Fair Without Leveling Down

Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, Mikaela Keller

Story

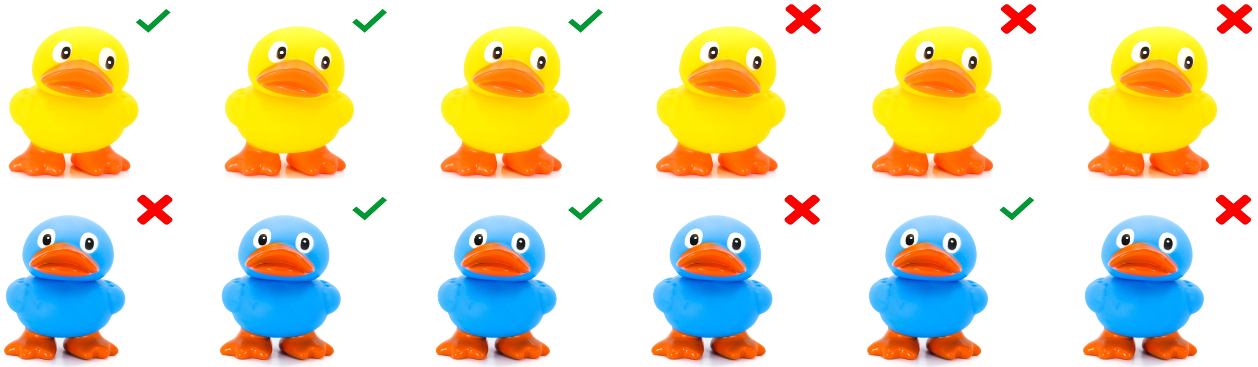
Let’s start with a story

Rubber Duck Recruitment

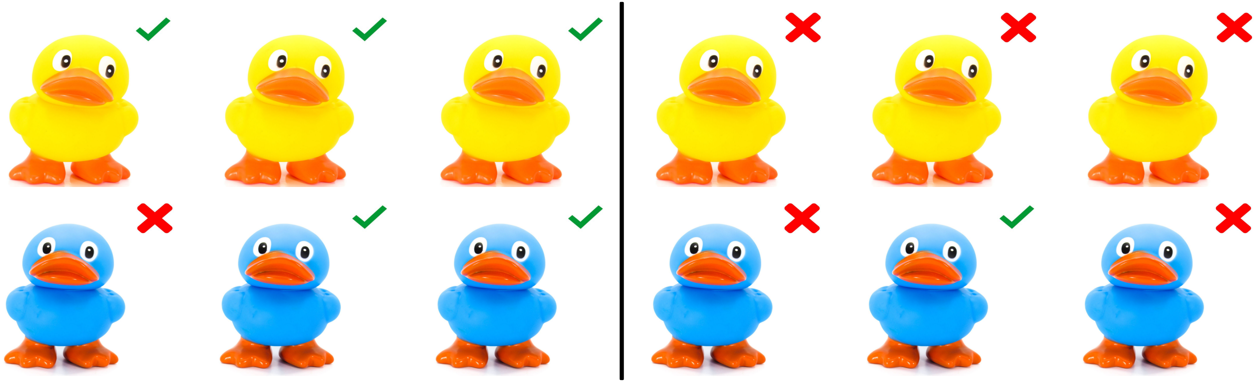
A company wants to automate the process of recruiting 🦆 to debug code.

[Rubber Duck Debugging]

Ducks

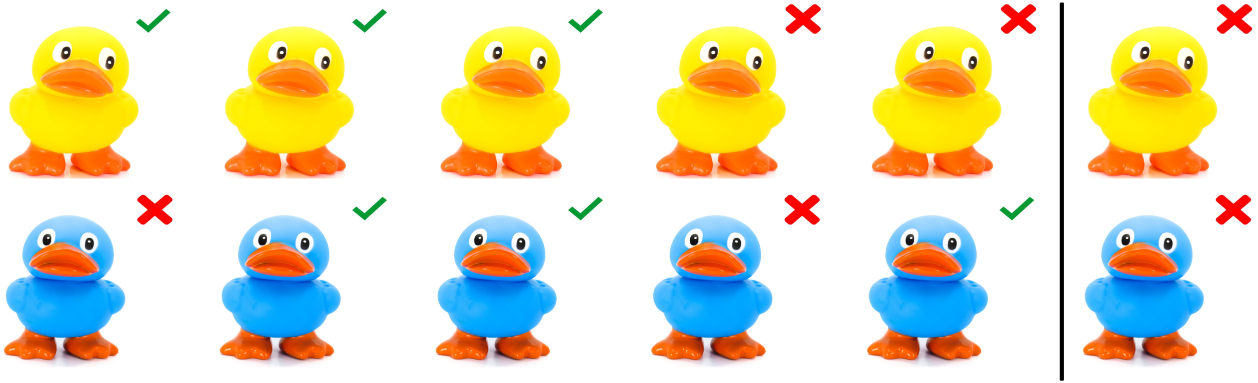


Learning A linear Classifier



```
Accuracy of Yellow Ducks - 6/6
Accuracy of Blue Ducks   - 4/6
Overall Accuracy Ducks   - 10/12
```

Learning A linear Classifier



```
Accuracy of Yellow Ducks   - 4/6
Accuracy of Blue Ducks     - 4/6
Overall Accuracy of Ducks - 8/12
```

Not Just Ducks

- Health Care
  - Skin disease detection [Kinyanjui et al., 2019]: A model learned on patients that mostly have light skin tone may be biased against patients that have darker skin tones.
- Natural Language Processing
  - Occupation prediction [De-Arteaga et al., 2019]: A model that learned to predict the profession of a person from its biography may perpetuate or even amplify existing gender biases in occupation classification.
- Justice
  - Recidivism prediction [Larson et al., 2016]: The COMPAS score was shown to be biased against black defendants, more often miss classifying them as having a high risk of recidivism than white defendants

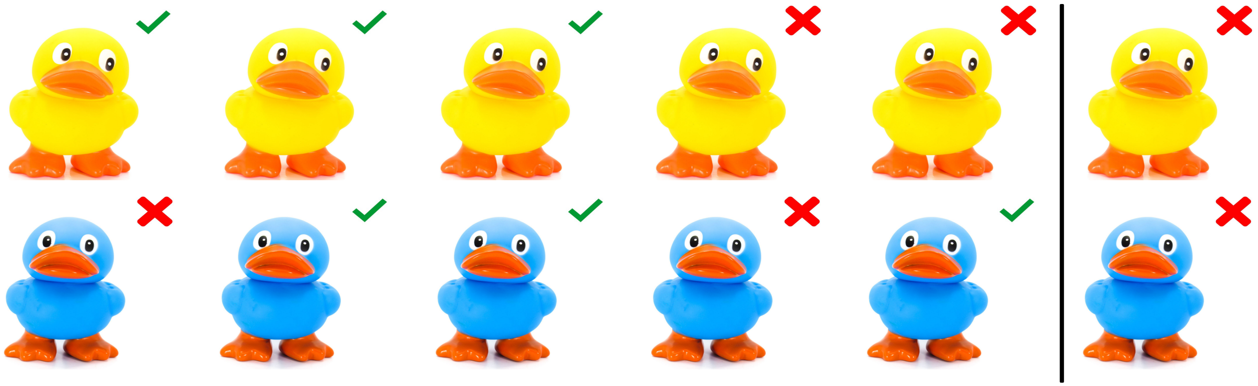
Objective - Fair Machine Learning

Learn models which are free from unjust behaviour

Group Fairness

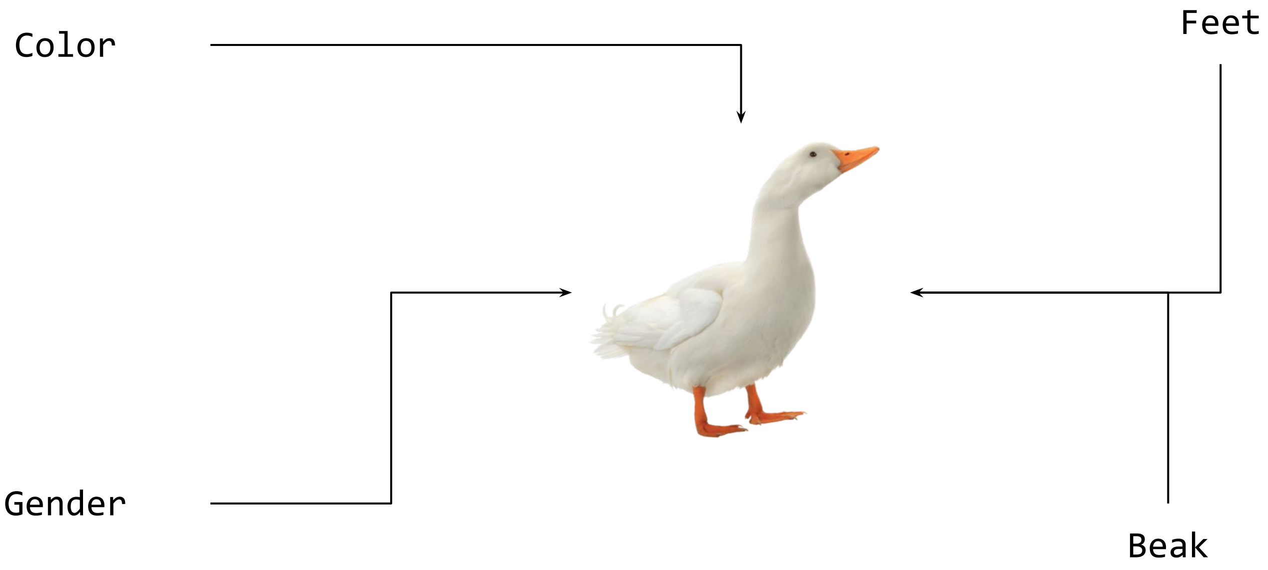
- Models which do not unjustly discriminate against a subgroup of population.
  - All subgroups accuracy should be similar.
  - All subgroup true positive rate should be similar.

Back to Our Example



Accuracy of Yellow Ducks   - 4/6
Accuracy of Blue Ducks   - 4/6
Overall Accuracy of Ducks - 8/12

Back to our setup



Contemporary Fairness Approaches

- Most contemporary fairness approaches only assume one sensitive axis.
    - For instance in the previous example, color was the sensitive axis.

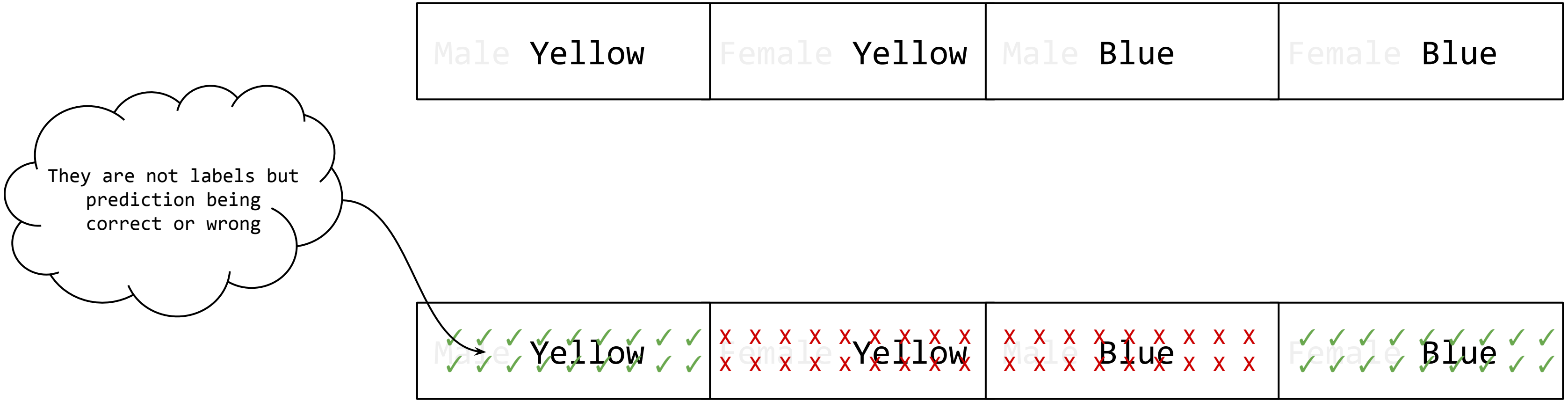
Contemporary Fairness Approaches

- Most contemporary fairness approaches only assume one sensitive axis.
  - For instance in the previous example, color was the sensitive axis.
- Even when they consider multiple axis say gender and race, they usually consider them independent i.e.
  - Be “fair” against color and be “fair” against gender.

However

- Being fair against color and against gender does not imply we are fair against color *x* gender

Yellow-Blue Accuracy Parity



Male-Female Accuracy Parity

Male Yellow	Female Yellow	Male Blue	Female Blue

Male Yellow ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓	Female Yellow x x x x x x x x x x x x x x x x x x	Male Blue x x x x x x x x x x x x x x x x x x	Female Blue ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓

All together now! - Intersectional Fairness

Male Yellow	Female Yellow	Male Blue	Female Blue

Simple Setup

Intuition of a typical Setup

- Input feature space $\mathcal{X}$, label space $\boldsymbol{Y}$, and sensitive space $\mathbf{Z}$
  - $\mathcal{X}$ - description of occupation or human faces or tweet
  - $\boldsymbol{Y}$ - occupation label or hate speech
  - $\mathbf{Z}$ - gender as the sensitive axis with {Male, Female, Non-binary} being its corresponding sensitive attribute, age with {young, old} being its corresponding sensitive attribute.
- We assume examples of the form - *(x,s,y)*
- A classifier h: $\mathcal{X}$ -> $\boldsymbol{Y}$

Setup - Group

- We define gender as any combination of the sensitive axis: Few such groups are
  - $g_{\{\text{Male, Black, A45}\}}$, $g_{\{\text{Female, Black, B45}\}}$, $g_{\{\text{Male, White, A45}\}}$, $g_{\{\text{Female, White, B45}\}}$
  - $g_{\{\text{Male, A45}\}}$, $g_{\{\text{Female, Black}\}}$, $g_{\{\text{ White, A45}\}}$, $g_{\{\text{Female. B45}\}}$
  - $g_{\{\text{Male}\}}$, $g_{\{\text{Female}\}}$, $g_{\{\text{ White}\}}$, $g_{\{\text{B45}\}}$

Fairness Measure

Statistical Parity Subgroup Fairness - Attempt 1*

Overall Stats of the dataset

Stats of the group

Weight of the group

$$\left|\boxed{P(h_\theta(x) = 1)} - \boxed{P(h_\theta(x) = 1 | g_i = 1)}\right| \times \boxed{P(g_i = 1)} \leq e^{\epsilon}$$

$$\forall g_i \in \mathcal{G}$$

*Preventing fairness gerrymandering: Auditing and learning for subgroup fairness.

Statistical Parity Subgroup Fairness - Attempt 1

Overall Stats of the dataset

Stats of the group

Weight of the group

$$\left| \boxed{P(h_\theta(x) = 1)} - \boxed{P(h_\theta(x) = 1 | g_i = 1)} \right| \times \boxed{P(g_i = 1)} \leq e^\epsilon \quad \forall g_i \in \mathcal{G}$$

Differential Fairness - Attempt 2*

$$e^{-\epsilon} \leq \frac{P(h_\theta(x) = 1 | g_i)}{P(h_\theta(x) = 1 | g_j)} \leq e^{\epsilon}$$

$$\forall g_i, g_j \in \mathcal{G}$$

Differential Fairness instantiated with demographic fairness

*An Intersectional Definition of Fairness

Differential Fairness - Attempt 2

$$\frac{\text{Best group metric}}{\text{Worst group metric}} \leq e^{\epsilon}$$

Differential Fairness instantiated with demographic fairness

Differential Fairness - Attempt 2

- Does not get affected by the weight of the class
- Obvious similarity in formulation to that of differential privacy
- Can be generalized to other notions of fairness such as Equal Opportunity, Equal Odds etc.
- It has few problems, which I will illustrate in a while!

Benchmarking!

- Dataset:
  - Celeb Multi Group (images encoded via pre-trained resnet18) - 4 binary sensitive axis resulting in 16 groups
- Methods:
  - Multiple Fairness Inducing Method
- Fairness Measure
  - False Positive Rate
- Model
  - Simple Non Linear

Benchmarking!

method	balanced accuracy	fairness
Unconstrained	0.8 +/- 0.01	1.49 +/- 0.19
Adversarial	0.8 +/- 0.0	1.45 +/- 0.19
FairGrad	0.77 +/- 0.01	1.0 +/- 0.06
INLP	0.8 +/- 0.0	1.28 +/- 0.09
Fair MixUp	0.8 +/- 0.0	1.31 +/- 0.14

Some hidden Aspects

Recall eps-fairness

$$\frac{\texttt{Best group metric}}{\texttt{Worst group metric}} \leq e^{\epsilon}$$

It is the ratio of best of group metric by worst of group metric.

Consider False Positive Rate

- Consider False positive rate as the metric
  - Higher it is worse it is for the group

Consider False Positive Rate

False Positive rate in sorted Order

0.2	0.21	0.38	0.38	0.38	0.39	0.39	0.4

- eps fairness is 0.4/0.2 = 2.
- For simplicity we ignore the log.

Improving fairness - 1st

2.0

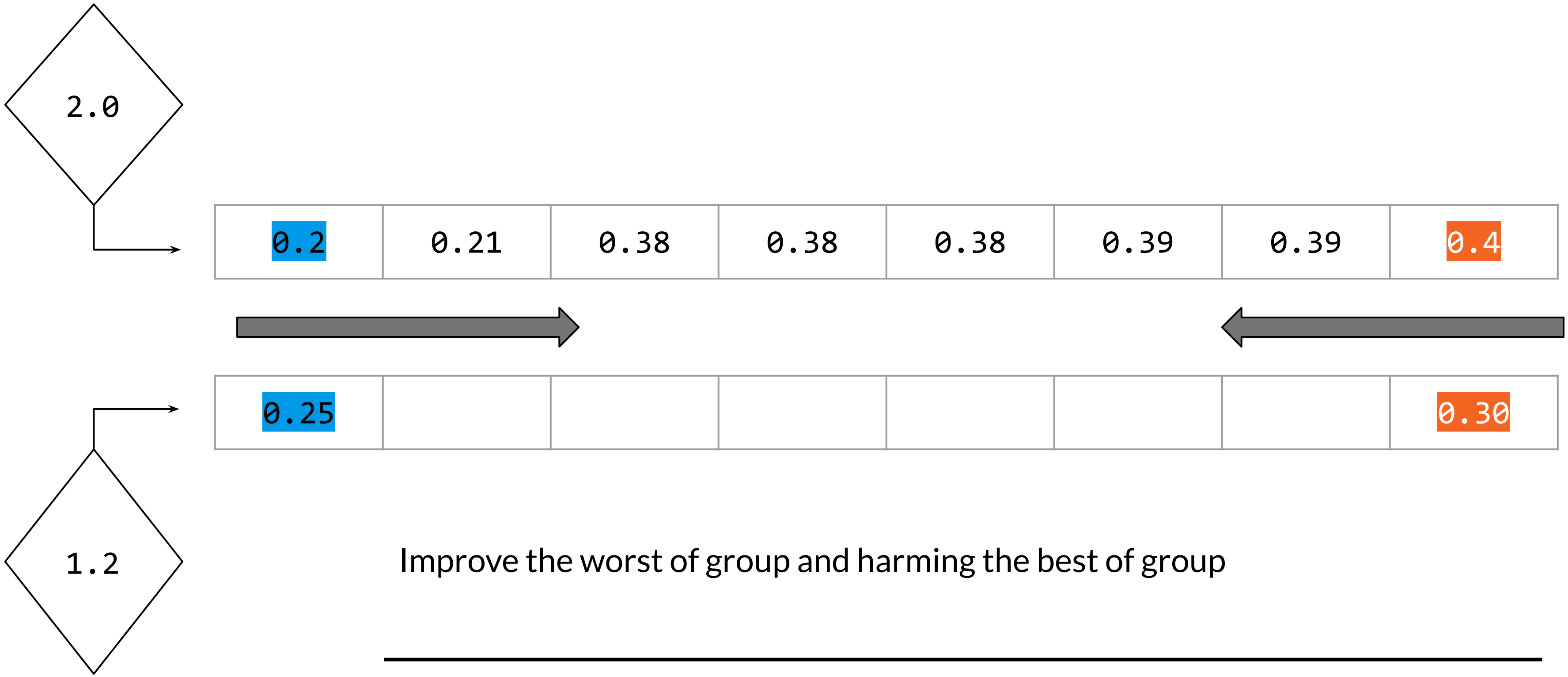
0.2	0.21	0.38	0.38	0.38	0.39	0.39	0.4

0.20							0.30

1.5

Improve the worst of group without harming the best of group

Improving fairness - 2nd



Improve the worst of group and harming the best of group

Improving fairness - 3rd

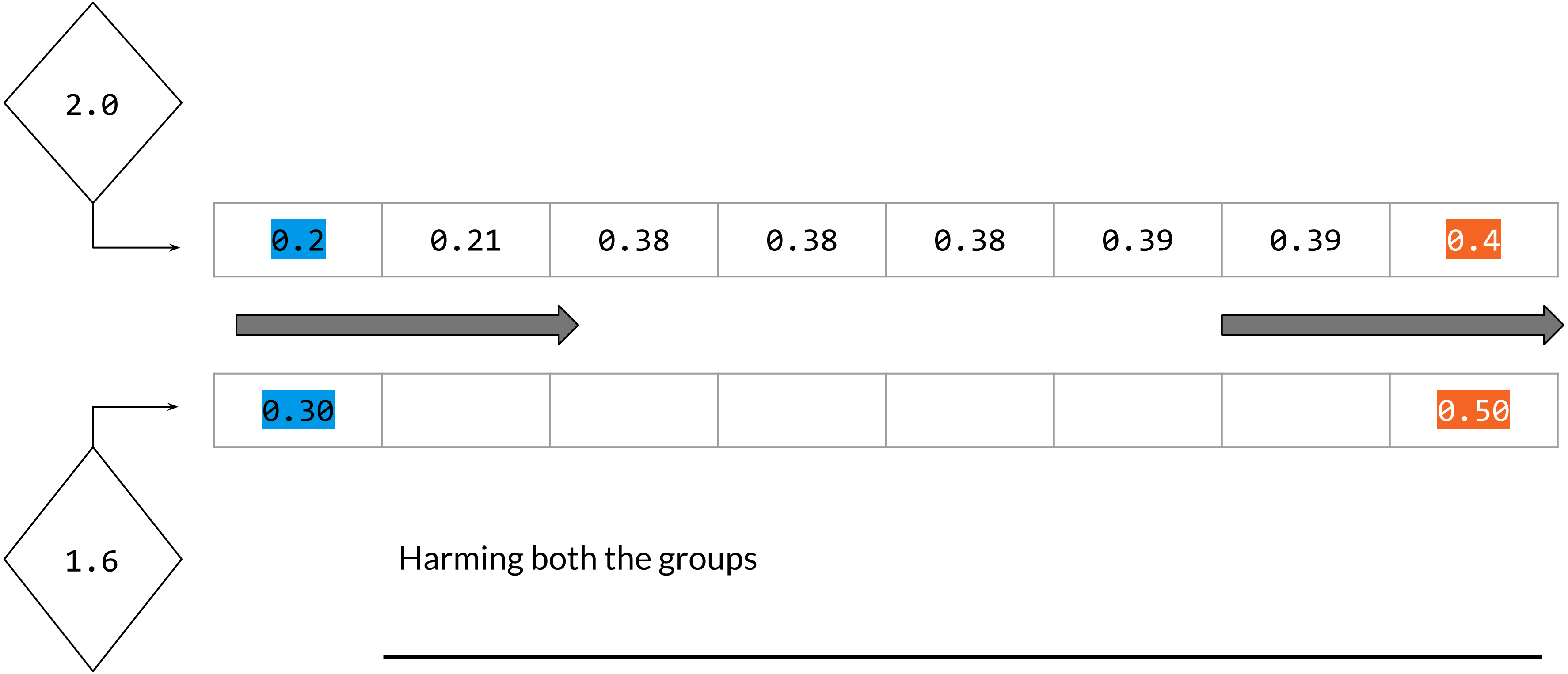
2.0

0.2	0.21	0.38	0.38	0.38	0.39	0.39	0.4
0.10							0.15

1.5

Improve the worst of group and also improving the best of group

Improving fairness - 4th



Harming both the groups

All together now

0.2	0.21	0.38	0.38	0.38	0.39	0.39	0.4

Harm the best of group and improve the worst of group

Improve the best of group and improve the worst of group

Harm the best of group and harm the best of group

Improve the worst of group while not improving the best

All have merits - maybe except one

0.2	0.21	0.38	0.38	0.38	0.39	0.39	0.4

Harm the best of group and improve the worst of group

Improve the best of group and improve the worst of group

Harm the best of group and harm the best of group

Improve the worst of group while not improving the best

Benchmarking!

method	balanced accuracy	fairness	min fair	max fair
Unconstrained	0.8 +/- 0.01	1.49 +/- 0.19	0.08 +/- 0.01	0.37 +/- 0.05
Adversarial	0.8 +/- 0.0	1.45 +/- 0.19	0.09 +/- 0.01	0.38 +/- 0.05
FairGrad	0.77 +/- 0.01	1.0 +/- 0.06	0.15 +/- 0.01	0.4 +/- 0.0
INLP	0.8 +/- 0.0	1.28 +/- 0.09	0.1 +/- 0.01	0.34 +/- 0.04
Fair MixUp	0.8 +/- 0.0	1.31 +/- 0.14	0.1 +/- 0.02	0.38 +/- 0.03

Some Observation

- Levelling down is even more evident in intersectional setting.
- And this leveling down was almost in all the approaches and dataset we explored.

Some Recommendations

How to Capture Intersectional Fairness

Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, Mikaela Keller

In this work, we tackle the problem of intersectional group fairness in the classification setting, where the objective is to learn discrimination-free models in the presence of several intersecting sensitive groups. First, we illustrate various shortcomings of existing fairness measures commonly used to capture intersectional fairness. Then, we propose a new framework called the α Intersectional Fairness framework, which combines the absolute and the relative performances between sensitive groups. Finally, we provide various analyses of our proposed framework, including the min-max and efficiency analysis. Our experiments using the proposed framework show that several in-processing fairness approaches show no improvement over a simple unconstrained approach. Moreover, we show that these approaches minimize existing fairness measures by degrading the performance of the best of the group instead of improving the worst.

Some Limitations

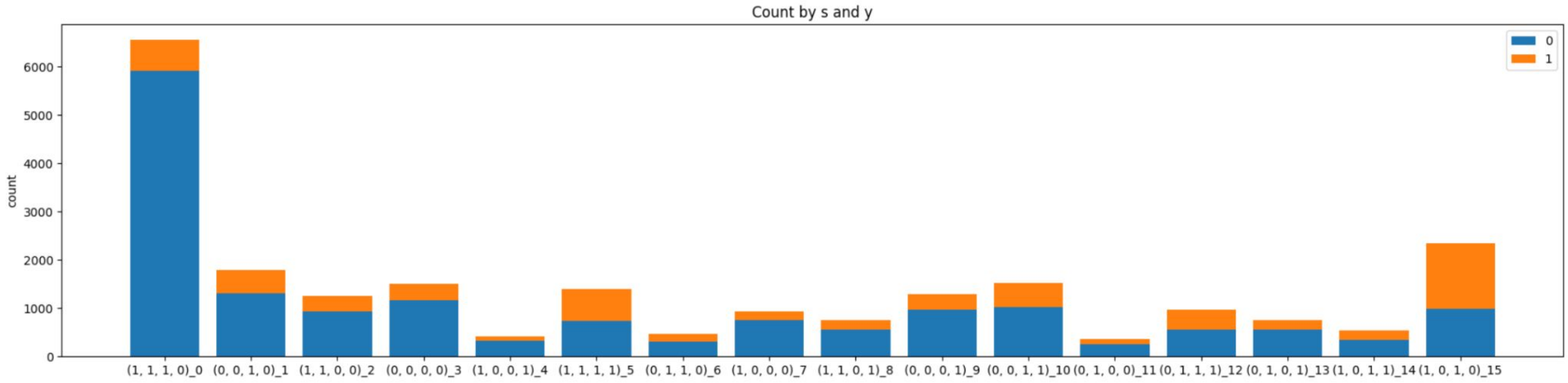
- A perfectly fair model might not be devoid of social harm.
    - if some socio-economic groups are not present in a given dataset, existing fairness-inducing approaches are likely to not have any positive impact.

Data

Limitations of data

- Limited number of datasets
  - Most of them are limited in size.
  - Have very skewed distribution.

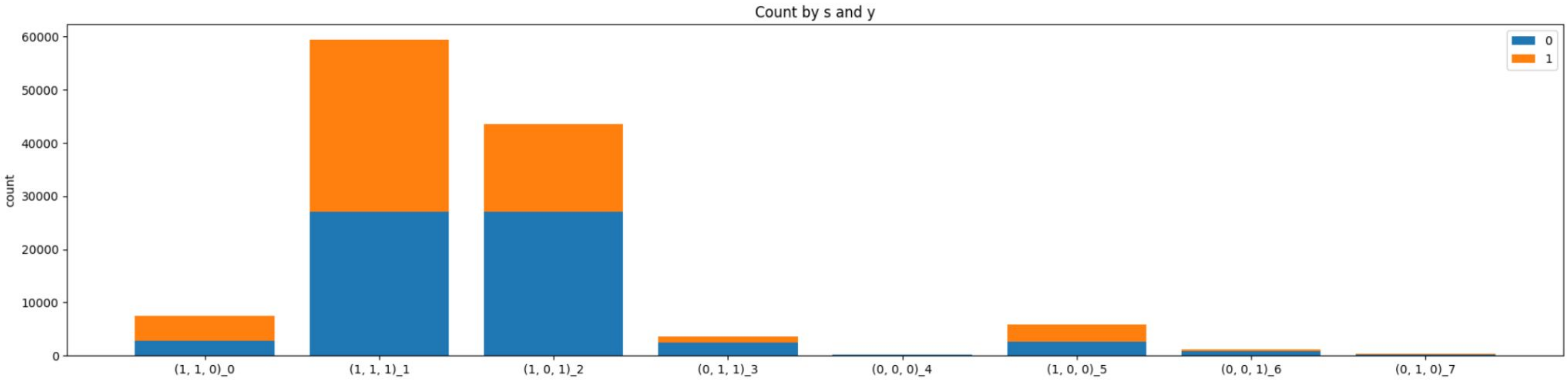
Twitter Hate Speech



Binary hate speech prediction with 4 binary sensitive axis.

Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition.

Celeb Multigroup (Artificial)



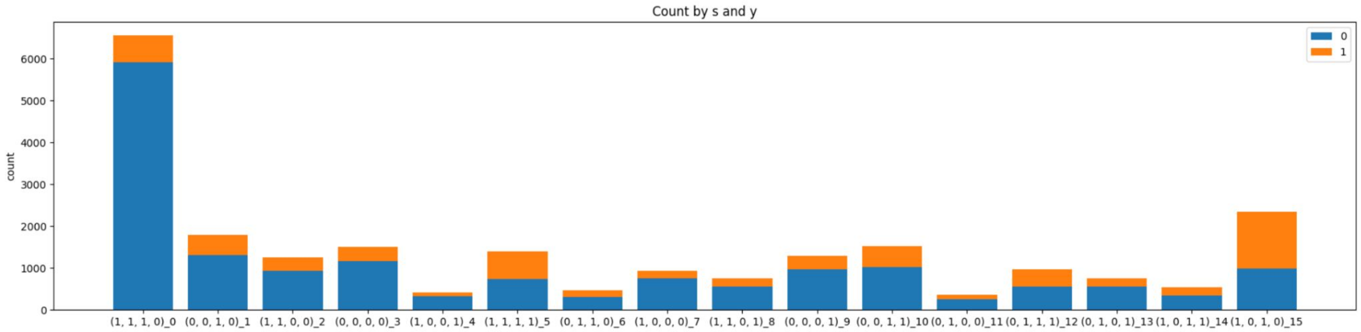
Binary “smiling” prediction with 3 binary sensitive axis.

Deep learning face attributes in the wild.

Data Generation - ongoing work!

Data Generation

- Use the data available in the larger groups to augment smaller group



Observations

- A group is composed of intersection of abstract group
    - $g_{\{\text{Male, Black, A45}\}} = g_{\{\text{X, Black, A45}\}} \cap g_{\{\text{Male, X, A45}\}} \cap g_{\{\text{Male, Black, X}\}}$

Observations

- A group is composed of intersection of abstract group
    - $g_{\{1,1,1\}} = g_{\{x,1,1\}} \cap g_{\{1,x,1\}} \cap g_{\{1,1,x\}}$
- By design, each of these abstract groups has more examples in them when compared to the given group

Observation

- A group is composed of intersection of abstract group
    - $g_{\{1,1,1\}} = g_{\{x,1,1\}} \cap g_{\{1,x,1\}} \cap g_{\{1,1,x\}}$
- By design, each of these abstract groups has more examples in them when compared to the given group
- Learn a transformation function which transforms examples from abstract group and spits out examples which looks similar to current group.

Objective

Learn a transformation function Gen_{Θ} which transforms input from the abstract groups to the required group

$$\mathcal{D}_{g_i} = Gen_{\theta}(\mathcal{D}_{abstract\_groups(g_i)}) \quad \forall g_i \in \mathcal{G}$$

Optimization Procedure

- We propose a Maximum Mean Discrepancy loss based mechanism which captures the difference between generated and real examples.

Exact training procedure

- Exact training procedure is a bit more involved but still easier to implement than GAN’s training loop.

Some Prelims Results

Generated data Quality

- A classifier which classifies if the example is from the real dataset or the generated dataset
  - Twitter hate Speech (text) - ~58%
  - Celeb Multi Group (images) - ~63%
  - Numeracy dataset (text) - ~62%
- However, subgroup accuracy varies quite a bit more.

Effect over various fairness metric

Method	Balanced Accuracy	Fairness	Best Performance	Worst Performance
Unconstrained	0.79 +/- 0.01	1.77 +/- 0.43	0.06 +/- 0.02	0.3 +/- 0.01
FairGrad	0.79 +/- 0.01	1.40 +/- 0.16	0.09 +/- 0.02	0.34 +/- 0.06
FairMixup	0.79 +/- 0.01	1.53 +/- 0.08	0.07 +/- 0.01	0.34 +/- 0.01
Adversarial	0.77 +/- 0.01	1.66 +/- 0.21	0.06 +/- 0.01	0.33 +/- 0.02
Unconstrained Augmented	0.78 +/- 0.01	1.98 +/- 0.34	0.04 +/- 0.01	0.27 +/- 0.05

Some other contributions

- Personalized model selection strategy
  - A separate snapshot of the same model for different groups
- Zero shot learning over few intersectional groups
  - No data for few groups
- Pre-print coming soon!


That's all Folks!

Reach out!

- If you want to talk about fairness or just about anything under the sun (Coffee specially), reach out to me at - https://gauravm.gitbook.io/about/