On Formalizing Fairness in Prediction with Machine Learning

Pratik Gajane [1] Mykola Pechenizkiy [2]

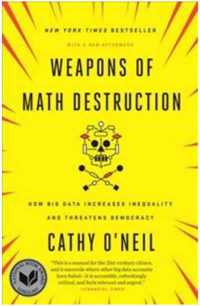
[1]Montanuniversität Leoben

[2]Eindhoven University of Technology

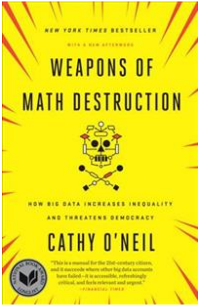
15$^{\text{th}}$ July 2018

Presentation at the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning

Introduction

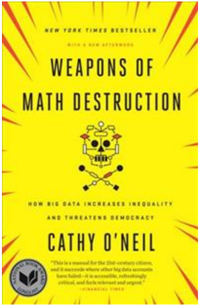


Introduction



➜ ICML 2018 Tutorial : Defining and designing fair algorithms by Sam Corbett-Davies and Sharad Goel.

Introduction



➜ ICML 2018 Tutorial : Defining and designing fair algorithms by Sam Corbett-Davies and Sharad Goel.

So long as I do not know what the just is, I shall hardly know whether it is a virtue or not.

Socrates

What? How? Why?

- Our Task: Analyze fairness formalizations considered in ML so far.

What? How? Why?

- Our Task: Analyze fairness formalizations considered in ML so far.
- Our Method: Juxtapose the formalizations in ML with their corresponding theories in Social Sciences.

What? How? Why?

- Our Task: Analyze fairness formalizations considered in ML so far.
- Our Method: Juxtapose the formalizations in ML with their corresponding theories in Social Sciences.
- Our Objective: Start a discussion and propose newer fairness formalizations in ML.

Mathematical Formulation

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

What is Fair?

- **Parity or preference?** : Statistical Parity or Social Preference?
- **Treatment or impact?** : A property of the process or of its results?

Table 1: The surveyed formalizations of fairness

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

Fairness through Unawareness

Definition

Protected attributes are not explicitly used in prediction.

- Not sufficient to avoid discrimination.
- $\sim$ "Blind" approach to counter discrimination.
- Various discriminatory practices following race-blind approach Bonilla-Silva (2013) [3], Taslitz (2007) [13].
- Race-blind approach is less efficient than race-conscious approach Fryer (2008) [5].
- Alternatively, some studies show a blind approach can work Glodin (2000) [6].

Counterfactual Measures

Definition

Changing A while holding attributes which are not causally dependent on A constant will not change the distribution of $\mathcal{H}$.

- Given $Z = z$ and $A = a$, for all y and $a \neq a'$, $\mathbb{P}\{\mathcal{H}_{A=a} = y | Z = z, A = a\} = \mathbb{P}\{\mathcal{H}_{A=a'} = y | Z = z, A = a\}$ where $\mathcal{H}_{A=a}$ = outcome of $\mathcal{H}$ if A had taken value a.
- Introduced by Kusner et al. [9]. Similar measure introduced independently by Kilbertus et al. [8].
- $\sim$ Counterfactual reasoning given by Lewis (1973) [10]
- Research to indicate that counterfactual reasoning is susceptible to hindsight bias and outcome bias.
- Some argue that counterfactual reasoning may negatively influence identifying causality.

Group Fairness

Definition

Prob. of an individual from one group getting a particular outcome $\approx$ *Prob. of an individual from another group getting same outcome.*

- Equivalent to statistical and demographic parity.
- Independent of "ground truth".
- $\sim$ *Collectivist egalitarianism* from distributive justice.
- Biggest implementation $=$ affirmative action.
- Arguments have been made for and against affirmative action Weisskopf (2004) [14].

Individual Fairness

Definition

Similar individuals get similar outcome.

- Mathematically, $\mathcal{H}(x_i) \approx \mathcal{H}(x_j) \mid d(x_i, x_i) \approx 0$ where is a distance metric for individuals.
- $\sim$ *Individualist egalitarianism* from distributive justice.
- Distance metric is critical to ensure non-discrimination.
- In some domains, reliable distance metric may be unavailable.

Equality of Opportunity

Definition

True positive rate should be the same for all the groups.

- $\mathbb{P}\{\mathcal{H}(x_i) = 1 \mid y_i = 1, x_i \in S\} = \mathbb{P}\{\mathcal{H}(x_j) = 1 \mid y_j = 1, x_j \in X \setminus S\}$
- *Disparate mistreatment* : Equivalence of misclassification rates across the groups.
- $\sim$ Equality of Opportunity by Rawls (1971) [11].
- Argument that it cannot deal with *stunted ambition* and *selection by bigotry* Arneson (1999) [1].
- Attributes like gender and race not deemed to be affecting an individual's life prospects while numerous surveys conclude otherwise.

Preference-based Fairness

Definition

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

Definition

(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.
- $\sim$ *envy-freeness* Arnsperger (1994) [2].
- Freedom from envy neither necessary nor sufficient for fairness (Holocombe 1977 [7])
- Envy-freeness formally expressed by *Pareto-efficiency*.
- Finding Pareto-efficient solutions computationally hard.

Equality of Resources

Definition

(Equality of resources) *Unequal distribution of benefits fair when it results from intentional decisions and actions. (Dworkin (1981) [4])*

- *Ambition-sensitive*
- *Endowment-insensitive*
- In the 2nd property, it differs from equality of opportunity.

Equality of Capability of Functioning

Definition

(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) [12])*

- 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.

Summary and Further Directions

- Juxtaposed ML fairness formlaizatioons with theories from distributive justice.
- Critique and analysis from the social sciences literature.
- Nominate two notions from the social sciences literature as prospective ML fairness formalizations.
- 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.

Thank you all.

References I

[1] Richard J. Arneson. Against rawlsian equality of opportunity. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 93(1):77–112, 1999.

[2] Christian Arnsperger. Envy-Freeness and Distributive Justice. *Journal of Economic Surveys*, 8(2):155–186, June 1994.

[3] Eduardo Bonilla-Silva. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States*. Rowman & Littlefield Publishers, 4th edition, 2013.

[4] Ronald Dworkin. What is equality? part 2: Equality of resources. *Philosophy and Public Affairs*, 10(4):283–345, 1981.

References II

[5] Roland Fryer, G. Loury, and T. Yuret. An economic analysis of color-blind affirmative action. *Journal of Law, Economics, and Organization*, 24(2):319–355, 2008.

[6] Claudia Goldin and Cecilia Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, September 2000.

[7] Randall G. Holcombe. Absence of envy does not imply fairness. *Southern Economic Journal*, 63(3):797–802, 1997.

References III

[8] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 656–666. Curran Associates, Inc., 2017.

[9] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[10] David Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

References IV

[11] John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

[12] Amartya Sen. *Inequality Reexamined*. Clarendon Press, Oxford, 1992. New York: Russell Sage Foundation, and Cambridge. MA: Harvard University Press, 1992; Italian translation: Il Mulino, 1994; French translation: Seuil, 2000.

[13] Andrew Taslitz. Racial blindsight: The absurdity of color-blind criminal justice. *Ohio State Journal of Criminal Law*, 2007.

[14] Thomas E. Weisskopf. *Affirmative action in the United States and India : a comparative perspective / Thomas E. Weisskopf*. Routledge London ; New York, 2004.