Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment (Zafar, Muhammad Bilal, et al.)

Tian Xie(tx2221), Yanchen Liu(yl4637), Zhiwen Huang (zh2387)

Github Link: [TianXie1999/Replicate-Paper: Replicate fairness paper (github.com)]

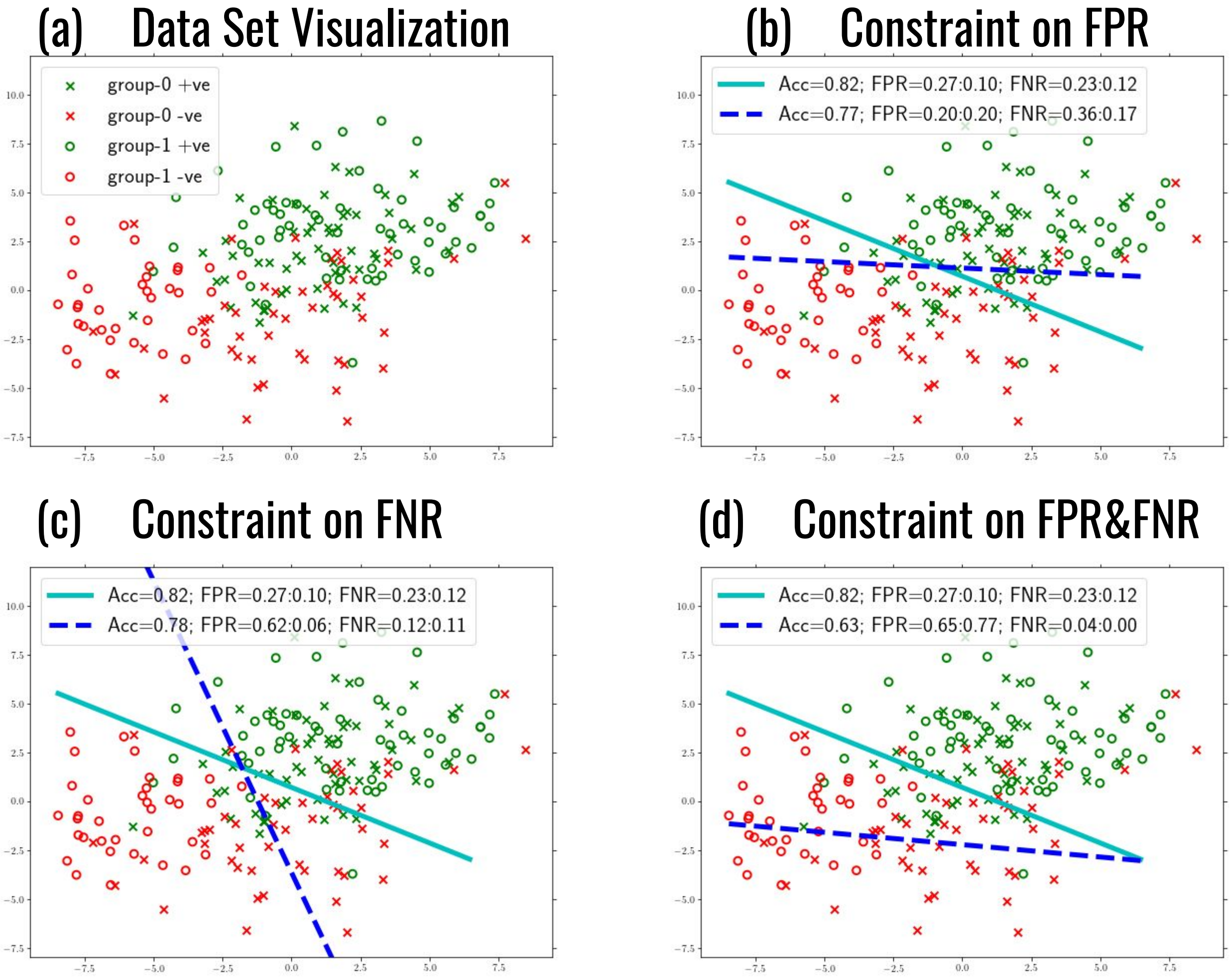
Columbia University
In the City of New York

Why and What?

valuable understandings on trade-off between fairness and accuracy

- **a fairness metric derived from ”disparate mistreatment”** : with respect to a sensitive attribute, the model would have misclassification rates differ for groups of people having different values of that sensitive attribute (e.g., blacks and whites)
- **Visualize the fairness-accuracy trade-off:** replicate several experiments on synthesized data and real-world data to visualize balance between fairness and accuracy

Synthesized Data



(a) Data Set Visualization

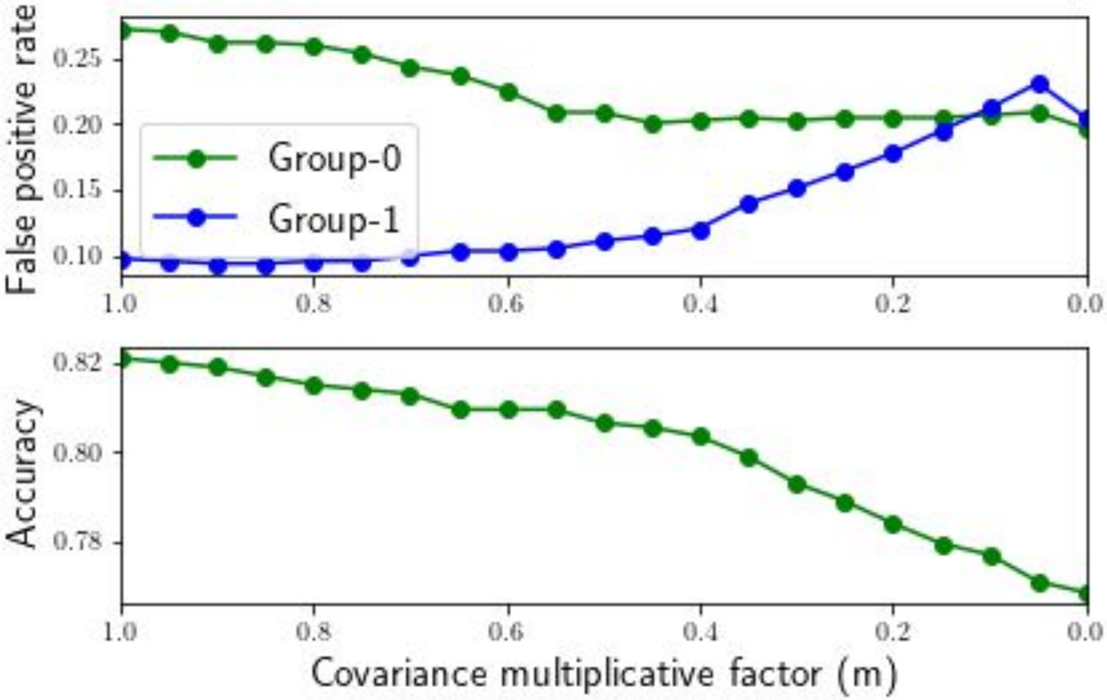
(b) Constraint on FPR

(c) Constraint on FNR

(d) Constraint on FPR&FNR

multivariate normal distribution: 2 non-sensitive features and 1 sensitive feature

Accuracy-Fairness Trade-off While relaxing the Threshold



Measure fairness using the **covariance** between the users sensitive attributes and the signed distance between the feature vectors of misclassified users and the classifier decision boundary

Real-world data:

ProPublica COMPAS data set

- Crime offenders for different demographic groups
- Logistic Regression
- **Sensitive attribute**: race
- **Y**: whether the individual recidivated

== Unconstrained (original) classifier ==

```
Accuracy: 0.671
||  s  || FPR. || FNR. ||
||  0  || 0.35 || 0.32 ||
||  1  || 0.15 || 0.59 ||
```

== Constraints on FPR ==

```
Accuracy: 0.653
||  s  || FPR. || FNR. ||
||  0  || 0.28 || 0.41 ||
||  1  || 0.24 || 0.51 ||
```

Credit Card data from Kaggle

- Credit status of client, personal and applicant information ->predict future defaults and credit card borrowing
- Logistic Regression
- **Sensitive attribute**: gender
- **Y**: whether the client has good credit (paid off that month)

== Unconstrained (original) classifier ==

```
Accuracy: 0.515
||  s  || FPR. || FNR. ||
||  0  || 0.38 || 0.58 ||
||  1  || 0.36 || 0.64 ||
```

== Constraints on FPR ==

```
Accuracy: 0.514
||  s  || FPR. || FNR. ||
||  0  || 0.39 || 0.57 ||
||  1  || 0.35 || 0.65 ||
```