A Fair Definition of Fairness

Giorgian Borca-Tasciuc

2022-03-02

1 Introduction

Intuitively, we hope that the some protected attribute of a person, such as race, should not affect the decision made by the neural network. A first-attempt approach to this might be something like:

> *A network is unfair if there is some set of individuals for which changing the class of the protected attribute leads to a different decision.*

This is a strong definition of unfairness, but further inspection shows that white this is a *sufficient* criterion for fairness, it is not a *necessary* one. It is still possible for the network to be "fair" under this criterion if it uses other variables to infer the race, while ignoring the race given in the input (or perhaps not given). A clear example of this is race and zip code. Consider the probability distribution on the input to a neural network given in table (1), along with the decision space characterized in table (2). Clearly, there is no region where changing the race leads to a different outcome. However, it is clear that there are variables *that can be used as a proxy for race*. We call such variables *race proxies*. In this case, this is the *zip code*, which is strongly correlated with the race.

Table 1: Input Distribution, conditioned on Race

$P(Z = Z_1 \mid R = \text{W}) = 0$	$P(Z = Z_1 \mid R = \text{B}) = 1$
$P(Z = Z_2 \mid R = \text{W}) = 1$	$P(Z = Z_2 \mid R = \text{B}) = 0$

Table 2: Decision Space

$D(Z = Z_1, R = \text{W}) = 0$	$P(Z = Z_1, R = \text{B}) = 0$
$D(Z = Z_2, R = \text{W}) = 1$	$P(Z = Z_2, R = \text{B}) = 1$

We might hope to address this by incorporating the *input probability distribution* into our definition of fairness. In this case, consider the following

two probability distributions $P(\vec{X}|\mathrm{W})$ and $P(\vec{X}|\mathrm{B})$. Consider $\rho(\mathrm{W})$, the region where the network "accepts" all individuals who are of the white race, and $\rho(\mathrm{B})$, the region where the network "accepts" all individuals who are of the black race. This allows us to define the follow expressions for the *advantage* $\mathrm{Adv}(R_1, R_2)$ of race R_1 with respect to race R_2:

$$\mathrm{Adv}(\mathrm{W}, \mathrm{B}) = \underbrace{\int_{\rho(\mathrm{W})} P(\vec{X}|\mathrm{W})dA}_{t_1} - \underbrace{\int_{\rho(\mathrm{W})\cap\rho(\mathrm{B})} P(\vec{X}|\mathrm{W})dA}_{t_2}$$

The term t_1 corresponds to the total probability that a white person is accepted, while the term t_2 corresponds to the probability that a white person who would have been accepted as a black person is accepted. Note that by definition, the *advantage* is always positive. Similarly, we can define the *advantage* of a black person over a white person:

$$\mathrm{Adv}(\mathrm{B}, \mathrm{W}) = \underbrace{\int_{\rho(\mathrm{B})} P(\vec{X}|B)dA}_{t_3} - \underbrace{\int_{\rho(\mathrm{W})\cap\rho(\mathrm{B})} P(\vec{X}|\mathrm{B})dA}_{t_4}$$

The term t_3 corresponds to the total probability that a white person is accepted, while the term t_4 corresponds to the probability that a white person who would have been accepted as a black person is accepted.

Now that we have defined the advantage, we are interested in whether it is capable of handling race proxies. Let us calculate the terms based on the toy examples from table (1) and (2). We will have to change the integrations to summations, but this does not affect the essence of our calculations, and the procedure for modifying the integrals to a mix of summations over integrals can be done in a methodical way, as will be explained later. First, we define our accept regions for each race:

$$\begin{aligned} Z &= \{Z_1, Z_2\} \\ \rho(\mathrm{W}) &= \{Z_2\} \\ \rho(\mathrm{B}) &= \{Z_2\} \\ \rho(\mathrm{W}) \cap \rho(\mathrm{B}) &= \{Z_2\} \end{aligned}$$

Then, we calculate the white's advantage of black:

$$\begin{aligned} t_1 &= \sum_{z\in\rho(\mathrm{W})} P(z|\mathrm{W}) = 1 \\ t_2 &= \sum_{z\in\rho(\mathrm{W})} P(z|\mathrm{W}) = 1 \\ \mathrm{Adv}(W, B) &= 0 \end{aligned}$$

And finally, we calculate black's advantage over white:

$$t_3 = \sum_{z \in \rho(\mathrm{W})} P(z|\mathrm{B}) = 0$$
$$t_4 = \sum_{z \in \rho(\mathrm{W})} P(z|\mathrm{B}) = 0$$
$$\mathrm{Adv}(B, W) = 0$$

Thus, integrating the probability distribution in such a straightforward way is not enough to determine fairness. The model is on-its-face "fair" in the sense that the declared race of the individual is irrelevant to the final decision. However, there is clearly something missing, because the model clearly demonstrates a "preference" for features that white people have over black people.

2 Fairness Criterions

2.1 Criterion Evaluation

To enable a further evaluation of our criterion for fairness, we will make explicit the axes along which we are evaluating our critera,

1. *Sufficiency*: A criterion is "sufficient" if all situations that satisfy the criterion are truly unfair.
2. *Necessity*: A criterion is "necessary" if all situations that are unfair satisfy the criterion.

2.2 Possible Terms and Metric Definitions

There are 6 possible terms one can coherently define on the acceptance region, as listed in table (3). From the possible terms, the following metrics are definable, as listed in table (4).

Investigating the criterions, we might say that for decisions that are race-blind, preference is a *necessary* metric, as any unfair model will demonstrates a positive preference, and advantage is a *sufficient* metric: any model that demonstrates advantage is demonstrably unfair.

We can now return to the example given in table (1) and table (2).

$$Z = \{Z_1, Z_2\}$$
$$\rho(\mathrm{W}) = \{Z_2\}$$
$$\rho(\mathrm{B}) = \{Z_2\}$$
$$\rho(\mathrm{W}) \cap \rho(\mathrm{B}) = \{Z_2\}$$

Table 3: Possible Terms

Term	Meaning
$P(\mathrm{A}\vert\mathrm{W}) = \int_{\rho(\mathrm{W})} P(\vec{X}\vert\mathrm{W})dX$	Probability of acceptance of a white person
$P(\mathrm{A}\vert\mathrm{B}) = \int_{\rho(\mathrm{B})} P(\vec{X}\vert\mathrm{B})dX$	Probability of acceptance of a black person
$P^{\mathrm{B}}(\mathrm{A}\vert\mathrm{W}) = \int_{\rho(\mathrm{B})} P(\vec{X}\vert\mathrm{B})dX$	Probability of acceptance of a white person under black criteria
$P^{\mathrm{W}}(\mathrm{A}\vert\mathrm{W}) = \int_{\rho(\mathrm{W})} P(\vec{X}\vert\mathrm{B})dX$	Probability of acceptance of a white person judged under black criteria
$P(\mathrm{JA}\vert\mathrm{W}) = \int_{\rho(\mathrm{W})\cap\rho(\mathrm{B})} P(\vec{X}\vert\mathrm{W})dX$	Probability of acceptance of a white person who would also have been accepted as black person
$P(\mathrm{JA}\vert\mathrm{B}) = \int_{\rho(\mathrm{W})\cap\rho(\mathrm{B})} P(\vec{X}\vert\mathrm{B})dX$	Probability of acceptance of a black person who would also have been accepted as a white person

Table 4: Possible Metrics

Name	Metric	Meaning
White Preference	$\mathrm{Pref}(W,B) = P(\mathrm{A}\vert\mathrm{W}) - P^{W}(\mathrm{A}\vert\mathrm{B})$	Quantifies how strongly white features are preferred
Black Preference	$\mathrm{Pref}(B,W) = P(\mathrm{A}\vert\mathrm{B}) - P^{B}(\mathrm{A}\vert\mathrm{W})$	Quantifies how strongly black features are preferred
White Advantage	$\mathrm{Adv}(W,B) = P(\mathrm{A}\vert\mathrm{W}) - P(\mathrm{JA}\vert\mathrm{W})$	Quantifies the amount of whites who would not have been accepted as blacks
Black Advantage	$\mathrm{Adv}(B,W) = P(\mathrm{A}\vert\mathrm{W}) - P(\mathrm{JA}\vert\mathrm{W})$	Quantifies the amount of blacks who would not have been accepted as whites

We calculate the model's preference for white features over black:

$$\begin{aligned}
\mathrm{Pref}(\mathrm{W},\mathrm{B}) &= P(\mathrm{A}\vert\mathrm{W}) - P^{W}(\mathrm{A}\vert\mathrm{B}) \\
P(\mathrm{A}\vert\mathrm{W}) &= \sum_{z\in\rho(\mathrm{W})} P(z\vert\mathrm{W})\cdot\frac{1}{|Z|} = \frac{1}{2} \\
P^{W}(\mathrm{A}\vert\mathrm{B}) &= \sum_{z\in\rho(\mathrm{W})} P(z\vert\mathrm{B})\cdot\frac{1}{|Z|} = 0 \\
\mathrm{Pref}(W,B) &= \frac{1}{2}
\end{aligned}$$

And we calculate the model's preference for black features over white:

$$\begin{aligned}
\text{Pref}(\text{B},\text{W}) &= P(\text{A}|\text{B}) - P^B(\text{A}|\text{W}) \\
P(\text{B}|\text{W}) &= \sum_{z\in\rho(\text{B})} P(z|\text{B}) \cdot \frac{1}{|Z|} = 0 \\
P^B(\text{A}|\text{W}) &= \sum_{z\in\rho(\text{W})} P(z|\text{W}) \cdot \frac{1}{|Z|} = 0 \\
\text{Pref}(B,W) &= 0
\end{aligned}$$

Thus, our metric detects that there is a preference for white features over black. Whether this is unfair requires investigation of the use of the model and the context in which it is applied. If the model was simply being used determine whether the individuals lived in a specific zip code, this would not be unfair.

2.3 Incorporating Valid Discrimination Factors into Preference

Sometimes, we would like our *preference* metric to ignore certain features that do defer across race, but it is not unfair to select based on that feature. This could be the case for a model in which applicants to a faculty position are automatically rejected if they do not have doctorate degrees. We define those features as $F = \{F_1, F_2, \dots F_N\}$. This leads to a redefinition of the following terms as shown in table (5), and to the following formulations of preference as shown in table (6).

Table 5: Feature-Sensitive Terms

Term
$P_F(\text{A}\|\text{W}) = \frac{1}{\|\|F_1\|\|\cdot\|\|F_2\|\|\cdot\ldots\cdot\|\|F_N\|\|} \int_F \int_{\rho(\text{W})} P(\vec{X}\|\text{W},\vec{F})dXdF$
$P_F(\text{A}\|\text{B}) = \frac{1}{\|\|F_1\|\|\cdot\|\|F_2\|\|\cdot\ldots\cdot\|\|F_N\|\|} \int_F \int_{\rho(\text{W})} P(\vec{X}\|\text{B},\vec{F})dXdF$
$P_F^{\text{B}}(\text{A}\|\text{W}) = \frac{1}{\|\|F_1\|\|\cdot\|\|F_2\|\|\cdot\ldots\cdot\|\|F_N\|\|} \int_F \int_{\rho(\text{B})} P(\vec{X}\|\text{B},\vec{F})dXdF$
$P_F^{\text{W}}(\text{A}\|\text{B}) = \frac{1}{\|\|F_1\|\|\cdot\|\|F_2\|\|\cdot\ldots\cdot\|\|F_N\|\|} \int_F \int_{\rho(\text{W})} P(\vec{X}\|\text{B},\vec{F})dXdF$

We can now use this feature-sensitive preference definition on a toy example. The input distribution and the decision space are characterized in table (7) and table (8). The variable E denotes the highest level of education. For simplication of the problem, we consider only two levels of education: Master's and Doctorate.

Table 6: Feature-Sensitive Preference

Name	Metric	Meaning
White Preference	$\text{Pref}_F(W, B) = P_F(\text{A}\|\text{W}) - P_F^W(\text{A}\|\text{B})$	Quantifies how strongly white features are preferred, ignoring the features in F
Black Preference	$\text{Pref}_F(B, W) = P_F(\text{A}\|\text{B}) - P_F^B(\text{A}\|\text{W})$	Quantifies how strongly black features are preferred, ignoring the features in F

Table 7: Input Distribution, conditioned on Race

$P(E = \text{Masters}, Z = Z_1 \| R = \text{W}) = 0$	$P(E = \text{Masters}, Z = Z_1 \| R = \text{B}) = 0.7$
$P(E = \text{Masters} Z = Z_2, \| R = \text{W}) = 0.5$	$P(E = \text{Masters}, Z = Z_2 \| R = \text{B}) = 0$
$P(E = \text{Doctorate}, Z = Z_1 \| R = \text{W}) = 0$	$P(E = \text{Doctorate} Z = Z_1 \| R = \text{B}) = 0.3$
$P(E = \text{Doctorate}, Z = Z_2 \| R = \text{W}) = 0.5$	$P(E = \text{Doctorate}, Z = Z_2 \| R = \text{B}) = 0$

Table 8: Decision Space

$D(E = \text{Masters}, Z = Z_1, R = \text{W}) = 0$	$D(E = \text{Masters}, Z = Z_1, R = \text{B}) = 0$
$D(E = \text{Masters}, Z = Z_2, R = \text{W}) = 0$	$D(E = \text{Masters}, Z = Z_2, R = \text{B}) = 0$
$D(E = \text{Doctorate}, Z = Z_1, R = \text{W}) = 1$	$D(E = \text{Doctorate}, Z = Z_1, R = \text{B}) = 1$
$D(E = \text{Doctorate}, Z = Z_2, R = \text{W}) = 1$	$D(E = \text{Doctorate}, Z = Z_2, R = \text{B}) = 1$

$$
\begin{aligned}
E &= \{\text{Masters}, \text{Doctorate}\} \\
Z &= \{Z_1, Z_2\} \\
\rho(\text{W}) &= \{(\text{Doctorate}\, Z_1), (\text{Doctorate}, Z_2)\} \\
\rho(\text{B}) &= \{(\text{Doctorate}\, Z_1), (\text{Doctorate}, Z_2)\} \\
\rho(\text{W}) \cap \rho(\text{B}) &= \{(\text{Doctorate}\, Z_1), (\text{Doctorate}, Z_2)\}
\end{aligned}
$$

We calculate the model's preference for white features over black, ignoring dif-

ferences in the education distribution:

$$
\begin{aligned}
\underset{\{E\}}{\text{Pref}}(\text{W}, \text{B}) &= P_{\{E\}}(\text{A}|\text{W}) - P^{W}_{\{E\}}(\text{A}|\text{B}) \\
P_{\{E\}}(\text{A}|\text{W}) &= \frac{1}{|E|} \cdot \sum_{e_1 \in E} \sum_{(z,e_2) \in \rho(\text{W})} P(z, e_2|\text{W}, e_1) \\
&= \frac{1}{|E|} \sum_{e_1 \in E} P(\text{Doctorate}, Z_1|\text{W}, e_1) + P(\text{Doctorate}, Z_2|\text{W}, e_1) \\
&= \frac{1}{|E|}(P(\text{Doctorate}, Z_1|\text{W}, \text{Masters}) + P(\text{Doctorate}, Z_2|\text{W}, \text{Masters}) \\
&+ P(\text{Doctorate}, Z_1|\text{W}, \text{Doctorate}) + P(\text{Doctorate}, Z_2|\text{W}, \text{Doctorate})) \\
&= \frac{1}{|E|}\left(P(\text{Doctorate}, Z_1\text{W}, \text{Doctorate}) + P(\text{Doctorate}, Z_2\text{W}, \text{Doctorate})\right) \\
&= \frac{1}{2} \cdot (0 + 1) \\
&= \frac{1}{2} \\
P^{W}_{\{E\}}(\text{A}|\text{B}) &= \frac{1}{|E|} \cdot \sum_{e_1 \in E} \sum_{(z,e_2) \in \rho(\text{W})} P(z, e_2|\text{B}, e_1) \\
&= \frac{1}{|E|} \sum_{e_1 \in E} P(\text{Doctorate}, Z_1|\text{B}, e_1) + P(\text{Doctorate}, Z_2|\text{B}, e_1) \\
&= \frac{1}{|E|}(P(\text{Doctorate}, Z_1|\text{B}, \text{Masters}) + P(\text{Doctorate}, Z_2|\text{B}, \text{Masters}) \\
&+ P(\text{Doctorate}, Z_1|\text{B}, \text{Doctorate}) + P(\text{Doctorate}, Z_2|\text{B}, \text{Doctorate})) \\
&= \frac{1}{|E|}\left(P(\text{Doctorate}, Z_1|\text{B}, \text{Doctorate}) + P(\text{Doctorate}, Z_2|\text{B}, \text{Doctorate})\right) \\
&= \frac{1}{2} \cdot (1 + 0) \\
&= \frac{1}{2} \\
\underset{\{E\}}{\text{Pref}}(W, B) &= \frac{1}{2} - \frac{1}{2} = 0
\end{aligned}
$$

Thus, our metric shows that the model has no preference for white non-educaiton features over black non-education features. A similar analysis will also demonstrate that the model has no preference for black non-education features over white non-education features.

2.4 Latent-Variable Discrimination

A technique often used to make the model "race-blind", for example, is to omit the race from the input feature. This precludes the ability to do race-specific

reachability analysis. However, using Bayes' law, one can still find the entire "accept" decision space. Then, the following probabilities can be calculated:

$$P(\mathrm{W}|A) = \int_{\rho(\mathrm{W})\cap\rho(\mathrm{B})} \frac{P(\vec{X}|\mathrm{W})P(\mathrm{W})}{P(\vec{X})} dX$$
$$P(\mathrm{B}|A) = \int_{\rho(\mathrm{W})\cap\rho(\mathrm{B})} \frac{P(\vec{X}|\mathrm{B})P(\mathrm{B})}{P(\vec{X})} dX$$

One can compare the observed white and black probabilities with the percentage of white and blacks in the input distribution: that is for a "fair" model, we can expect $P(\mathrm{W}|A) \approx P(\mathrm{W})$ and $P(\mathrm{B}|A) \approx P(\mathrm{B})$: although this is somewhat shallow definition of fairness that is not acceptable in all cases. One can imagine training a neural network to detect the presence of a certain disease: if one race is more susceptible to a certain disease than others, we expect $P(R|A) > P(R)$.

3 Integration Across Mixed-Type Probability Distributions

Assume we have a probability distribution defined on $P(X_1, X_2, \ldots X_N)$, where some X_i are discrete and other X_i are continuous. To aggregate across $X = \begin{bmatrix} X_1 & X_2 & \ldots & X_N \end{bmatrix}^T$ we define the following expression: