Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Md Musa
Leibniz University of Hanover

Index

- Introduction
- Background
  - Disparate treatment
  - Disparate Impact
  - Why we need Disparate Mistreatment
  - Case Studies
- Classification without disparate mistreatment
- Experiment on synthetic data & real World Dataset
- Pros & Cons
- Conclusion & Future work

Introduction

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Author:
Muhammad Bilal Zafar
Isabel Valera
Manuel Gomez Rodriguez
Krishna P. Gummadi

-- Max Planck Institute for Software Systems (MPI-SWS)

Mentors:

Prof. Dr. techn. Wolfgang Nejdl- Leibniz Universität Hannover

Background

Fairness is not a technological problem but unfair behavior can be omitted using automated data-driven decision making system.

- Disparate treatment
- Disparate Impact
- Disparate Mistreatment

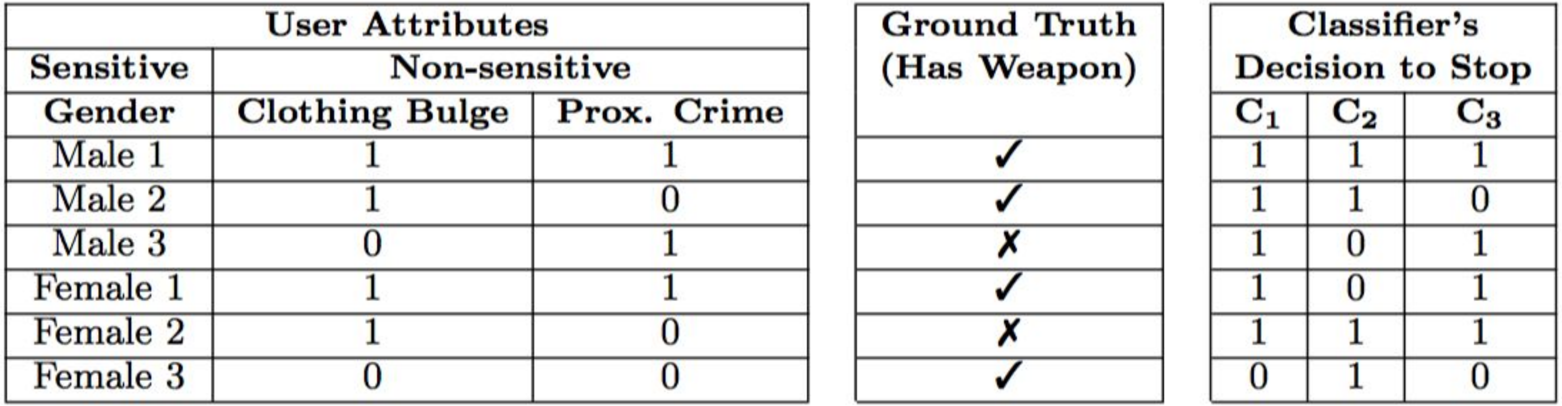
Sensitive Attributes (race, gender, color)

Why disparate mistreatment

- Tackle unfairness in decision
- Historical decision in training data are bases & sensitive attributes
- On info about ground truth

Then we need disparate mistreatment, where ground truth is available for historical decision used during training stage.

Case Studies

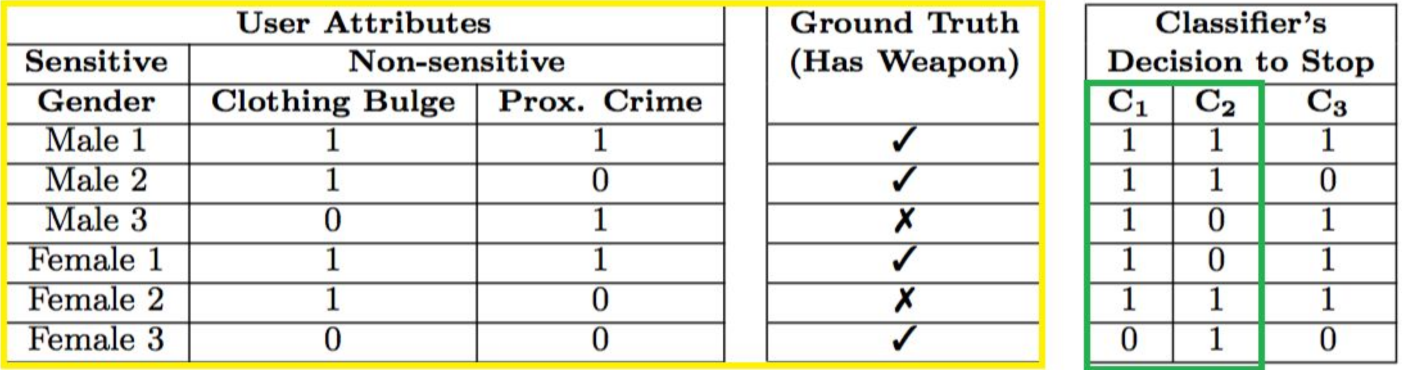


User Attributes			Ground Truth	Classifier's		
Sensitive	Non-sensitive		(Has Weapon)	Decision to Stop		
Gender	Clothing Bulge	Prox. Crime		C_1	C_2	C_3
Male 1	1	1	✓	1	1	1
Male 2	1	0	✓	1	1	0
Male 3	0	1	✗	1	0	1
Female 1	1	1	✓	1	0	1
Female 2	1	0	✗	1	1	1
Female 3	0	0	✓	0	1	0

Figure: Decisions of three fictitious classifiers (C1, C2 and C3) on whether (1) or not (0) to stop a pedestrian on the suspicion of possessing an illegal weapon.

- Gender is a sensitive attribute whereas the other two attributes are non-sensitive.
- Ground truth on whether the person is actually in possession of an illegal weapon is also shown.

Disparate Mistreatment

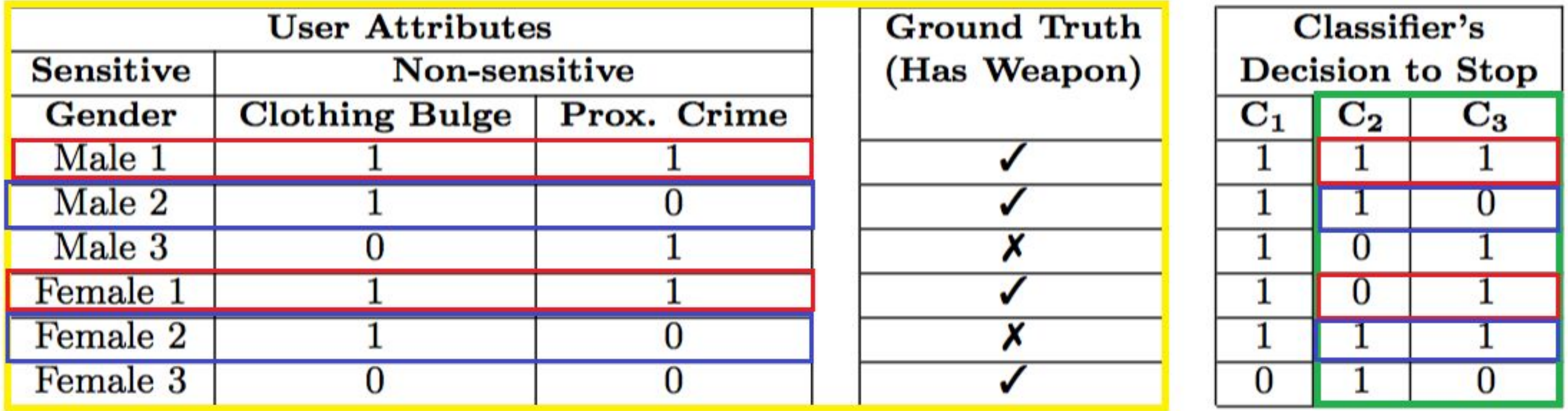


User Attributes			Ground Truth	Classifier's		
Sensitive	Non-sensitive		(Has Weapon)	Decision to Stop		
Gender	Clothing Bulge	Prox. Crime		C_1	C_2	C_3
Male 1	1	1	✓	1	1	1
Male 2	1	0	✓	1	1	0
Male 3	0	1	✗	1	0	1
Female 1	1	1	✓	1	0	1
Female 2	1	0	✗	1	1	1
Female 3	0	0	✓	0	1	0

C1 has different false negative rates for males and females

C2 has different false positive rates and different false negative rates for males and females.

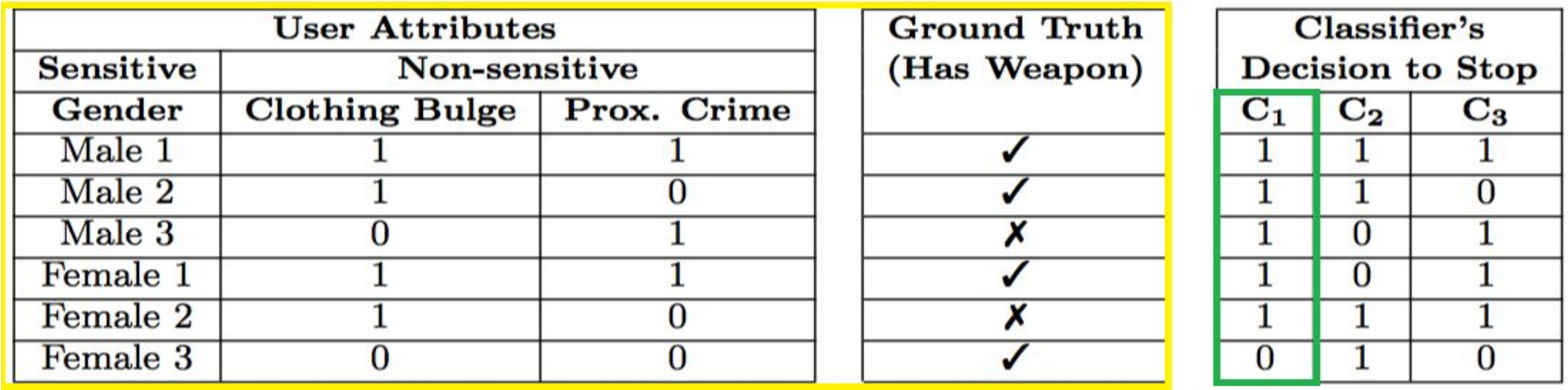
Disparate Treatment



User Attributes			Ground Truth (Has Weapon)	Classifier's Decision to Stop		
Sensitive	Non-sensitive					
Gender	Clothing Bulge	Prox. Crime		C_1	C_2	C_3
Male 1	1	1	✓	1	1	1
Male 2	1	0	✓	1	1	0
Male 3	0	1	✗	1	0	1
Female 1	1	1	✓	1	0	1
Female 2	1	0	✗	1	1	1
Female 3	0	0	✓	0	1	0

Disparate treatment arises when a decision-making system provides different outputs for groups of people with the same values of non-sensitive attributes but different values of sensitive attributes.

Disparate Impact



User Attributes			Ground Truth (Has Weapon)	Classifier's Decision to Stop		
Sensitive	Non-sensitive					
Gender	Clothing Bulge	Prox. Crime		C_1	C_2	C_3
Male 1	1	1	✓	1	1	1
Male 2	1	0	✓	1	1	0
Male 3	0	1	✗	1	0	1
Female 1	1	1	✓	1	0	1
Female 2	1	0	✗	1	1	1
Female 3	0	0	✓	0	1	0

Disparate impact arises when a decision making system provides outputs that benefit (hurt) a group of people sharing a value of sensitive attribute more frequently than other groups of people.

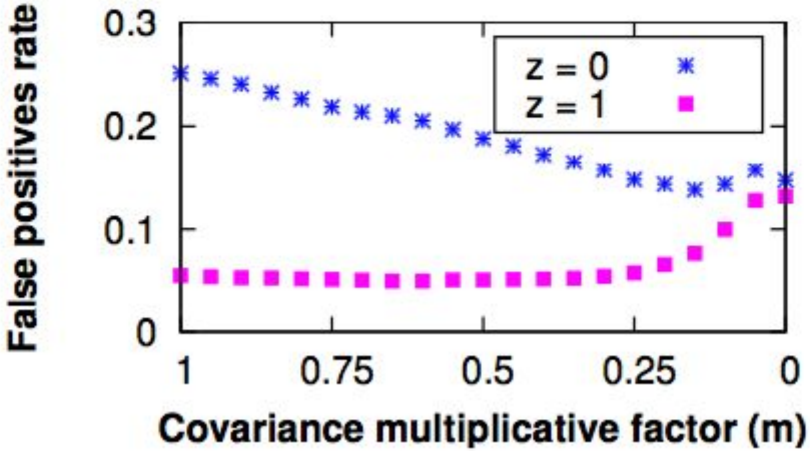
Classification without disparate mistreatment

- Train decision boundary-based classifiers.
- Tractable effective proxy for fairness.
- Convert constraints into Disciplined Convex-Concave Program
- Not required sensitive attributes (keeping the feature disjoint from sensitive attribute)
- False positive rate and false negative rate where close the values of both to 0 and even lower value is will mistreatment.

Experiments on synthetic data

Only false positive rate or false negative rate

- Decreasing the covariance threshold causes the false positive rates for both groups to become similar.

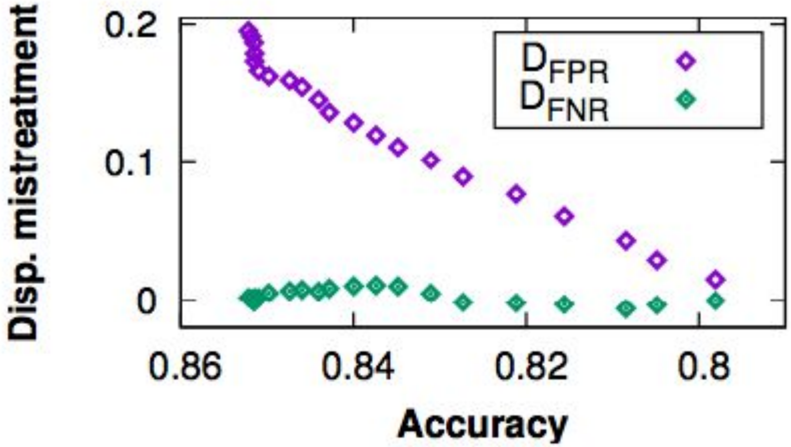


Covariance vs false positive rate

Experiments on synthetic data

Only false positive rate or false negative rate

- Increasing degree of fairness corresponds to a steady decrease in accuracy

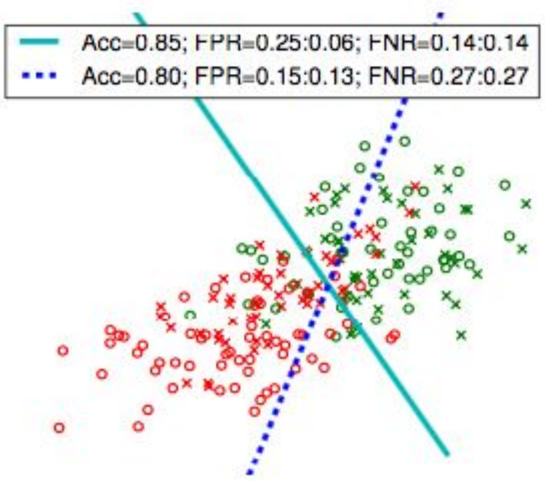


Fairness vs Accuracy

Experiments on synthetic data

Only false positive rate or false negative rate

Decision boundary (solid line) and fair decision boundary (dashed line), along with corresponding accuracy and false positive rates for groups z = 0 (crosses) and z = 1 (circles).

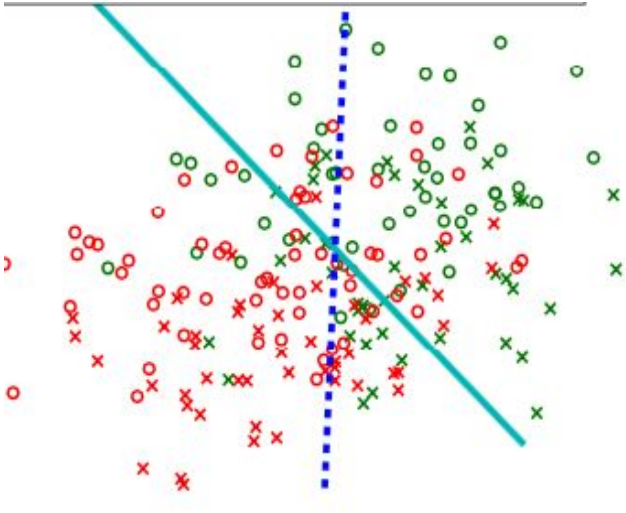


Boundaries

Experiments on synthetic data

Both false positive rate or false negative rate

Removing disparate mistreatment on both at the same time leads to very similar results.

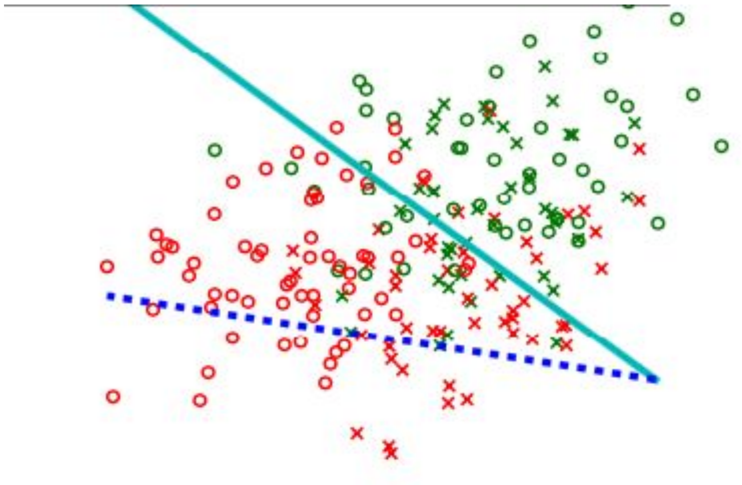


Opposite Signs

Experiments on synthetic data

Both false positive rate or false negative rate

Removing disparate mistreatment on both at the same time causes a larger drop in accuracy.



Same Signs

Real world dataset

- Ground truth available
- Train logistic regression classifiers with three types of constraints
  - constraints on false positive rate
  - false negative rate
  - on both

- ProPublica COMPAS
- consider a subset of offenders whose race was either black or white.

Pros & Cons

- Avoid disparate mistreatment
- Didn’t consider sensitive attribute
- Accuracy rate same as Hardt et al method.

- Large dataset
- Simultaneously remove & accuracy drop ~5%.

Conclusion & Future work

Disparate mistreatment for decision boundary-based classifiers, which can be easily incorporated into their formulation as convex-concave constraints. Experiments on synthetic as well as real-world datasets show that our methodology is effective at avoiding disparate mistreatment, often at a small cost in terms of accuracy.

- Not a convex program
- Optimization and smaller datasets.

Thanks For Listening

Any Question?

Thanks For Listening

Any Question?

No

Great!