Ensuring Fairness with Transparent Auditing of Quantitative Bias in AI Systems

袁至誠
Chih-cheng Rex Yuan
rexyuan.com

Institute of Information Science, Academia Sinica

Friday 30th August, 2024

AI Bias

- AI is widely used in decision making:
  - School admission.
  - Loan approval.
  - Hiring.
  - Policing.
  - Censorship.
  - etc..
- Decision making by AI may be biased.
- Bias can come from several sources:
  - Biased data. ML is deisgned to replicate this.
  - Missing data. The datasets might not be representative.
  - Biased algorithms. The objective functions might introduce bias.
  - Sensitive attributes: Age, Gender, ..., etc..

Protected Attributes

What are the protected(sensitive) attributes?

Age	Gender	Occupation	Income	Education
28	M	Engineering	$80,000	Master
28	F	Engineering	$65,000	Master
45	M	Medicine	$100,000	Doctorate
40	F	Legal	$150,000	Law Degree
32	M	Education	$55,000	Bachelor

Table: Example Dataset

Fairness Through Unawareness

The most straightforward solution to fairness seems to be that just simply dropping all the protected columns.

- This is called fairness through unawareness.
- Formally it's
$$X_i = X_j \to \hat{Y}_i = \hat{Y}_j$$
where i, j are individuals; X is the set of attributes except protected attributes; and $\hat{Y}$ is the prediction.
- Also known as fairness through blindness and anti-classification.

Fairness Through Unawareness

The downside of this is there could still be "proxy" attributes that correlate with protected attributes: like Occupation still correlates with Income.

~~Age~~	~~Gender~~	Occupation	~~Income~~	Education
~~28~~	~~M~~	Engineering	~~$80,000~~	Master
~~28~~	~~F~~	Engineering	~~$65,000~~	Master
~~45~~	~~M~~	Medicine	~~$100,000~~	Doctorate
~~40~~	~~F~~	Legal	~~$150,000~~	Law Degree
~~32~~	~~M~~	Education	~~$55,000~~	Bachelor

Table: Example Dataset

Demographic Parity

Formally, it requires that

$$|P[\hat{Y} = 1|S = 1] - P[\hat{Y} = 1|S \neq 1]| \leq \epsilon$$

where $\hat{Y} = 1$ represents acceptance(positive); $S = 1$ represents privileged group; $S \neq 1$ represents unprivileged group where S is some protected attributes.

Let's set $\epsilon = 0.2$. If for some job opening, male is the privileged group and there are 10 female applicants and 100 male applicants, and there are 8 accepted females and 90 accepted males:

$$|8/10 - 90/100| = 0.1 < \epsilon \quad \text{so this is fair}$$

while if there were 1 accepted females and 50 accepted males:

$$|1/10 - 50/100| = 0.4 > \epsilon \quad \text{so this is unfair}$$

Disadvantages of Demographic Parity

- A fully accurate classifier may be considered unfair.
- The notion permits that we accept the qualified applicants in one demographic, but random individuals in another, so long as the percentages of acceptance match.
- For example, the case with 90 qualified males and only 1 qualified females.

Equalized Odds

- Equalized odds is designed to address the downsides of the previous two by taking into accounts the "ground truths" and consider the difference between false-positive rates and true-positive rates of the groups.
- Formally, it requires that

$$|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]| \leq \epsilon$$
$$|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]| \leq \epsilon$$

  where Y represents ground truths.
- A fully accurate classifier will necessarily satisfy the two equalized odds constraints.

Equal Opportunity

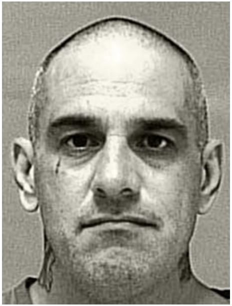
- It is a relaxation of equalized odds.
- Formally, it requires that

$$|P[\hat{Y} = 1 | S = 1, Y = 1] - P[\hat{Y} = 1 | S \neq 1, Y = 1]| \leq \epsilon$$

where Y represents ground truths.

COMPAS

recidivism *noun*
the tendency of a convicted criminal to reoffend.



- Prior Offenses: 2 armed robberies, 1 attempted armed robbery
- Subsequent Offenses: 1 grand theft
- Risk Score: 3

- Prior Offenses: 4 juvenile misdemeanors
- Subsequent Offenses: None
- Risk Score: 8

COMPAS

- COMPAS is an algorithm used by U.S. courts for predicting recidivism based on a questionaire and background information.
- In 2016, ProPublica found that the algorithm is biased.

  > *Black defendants were often predicted to be at a higher risk of recidivism than they actually were. White defendants were often predicted to be less risky than they were.*

- The false-positive rates vary significantly across black people and white people, violating equalized odds.
- Supreme Court ruled that it can be considered by judges during sentencing, but there must be warnings about the tool's "limitations and cautions."

---

[1] (Link) ProPublica - How We Analyzed the COMPAS Recidivism Algorithm
[2] (Link) Vsauce2 - The Dangerous Math Used To Predict Criminals

Other Measures

Fairness Measure	Definition
Disparate Impact	$\frac{P[\hat{Y}=1\|S\neq 1]}{P[\hat{Y}=1\|S=1]} \geq 1-\epsilon$
Demographic Parity	$\|P[\hat{Y}=1\|S=1]-P[\hat{Y}=1\|S\neq 1]\| \leq \epsilon$
Conditional Statistical Parity	$\|P[\hat{Y}=1\|S=1,L=l]-P[\hat{Y}=1\|S\neq 1,L=l]\| \leq \epsilon$
Overall Accuracy Equality	$\|P[Y=\hat{Y}\|S=1]-P[Y=\hat{Y}\|S\neq 1]\| \leq \epsilon$
Mean Difference	$\|E[\hat{Y}\|S=1]-E[\hat{Y}\|S\neq 1]\| \leq \epsilon$
Equalized Odds	$\|P[\hat{Y}=1\|S=1,Y=0]-P[\hat{Y}=1\|S\neq 1,Y=0]\| \leq \epsilon$
	$\|P[\hat{Y}=1\|S=1,Y=1]-P[\hat{Y}=1\|S\neq 1,Y=1]\| \leq \epsilon$
Equal Opportunity	$\|P[\hat{Y}=1\|S=1,Y=1]-P[\hat{Y}=1\|S\neq 1,Y=1]\| \leq \epsilon$
Predictive Equality	$\|P[\hat{Y}=1\|S=1,Y=0]-P[\hat{Y}=1\|S\neq 1,Y=0]\| \leq \epsilon$
Conditional Use Accuracy Equality	$\|P[Y=1\|S=1,\hat{Y}=1]-P[Y=1\|S\neq 1,\hat{Y}=1]\| \leq \epsilon$
	$\|P[Y=0\|S=1,\hat{Y}=0]-P[Y=0\|S\neq 1,\hat{Y}=0]\| \leq \epsilon$
Predictive Parity	$\|P[Y=1\|S=1,\hat{Y}=1]-P[Y=1\|S\neq 1,\hat{Y}=1]\| \leq \epsilon$
Equal Calibration	$\|P[Y=1\|S=1,\hat{V}=v]-P[Y=1\|S\neq 1,\hat{V}=v]\| \leq \epsilon$
Positive Balance	$\|E[\hat{V}\|Y=1,S=1]-E[\hat{V}\|Y=1,S\neq 1]\| \leq \epsilon$
Negative Balance	$\|E[\hat{V}\|Y=0,S=1]-E[\hat{V}\|Y=0,S\neq 1]\| \leq \epsilon$

Table: Fairness measures.

Framework

We made a tool to calculate these measures.
Our framework consists of 3 parties: data provider, model maker, and auditor.

- Data provider has access to the real world data. For example, a census bureau.
- Model maker designs AI models that is to be used on data. They can use our tool on their in-house data to test their models.
- Auditor is some 3rd-party that takes real data and a model or model result. They can use our tool to determine the fairness of the model.

Predicate

We first introduce the idea of a predicate.

- A predicate is a decider that determines if some condition holds for some given input.
- For example, an "adult(x)" predicate might take a person x's data and returns if they are over 18 year old.

Predicate

$$\begin{aligned}\text{adult}(r_i) &:= r_i(\text{“Age”}) > 18\\ \text{rich}(r_i) &:= r_i(\text{“Income”}) > 3000\end{aligned}$$

	Age	Income	adult(r_i)	rich(x)
r_1	14	\$10000	False	True
r_2	17	\$1000	False	False
r_3	45	\$1500	True	False
r_4	40	\$2000	True	False
r_5	32	\$5000	True	True

Table: Example Dataset with Predicates

Abstraction

Our framework tool abstracts the fairness measures:

- A row is a lookup table or dictionary. $r_n(\text{"sex"}) = \text{"Female"}$ means r_n's sex is female.
- A privileged predicate R takes a row and determines if it belongs to the privileged group. For example, $R(r_i) := r_i(\text{"race"}) == \text{"Caucasian"}$ means the privileged group is those with race being Caucasian.
- A positive predicate $\hat{P}$ takes a row and determines if its prediction is positive. For example, $\hat{P}(r_i) := r_i(\text{"score"}) > 7$ means a row's prediction is positive if its score is greater than 7.
- A ground truth predicate T takes a row and gives the ground truth of the result. For example, $T(r_i) := r_i(\text{"recid"}) ==$ True indicates a row's actual recidivism.

Definition

We can use these abstractions to define fairness measures. For equal opportunity, recall its formal definition:

$$|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]| \leq \epsilon$$

To model Y, $\hat{Y}$, and S, we define the corresponding T, $\hat{P}$, and R.

- $Y = 1$ if and only if T is true
- $\hat{Y} = 1$ if and only if $\hat{P}$ is true
- $S = 1$ if and only if R is true
- and vice versa

This way, we can calculate equal opportunity as a function:

$$\text{equal\_opportunity}(\epsilon, R, \hat{P}, T)$$

Definition

$$\text{equal\_opportunity}(\epsilon, R, \hat{P}, T)$$

This is implemented in our tool as a Python function:

```
def equal_opportunity(
    ratio,
    privileged_predicate,
    positive_predicate,
    truth_predicate):
    ...
```

Application

For demonstration, we analyzed the COMPAS dataset. Let's set the predicates:

$$R(r_i) := r_i(\text{“race”}) \neq \text{“African-American”}$$
$$\hat{P}(r_i) := r_i(\text{“score\_text”}) \in \{\text{“Medium”}, \text{“High”}\}$$
$$T(r_i) := r_i(\text{“two\_year\_recid”}) == \text{True}$$

These predicates can be published for transparency.

Application

$$R(r_i) := r_i(\text{“race”}) \neq \text{“African-American”}$$
$$\hat{P}(r_i) := r_i(\text{“score\_text”}) \in \{\text{“Medium”}, \text{“High”}\}$$
$$T(r_i) := r_i(\text{“two\_year\_recid”}) == \text{True}$$

These predicates can then be encoded with Python functions:

```
equal_opportunity(
    0.2,
    lambda row: row['race'] != 'African-American',
    lambda row: row['score_text']
      in {'Medium', 'High'},
    lambda row: row['is_recid'] == '1')
```

Application

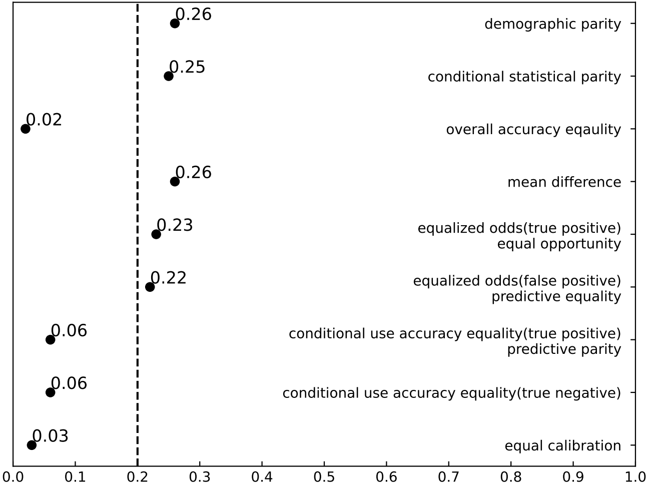


Figure: Unprivileged Group: African-American

Application

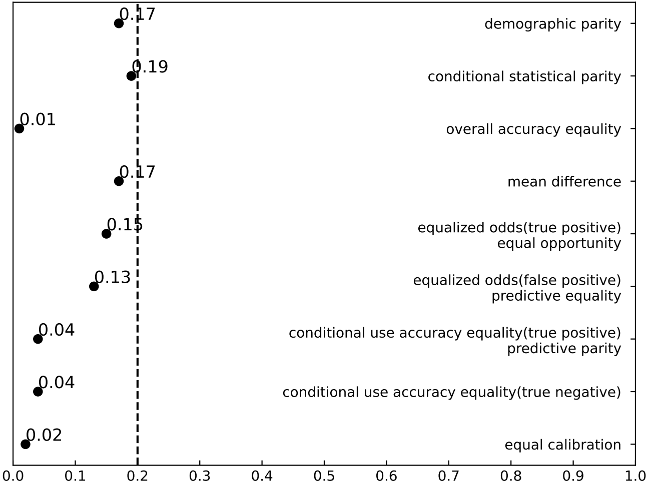


Figure: Unprivileged Group: Caucasian

Application

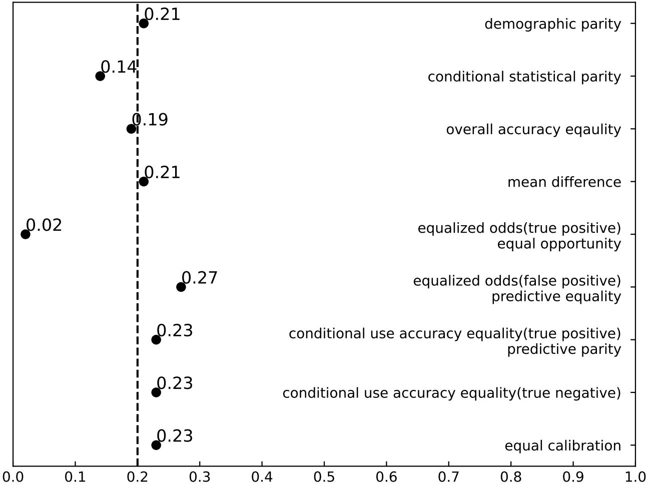


Figure: Unprivileged Group: Asian (32 rows)

Conclusion

- Decision making by AI may be biased.
- With our framework tool, auditors can comprehensively review the fairness of an AI system.

Slides: `https://github.com/RexYuan/Shu`.

Package: `https://pypi.org/project/fairness-checker`.