Algorithmic Fairness and Ethics

Srujana,Vishakha,Kharishma

June 3, 2024

Scenarios of algorithms being unfair:

- It was shown that Google's ad-targeting algorithm had proposed higher-paying executive jobs more for men than for women
- It was recently exposed that Amazon discovered that their AI hiring system was discriminating against female candidates

Causes of unfairness:

- Biases in datasets
- Biases due to missing data
- An algorithm which is intended to support majority over minority for the objective of minimizing prediction errors
- Proxy attribute

Fairness definition and measures:

Legally there are 2 types of discrimination:

- **Disparate treatment :** intentionally affecting a class of people
- **Disparate impact :** unintentionally a class of people being affected algorithms that are not given acess to sensitive data may not come under realm of disparate treatment but they show disparate impact

Fairness definition and measures:

Fairness in machine learning refers to the various attempts at correcting algorithmic bias in automated decision processes based on machine learning models.

Group notions of fairness:

- **Disparate impact :** the proportion of the positive predictions is similar across groups the same is represented as

$$\frac{P(Y' \mid S \neq 1)}{P(Y' \mid S = 1)} \geq 1 - \epsilon$$

- **Demographic parity/statistical parity :** same like disparate impact difference instead of ratio represented as

$$\frac{P(Y' \mid S \neq 1)}{P(Y' \mid S = 1)} \leq 1 - \epsilon$$

Group notions of fairness:

- **Equalized odds :** considers the difference between false positive rates(fpr) and true positive rates(tpr) of both groups to be bounded to epsilon

$$P(\hat{Y}=1 \mid S\neq 1, Y=0) - P(\hat{Y}=1 \mid S\neq 1, Y=0) \leq \epsilon$$

$$P(\hat{Y}=1 \mid S\neq 1, Y=1) - P(\hat{Y}=1 \mid S\neq 1, Y=1) \leq \epsilon$$

- **Equal opportunity :**

$$P(\hat{Y}=1 \mid S\neq 1, Y=1) - P(\hat{Y}=1 \mid S\neq 1, Y=1) \leq \epsilon$$

Individual Fairness

- $$P(Y\hat{(}i) = Y \mid X\hat{(}i), S\hat{(}i)) - P(Y\hat{(}j) = Y \mid X\hat{(}j), S\hat{(}j)) \leq \epsilon$$

  if d(i,j) $\approx$ 0

  where i and j denote two values ,s(.)refers to th individual's sensitive attriubtes and x(.) refers to their associated features .d(i,j) is the distace matric betwen individuals that can be defined depeding on the domain such that similarity is measurd according to an intended

Tradeoffs

- **Trade-Off:** Achieving higher fairness without significantly compromising accuracy
- **Goal:** Develop a fairness-aware algorithm that balances accuracy and fairness effectively.

Mechanisms to enhance fairness:

There are three types of mechanisms to enhance fairness :

- Preprocess
- Inprocess
- Postprocess

Preprocess mechaisms:

Modifies training data to remove bias before training.
Techniques:

- Re-weighting.
- Correlation Remover.

Advantages:

- Model-agnostic.
- Ensures fair training data.

Disadvantages

- May lose information
- Doesn't address model-specific biases

Inprocess mechanisms:

Incorporates fairness constraints directly into the learning algorithm. Techniques:

- Fairness-Constraint Optimization.
- Adversal Debiasing.

Advantages

- Directly ensures fairness in the model.
- Optimizes accuracy and fairness simultaneously.

Disadvantages

- Requires modification of the algorithm,
- More complex

Post process mechanisms :

Adjusts model predictions to ensure fairness after training.
Techniques

- Threshold Adjustment.
- Calibrated Equalized Odds.

Advantages

- Applicable for any trained model.
- Simple implementation.

Disadvantages

- Might reduce accuracy.
- Limited fairness based on the model

Observation

Slight decrease of accuracy when applying fairness mitigation techniques
Reasons:

- Trade-off Between Fairness and Accuracy.
- Alteration of Training Data (Pre-processing Techniques).
- Fairness Constraints in Learning (In-processing Techniques).
- Adjustment of Predictions (Post-processing Techniques).
- Complexity of Fairness Constraints:

Algorithmic Ethics

Algorithmic ethics involves ensuring that computer programs and algorithms make fair, unbiased, and responsible decisions. It emphasizes the development and deployment of algorithms in ways that are ethical, transparent, and respectful of individuals' rights.

Basic Principles of Algorithmic Ethics

- **Fairness**: AI systems should be impartial, ensuring equal treatment for all individuals and avoiding discrimination.
- **Transparency**: The processes and decisions made by AI should be clear, understandable, and open to scrutiny.
- **Accountability**: There should be clear guidelines and accountability for the outcomes of AI systems, with developers and organizations held responsible for their actions.
- **Privacy**: Safeguarding personal data and ensuring it is used ethically and in compliance with privacy laws is crucial.
- **Safety**: AI systems should be designed to operate safely, without causing harm, and should be reliable and robust.

Addressing Ethical Concerns in Algorithmic Decision-Making

Ethical concerns that arise in the development and deployment of AI systems. Let's explore some of these concerns:

- **Inconclusive evidence leading to unjustified actions**
- **Inscrutable evidence leading to opacity**
- **Misguided evidence leading to bias**
- **Unfair outcomes**
- **Transformative effects**
- **Traceability**

Conclusion

A fair algorithm should maintain accuracy while enhancing fairness, striving for a balance that maximizes both utility and fairness in practical applications.

Thank You