- Minimizing ***L(W)***, does not guarantee ***L(W)*** and ***L (W)*** are equally minimized
  - Stripping sensitive attributes does not help!

Origins of disparate mistreatment

	~~SA_1~~	NSA_2	…	NSA_m
$User_1$	~~$x_{1,1}$~~	$x_{1,2}$	…	$x_{1,m}$
$User_2$	~~$x_{2,1}$~~			$x_{2,m}$
$User_3$	~~$x_{3,1}$~~			$x_{3,m}$
…	…			…
$User_n$	~~$x_{n,1}$~~	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

- Minimizing ***L(W)***, does not guarantee ***L(W)*** and ***L(W)*** are equally minimized
  - To avoid disp. mistreatment, we need ***L(W) = L(W)***

Origins of disparate mistreatment

	~~SA_1~~	NSA_2	…	NSA_m
$User_1$	~~$x_{1,1}$~~	$x_{1,2}$	…	$x_{1,m}$
$User_2$	~~$x_{2,1}$~~			$x_{2,m}$
$User_3$	~~$x_{3,1}$~~			$x_{3,m}$
…	…			…
$User_n$	~~$x_{n,1}$~~	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

- Minimizing ***L(W)***, does not guarantee ***L(W)*** and ***L (W)*** are equally minimized
  - Put differently, we need: $P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1)$

Origins of disparate impact

	SA_1	NSA_2	…	NSA_m
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
…	…			…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Reject
…
Accept

Origins of disparate impact

	SA_1	NSA_2	…	NSA_m
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
…	…			…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Accept
…
Reject

- Suppose training data has biased labels!

Origins of disparate impact

	SA_1	NSA_2	…	NSA_m
$User_1$	$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$
$User_2$	$x_{2,1}$			$x_{2,m}$
$User_3$	$x_{3,1}$			$x_{3,m}$
…	…			…
$User_n$	$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Accept
…
Reject

- Suppose training data has biased labels!
- Classifier will learn to make biased decisions
  - Using sensitive attributes (SAs)

Origins of disparate impact

	~~SA_1~~	NSA_2	…	NSA_m
$User_1$	~~$x_{1,1}$~~	$x_{1,2}$	…	$x_{1,m}$
$User_2$	~~$x_{2,1}$~~			$x_{2,m}$
$User_3$	~~$x_{3,1}$~~			$x_{3,m}$
…	…			…
$User_n$	~~$x_{n,1}$~~	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Accept
…
Reject

- Suppose training data has biased labels!
- Stripping SAs does not fully address the bias

Origins of disparate impact

	~~SA_1~~	NSA_2	…	NSA_m
$User_1$	~~$x_{1,1}$~~	$x_{1,2}$	…	$x_{1,m}$
$User_2$	~~$x_{2,1}$~~			$x_{2,m}$
$User_3$	~~$x_{3,1}$~~			$x_{3,m}$
…	…			…
$User_n$	~~$x_{n,1}$~~	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Accept
…
Reject

- Suppose training data has biased labels!
- Stripping SAs does not fully address the bias
  - NSAs correlated with SAs will be given more / less weights
  - Learning tries to compensate for lost SAs

Analogous to 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 without ground truth

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 detection policy

- 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!

Origins of disparate impact

	~~SA_1~~	NSA_2	…	NSA_m
$User_1$	~~$x_{1,1}$~~	$x_{1,2}$	…	$x_{1,m}$
$User_2$	~~$x_{2,1}$~~			$x_{2,m}$
$User_3$	~~$x_{3,1}$~~			$x_{3,m}$
…	…			…
$User_n$	~~$x_{n,1}$~~	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Accept
…
Reject

- Suppose training data has biased labels!
- Stripping SAs does not fully address the bias

Origins of disparate impact

	~~SA_1~~	NSA_2	…	NSA_m
$User_1$	~~$x_{1,1}$~~	$x_{1,2}$	…	$x_{1,m}$
$User_2$	~~$x_{2,1}$~~			$x_{2,m}$
$User_3$	~~$x_{3,1}$~~			$x_{3,m}$
…	…			…
$User_n$	~~$x_{n,1}$~~	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Accept
…
Reject

- Suppose training data has biased labels!
- Stripping SAs does not fully address the bias
- What if we required proportional outcomes?

Origins of disparate impact

	~~SA_1~~	NSA_2	…	NSA_m
$User_1$	~~$x_{1,1}$~~	$x_{1,2}$	…	$x_{1,m}$
$User_2$	~~$x_{2,1}$~~			$x_{2,m}$
$User_3$	~~$x_{3,1}$~~			$x_{3,m}$
…	…			…
$User_n$	~~$x_{n,1}$~~	$x_{n,2}$	…	$x_{n,m}$

Decision
Accept
Reject
Accept
…
Reject

- Suppose training data has biased labels!
- Stripping SAs does not fully address the bias
- Put differently, we need: $P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$

Summary: 3 notions of discrimination

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 training data is biased
   - To avoid: $P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$

3. Disparate mistreatment: Specific to machine learning
   - To avoid: $P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1)$

Learning to avoid discrimination

- Idea: Discrimination notions as constraints on learning

- Optimize for accuracy under those constraints

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

A few observations

- No free lunch: Additional constraints lower accuracy
  - Tradeoff between accuracy & discrimination avoidance
- 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
    - i.e., by using sensitive attributes

Key challenge

- How to learn efficiently under these constraints?

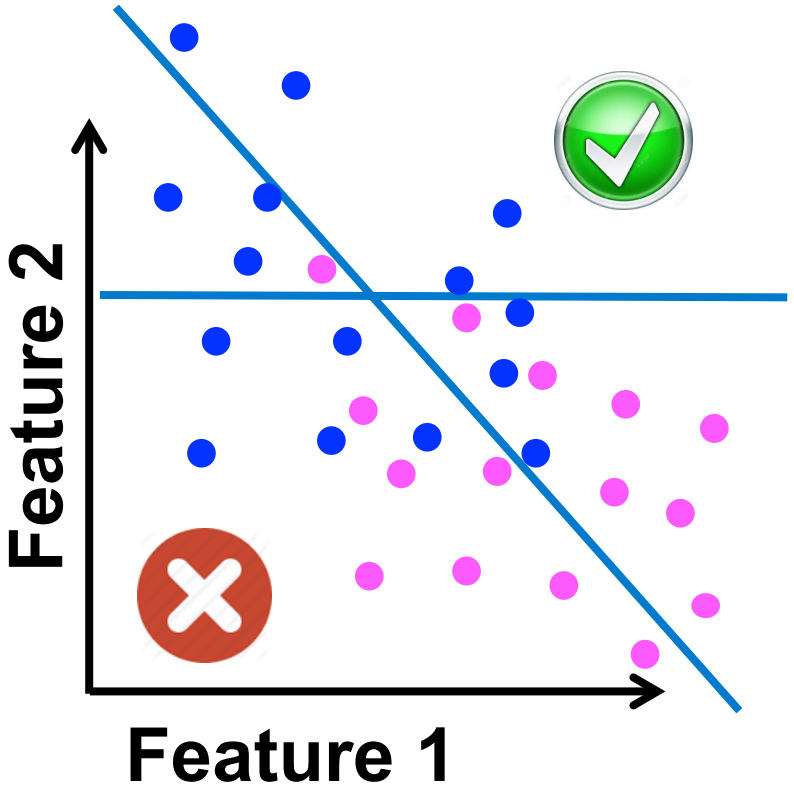
$$\begin{aligned} &minimize\ L(\mathbf{w}) \\ &subject\ to \quad P(\hat{y}=1|z=0) = P(\hat{y}=1|z=1) \end{aligned}$$

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

- 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

Disparate impact constraints: Intuition

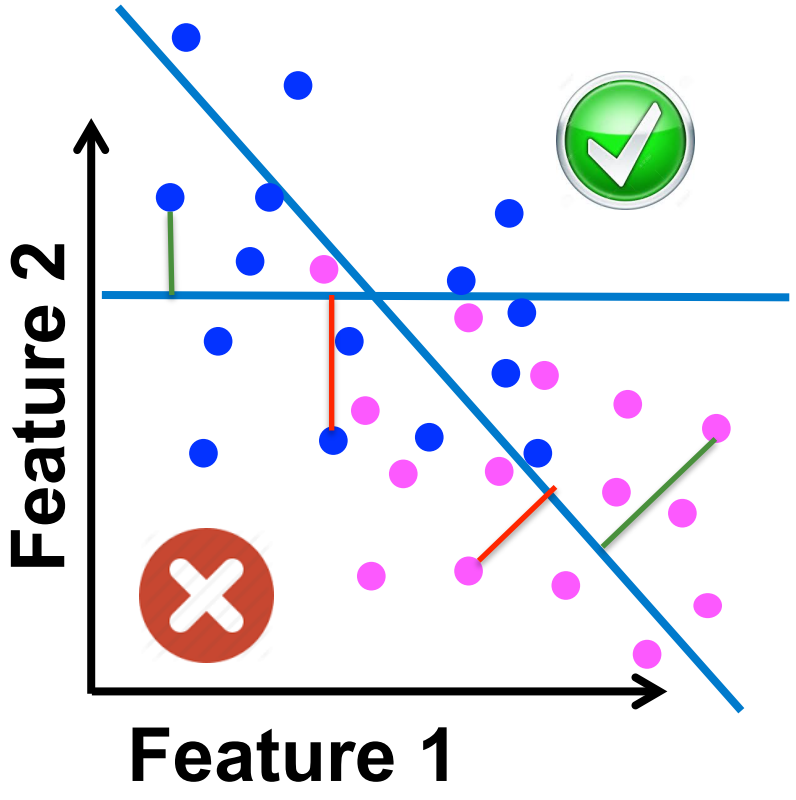


Males
Females

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

Limit the differences in the acceptance (or rejection) ratios across members of different sensitive groups

Disparate impact constraints: Intuition



Males
Females

A proxy measure for $P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$

Limit the differences in the average strength of acceptance and rejection across members of different sensitive groups

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

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

Learning classifiers w/o disparate impact

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

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

- New formulation: Convex, easy-to-learn

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

A few observations

- Our formulation can be applied to a variety of decision boundary classifiers (& loss functions)
  - hinge-loss, logistic loss, linear and non-linear SVM
- Works well on test data-sets
  - Achieves proportional outcomes with low loss in accuracy
- Can easily change our formulation to optimize for fairness under accuracy constraints
- Feasible to achieve disp. treatment & impact simultaneously

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

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

A few observations

- Our formulation can be applied to a variety of decision boundary classifiers (& loss functions)
- Can constrain for all misclassifications or for false positives & only false negatives separately
- Works well on a real-world recidivism risk estimation data-set
  - Addressing a concern raised about COMPASS, a commercial tool for recidivism risk estimation

Summary: Discrimination through computational lens

- Defined three notions of discrimination
  - disparate treatment / impact / mistreatment
  - They are applicable in different contexts
- Proposed mechanisms for mitigating each of them
  - Formulate the notions as constraints on learning
  - Proposed measures that can be efficiently learned

Future work: Beyond binary classifiers

- How to learn
  - Non-discriminatory multi-class classification
  - Non-discriminatory regression
  - Non-discriminatory set selection
  - Non-discriminatory ranking

Fairness beyond discrimination

- Consider today's recidivism risk prediction tools
  - They use features like personal criminal history, family criminality, work & social environment
  - Is using family criminality for risk prediction fair?
  - How can we reliably measure a social community's sense of fairness of using a feature in decision making?
  - How can we account for such fairness measures when making decisions?

Beyond fairness: FATE of Algorithmic Decision Making

- Fairness: The focus of this talk
- Accountability: Assigning responsibility for decisions
  - Helps correct and improve decision making
- Transparency: Tracking the decision making process
  - Helps build trust in decision making
- Explainability: Interpreting (making sense of) decisions
  - Helps understand decision making

Our works

- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez and Krishna P. Gummadi. *Fairness Constraints: A Mechanism for Fair Classification*. In FATML, 2015.
- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez and Krishna P. Gummadi. *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment*. In FATML, 2016.
- Miguel Ferreira, Muhammad Bilal Zafar, and Krishna P. Gummadi. *The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems*. In FATML, 2016.
- 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.

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.