Obtaining fairness using optimal transport theory

Philippe Besse, and Jean-Michel Loubes[1]
Institut de Mathématiques de Toulouse

July 10, 2018

1 Introduction

Along the last decade, Machine Learning methods have become more popular to build decision algorithms. Originally meant for recommendation algorithms over the Internet, they are now widely used in a large number of very sensitive areas such as medicine, human ressources with hiring policies, banks and insurance (lending), police, and justice with criminal sentencing, see for instance in [?] [?] or [?] and references therein. The decisions made by what is known referred to as IA have a growing impact on human's life. The whole machinery of these technics relies on the fact that a decision rule can be learnt by looking at a set of labeled examples called the learning sample and then this decision will be applied for the whole population which is assumed to follow the same underlying distribution. So the decision is highly influenced by the choice of the learning set.

In some cases this learning sample may present some bias or discrimination that could possibly be learnt by the algorithm and then propagated to the entire population by automatic decisions and, even worse, providing a mathematical legitimacy for this unfair treatment. When giving algorithms the power to make automatic decisions, the danger may come that the reality may be shaped according to their prediction, thus reinforcing their beliefs in the model which is learnt. Classification algorithms are one particular locus of fairness concerns since classifiers map individuals to outcomes.

Hence, achieving fair treatment is one of the growing field of interest in machine learning. We refer for instance to [?] or [?] for a recent survey on this topic. For this, several definitions of fairness have been considered. In this paper we focus on the notion of disparate impact for protected variables introduced in [?]. Actually, some variables, such as sex, age or ethnic origin, are potentially sources of unfair treatment since they enable to create information that should not be processed out by the algorithm. Such variables are called in the literature protected variables. An algorithm is called fair with respect to these attributes when its outcome does not allow to make inference on the information they convey. Of course the naive solution of ignoring these attributes when learning the classifier does not ensure this, since the protected variables may be closely correlated with other features enabling a classifier to reconstruct them.

Two solutions have been considered in the machine learning literature. The first one consists in changing the classifier in order to make it not correlated to the protected attribute. We refer for instance to [?] or [] and references therein. Yet changing the way a model is built or explaining how the classifier is chosen may be seen too intrusive for many companies or some may not be able to change the way they build the model. Hence a second solution consists in changing the input data so that predictability of the protected attribute is impossible. The data will be blurred in order to obtain a fair treatment of the protected class. This point of view has been proposed in [?], [?] or [?] for instance.

In the following we first provide a statistical analysis of the Disparate Impact definition and recast some of the ideas developed in [?] to stress the links between fairness, predictability and the distance between the distributions of the variables given the protected attribute.

2 Fairness using Disparate Impact assessment

Consider the probability space $(\Omega, \mathcal{B}, \mathbb{P})$, with $\mathcal{B}$ the Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geqslant 1$. In this paper, we tackle the problem of forecasting a binary variable $Y : \Omega \to \{0,1\}$, using observed covariates $X : \Omega \to \mathbb{R}^d$, $d \geqslant 1$. We assume moreover that the population can be divided into two categories that represent a bias, modeled by a variable $S : \Omega \to \{0,1\}$. This variable is called the protected attribute, which takes the values $S = 0$ for the "minority"class and supposed to be the unfavored class; and $S = 1$ for the "default", and usually favored class. We also introduce also a notion of positive prediction in the sense that $Y = 1$ represents a success while $Y = 0$ is a failure.
Hence the classification problem aims at predicting a success using the variables X, using a family of binary classifiers $g \in \mathcal{G} : \mathbb{R}^d \to \{0,1\}$. For every $g \in \mathcal{G}$, the outcome of the classification will be the prediction $\hat{Y} = g(X)$. We refer for instance to [?] for a complete description of classification problems in statistical learning.

In this framework, discrimination or unfairness of the classification procedures, appears as soon as the prediction and the protected attribute are too closely related, in the sense that statistical inference on Y may lead to learn the distribution of the protected attribute S. This issue has received lots of interest among the last years and several ways to quantify this *discrimination bias* have been given. We highlight two of them, whose interest depends on the particular problem. More precisely, we can deal with two situations, depending whether the true distribution of the label Y is available. If it is known, the definition introduced in [?], defines that a classifier $g : \mathbb{R}^d \to \{0,1\}$ achieves *Overall Accuracy Equality*, with respect to the joint distribution of (X, S, Y), if

$$\mathbb{P}(g(X) = Y \mid S = 0) = \mathbb{P}(g(X) = Y \mid S = 1). \tag{2.1}$$

This entails that the probability of a correct classification is the same across groups and, hence, the classification error is independent of the group. This idea can be also found in the [?] as the condition of g having *Disparate Mistreatment*, which happens when the probability of error is different for each group as in (2.1).

Nevertheless, in many problems, the true Y is not available (this data may be very sensitive and the owner of the data may not want to make it available), or the classification methodology can not be changed, so the study of fairness must be based on the outcome $\hat{Y}$. In this situation, following [?] or [?], a classifier $g : \mathbb{R}^d \to \{0,1\}$ is said to achieve *Statistical Parity*, with respect to the joint distribution of (X, S), if

$$\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1). \tag{2.2}$$

This means that the probability of a successful outcome is the same across the groups. For instance, if we consider that the protected variable represents gender, the value $S = 0$ would be assigned to "female"and $S = 1$ to "male", we would say that the algorithm used by a company achieves *Statistical Parity* if a man and a woman have the same probability of success (for instance being hired or promoted).
We will use the following notations

$$a(g) := \mathbb{P}(g(X) = 1 \mid S = 0), \quad b(g) := \mathbb{P}(g(X) = 1 \mid S = 1).$$

In the following we consider classifiers g such that $a(g) > 0$ and $b(g) > 0$, which means that the classifier is not totally fair or unfair in the sense that the classifier does not predict the same outcome for a whole population according to the protected attribute.

The independency described in (2.2) is difficult to achieve and may not exist in the real data, hence, to assess this kind of fairness, an index called *Disparate Impact of the classifier g with respect to* (X, S) , has been introduced in [**?**] as

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 \mid S = 0)}{\mathbb{P}(g(X) = 1 \mid S = 1)}. \tag{2.3}$$

The ideal scenario where g achieves *Statistical Parity* is equivalent to $DI(g, X, S) = 1$. Statistical Parity is often unrealistic and is often relaxed into achieving a certain level of fairness as described in the following definition.

Definition 2.1. The classifier $g : \mathbb{R}^d \to \{0, 1\}$ has Disparate Impact at level $\tau \in (0, 1]$, with respect to (X, S), if $DI(g, X, S) \leq \tau$.

Note the Disparate Impact of a classifier measures its level of fairness: the smaller the value of τ, the less fair it is. The classification rules considered in this framework are such that $b(g) \geqslant a(g) > 0$, because we are assuming that the default class $S = 1$ is more likely to have a successful outcome. Thus, in the definition, the level of fairness τ takes values $0 < \tau \leq 1$. We point out that the value $\tau_0 = 0.8 = 4/5$, which is also known in the literature as the 80% *rule* has been cited as a legal score to decide whether the discrimination of the algorithm is acceptable or not (see for instance in [**?**]). This rule can be explained as "for every 5 individuals with successful outcome in the majority class, 4 in the minority class will have a successful outcome too".

In what follows, to promote fairness, it will be useful to state the definition in the reverse sense. A classifier does not have Disparate Impact at level τ, with respect to (X, S), if $DI(g, X, S) > \tau$.

Finally another definition has been given in the statistical literature on fair learning. Given a classifier $g \in \mathcal{G}$, its Balanced Error Rate (BER) with respect to the joint distribution of the random vector (X, S) is defined as the average class-conditional error

$$BER(g, X, S) = \frac{\mathbb{P}\left(g(X) = 0 \mid S = 1\right) + \mathbb{P}\left(g(X) = 1 \mid S = 0\right)}{2}. \tag{2.4}$$

Notice that $BER(g, X, S)$ is the general misclassification error of $g \in \mathcal{G}$ in the particular case when we have $\mathbb{P}(S = 0) = \mathbb{P}(S = 1) = 1/2$, which consists in the ideal situation when both protected classes have the same probability of occurence. This quantity enables to define the notion of $\varepsilon-$predictability of the protected attribute. S is said to be $\varepsilon-$predictable from X if there exists a classifier $g \in \mathcal{G}$ such that

$$BER(g, X, S) \leq \varepsilon.$$

Equivalently, S is said not to be $\varepsilon-$predictable from X if $BER(g, X, S) > \varepsilon$, for all classifier g chosen in the class $\mathcal{G}$. Thus, if the minimum of this quantity is achieved by a classifier g^*,

$$\min_{g \in \mathcal{G}} BER(g, X, S) = BER(g^*, X, S) = \varepsilon^*$$

then it is clear that S is not $\varepsilon-$predictable from X for all $\varepsilon \leq \varepsilon^*$.

In the following, we recast previous notions of fairness and provide a probabilistic framework do highlight the relationships between the distribution of the observations and the fairness of the classification problem.

The following theorem generalizes the result in [**?**], showing the relationship between predictability and Disparate impact.

Theorem 2.1. *Given random variables $X \in \mathbb{R}^d$, $S \in \{0,1\}$, the classifier $g \in \mathcal{G}$ has Disparate Impact at level $\tau \in [0,1]$, with respect to (X,S), if and only if $BER(g,X,S) \leq \frac{1}{2} - \frac{\alpha(g)}{2}(\frac{1}{\tau} - 1)$, where $\alpha(g) = \mathbb{P}(g(X) = 1|S = 0)$.*

The following theorem establishes the relationship between ε^* the minimum Balance Error Rate and distance in Total Variation between the two conditional distributions $\mathcal{L}\left(X|S=0\right)$ and $\mathcal{L}\left(X|S=1\right)$.

Theorem 2.2. *Given the variables $X : \Omega \to \mathbb{R}^d$, $d \geqslant 1$, and $S : \Omega \to \{0,1\}$,*

$$\min_{g \in \mathcal{G}} BER(g,X,S) = \frac{1}{2}\left(1 - d_{TV}\left(\mathcal{L}\left(X|S=0\right), \mathcal{L}\left(X|S=1\right)\right)\right).$$

Hence we can see that fairness expressed through the notion of Disparate Impact depends highly on the conditional distributions of the variables X conditionally to the protected attribute, $\mathcal{L}\left(X|S=0\right)$ and $\mathcal{L}\left(X|S=1\right)$.

Actually Theorem 2.2 implies that S is not $\varepsilon-$predictable from X if, and only if,

$$d_{TV}\left(\mathcal{L}\left(X|S=0\right), \mathcal{L}\left(X|S=1\right)\right) < 1 - 2\varepsilon. \tag{2.5}$$

Hence the smaller the Total Variation distance, the greater is the value ε that we could find satisfying Equation (2.5) and thus, the less predictable using the variables X will be S . The best case happens when $d_{TV}\left(\mathcal{L}\left(X|S=0\right), \mathcal{L}\left(X|S=1\right)\right) = 0$, which is equivalent to the equality of both conditional distributions $\mathcal{L}\left(X|S=0\right) = \mathcal{L}\left(X|S=1\right)$. In this situation, we will have that S is not $\varepsilon-$predictable from X, for all $\varepsilon \leq \frac{1}{2}$, and that X and S are independent random variables. Note that clearly $\varepsilon = 1/2$ non predictability is the best that can be achieved.

3 Removing disparate impact using Optimal Transport

Some classification procedures exhibit a discrimination bias quantified through a potential Disparate Impact in the classification outcome $\hat{Y} = g(X)$, with respect to the joint distribution of (X,S). To get rid of the possible discrimination committed by a classifier g, two main strategies can be used, either modifying the classifiers or modifying the input data. In this work, we are facing the problem where we have no access to the values Y of the learning sample, hence we focus on the methodologies that intend to modify the data in order to achieve fairness.
The main idea is to change the data in order to break their relationship with the protected attribute. This transformation is called repairing the data. We propose to map the conditional distributions to a common distribution in order to achieve statistical parity as described in (2.2). The choice of the common distribution in one dimension is described as the distribution obtained by taking the mean of the quantile functions. A total repair of the data amounts to modify the input variables X building a repaired version, denoted by $\tilde{X}$, such that any classifier g will have Disparate Impact at level $\tau = 1$, with respect to $(\tilde{X}, S)$. This means that every classifier g used to predict the target class Y from the new variable $\tilde{X}$ will achieve *Statistical Parity* with respect to $(\tilde{X}, S)$. As a counterpart, it is clear that the choice of the distribution to whom the

original variables are mapped should convey as much as information possible on the original variable, otherwise it would hamper the accuracy of the new classification. This constraint led some authors to recommend the use of the so-called Wasserstein barycenter.

We now present some statistical justifications for this choice and provide some comments on the way to repair the data to obtain fair enough classification rules without modifying too much the original data set.

Achieving Statistical Parity amounts to modify the original data into a new random variable $\tilde{X}$ such that the conditional distribution with respect to the protected attribute S is the same for all groups, namely

$$\mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right). \tag{3.1}$$

In this case, any classifier g built with such information will be such that

$$\mathcal{L}\left(g(\tilde{X}) \mid S=0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S=1\right),$$

which implies that $DI(g, \tilde{X}, S) = 1$ and so this transformation promotes full fairness of the classification rule.

To achieve this transformation, the solution detailed in many papers is to map both conditional distributions $\mu_0 := \mathcal{L}(X|S=0)$ and $\mu_1 := \mathcal{L}(X|S=1)$ onto a common distribution ν. Actually, the distribution of the original variables X is transformed using a map T_S which depends on the value of the protected attribute S

$$\begin{array}{rccc} T_S: & \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ & X & \longmapsto & \tilde{X} = T_S(X), \end{array}$$

and such that

$$\mathcal{L}\left(T_0(X) \mid S=0\right) = \mathcal{L}\left(T_1(X) \mid S=1\right). \tag{3.2}$$

Note that the function T_S is random because of its dependence on the binary random variable S.

In this framework, the problem of achieving Statistical Parity is the same as the problem of finding a (random) function T_S such that (3.2) holds. As it is represented in Figure 1, if we denote by $\mu_S \sim X \mid S$, our goal is to map these two distributions to a common law $\nu = \mu_S \circ T_S^{-1}$. Consequently, two different problems arise

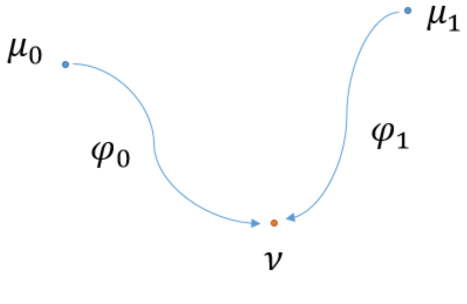

Figure 1

- First of all, the choice of the distribution ν should be as similar as possible to both distributions μ_0 and μ_1 at the same time, in order to reduce the amount of information lost with this transformation and thus still enabling the prediction task using the modified variable $\tilde{X} \sim \nu$ instead of the original X.

- On the other hand, once we have selected the distribution ν, we have to find the optimal way of transporting μ_1 and μ_0 to this new distribution ν.

From Section 2, the natural distance related to fairness between the two conditional distributions is the total variation distance and that should be used. However, this distance is computationally difficult to handle, hence previous works promote the use of Wasserstein distance which appears as a natural distance to move distributions.
In our particular problem, where we have $J = 2$, the two conditional distributions of the random variable X by the protected attribute S are going to be transformed into the distribution of the Wasserstein barycenter μ_B between μ_0 and μ_1, with weights π_0 and π_1, defined as

$$\mu_B \in argmin_{\nu\in\mathcal{P}_2} V_2^2(\mu_0,\mu_1;\pi_0,\pi_1) = argmin_{\nu\in\mathcal{P}_2}\left\{\pi_0 W_2^2(\mu_0,\nu) + \pi_1 W_2^2(\mu_1,\nu)\right\}.$$

Let $\tilde{X}$ be the transformed variable with distribution μ_B. For each $S = s$, the deformation will be performed through the optimal transport map $T_s : \mathbb{R}^d \to \mathbb{R}^d$ pushing each μ_s towards the weighted barycenter μ_B, whose existence is guaranteed as soon as μ_s are absolutely continuous with respect to Lebesgue measure which satisfies

$$\mathbb{E}\left(\|X - T_s(X)\|^2 \mid S = s\right) = W_2^2(\mu_s,\mu_B). \tag{3.3}$$

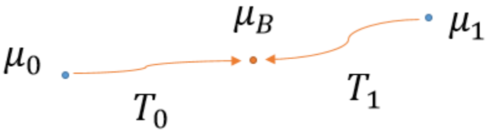

Figure 2

To understand the use of the Wasserstein barycenter distribution as the target distribution for μ_0 and μ_1, we quantify the amount of information lost by replacing the distribution of X by the distribution of $\tilde{X}$ obtained by transporting these two distributions. Set the random transport plan $T_S : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, and the modified variable $\tilde{X} = T_S(X)$. We point out that choosing the distribution of $\tilde{X}$ amounts to choose the transportation plans T_0 and T_1.

Hence the amount of information lost coming from changing the data is given by the following theorem.

Theorem 3.1. *Set $\eta_s(x) = \mathbb{P}(Y = 1 \mid X = x, S = s)$. Consider $X \in \mathbb{R}^d$ and $S \in \{0,1\}$. Let $T_S : \mathbb{R}^d \to \mathbb{R}^d$, $d \geqslant 1$ be a random transformation of X such that $\mathcal{L}(T_0(X) \mid S = 0) = \mathcal{L}(T_1(X) \mid S = 1)$, and consider the transformed version $\tilde{X} = T_S(X)$. For each $s \in \{0,1\}$, assume that the function $\eta_s(X)$ is Lipschitz with constant $K_s > 0$ Then, if $K = \max\{K_0, K_1\}$,*

$$\mathcal{E}(\tilde{X}) \leq 2\sqrt{2}K\left(\sum_{s=0,1}\pi_s W_2^2(\mu_s,\mu_{s\sharp}T_s)\right)^{\frac{1}{2}}. \tag{3.4}$$