Measuring Fairness in Ranked Outputs

Ke Yang and Julia Stoyanovich

Drexel University, Philadelphia, PA, USA

Motivation

Example: College admissions

Applicant data

age	gender	GPA	…	income
20	F	4.0	…	50K
19	M	3.5	…	35K
21	F	3.2	…	30K
17	M	3.8	…	40K
20	F	3.6	…	60K
…	…	…	…	…

Ranker

scoring function
learned model
…

Ranking

Disparate treatment is the illegal practice of treating an entity, such as a creditor or employer, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

Our goal: detect and mitigate the effects of disparate treatment in applications that output **ranked results.**

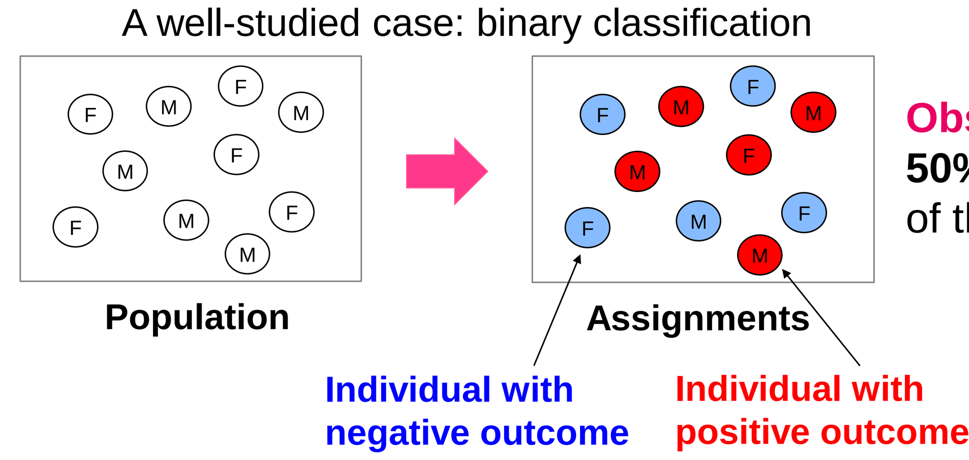
Fairness: Assigning outcomes to individuals

Assumptions

- Protected group membership is **binary** (e.g. F / M)
- Item quality is **independent** of protected group membership

Fairness is concerned with how outcomes are assigned to individuals, and, for a specific formulation, to members of a **protected group**.

A well-studied case: binary classification

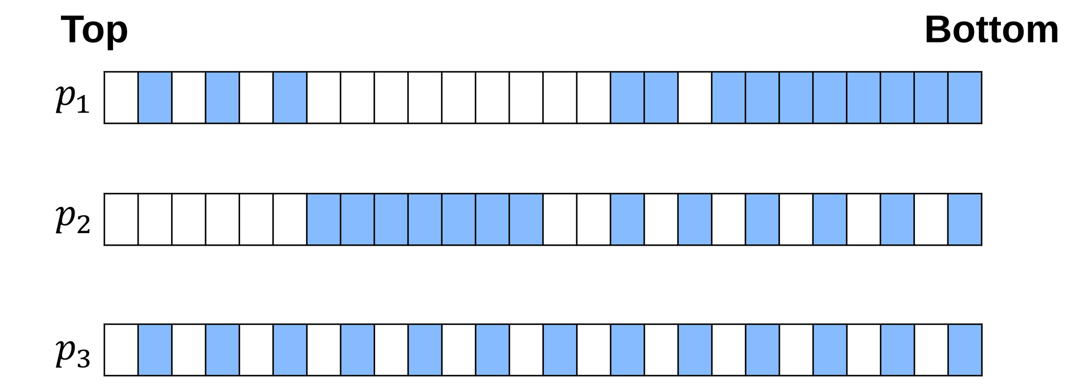


Observe that protected group (F) is **50%** of the population but only **20%** of the positive outcomes.

Positive Outcomes	Negative Outcomes
offered employment	denied employment
accepted to school	rejected from school
offered a loan	denied a loan
offered a discount	not offered a discount

What is a positive outcome in a ranking?

A ranking is relative. Being in the top-10 is better than in the top-20. Being in the top-20 is better than in the top-100, etc.



Is p_1 fair? Is p_2 fair? Is p_3 fair?

Intuition: Just as it is better (for an individual) to be ranked higher, it is more important to be fair (to groups) at higher ranks.

Idea: Look at a ranking at multiple cut-off points. Compute **set-wise** fairness at each point. Compound set-wise fairness progressively, in **a rank-aware** manner.

Our Approach

Set-wise fairness measures

Database $I(\underline{k}, s, x_1, \ldots, x_m)$ with N items, s denotes membership in protected group, $x_1, \ldots, x_m$ are descriptive attributes.

- $S^+ \subseteq I$ protected group, $S^- \subseteq I \setminus S^+$ remaining items
- c cut-off point in a ranking

Inspired by group fairness measures in binary classification

KL divergence $\quad \boldsymbol{KL} = D_{KL}(P_c||Q_N) \quad P_c = \left(\frac{S_c^+}{c}, \frac{S_c^-}{c}\right) \quad Q_N = \left(\frac{S_N^+}{N}, \frac{S_N^-}{N}\right)$

Normalized difference $\quad \boldsymbol{ND} = \left|\frac{S_c^+}{c} - \frac{S_N^+}{N}\right|$

Ratio difference $\quad \boldsymbol{RD} = \left|min\left(\frac{S_c^+}{S_c^-}, \frac{S_N^+}{S_N^-}\right) - \frac{S_N^+}{S_N^-}\right|$

Rank-aware fairness

SF_c: Set-wise fairness at cut-off c, one of KL, ND or RD

Z: Normalizer

$$GF = \frac{1}{Z}\sum_{c=10,20,\ldots}^{N} \frac{SF_c}{log_2 c}$$

$GF \in [0,1]$ 0 is worst, 1 is best.

Inspired by Normalized Discounted Cumulative Gain (NDCG) in IR

An example

Ranking: M, F, M, M, F, F

$\boldsymbol{c = 2}$

$$\boldsymbol{SF_{c=2}} \begin{cases} \boldsymbol{KL_{c=2}} = D_{KL}(P_c||Q_N) = 0 \\ P_c = \left(\frac{1}{2}, \frac{1}{2}\right) Q_N = \left(\frac{3}{6}, \frac{3}{6}\right) \\ \boldsymbol{ND_{c=2}} = \left|\frac{S_c^+}{c} - \frac{S_N^+}{N}\right| = \left|\frac{1}{2} - \frac{3}{6}\right| = 0 \\ \boldsymbol{RD_{c=2}} = \left|\frac{S_c^+}{S_c^-} - \frac{S_N^+}{S_N^-}\right| = \left|\frac{1}{1} - \frac{3}{3}\right| = 0 \end{cases}$$

$\boldsymbol{c = 4}$

$$\boldsymbol{SF_{c=4}} \begin{cases} \boldsymbol{KL_{c=4}} = D_{KL}(P_c||Q_N) = 0.19 \\ P_c = \left(\frac{1}{4}, \frac{3}{4}\right) Q_N = \left(\frac{3}{6}, \frac{3}{6}\right) \\ \boldsymbol{ND_{c=4}} = \left|\frac{S_c^+}{c} - \frac{S_N^+}{N}\right| = \left|\frac{1}{4} - \frac{3}{6}\right| = 0.25 \\ \boldsymbol{RD_{c=4}} = \left|\frac{S_c^+}{S_c^-} - \frac{S_N^+}{S_N^-}\right| = \left|\frac{1}{3} - \frac{3}{3}\right| = 0.66 \end{cases}$$

$\boldsymbol{c = 6}$

$$\boldsymbol{SF_{c=6}} \begin{cases} \boldsymbol{KL_{c=6}} = D_{KL}(P_c||Q_N) = 0 \\ P_c = \left(\frac{3}{6}, \frac{3}{6}\right) Q_N = \left(\frac{3}{6}, \frac{3}{6}\right) \\ \boldsymbol{ND_{c=6}} = \left|\frac{S_c^+}{c} - \frac{S_N^+}{N}\right| = \left|\frac{3}{6} - \frac{3}{6}\right| = 0 \\ \boldsymbol{RD_{c=6}} = \left|\frac{S_c^+}{S_c^-} - \frac{S_N^+}{S_N^-}\right| = \left|\frac{3}{3} - \frac{3}{3}\right| = 0 \end{cases}$$

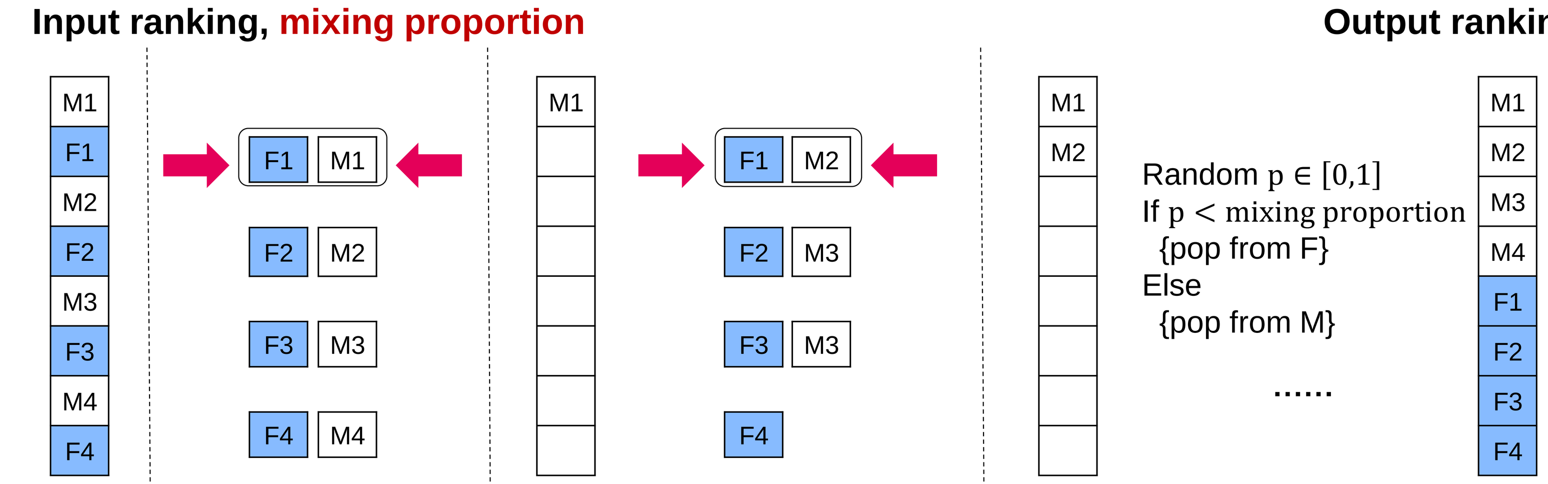
Compound:

$$GF = \frac{1}{Z}\left(\frac{SF_2}{log_2 2} + \frac{SF_4}{log_2 4} + \frac{SF_6}{log_2 6}\right) = \begin{cases} 0.13, SF = KL \\ 0.29, SF = ND \\ 0.22, SF = RD \end{cases}$$

$Z_{KL} = 0.705 \quad Z_{ND} = 0.423 \quad Z_{RD} = 1.468$

Synthetic data generator and optimization framework

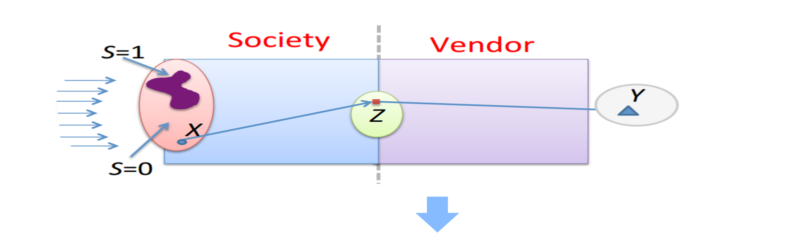
Input ranking, mixing proportion — **Output ranking**



Optimization Framework

Goal: Mitigate lack of fairness

Method: Learn a model to minimize the loss function L



$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

Applied **L-BFGS** algorithm to minimize L

Performed a simple **grid search** to find a good set of hyper-parameters A_x, A_y, A_z

Data and Evaluation

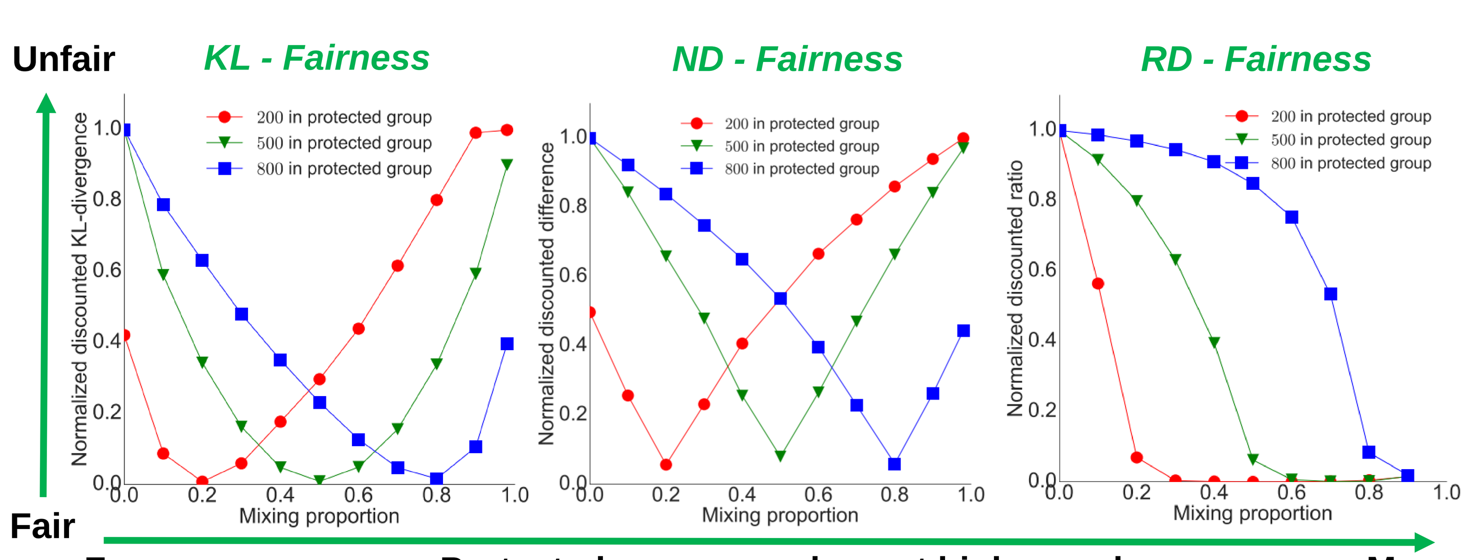
Synthetic data

Generate synthetic rankings

Protected group ratio: 20%, 50%, and 80%

Item number: 1000

Mixing proportion: [0,1]



Real data

ProPublica / COMPAS

Close to **7,000** criminal defendant records.

Racial & gender bias in predictions of future criminal activity.

Score: recidivism score, violent recidivism score, number of prior arrests.

Sensitive attributes: **race** (51% black) **gender** (19% female)

German credit

Financial information about **1,000** individuals applying for loans.

Ranking attributes: duration (month), credit amount, status of existing account, employment length.

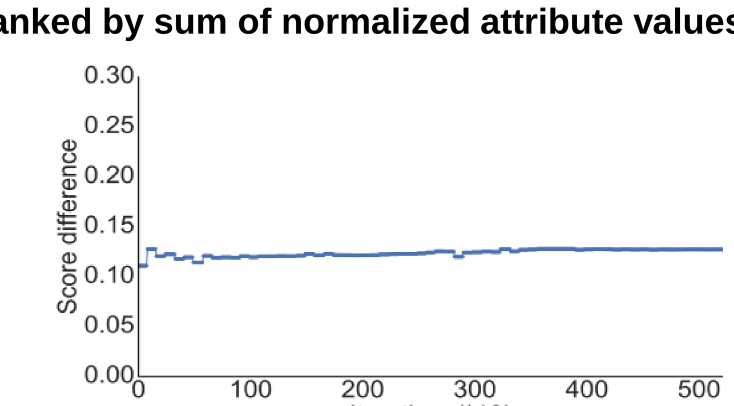
Score: duration (months), credit amount, score by summation of all attributes.

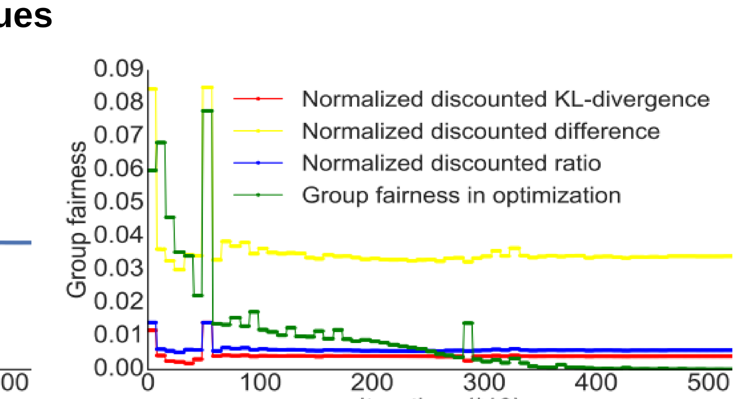
Sensitive attributes: **age** (15% younger than 25) **gender** (69% female)

Data	Sensitive Attribute	Score	KL	ND	RD
ProPublica	Race	Recidivism	0.17	0.44	0.58
		Violent recidivism	0.18	0.44	0.57
		Prior arrests	0.04	0.23	0.36
	Gender	Recidivism	0.02	0.15	0.02
		Violent recidivism	0.01	0.12	0.01
		Prior arrests	0.01	0.12	0.00
German Credit	Gender	Credit Amount	0.03	0.16	0.32
		Duration Month	0.01	0.09	0.00
		Score	0.01	0.11	0.00
	Age<25	Credit Amount	0.03	0.11	0.00
		Duration Month	0.02	0.06	0.06
		Score	0.02	0.11	0.27

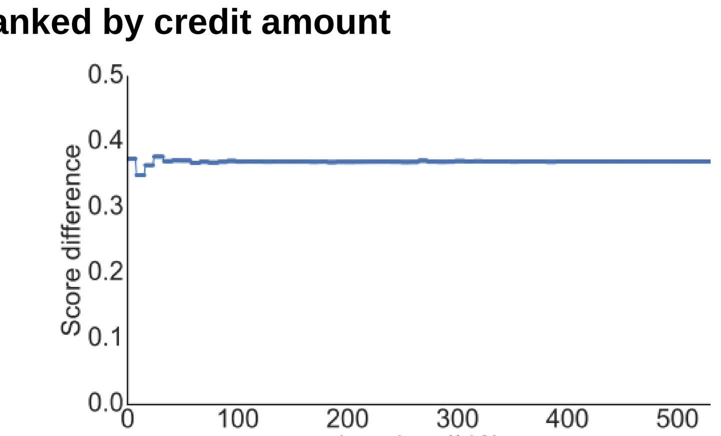
Performance of optimization on German credit

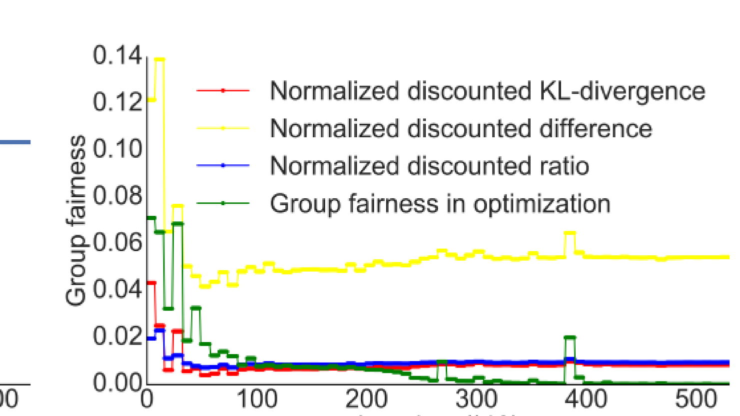
Ranked by sum of normalized attribute values





Ranked by credit amount





References

1. Ke Yang and Julia Stoyanovich. Measuring Fairness in Ranked Outputs. http://arxiv.org/abs/1610.08559
2. More information: http://dataresponsibly.com, https://www.cs.drexel.edu/dbgroup/

Acknowledgements

This research was supported in part by NSF Grants No. 1464327 & 1539856, and BSF Grant No. 2014391.