Human-Centered Computing
Research Group

Freie Universität Berlin

«Human-Centered Data Science»

Exercise 8

Lars Sipos

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

14.06.2022

Introductory Questions

- How does bias relate to fairness?
- Individual vs. Group Fairness?
- Disparate treatment vs. Disparate impact?

Fairness Metrics

Why should we care?

Fairness Metrics

Task: Provide a **definition** and **example** for your chosen metric.

Present your results to the rest of the class.

Metrics:

- Average odds difference
- Disparate impact ratio
- [Optional] Equalized odds
- [Optional] Conditional use accuracy equality

You have time until: XX

Group Fairness

Definition

A classifier satisfies the definition if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class.

Example

Probability for male and female applicants to have good predicted credit score:

$$P(d = 1|G = m) = P(d = 1|G = f)$$

Results from the classifier:

- $P(d = 1|G = m) = 0.81$
- $P(d = 1|G = f) = 0.75$

Average Odds Difference

Definition

Average of difference in false positive rates and true positive rates between unprivileged and privileged groups.

$$\frac{(FPR_{G=\text{unprivileged}} - FPR_{G=\text{privileged}}) + (TPR_{G=\text{unprivileged}} - TPR_{G=\text{privileged}})}{2}$$

Example

Results from the classifier:

» $FPR_{G=f} = 0.2$ and $FPR_{G=m} = 0.14$

» $TPR_{G=f} = 0.86$ and $TPR_{G=m} = 0.93$

Average Odds Difference:

» $$\frac{(0.2 - 0.14) + (0.86 - 0.93)}{2} = 0.005$$

Disparate Impact Ratio

Definition

The ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.

$$\frac{P(d = \text{pos\_label} \mid G = \text{unprivileged})}{P(d = \text{pos\_label} \mid G = \text{privileged})}$$

Example

Results from the classifier:

» $P(d = 1 \mid G = f) = 0.73$

» $P(d = 1 \mid G = m) = 0.74$

Disparate Impact Ratio:

» $\frac{0.73}{0.74} \approx 0.986$

Bias Mitigation Algorithms

Task: Explain your chosen algorithm to the rest of the class.

Algorithms:

- Optimized Pre-Processing *(Pre-processing)*
- Disparate impact remover *(Pre-processing)*
- Adversarial Debiasing *(In-processing)*
- Reject option classification *(Post-processing)*

You have time until: XX

Next Time

you will have …

1. actively participated in the lecture
2. submitted the fourth programming assignment

Have fun!