On Formalizing Fairness in Prediction with ML

Pratik Gajane[†], Mykola Pechenizkiy[⋆]

pratik.gajane@unileoben.ac.at, m.pechenizkiy@tue.nl

Montanuniversitat Leoben[†], Eindhoven University of Technology[⋆]

What is Fair?

- **Parity or preference?**
- **Treatment or impact?**

Table 1: The surveyed formalizations of fairness

	Parity	Preference
Treatment	1) Unawareness 2) Counterfactual measures	6) Preferred treatment
Impact	3) Group fairness 4) Individual fairness 5) Equality of opportunity	7) Preferred impact

Mathematical Formulation

- X : Set of individuals i.e. *population*
  Y : Set of outcomes
  A : *Protected* attributes e.g. race, gender etc
  Z : Remaining attributes
- For individual $x_i \in X$, let true outcome (label) be $y_i \in Y$.
- Predictor $\mathcal{H} : X \to Y$ such that $\mathcal{H}(x_i)$ is the predicted outcome.
- Group-conditional predictor $\mathcal{H} = \{\mathcal{H}_S\}$ where $S \subset X$.

Fairness through Unawareness ∼ "Blind" Approach to Fairness

Definition 1. *Protected attributes are not explicitly used in prediction.*

- Not a sufficient condition to avoid discrimination.
- Discriminatory practices following race-blind approach in education, housing, credit, criminal justice system etc (Bonilla-Silva (2013)).
- Some studies show a blind approach works for few specific tasks.

Counterfactual Measures ∼ Counterfactual Reasoning

Definition 2. *Given $Z = z$ and $A = a$, for all y and $a \neq a'$,*

$$\mathbb{P}\{\mathcal{H}_{A=a} = y \mid Z = z, A = a\} = \mathbb{P}\{\mathcal{H}_{A=a'} = y \mid Z = z, A = a\}$$

- $\mathcal{H}_{A=a}$ = outcome of $\mathcal{H}$ if A had taken value a.
- Research to indicate that counterfactual reasoning is susceptible to hindsight bias and outcome bias (Petrocelli (2010).
- Some argue that counterfactual reasoning may negatively influence identifying causality (Roese (1997)).

Group Fairness ∼ Collectivist Egalitarianism

Definition 3. *Group fairness with bias ϵ with respect to groups $S, T \subseteq X$ and $O \subseteq A$ being any subset of outcomes iff*

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

- Equivalent to statistical and demographic parity.
- Biggest implementation = affirmative action.
- See Weisskopf (2004) for arguments made for and against affirmative action.

Individual Fairness ∼ Individualistic Egalitarianism

Definition 4. *$\mathcal{H}(x_i) \approx \mathcal{H}(x_j) \mid d(x_i, x_i) \approx 0$ where $d : X \times X \to \mathbb{R}$ is a distance metric for individuals.*

- Distance metric critical for ensuring fairness.
- In some domains, reliable and non-discriminating distance metric may be unavailable.

Equality of Opportunity ∼ Equality of Opportunity

Definition 5. $\mathbb{P}\{\mathcal{H}(x_i) = 1 \mid y_i = 1, x_i \in S\} = \mathbb{P}\{\mathcal{H}(x_j) = 1 \mid y_j = 1, x_j \in X \backslash S\}$

- Keeps true positive rate same for all the groups.
- Argument that it cannot deal with *stunted ambition* and *selection by bigotry*.
- Attributes like gender and race not deemed to be affecting an individual's life prospects while numerous surveys conclude otherwise.

Preference-based Fairness ∼ Envy-freeness

Definition 6. *(*Preferred treatment*) A group-conditional predictor in which each group receives more benefit from their respective predictor.*

Definition 7. *(*Preferred impact*) $\mathcal{H}$ has preferred impact compared to $\mathcal{H}'$ if $\mathcal{H}$ offers at-least as much benefit as $\mathcal{H}'$ for all the groups.*

- In certain domains, no single universally accepted beneficial outcome.
- Freedom from envy is neither necessary nor sufficient for fairness.
- Envy-freeness from-ally expressed by *Pareto-efficiency*.
- Finding Pareto-efficient solutions computationally very hard.

Prospective Notions of Fairness

Definition 8. *(*Equality of resources*) Unequal distribution of benefits fair when it results from intentional decisions and actions. (Dworkin (1981))*

- *Ambition-sensitive* and *endowment-insensitive*.
- Being endowment-insensitive differentiates equality of resources from equality of opportunity.

Definition 9. *(*Equality of capability of functioning*) In order to equalize capabilities, people should be compensated for their unequal powers to convert opportunities into functionings. (Sen (1992))*

- Functionings = various states of existence and activities that an individual can undertake.
- Calls for addressing inequalities due to social endowments (e.g. gender) as well as natural endowments (e.g. sex).
- Used in the foundations of human development paradigm by the United Nations.
- High informational requirement and difficult to express mathematically.

Further Directions

- Use of social science literature while choosing fairness formalizations in particular domains.
- Fair prediction cannot be achieved without considering social issues such as unequal access to resources and social conditioning.
- Acknowledge their impact and attempt to incorporate them in fairness formalizations.

Key references

[Weisskopf, Thomas (2004)] Affirmative action in the United States and India : a comparative perspective

[Bonilla-Silva, Eduardo (2013)] Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States

[Petrocelli, John (2010)] Event detail and confidence in gambling: The role of counterfactual thought reactions

[Roese, Neal and Olson, James (1997)] Counterfactual Thinking: The Intersection of Affect and Function

[Dworkin, Ronald (1981)] What is Equality? Part 2: Equality of Resources

[Sen, Amartya (1992)] Inequality Reexamined