On Formalizing Fairness in Prediction with Machine Learning

Pratik Gajane, Mykola Pechenizkiy

Fairness: From Social Science to Machine Learning

- **Machine learning**

  → Critical decision making affecting **human** lives

- ML algorithms should be prevented from systematic discrimination

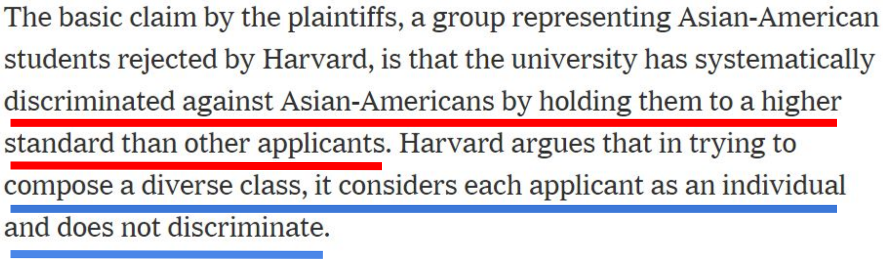
“ML with Fairness”

Fairness: From Social Science to Machine Learning

- However, people might have **diverse understandings** of fairness…

The New York Times

The Harvard Bias Suit by Asian-Americans: 5 Key Issues



The basic claim by the plaintiffs, a group representing Asian-American students rejected by Harvard, is that the university has systematically discriminated against Asian-Americans by holding them to a higher standard than other applicants. Harvard argues that in trying to compose a diverse class, it considers each applicant as an individual and does not discriminate.

https://www.nytimes.com/2018/12/20/us/harvard-asian-american-students-discrimination.html

Fairness: From Social Science to Machine Learning

This paper…

❑ Introduce **different notions of fairness** and how they are **formalized** in machine learning literature

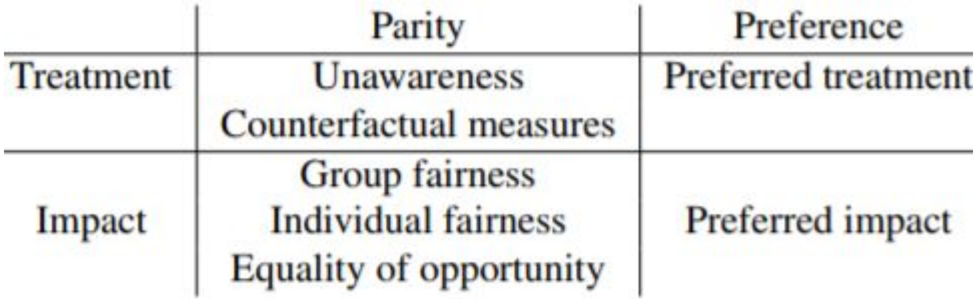
❑ Provide theoretical and empirical **critiques** of each notion from **social sciences**

❑ Determine the **suitability** of each formalization of fairness in the context of **machine learning**

Formalizations of Fairness: Taxonomy

- **Parity or preference?** : whether fairness means achieving parity or satisfying the preferences.
- **Treatment or impact?** : whether fairness is to be maintained in treatment or impact (results).

7 existing notions of fairness in ML literature



	Parity	Preference
Treatment	Unawareness Counterfactual measures	Preferred treatment
Impact	Group fairness Individual fairness Equality of opportunity	Preferred impact

What is Fair: Fairness through Unawareness

❑ Definition 1

A predictor is said to achieve ***fairness through unawareness*** *if* *protected attributes* *are not explicitly used in the prediction process.*

Social science (SS) notion: being **“blind”** to counter discrimination

Protected Attributes

certain demographic attributes protected by law against discrimination (e.g. sex, gender, race, etc.)

What is Fair: Fairness through Unawareness

- ❑ **Critiques**
  - ❑ Protected attributes may be no longer blind when **additional information** is available
  - ❑ Discriminatory practices have been observed following race-blind approach in SS studies
- ❑ **Suitability**
  - ❑ problematic for domains in which protected attributes can be **deducted from easily available non-protected attributes**

Protected Attributes

certain demographic attributes protected by law against discrimination (e.g. sex, gender, race, etc.)

What is Fair: Counterfactual Measures

❑ Definition 2

A predictor $\mathcal{H}$ is ***counterfactually fair****, given protected attributes A = a and non-protected attributes Z = z, iff for all outcome y and a ≠ a′,*

$$\mathbb{P}\{\mathcal{H}(A,Z) = y | A = a, Z = z\} = \mathbb{P}\{\mathcal{H}(A,Z) = y | A = a', Z = z\}$$

SS notion: **“counterfactual reasoning”**

➔ The outcome still remains the same even if the protected attributes were flipped

What is Fair: Counterfactual Measures

- ❏ **Critiques**
  - ❏ Hindsight bias & Outcome bias
  - ❏ Negatively influence the process of causality
- ❏ **Suitability**
  - ❏ problematic for domains where **the above mentioned biases are frequently observed**, e.g., health-care and judicial systems

Hindsight Bias
the tendency for people to perceive events that have already occurred as having been more predictable than they actually were

Outcome Bias
evaluating the quality of a decision when the outcome of that decision is already known

What is Fair: Group Fairness

❑ Definition 3

A predictor $\mathcal{H}: X \to Y$ achieves ***group fairness*** *with bias* ϵ *with respect to groups $S, T \subset X$ and $O \subseteq Y$ being any subset of outcomes iff*

$$|\mathbb{P}\{\mathcal{H}(x_i) \in O | x_i \in S\} - \mathbb{P}\{\mathcal{H}(x_j) \in O | x_j \in T\}| \leq \epsilon$$

SS notion: **“collectivist egalitarianism”**

➔ Affirmative Action Policies (US, India, etc.)

Affirmative Action

the policy of promoting the education and employment of members of groups that are known to have previously suffered from discrimination

What is Fair: Group Fairness

Critiques

- It is not **meritocratic**
  - Group fairness is blind to “ground truth” → discrimination against **“qualified“** candidates
  - The predictor can select anyone **within a group** as long as it maintains statistical parity
- It reduces **efficiencies**

Suitability

- The controversies above limits its applicability

Meritocracy

certain things, like economic goods or power, should be vested in individuals on the basis of talent, effort, and achievement

What is Fair: Individual Fairness

❑ Definition 4

*A predictor achieves **individual fairness** iff*

$$D(\mathcal{H}(x_i)_Y, \mathcal{H}(x_j)_Y) \approx 0 \,|\, d(x_i, x_j) \approx 0$$

where d : X × X → R is a distance metric for individuals and D is a distance measure for distributions.

SS notion: **“individualist egalitarianism”**

➔ Similar outputs for similar individuals

What is Fair: Individual Fairness

- **Critiques**
  - How to define the similarity of individuals?
  - If the distance metric **uses the protected attributes directly or indirectly**, a predictor satisfying Definition 4 could still be discriminatory
- **Suitability**
  - not suitable for domains where **reliable and non-discriminating distance metric** is not available

What is Fair: Equality of Opportunity

❑ Definition 5

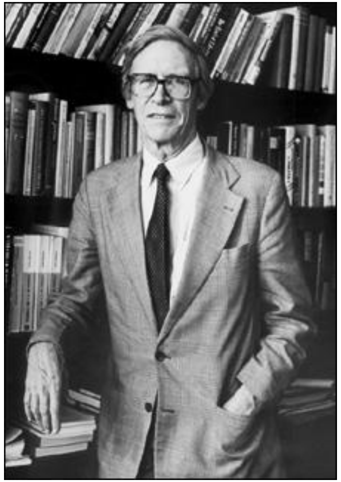
A predictor $\mathcal{H}$ is said to satisfy ***equal opportunity*** *with respect to group $S \subset X$ iff (here y denotes the true label)*

$$\mathbb{P}\{\mathcal{H}(x_i) = 1 | y_i = 1, x_i \in S\} = \mathbb{P}\{\mathcal{H}(x_j) = 1 | y_j = 1, x_j \in X \backslash S\}$$

"equivalence of true positive rate across groups"

SS notion: **John Rawls' *A theory of Justice* (1971)**

➔ People with "the same native talent and the same ambition" have the same prospects of success



John B. Rawls (1921 - 2002)

What is Fair: Equality of Opportunity

- ❑ Critiques
  - ❑ “**Stunted ambition**” & “**Selection by bigotry**”
  - ❑ Not considering the effect of discrimination due to protected attributes which essentially affect one’s **access to opportunities** (“structural barriers”)
- ❑ Suitability
  - ❑ problematic for domains in which there exists vast evidence that **protected attributes do indeed affect one’s prospects**

What is Fair: Preference-based Fairness

Definition 6

(Preferred treatment) A ***group-conditional*** *predictor is said to satisfy* ***preferred treatment*** *if each group receives more* *group benefit* *from their respective predictor than they would have received from any other predictor i.e.*

$$\mathbb{B}_S(\mathcal{H}_S) \geq \mathbb{B}_S(\mathcal{H}_T) \qquad \text{for all } S, T \subset X$$

Group Benefit

The expected proportion of individuals in the group for whom the predictor predicts the beneficial outcome.

(Alternate def: The expected proportion of individuals from the group who receive the beneficial output for whom the true label is the same.)

What is Fair: Preference-based Fairness

❑ Definition 7

(Preferred impact) A predictor $\mathcal{H}$ is said to have ***preferred impact*** *as compared to another predictor $\mathcal{H}'$ if $\mathcal{H}$ offers* ***at-least as much benefit as*** *$\mathcal{H}'$ for all the groups.*

$$\mathbb{B}_S(\mathcal{H}) \geq \mathbb{B}_S(\mathcal{H}') \qquad \text{for all } S \subset X$$

- ❑ Individuals in one group may prefer another outcome than the one preferred by the majority of the group.

SS notion: “**Envy-Freeness**”

➔ It can be defined in terms of ordinal preference relations of the utility values of the predictors.

Envy-Freeness

In an envy-free division, every agent feels that their share is at least as good as the share of any other agent, and thus no agent feels envy.

What is Fair: Preference-based Fairness

- **Critiques**
    - Freedom from envy is neither necessary nor sufficient for fairness.
    - “**Pareto-efficiency**”
    - Deciding whether there is a Pareto-efficient envy-free allocation is computationally very hard even with simple additive preferences.
- **Suitability**
    - Limited to the domains where such an effective and envy-free allocation can be computed easily.

Pareto-Efficiency

An allocation is ‘Pareto efficient’ if there is no other allocation in which some other individual is better off and no individual is worse off.

Prospective notions of fairness: Equality of Resources

NEW!

Definition 1

Unequal distribution of social benefits is only considered fair when it results from the ***intentional decisions and actions*** *of the concerned individuals.*

- *Ambition-sensitive*: Each individual’s ambitions and choices that follow them ascertains their benefits.
- *Endowment-insensitive*: Each individual’s unchosen circumstances including the ***natural endowments*** should be offset.

Prospective notions of fairness: Equality of Capability of Functioning

NEW!

❑ Definition 2

People should not be held responsible for attributes they had no say in to include personal attributes which cause difficulty in developing ***functionings****.*

- ❑ In order to equalize capabilities, people should be compensated for their unequal powers to convert opportunities into functionings.
- ❑ Flexible and widely used in many ways
- ❑ Difference between resource equality and capability equality
  - ❑ **Social endowment & Natural endowment**
  - ❑ **what we can get vs what we can do**

Functionings

“being and doing”: various states of existence and activities that an individual can undertake.

Prospective notions of fairness

- ## Critiques
  - To **Def 2**: The failure to identify of valuable capabilities
  - To **Def 1 and 2**: The informational requirement of this approach can be very high
    - Difficult to make exact mathematical formalizations
- ## Suitability
  - Makes the open problem of formalizing them worthwhile.

Discussion and Summary

- Fair prediction cannot be addressed without considering **social issues** such as unequal access to resources and social conditioning.
- It is important to acknowledge their impact and attempt to incorporate them in **fairness formalizations**.
- **Seven existing notions** in ML society: Fairness through Unawareness, Counterfactual Measures, Group Fairness, Individual Fairness, Individual Fairness, Equality of Opportunity, and Preference-based Fairness (Preferred treatment and Preferred impact)
- **Two new notions** in ML society: Equality of resources and Equality of capability of functioning.
- Short but dense, read references to better understand concept.