Counterfactual Fairness Using Principal Stratification

Abstract

The recent years have witnessed an increasing reliance on algorithmic decision-making in every aspect of our daily life. It is important to ensure that such algorithms do not discriminate certain group of individuals with, for example, a particular race, gender, or sexual orientation. Using the concept of principal stratification from causal inference, we propose a new definition of counterfactual fairness for decisions or predictions made by machine learning algorithms. The key idea is that a fair machine learning algorithm should not discriminate among individuals whose outcome of interest is not affected by the protected attribute. Our definition differs from the existing ones of counterfactual fairness in that the target individuals are defined by the potential values of the outcome, which explicitly take into account the influence of the protected attribute on the outcome. Our definition is motivated by a life insurance example, where the insurance company reasonably charges less premium for women than that for men due to the fact that women live longer than men. We connect our definition with existing counterfactual fairness and show how to empirically evaluate the definition. We also provide a post-processing method to adjust an unfair prediction and show that the post-processed prediction is approximately fair. Finally, we illustrate our methodology in a real dataset.

1 INTRODUCTION

Machine learning plays an important role in an increasing number of domains in the big data era. It has changed our daily life in every aspect, including credit scoring, healthcare and education. Traditionally, machine learning prediction typically focuses on the accuracy [Angwin and Larson, 2016, Angwin et al., 2016]. Using massive training data and modern machine learning methods, the algorithm often outperforms human in prediction accuracy. However, researchers have raised the concern that algorithmic prediction may be unfair with respect to certain social groups. In many fields such as credit scoring and hiring, fairness is one of the non-negligible aspects in decision making.

There are many works on the definitions of fairness. Some fairness notions are association-based [Dwork et al., 2012, Feldman et al., 2015, Corbett-Davies et al., 2017, Hardt et al., 2016], defined as the independence or conditional independence between the protected attribute and the machine learning prediction of the outcome of interest. Recently, researchers have proposed several causality-based definitions for fairness [Kusner et al., 2017, Zhang and Bareinboim, 2018, Nabi and Shpitser, 2018, Zhang et al., 2017, Chiappa, 2019, Wang et al., 2019, Coston et al., 2020, Imai and Jiang, 2020], which rely on the hypothetical prediction that would be generated by the machine learning algorithm if the protected attribute were set to a fixed value. The idea is that the intervention on the protected attribute should not change the distribution of the machine learning predictions, with different definitions targeting different causal pathways. Such definitions often target the sensitive features such as gender, race and ethnicity, and rely upon the belief that the outcome is not affected by the protected attribute. Although this belief is reasonable in many studies such as the prediction of recidivism, it might be violated, or at least hard to justify in some other studies. For example, consider that a life insurance company uses machine learning algorithms to predict whether a high or low premium should be charged for each individual. Following the fairness definitions in the literature, a fair prediction with respect to gender should not be affected by the hypothetical intervention on people's gender. However, it is known that women live longer than men, and thus the premium for women should reasonably be lower than that for men. Therefore, the current fairness definitions might not be applicable to scenarios in which the protected attribute does have an effect on the predicted outcome.

The suspicion of the belief becomes more severe if we would like to generalize the definitions of fairness to take into account more attributes. This is pointed out by Chouldechova and Roth [2020], asking "what features should be protected? Should these only be attributes that are sensitive on their own, like race and gender, or might attributes that are innocuous on their own correspond to groups we wish to protect."

We propose a new definition of counterfactual fairness using the concept of principal stratification from the causal inference literature [Frangakis and Rubin, 2002]. Unlike the existing counterfactual fairness definitions, ours takes into account the effect of the protected attribute on the outcome. The idea is that the protected attribute should not affect the machine learning prediction only for the people whose outcome is not changed by the intervention on the protected attribute. For the other people whose outcome is different under different intervention on the attribute, we allow the prediction to be different. The belief that the outcome is not affected by the protected attribute is formalized using principal stratification. As a result, our definition of fairness could be more widely applied to settings where the null effect of the protected attribute on the outcome is controversial.

We compare our definition with other definitions of counterfactual fairness in the literature. Under the assumption that the protected attribute does not change the outcome for all individuals, we show the equivalence relationship between our definition and the counterfactual fairness in [Kusner et al., 2017]. We propose two approaches to evaluate our proposed definitions. We also provide a post-processing method to adjust an unfair prediction and show that the post-processed prediction is approximately fair the sense that it tends to be fair as the sample size goes to infinity. The proposed methods are applied to a real dataset as a demonstration.

2 COUNTERFACTUAL FAIRNESS USING PRINCIPAL STRATIFICATION

Let A_i be the binary protected attribute (e.g., gender and race) and Y_i be the binary outcome of interest. Let S_i be the binary predicted value for Y_i from a machine learning algorithm or a binary decision made by human or machine, possibly depends on and/or related to Y_i. For simplicity, we will refer to S_i as a machine learning prediction. Following the causal inference literature [Neyman, 1923, Imbens and Rubin, 2015], let $Y_i(a)$ and $S_i(a)$ be the potential values of the outcome and the prediction if the protected attribute were set to $A_i = a$. Although some protected attributes are not manipulable, we can consider the intervention on the perceived attribute, similar to Bertrand and Mullainathan [2004]'s study on the effect of gender. The observed Y_i and S_i are given by $Y_i = A_i \cdot Y_i(1) + (1 - A_i) \cdot Y_i(0)$ and $S_i = A_i \cdot S_i(1) + (1 - A_i) \cdot S_i(0)$. We assume that $\{A_i, Y_i(1), Y_i(0), S_i(1), S_i(0)\}_{i=1,\ldots,n}$ are independently and identically sampled from a super-population, and thus the observed $\{A_i, Y_i, S_i\}_{i=1,\ldots,n}$ are also independently and identically distributed. As a result, we suppress the subscript i for simplicity.

The principal stratification is defined as the joint potential values of Y, i.e. $(Y(1), Y(0))$. Unlike the observed outcome, the principal stratum of an individual is not affected by the protected attribute. Each principal stratum characterizes how one's outcome is affected by the protected attribute. For instance, in the life insurance example, let $A = 1$ $(A = 0)$ denote men (women), $Y = 1$ $(Y = 0)$ denote having a lifetime more (less) than 65 years, and $S = 1$ $(S = 0)$ denote the indicator of being charged a high insurance premium. Then, $(Y(1) = 1, Y(0) = 1)$ represents the people who would live more than 65 years regardless of their gender, and $(Y(1) = 0, Y(0) = 1)$ represents the people who would not live more than 65 years if they were a man. Since we cannot simultaneously observe $Y(1)$ and $Y(0)$ for a particular individual, we do not know which principal stratum an individual belongs to.

Intuitively, in the life insurance example, if the life expectancy of an individual is not affected by gender, then the machine suggested premium should not be different if the gender of an individual were altered. This motivates our definition of counterfactual fairness, which is based on the principal stratification.

Definition 1 *A machine learning prediction S is fair with respect to the outcome Y and the protected attribute A if the following equations hold:*

$$\begin{aligned} &\mathrm{pr}(S(1) = 1 \mid Y(1) = Y(0) = 1) \\ &\qquad = \mathrm{pr}(S(0) = 1 \mid Y(1) = Y(0) = 1), \end{aligned} \tag{1}$$

and

$$\begin{aligned} &\mathrm{pr}(S(1) = 1 \mid Y(1) = Y(0) = 0) \\ &\qquad = \mathrm{pr}(S(0) = 1 \mid Y(1) = Y(0) = 0). \end{aligned} \tag{2}$$

Definition 1 requires the specification of both the outcome and the protected attribute. A machine learning prediction might be fair with respect to one protected attribute (outcome) but not fair with respect to another. We can include observed covariates X and the protected attribute A in the conditioning set, e.g.,

$$\begin{aligned} &\mathrm{pr}(S(1) = 1 \mid Y(1) = Y(0) = 1, X, A = 1) \\ &\qquad = \mathrm{pr}(S(0) = 1 \mid Y(1) = Y(0) = 1, X, A = 0). \end{aligned}$$

which means that the machine learning prediction is fair conditional on X and A. For simplicity, we will omit the conditioning on X and A. We note that, although Definition 1 seems to be similar to Definition 1 in Imai and Jiang [2020],

Table 1: Numerical example of Definition 1. Each cell represents a principal stratum $(Y(1), Y(0))$. The two numbers within the cell represent the number of individuals being charged high premium $(S(a) = 1)$ and low premium $(S(a) = 0)$. In this example, the distributions of $S(1)$ and $S(0)$ are the same within the principal strata $(Y(1) = 1, Y(0) = 1)$ and $(Y(1) = 0, Y(0) = 0)$.

		$S(1)$		$S(0)$	
		$Y(0)=1$	$Y(0)=0$	$Y(0)=1$	$Y(0)=0$
$Y(1)=1$	High	50	30	50	70
	Low	150	70	150	30
$Y(1)=0$	High	170	150	30	150
	Low	30	50	170	50

they are completely different, as the conditioning set of Definition 1 is based on the potential outcome $Y(A = a)$ instead of $Y(S = s)$.

In Definition 1, the two principal strata $(Y(1) = 1, Y(0) = 1)$ and $(Y(1) = 0, Y(0) = 0)$, represent the individuals whose outcome is not affected by the protected attribute. Definition 1 also requires $S(1)$ and $S(0)$ to be equally distributed in each of the two basic principal strata with $Y(1) = Y(0)$. The number of such principal strata will increase as the number of values of Y increases. Moreover, when Y is continuous, Definition 1 requires the equality holds in infinitely many basic principal strata. Therefore, to simplify the generalization for non-binary outcomes, we introduce a weaker version of Definition 1.

Definition 2 *A machine learning prediction S is fair with respect to the outcome Y and the protected attribute A if the following condition holds:*

$$\begin{aligned} &\mathrm{pr}(S(1) = 1 \mid Y(1) = Y(0)) \\ &\qquad = \mathrm{pr}(S(0) = 1 \mid Y(1) = Y(0)). \end{aligned} \tag{3}$$

Obviously, Definition 1 implies Definition 2. Table 1 presents a numerical example in the context of life insurance scenario to illustrate Definitions 1 and 2. In this example, $(Y(1) = 1, Y(0) = 1)$ represents the 'healthy' people who would live a life more than 65 years regardless of their gender, and $(Y(1) = 0, Y(0) = 0)$ represents the 'weak' people who would live a life less than 65 years regardless of their gender. Because the gender does not affect the lifetime of the people in these two principal strata, the insurance company would charge similar premiums for them under different genders. This is reflected in Table 1 by

$$\begin{aligned} &\mathrm{pr}(S(1) = 1 \mid Y(1) = Y(0) = 1) \\ &\qquad = \mathrm{pr}(S(0) = 1 \mid Y(1) = Y(0) = 1) = 0.75, \end{aligned}$$

and

$$\begin{aligned} &\mathrm{pr}(S(1) = 1 \mid Y(1) = Y(0) = 0) \\ &\qquad = \mathrm{pr}(S(0) = 1 \mid Y(1) = Y(0) = 0) = 0.25. \end{aligned}$$

Table 2: The relationship between the observed stratum and principal stratum. Each cell represents one observed stratum defined by (A, Y), which contains two principal strata that are compatible with the observed stratum. For example, in the two principal strata of the first cell, if the observed attribute $A = 1$, then we know the observed outcome $Y = 1$.

	$A = 1$	$A = 0$
$Y = 1$	$(Y(1) = 1, Y(0) = 1)$ $(Y(1) = 1, Y(0) = 0)$	$(Y(1) = 1, Y(0) = 1)$ $(Y(1) = 0, Y(0) = 1)$
$Y = 0$	$(Y(1) = 0, Y(0) = 1)$ $(Y(1) = 0, Y(0) = 0)$	$(Y(1) = 1, Y(0) = 0)$ $(Y(1) = 0, Y(0) = 0)$

As a result, this example satisfies both Definitions 1 and 2. It is important to note that Definitions 1 and 2 have no requirement on the other two principal strata. For example, people in the stratum $(Y(1) = 0, Y(0) = 1)$ would not live as long as they would if their gender were male, possibly due to biological or behavioral reasons. Therefore, if the insurance company knows the gender information to be male, then it will be more likely for the company to charge high premium for people in this stratum, e.g. $\mathrm{pr}(S(1) = 1 \mid Y(1) = 0, Y(0) = 1) = 0.7$ in Table 1. In contrast, if the insurance company knows the gender information to be female, then it will be less likely for the company to charge high premium, e.g. $\mathrm{pr}(S(0) = 1 \mid Y(1) = 0, Y(0) = 1) = 0.3$ in Table 1. Similar argument applies to principal stratum $(Y(1) = 1, Y(0) = 0)$. Therefore, the distributions of $S(1)$ and $S(0)$ are different in these two principal strata.

It is important to note that conditioning on the principal stratum is different from conditioning on the observed outcome and the protected attribute. Table 2 illustrates the relationship between the observed stratum defined by (A, Y) and the principal stratum. In the table, each observed stratum is a mixture of two principal strata. Because Definition 1 only requires $S(1)$ and $S(0)$ are equally distributed within two of the four principal strata, it does not imply that they are equally distributed within each observed stratum. In other words, it is possible that $\mathrm{pr}(S(1) \mid A, Y) \neq \mathrm{pr}(S(0) \mid A, Y)$ for S that satisfies Definition 1.

Similar to Lemma 1 in Kusner et al. [2017], we have the following implication.

Lemma 1 *If the protected attribute A has no individual causal effect on S, then S is fair with respect to both Definitions 1 and 2.*

Indeed, the condition of Lemma 1 is a strong guarantee of fairness. In fact, we can use this lemma to make fair decisions. However, the converse of Lemma 1 is not true. The evaluation of fairness will be discussed in Section 3.

To end this section, we compare Definitions 1 and 2 with the counterfactual fairness and counterfactual equalized odd.

The counterfactual fairness in Kusner et al. [2017] requires the distribution of the potential value of the decision to be the same under different values of the protected attribute.

Definition 3 (Counterfactual Fairness) *A machine learning prediction S satisfies the counterfactual fairness with respect to the protected attribute A if the following condition holds:*

$$\mathrm{pr}(S(1)=1)=\mathrm{pr}(S(0)=1). \tag{4}$$

We omit the conditioning on X and A in the original definition of Kusner et al. [2017]. In the life insurance example, Definition 3 means that the machine predicted premium for men and women should be equally distributed. Unlike Definitions 1 and 2, it does not require the specification of the outcome of interest. The main difference between Definition 3 and Definition 1 (Definition 2) lies in the target population. Definition 3 focuses on the people in the whole population regardless of their outcome, while Definition 1 focuses only on the people whose outcome is not affected by the protected attribute. From the example in Table 1, we can obtain

$$\mathrm{pr}(S(1)=1)=4/7, \quad \mathrm{pr}(S(1)=1)=3/7.$$

Therefore, although the example satisfies Definition 1, it fails to meet the criterion of counterfactual fairness. The following theorem establishes the equivalence relationship between Definition 2 and Definition 3.

Theorem 1 *If $Y_i(1)=Y_i(0)$ for all i, then Definition 2 is equivalent to Definition 3.*

The condition $Y_i(1)=Y_i(0)$ eliminates the principal strata $(Y(1)=1, Y(0)=0)$ and $(Y(1)=0, Y(0)=1)$. It means that the protected attribute does change the outcome for any individual, i.e., the individual causal effect of the protected attribute on the outcome is zero. Although one may think this condition is reasonable in scenarios where fairness is of concern, it is too restrictive to assume it holds for all individuals, as in the life insurance example. Therefore, our definitions of counterfactual fairness focus only on the individuals that satisfy this condition. We view $Y(1)=Y(0)$ as the belief that underlies counterfactual fairness — there is no reason to require a machine learning prediction to be fair with respect to a protected attribute if the protected attributed can influence the target outcome. Therefore, our proposed definitions can be more widely applied to settings when $Y(1)=Y(0)$ might not hold for all the individuals. Moreover, our definitions answer the question, "what features should be protected?" posed by Chouldechova and Roth [2020] — only the attribute that satisfies $Y(1)=Y(0)$ should be protected.

Another similar notion is counterfactual equalized odd [Coston et al., 2020]. Because the potential outcome $Y(0)$ is a baseline covariate variable, counterfactual equalized odd can be regarded as a special case of counterfactual fairness. Moreover, as shown in Table 3, counterfactual equalized odd conditions on groups **B** and **D**, while Definition 1 conditions on groups **A** and **D** separately, and Definition 2 conditions on groups **A** and **D**.

Definition 4 (Counterfactual Equalized Odd) *A machine learning prediction S satisfies the counterfactual equalized odd with respect to the protected attribute A if the following condition holds:*

$$\mathrm{pr}(S(1)=1 \mid Y(0)=1)=\mathrm{pr}(S(0)=1 \mid Y(0)=1). \tag{5}$$

Table 3: The four principal strata

	$Y(0)=0$	$Y(0)=1$
$Y(1)=0$	**A**	**B**
$Y(1)=1$	**C**	**D**

3 EVALUATION AND ESTIMATION

3.1 STATISTICAL BOUNDS

We consider the empirical evaluation of our definitions of fairness in this section. We begin by introducing the unfoundedness assumption.

Assumption 1 *There exist a set of covariates X, such that $A \perp \{Y(1), Y(0), S(1), S(0)\} \mid X$.*

Assumption 1 requires X to include all the common causes for the protected attribute and the outcome of interest (machine learning prediction). It hold automatically in randomized experiments of A, such as that in Bertrand and Mullainathan [2004]. However, in most cases, it is hard to conduct randomized experiments for the protected attribute. As a result, we need to collect sufficient covariates to ensure Assumption 1.

Evaluating Definitions 1 and 2 requires the identification of $\mathrm{pr}(S(a) \mid Y(1), Y(0))$ for $a=0,1$. However, since the principal strata are not directly observable, these probabilities are not identifiable even with Assumption 1. Therefore, we propose an approach to detect unfairness from the observed data, without relying on restrictive identification assumptions. For simplicity, we omit X in Assumption 1 and implicitly condition on X in the following discussion.

We focus on Definition 1 first and consider the bounds on

$$\begin{aligned} &\mathrm{pr}(S(1)=1 \mid Y(1)=Y(0)=1) \\ &\qquad -\mathrm{pr}(S(0)=1 \mid Y(1)=Y(0)=1), \end{aligned} \tag{6}$$

$$\begin{aligned} &\mathrm{pr}(S(1)=1 \mid Y(1)=Y(0)=0) \\ &\qquad -\mathrm{pr}(S(0)=1 \mid Y(1)=Y(0)=0). \end{aligned} \tag{7}$$

If the bounds on (6) or (7) do not cover zero, then we conclude that Definition 1 is violated. Under Assumption 1, we can write (6) and (7) as

$$\begin{aligned}\tau_1 =& \mathrm{pr}(S=1 \mid A=1, Y(1)=Y(0)=1) \\ &- \mathrm{pr}(S=1 \mid A=0, Y(1)=Y(0)=1), \\ \tau_0 =& \mathrm{pr}(S=1 \mid A=1, Y(1)=Y(0)=0) \\ &- \mathrm{pr}(S=1 \mid A=0, Y(1)=Y(0)=0).\end{aligned}$$

Denote $p_{ay} = \mathrm{pr}(Y=y \mid A=a)$ and $q_{ay} = \mathrm{pr}(S=1 \mid A=a, Y=y)$, which can be calculated from the observed data. The following proposition gives the sharp bounds on τ_1 and τ_0.

Proposition 1 *Under Assumption 1, the sharp bounds on τ_1 are*

$$\begin{aligned}&\mathrm{Lower}(\tau_1) \\ &= \max\left\{0, 1-\frac{(1-q_{11})p_{11}}{p_{01}-p_{10}}\right\} - \min\left\{1, \frac{q_{01}p_{01}}{p_{01}-p_{10}}\right\}, \\ &\mathrm{Upper}(\tau_1) \\ &= \min\left\{1, \frac{q_{11}p_{11}}{p_{01}-p_{10}}\right\} + \min\left\{0, \frac{(1-q_{01})p_{01}}{p_{01}-p_{10}} - 1\right\}.\end{aligned}$$

The sharp upper and lower bounds on τ_0 are

$$\begin{aligned}&\mathrm{Lower}(\tau_0) \\ &= \max\left\{0, 1-\frac{(1-q_{10})p_{10}}{p_{10}-p_{01}}\right\} - \min\left\{1, \frac{q_{00}p_{00}}{p_{10}-p_{01}}\right\}, \\ &\mathrm{Upper}(\tau_0) \\ &= \min\left\{1, \frac{q_{10}p_{10}}{p_{10}-p_{01}}\right\} + \min\left\{0, \frac{(1-q_{00})p_{00}}{p_{10}-p_{01}} - 1\right\}.\end{aligned}$$

The proof of Proposition 1 follows directly from Proposition A.4 in Jiang et al. [2016]. These bounds are sharp in the sense that they are attainable. Therefore, they are the narrowest bounds on τ_1 and τ_0 given the observed distribution. In Proposition 1, if $p_{01} > p_{10}$, then the bounds on τ_0 are $[-1,1]$; if $p_{10} > p_{01}$, then the bounds on τ_1 are $[-1,1]$. As a result, the bounds are informative for at most one of τ_1 and τ_0. Based on Proposition 1, we can obtain the conditions under which Definition 1 is violated.

Corollary 1 *Under Assumption 1, the equality in* (1) *is violated if either of the following two inequalities hold:*

$$\begin{aligned} q_{01}p_{01} + (1-q_{11})p_{11} &< p_{01} - p_{10}, \\ q_{11}p_{11} + (1-q_{01})p_{01} &< p_{01} - p_{10}.\end{aligned}$$

The equality in (2) *is violated if either of the following two inequalities hold:*

$$\begin{aligned} q_{00}p_{00} + (1-q_{10})p_{10} &< p_{10} - p_{01}, \\ q_{10}p_{10} + (1-q_{00})p_{00} &< p_{10} - p_{01}.\end{aligned}$$

If either of the four inequalities in Corollary 1 fails, then we can conclude that Definition 1 is violated.

3.2 EVALUATION USING OPTIMIZATION

Although the bounds in Proposition 1 are sharp for τ_1 and τ_0 separately, they may not be sharp for them simultaneously. That is, we do not know whether the bounds on τ_1 and τ_0 are attainable at the same time. As a result, we might lose power by using the bounds in Theorem 1 for detecting unfairness. Below, we propose an alternative approach to evaluate Definition 1.

Denote $w_{s_0 s_1 y_0 y_1} = \mathrm{pr}(S(0)=s_0, S(1)=s_1, Y(0)=y_0, Y(1)=y_1)$ and $p'_{sya} = \mathrm{pr}(S=s, Y=y|A=a)$. We can write τ_1 and τ_0 as

$$\begin{aligned}\tau_1 &= \frac{\tau_1^{'}}{w_{1111}+w_{1011}+w_{0111}+w_{0011}}, \\ \tau_0 &= \frac{\tau_0^{'}}{w_{1100}+w_{1000}+w_{0100}+w_{0000}},\end{aligned}$$

where $\tau_1^{'} = w_{0111} - w_{1011}$ and $\tau_0^{'} = w_{0100} - w_{1000}$. Under Assumption 1, we have

$$\begin{aligned} p'_{sy0} &= \sum_{a,b} w_{sayb}, \\ p'_{sy1} &= \sum_{a,b} w_{asby},\end{aligned}$$

for $s, y = 0, 1$. The terms on the left-hand side of the above equations can be calculated from the observed data. Therefore, the above equations impose constraints on w's. In addition, because w's are probabilities, they need to be within $[0,1]$. Therefore, we obtain the following constraints on w's:

$$\left\{\begin{array}{rcl} \sum_{a,b} w_{s_0,a,y_0,b} &=& p'_{s_0 y_0 0}, \\ \sum_{a,b} w_{a,s_1,b,y_1} &=& p'_{s_1 y_1 1}, \\ w_{s_0 s_1 y_0 y_1} &\geq& 0, \end{array}\right. \tag{8}$$

for $s_0, s_1, y_0, y_1 = 0, 1$. Because the signs of τ_1 and τ_0 are the same as τ_1' and τ_0' respectively, we can calculate the bounds on τ_1' and τ_0' under the constraints in (8) to evaluate Definition 1.

Note that, we can include the constraint $\tau_1' = 0$ when calculating the bounds of τ_0', thereby avoiding the problem in Proposition 1 that the bounds might not be attainable simultaneously.

Similarly, when we evaluation Definition 2, we only need to calculate the bound of $\tau = w_{1011} + w_{1000} - w_{1000} - w_{1011}$ or include a new constraint $\tau = 0$ in (8). We present a numerical example to illustrate the evaluation of Definition 2. Suppose that the estimation of $\mathrm{pr}(S, Y \mid A)$, which is denoted by $\hat{\mathrm{pr}}(S, Y \mid A)$, is presented in Table 4. Let $\mathscr{D} \subset \mathbb{R}^{16}$ be the feasible region determined by the following constraints:

$$\left\{\begin{array}{l} w_{s0y0} + w_{s0y1} + w_{s1y0} + w_{s1y1} = \hat{\mathrm{pr}}(S=s, Y=y \mid A=0), \\ w_{0s0y} + w_{0s1y} + w_{1s0y} + w_{1s1y} = \hat{\mathrm{pr}}(S=s, Y=y \mid A=1), \\ w_{1011} + w_{1000} - w_{1000} - w_{1011} = 0, \\ w_{abcd} \geq 0, \text{ for } a, b, c, d, s, y \in \{0, 1\}. \end{array}\right.$$

Table 4: Numerical example of the evaluation of Definition 2. The values of the conditional probability $\hat{\text{pr}}(S,Y \mid A)$ are given in the table.

	$A=0$		$A=1$	
	$Y=0$	$Y=1$	$Y=0$	$Y=1$
$S=0$	0.53	0.04	0.90	0.02
$S=1$	0.11	0.32	0.03	0.05

Then, one can verify that $\mathscr{D}=\emptyset$, as the upper bound on $w_{1011}+w_{1000}+w_{1000}+w_{1011}$ is strictly below zero assuming other constraints hold. Thus, Definition 2 is violated.

3.3 POST-PROCESSING METHOD AND ESTIMATION

In applications, if we have already obtained an unfair prediction, then it is efficient to obtain a fair prediction by adjusting the unfair one. Such an approach for obtaining a fair prediction is usually called post-processing method [Kim et al., 2019, Hardt et al., 2016]. The advantage of the post-processing method is that it can be applied to models that are already in use but evaluated unfair.

Inspired by Mishler and Kennedy [2020], if a prediction is evaluated unfair, we can balance it through an intervention. Consider a set of non-negative parameters $\varepsilon=\{\varepsilon_{00},\varepsilon_{01},\varepsilon_{10},\varepsilon_{11}\}$, where each parameter ε_{ab} represents probability that force $S=b$ of an individual with $A=a$. It is clear that $\varepsilon_{ab}+\varepsilon_{a(1-b)}\leq 1$. Denote by S' the final decision after the post-processing, then the probability of $(S'=1,Y=y)$ can be derived by:

$$\begin{aligned}
&\text{pr}(S'=s,Y=y \mid A=a)\\
=\;&\varepsilon_{as}\cdot\text{pr}(Y=y \mid A=a)\\
&+(1-\varepsilon_{a0}-\varepsilon_{a1})\cdot\text{pr}(S=s,Y=y \mid A=a)\\
=\;&\varepsilon_{as}(p'_{0ya}+p'_{1ya})+(1-\varepsilon_{a0}-\varepsilon_{a1})p'_{sya}\\
=\;&\varepsilon_{as}p'_{(1-s)ya}+(1-\varepsilon_{a(1-s)})p'_{sya}
\end{aligned}$$

Our goal is to obtain a fair prediction S' while changing the original prediction S as little as possible. Thus, we propose the following programming problem:

$$\begin{array}{ll}
& \varepsilon^*=\arg\min_{\varepsilon}\sum_{a,b}\varepsilon_{ab},\\
\text{s.t.} & \sum w_{s_0,a,y_0,b}=\varepsilon_{0s_0}p'_{(1-s_0)y_00}+(1-\varepsilon_{0(1-s_0)})p'_{s_0y_00},\\
& \sum w_{a,s_1,b,y_1}=\varepsilon_{1s_1}p'_{(1-s_1)y_11}+(1-\varepsilon_{1(1-s_1)})p'_{s_1y_11},\\
& f(w)=0,\\
& w_{s_0s_1y_0y_1}\geq 0,\\
& \varepsilon_{ab}\geq 0,\\
& \varepsilon_{ab}+\varepsilon_{a(1-b)}\leq 1,\\
& \text{for } a,b,s_0,s_1,y_0,y_1\in\{0,1\}.
\end{array} \tag{9}$$

Here we use w to denote $\{w_{abcd}\}$, and use $f(w)$ to denote the fairness constraint: for Definition 1, $f(w)=0$ means

$$w_{0100}-w_{1000}=0 \quad , \quad w_{0111}-w_{1011}=0,$$

and for Definition 2, $f(w)=0$ means

$$w_{0100}+w_{0111}-w_{1000}-w_{1011}=0.$$

If the solution $\varepsilon^*=0$, then S satisfies the corresponding definition of fairness. The larger $\sum_{a,b}\varepsilon^*_{ab}$ is, the more unfair S is. Hence, ε^* is not only a post-processing parameter, but also a measure of unfairness. In particular, $\sum_{a,b}\varepsilon^*_{ab}=0$ indicates the prediction S is fair corresponding to the evaluation method in section 3.2.

In order to solve the above programming problem and obtain ε^*, we need to know the probabilities $\{p'_{sya}\}$. According to the definition of p'_{sya}, this quantity can be estimated from observational data using standard statistical probability estimation techniques. For example, in the discrete case, we can simply use the observed frequencies as estimates for the corresponding probabilities. Therefore, we propose to use $\hat{p}'_{sya}$, which is the estimated probability of p'_{sya}, as an approximation of p'_{sya}. Substituting p'_{sya} by $\hat{p}'_{sya}$ in the original programming problem leads to the following new problem:

$$\begin{array}{ll}
& \hat{\varepsilon}^*=\arg\min_{\varepsilon}\sum_{a,b}\varepsilon_{ab},\\
\text{s.t.} & \sum_{a,b} w_{s_0,a,y_0,b}=\varepsilon_{0s_0}\hat{p}'_{(1-s_0)y_00}+(1-\varepsilon_{0(1-s_0)})\hat{p}'_{s_0y_00},\\
& \sum_{a,b} w_{a,s_1,b,y_1}=\varepsilon_{1s_1}\hat{p}'_{(1-s_1)y_11}+(1-\varepsilon_{1(1-s_1)})\hat{p}'_{s_1y_11},\\
& f(w)=0,\\
& w_{s_0s_1y_0y_1}\geq 0,\\
& \varepsilon_{ab}\geq 0,\\
& \varepsilon_{ab}+\varepsilon_{a(1-b)}\leq 1,\\
& \text{for } a,b,s_0,s_1,y_0,y_1\in\{0,1\}.
\end{array}$$

The following result shows that, if the estimations $\{\hat{p}'_{sya}\}$ are close to $\{p'_{sya}\}$, then the substitution will have little influence on the optimal ε^*.

Theorem 2 *For p_{sya} and its estimator $\hat{p}'_{sya}$, if $||p'_{sya}-\hat{p}'_{sya}||=O_P(1/\sqrt{n})$, where n denotes the sample size, then*

$$||\varepsilon^*-\hat{\varepsilon}^*||=O_P(1/\sqrt{n}).$$

We note that, the requirement that $||p_{sya}-\hat{p}_{sya}||=O_P(1/\sqrt{n})$ in Theorem 2 is easily satisfied with some mild assumptions, see, e.g. Györfi et al. [2006].

After obtaining the solution $\hat{\varepsilon}^*$, we can have a new prediction S'. We intend to show that S' is approximately fair, in the sense that S' tends to be fair as $n\rightarrow\infty$. To this end, consider the following programming problem:

$$\begin{array}{ll}
& \hat{\alpha}^*=\min_{w} g(w),\\
\text{s.t.} & \sum_{a,b} w_{s_0,a,y_0,b}=\hat{\varepsilon}_{0s_0}p'_{(1-s_0)y_00}+(1-\hat{\varepsilon}_{0(1-s_0)})p'_{s_0y_00},\\
& \sum_{a,b} w_{a,s_1,b,y_1}=\hat{\varepsilon}_{1s_1}p'_{(1-s_1)y_11}+(1-\hat{\varepsilon}_{1(1-s_1)})p'_{s_1y_11},\\
& w_{s_0s_1y_0y_1}\geq 0,\\
& \text{for } a,b,s_0,s_1,y_0,y_1\in\{0,1\}.
\end{array}$$

Here we use $g(w)$ to evaluation fairness. For Definition 1,

$$g(w) = (w_{1011} - w_{0111})^2 + (w_{1000} - w_{0100})^2,$$

and for Definition 2,

$$g(w) = (w_{1011} + w_{1000} - w_{0111} - w_{0100})^2.$$

Then we have the following Theorem:

Theorem 3 *For ε^* and its estimator $\hat{\varepsilon}$. If $||p'_{sya} - \hat{p}'_{sya}|| = O_P(1/\sqrt{n})$, we have:*

$$||\hat{\alpha}^*||_1 = O_P(1/\sqrt{n}).$$

Theorem 3 means that, the post-processed prediction is approximately fair.

4 EXPERIMENTS

We illustrate our method in a student performance dataset on students' achievement in secondary education of two Portuguese schools Cortez and Silva [2008], available at `http://archive.ics.uci.edu/ml/datasets/Student+Performance`. This data set contains 33 attributes. Two data files are provided regarding the performance in two distinct subjects: Mathematics (mat-data for short) and Portuguese (por-data for short). The mat-data contain 395 observations and the por-data contain 649 observations. The target outcome Y is the indicator of top performance in mathematics for the mat-data and in Portuguese for the por-data, with $Y = 1$ if the first period grade G1 of the student is in the top 1/3 of the cohort and $Y = 0$ otherwise. We use naive Bayesian classifiers to predict the outcome in both data files. Our goal is to evaluate the fairness of the naive Bayesian classifiers with respect to three protected attributes.

The experiments are run on a computer with 2.50GHz Intel i5 CPU and 8GB memory. All algorithms are implemented using `R 3.6.0`. The Bayesian classifiers are trained using R package `e1071`, the linear programming problems are solved using R package `Rglpk`.

Pre-processing and modeling Although there are 33 variables in the data sets, we only use 28 variables when predicting G1. The dropped variables are school (students' school), failures (number of past class failures), G2 (second period grade), and G3 (final year grade). Variable school was removed since the observations are unbalanced between the two levels of variable school; variable failures was removed since variable failures contains too many levels; variable G2 and G3 were removed since logically they are consequences of G1. We dichotomize the variables in the data set, with detailed rules shown in the supplementary material. After pre-processing, we use naive Bayesian classifiers to predict the outcome in both data files. The prediction accuracy is 83.03% in mat-data and 80.43% in por-data.

Evaluation of fairness We consider three binary attributes,

- gender: $A = 1$ for male and $A = 0$ for female;
- address: $A = 1$ if student lives in an urban area and $A = 0$ if student lives in a rural area;
- higher: $A = 1$ if student wants to take higher education and $A = 0$ otherwise.

For gender, we invoke Assumption 1 with $X = \emptyset$. It is plausible because gender is determined by genotypes randomly passed from parents to offsprings during meiosis. For address, we invoke Assumption 1 with $X =$ gender; for higher, we invoke Assumption 1 with $X = \{\text{gender}, \text{address}\}$. The unfoundedness assumption might be violated for address and higher. For example, the outcome and address might be both affected by social and economic status of the student's family. However, we do not have these variables in the dataset. Therefore, we control for the selected covariates and use the experiments for illustration purposes only.

We evaluate Definitions 1 and 2 for the three attributes by calculating the bounds for τ'_0, τ'_1 and τ' using linear programming. When evaluating the bounds of τ'_0, we use constraints defined by Equation (8), for evaluating the bounds of τ'_1, we also add an additional constraint $\tau'_0 = 0$. For address and higher, we calculate the bounds within each subgroup defined by the covariates. Thus, we evaluate Definitions 1 and 2 for these two attributes within each subgroup.

Table 5 presents the result. In the mat-data, all the bounds with respect to gender and address cover zero. Therefore, the data do not provide evidence of the violation of Definitions 1 and 2 in the mat-data. In the por-data, the bounds with respect to gender cover zero, indicating no evidence of the violation of Definitions 1 and 2 with respect to the math performance. For address, the bounds on τ'_1 and τ' do not cover zero for both males and females. Therefore, we can conclude that the naive Bayesian classifier discriminates among the students living in different areas with respect to Portuguese performance. For higher, the bounds on τ'_1 and τ' do not cover zero in all the four subgroups. Therefore, we can conclude that the naive Bayesian classifier discriminates among the students with different career tendencies with respect to Portuguese performance.

Moreover, The sign of the bounds could inform us of the direction of the discrimination. In the por-data, the bounds on τ'_1 and τ' are all positive except for gender attribute. And we take the attribute address for example to explain the implication. The bounds on τ' with respect to address are positive, implying $\text{pr}\{S(1)|Y(1) = Y(0)\} > \text{pr}\{S(0)|Y(1) = Y(0)\}$. This means that for students whose performance is not influenced by whether living in an urban or rural area, the Bayesian classifier tends to predict a higher grade for students living in an urban area. Similar explanations could

Table 5: Evaluation of Definitions 1 and 2 for the naive Bayesian classifier in the mat-data and the por-data. The bounds on τ_0' and τ_1' are in the fourth and fifth columns, and the bounds on τ' are in the sixth column.

Subject	Attribute	Subgroup	τ_0'	τ_1'	τ'
Mathematics	gender	-	[-0.160 0.134]	[-0.091, 0.101]	[-0.251 0.235]
	address	female	[-0.146 0.182]	[-0.091, 0.136]	[-0.238, 0.318]
		male	[-0.126, 0.133]	[-0.077, 0.136]	[-0.199, 0.269]
	higher	female	[0.000, 0.211]	[0.000, 0.039]	[0.000, 0.461]
		male	[-0.062, 0.140]	[-0.021, 0.297]	[-0.021, 0.453]
		rural	[0.000, 0.159]	[0.000, 0.175]	[0.000, 0.492]
		urban	[-0.071, 0.184]	[-0.071, 0.244]	[-0.090, 0.470]
Portuguese	gender	-	[-0.052, 0.068]	[-0.135, 0.060]	[-0.188, 0.060]
	address	female	[-0.053, 0.042]	**[0.004, 0.133]**	**[0.004, 0.175]**
		male	[-0.052, 0.074]	**[0.048, 0.286]**	**[0.048, 0.360]**
	higher	female	[-0.029, 0.055]	**[0.231, 0.371]**	**[0.231, 0.426]**
		male	[-0.065, 0.065]	**[0.053, 0.294]**	**[0.053, 0.359]**
		rural	[0.000, 0.077]	**[0.085, 0.429]**	**[0.085, 0.505]**
		urban	[-0.034, 0.051]	**[0.154, 0.268]**	**[0.154, 0.319]**

Table 6: Post-processing result of the naive Bayesian classifier for τ' in the por-data.

Subject	Attribute	Subgroup	ε_{00}	ε_{01}	ε_{10}	ε_{11}
Mathematics	gender	-	0	0	0	0
	address	female	0	0	0	0
		male	0	0	0	0
	higher	female	0	0	0	0
		male	0	0	0	0
		rural	0	0	0	0
		urban	0	0	0	0
Portuguese	gender	-	0	0	0.000	0
	address	female	0	0	**0.005**	0
		male	0	0	**0.062**	0
	higher	female	0	0	**0.268**	0
		male	0	0	**0.070**	0
		rural	0	0	**0.130**	0
		urban	0	0	**0.174**	0

be made for the bounds on τ_1' and τ' with respect to attribute address and higher .

We post-process the naive Bayesian classifier in both the mat-data and the por-data with definition 2, and the post-processing results are shown in Table 6. As the evaluation shows that the data do not provide evidence of the violation of Definitions 1 and 2 in the mat-data, the post-processing also shows that we need not to change the result of the Bayesian classifier. While in the por-data, the evaluation suggests that there exists strong unfairness. Therefore, As we can see, the post-processing suggests that we only change the Bayesian classifier to 0 with probability ε_{10} when the attribute equals 1, which can eliminate the positive effect of the corresponding attribute, this coincide with the explanation for the result of the fairness evaluation above.

5 CONCLUSION

In this paper, we propose a new definition of counterfactual fairness based on principal stratification from causal inference literature. The definition takes into account how the protected attribute affects the outcome of interest. It requires that the protected attribute does not affect the machine algorithm prediction for people whose outcome is not affected by the attribute.

We propose two approaches to evaluate the proposed definition, which are able to conclude the violation of the fairness definitions based on the observed data. Because we can only obtain bounds on the quantities in our definitions, the equalities required by the definitions can only be falsified. That is, if the bounds do not cover zero, we can conclude the definition is violated. However, if the bounds cover zero, we cannot conclude the definition is satisfied.

We only considered the binary protected attribute and outcome. For other types of attributes and outcomes, the proposed definitions will become more complicated, as the number of principal strata increases drastically with the number of the values of the attribute and the outcome. Extension of the proposed methodology to the case of general attributes and outcomes is interesting for future research.

Using principle stratification, we have divided people in four groups in Table 3. As we mentioned in past sections, the definition 1 and definition 2 focus on people in group **A** and **D**. It would be interesting to consider the property in the group **B** and **C**. For example,we can add a constraint $w_{1010} \geq w_{0110}$ to protect the people with $A = 0$. And have similar theoretical property. As far as we know, there are almost no articles discussing such protective property. But we believe it can be widely used in future research.

References

Julia Angwin and Jeff Larson. Bias in criminal risk scores is mathematically inevitable, researchers say. *Propublica, available at: https://goo. gl/S3Gwcn (accessed 5 March 2018)*, 2016.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23(2016):139–159, 2016.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

Silvia Chiappa. Path-specific counterfactual fairness. In *Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. In *Proceedings of the 5th Future Business Technology Conference*, page 5–12, 2008.

Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 582–593, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Constantine Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

Kosuke Imai and Zhichao Jiang. Principal fairness for human and algorithmic decision-making. *arXiv e-prints*, May 2020.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Zhichao Jiang, Peng Ding, and Zhi Geng. Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):829–848, 2016.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Alan Mishler and Edward H Kennedy. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. *arXiv preprint arXiv:2009.02841*, 2020.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

J. Neyman. On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science*, 5:465–480, 1923.

Yixin Wang, Dhanya Sridhar, and David M Blei. Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870*, 2019.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making – the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, , IJCAI-17*, pages 3929–3935, 2017.