A short note on the algorithmic fairness: demographic parity

Yen-Chi Chen
University of Washington
February 24, 2022

Reference:

> [ZDC2022] Xianli Zeng, Edgar Dobriban, Guang Cheng."Bayes-Optimal Classifiers under Group Fairness". arXiv:2202.09724

This short note is a gentle introduction to the paper [ZDC2022], which the authors derive an elegant form of Bayes classifier under the demographic parity.

1 Setup

Let (X,A) be the feature/covariates and A is the protected variable. We consider a binary classification problem where the class label $Y \in \{0,1\}$. For simplicity, we assume that $A \in \{0,1\}$ and the other covariate X is a continuous random vector. In a typical classification problem, we have a new observation (X',A') and we will make a prediction based on the classifier $c(X',A') \in \{0,1\}$. In this case, we may write $\widehat{Y}_c = c(X',A')$.

The *demographic parity (DP)* considers the discrepancy between predicting $\widehat{Y}_c = 1$ under $A=0$ and $A=1$. Specifically, a classifier reached DP if

$$P(\widehat{Y}_c|A=1) = P(\widehat{Y}_c|A=0).$$

In reality, it is pretty restrictive to have the equality constraint. The performance of classification could be very bad. So instead, we consider the difference in demographic parity (DDP)

$$DDP(c) = P(\widehat{Y}|A=1) - P(\widehat{Y}|A=0).$$

As long as DDP is less than some user-specified threshold δ, the classifier is considered as a fair classifier.

When there is no constraint, it is well-known that the Bayes classifier is of the form

$$c^*(x,a) = I\left(m_a(x) \geq \frac{1}{2}\right), \qquad m_a(x) = P(Y=1|X=x,A=a).$$

Note that $m_a(x)$ is also the regression function, i.e., $\mathbb{E}(Y|X=x,A=a) = m_a(x)$.

When the Bayes classifier satisfies $|DDP(c^*)| \leq \delta$, there is nothing we need to do to modify it. However, when $|DDP(c^*)| > \delta$, we have to make some modifications so that the modified classifier is fair. For simplicity, we denote $D^* = DDP(c^*)$.

Suppose $m_A(X)$ is a continuous random variable, [ZDC2022] showed that a Bayes classifier with DDP being controlled at δ is like

$$c^*_\delta(x,a) = I\left(m_a(x) \geq \frac{1}{2} + (2a-1)\cdot\frac{D^*}{|D^*|}\cdot t^*_\delta\right), \tag{1}$$

where t^*_δ is from the following equation:

$$S_1\left(\frac{1}{2}+t^*_\delta\right) = S_0\left(\frac{1}{2}-t^*_\delta\right) + \frac{D^*}{|D^*|}\delta, \quad S_a(\tau) = P(m_a(X) > \tau). \tag{2}$$

In what follows, we will derive how this classifier is constructed. Note that this expression is not identical to [ZDC2022], we will explain the difference later.

2 Constructing a classifier with demographic parity

For simplicity, we first consider the case $D^* > 0$ and $|D^*| > \delta$.

2.1 Changing the threshold

The fact that $D^* > 0$ implies that

$$D^* = P(\widehat{Y}_{c^*} = 1|A=1) - P(\widehat{Y}_{c^*} = 1|A=0) > 0 \Rightarrow P(\widehat{Y}_{c^*} = 1|A=1) > P(\widehat{Y}_{c^*} = 1|A=0).$$

Thus, one way to decrease the DDP is the change the threshold in our classifier. Originally, we compare the regression function $m_a(x)$ versus the threshold $\frac{1}{2}$. The DDP $D^* > 0$ implies that this threshold will favor the case of $A = 1$ over $A = 0$. To decrease DDP, we consider

$$I\left(m_1(x) \geq \frac{1}{2}+t\right), \quad I\left(m_0(x) \geq \frac{1}{2}-t\right). \tag{3}$$

Namely, the new classifier will make it easier to predict $\widehat{Y} = 1$ for the case of $A = 0$. When $t = 0$, we recover the original Bayes classifier. On the extreme case $t \to \infty$, $P(\widehat{Y} = 1|A=0) \to 1$ while $P(\widehat{Y} = 1|A=1) \to 0$, so we can increase t to decrease the DDP.

Note that the classifier of the form of equation (3) is from a *symmetric deviation approach*: we decrease the threshold of $A = 1$ and increase the threshold of $A = 0$ by the same amount.

We denote $c^*_t(x,a)$ to be the classifier

$$c^*_t(x,a) = I\left(m_1(x) \geq \frac{1}{2}+(2a-1)t\right). \tag{4}$$

Equation (4) is equivalent to (3). We further denote $D(t) = DDP(c^*_t)$ to be the DDP of the classifier c^*_t.

With the above analysis, one can easily see that $D(t)$ is a non-increasing function with respect to t and $D(0) = D^*$ is the DDP under the Bayes classifier.

2.2 Feasible range of t

Given that $D(t)$ is a decreasing function with respect to t and we have $D(0) = D^* > \delta$ and $D(\infty) = -1$. The feasible region

$$\mathcal{T}_\delta = \{t > 0 : D(t) \in [0,\delta]\} = [a^*_\delta, b^*], \tag{5}$$

where b^* is from solving (case of DDP=0)

$$S_1\left(\frac{1}{2}+b^*\right)=S_0\left(\frac{1}{2}-b^*\right)$$

and a^* is from solving (case of DDP=δ)

$$S_1\left(\frac{1}{2}+a^*_\delta\right)=S_0\left(\frac{1}{2}-a^*_\delta\right)+\delta, \tag{6}$$

where $S_a(\tau)=P(m_a(X)>\tau)$. You may already notice the similarity between equations (6) and (2).

2.3 Risk minimization

To construct the Bayes classifier, we need to consider risk minimization. For the classifier c^*_t, clearly the risk we are taking compared to the original Bayes classifier will be increasing when t increases. So we want to choose t as small as possible. Therefore, we will choose the minimal value of $t\in\mathcal{T}_\delta$. Thus, the optimal choice will be $t^*_\delta=a^*_\delta$ from solving the equation (6).

2.4 The symmetric argument

Finally, note that all the above analysis assumes that $D^*>0$. When $D^*<0$, we will swap case $A=0$ and $A=1$ and consider the classifier

$$c^*_t(x,a)=I\left(m_1(x)\geq\frac{1}{2}-(2a-1)t\right)$$

so that $t\geq 0$ is a feasible choice.

Thus, a simple way to obtain a unified form of the classifier regardless of the sign of D^* is to make good use of $\frac{D^*}{|D^*|}$, in which it will be $+1$ or -1 depending on the situation we are considering. As a result, we modify equation (6) to

$$S_1\left(\frac{1}{2}+t^*_\delta\right)=S_0\left(\frac{1}{2}-t^*_\delta\right)+\frac{D^*}{|D^*|}\delta$$

and choose t^*_δ by the above equation. The final classifier will then be

$$c^*_\delta(x,a)=I\left(m_1(x)\geq\frac{1}{2}+(2a-1)\cdot\frac{D^*}{|D^*|}\cdot t^*_\delta\right),$$

which is equation (2).

3 The optimal classifier of [ZDC2022]

In [ZDC2022], their optimal classifier is slightly different from equation (1). Instead of using equation (3), they consider the classifiers of the form

$$I\left(m_1(x)\geq\frac{1}{2}+\frac{t}{p_1}\right),\quad I\left(m_0(x)\geq\frac{1}{2}-\frac{t}{p_0}\right), \tag{7}$$

where $p_1 = P(A=1)$ and $p_0 = P(A=0)$.

For classifiers of the above form, the optimal t will be chosen from

$$S_1\left(\frac{1}{2}+\frac{t_\delta^\dagger}{p_1}\right) = S_0\left(\frac{1}{2}-\frac{t_\delta^\dagger}{p_0}\right)+\delta$$

and the optimal classifier is

$$c_\delta^*(x,a) = I\left(m_a(x) \geq \frac{1}{2}+(2a-1)\cdot\frac{D^*}{|D^*|}\cdot\frac{t_\delta^\dagger}{p_a}\right) \tag{8}$$

by the same argument as the previous secion.

4 The real Bayes classifier

While equations (1) and (8) provide Bayes classifiers under certain classes, they may not be the Bayes classifier among all classes.

To derive the real Bayes classifier, we should consider all classifiers of the form (in contrast to equations (3) and (7))

$$I\left(m_1(x) \geq \frac{1}{2}+t_1\right), \quad I\left(m_0(x) \geq \frac{1}{2}-t_0\right), \tag{9}$$

where both t_0, t_1 can change independently from each other. We may express the classifier $c_{t_0,t_1}(x,a)$ as

$$c_{t_0,t_1}(x,a) = aI\left(m_1(x) \geq \frac{1}{2}+t_1\right)+(1-a)I\left(m_0(x) \geq \frac{1}{2}-t_0\right).$$

In the continuous case (of $\eta_1(X)$ and $\eta_0(X)$), controlling DDP to be δ will introduce the constraint

$$S_1\left(\frac{1}{2}+t_1\right) = S_0\left(\frac{1}{2}-t_0\right)+\delta.$$

Thus, the feasible set

$$\left\{(t_0,t_1): \quad S_1\left(\frac{1}{2}+t_1\right) = S_0\left(\frac{1}{2}-t_0\right)+\delta\right\}$$

forms a one-dimensional manifold and can be indexed by either t_0 or t_1. WLOG, we index the feasible choice by t_1 and set

$$t_0^*(t_1) = S_0^{-1}\left(S_1\left(\frac{1}{2}+t_1\right)-\delta\right)+\frac{1}{2}. \tag{10}$$

The challenge is that when t_1 increases, $t_0^*(t_1)$ decreases, so it was not immediately clear how to optimally balance between the two.

To define the fair Bayes classifier, we analyze the *excess risk* relative to the original Bayes classifier. For any classifier c, the excess risk is

$$R(c) - R(c^*) = R(c) - \min_{c'} R(c'),$$

where c^* is the original Bayes classifier. Because the Bayes classifier can be written as

$$c^*(x,a) = I\left(m_a(x) \geq \frac{1}{2}\right),$$

the excess risk of the classifier c_{t_0,t_1} will be

$$\begin{aligned} R(c_{t_0,t_1}) - R(c^*) &= p_1 \cdot \int 2\left(m_1(x) - \frac{1}{2}\right) I\left(m_1(x) \in \left[\frac{1}{2}, \frac{1}{2} + t_1\right]\right) p(x|A=1)dx \\ &\quad + p_0 \cdot \int 2\left(\frac{1}{2} - m_0(x)\right) I\left(m_0(x) \in \left[\frac{1}{2} - t_0, \frac{1}{2}\right]\right) p(x|A=0)dx \\ &= 2p_1 \int_{w_1=0}^{t_1} w_1 \cdot F_1\left(\frac{1}{2} + dw_1\right) + 2p_0 \int_{w_0=0}^{t_0} w_0 \cdot F_0\left(\frac{1}{2} - dw_0\right) \\ &= 2p_1 \mathbb{E}(W_1 I(0 \leq W_1 \leq t_1)) + 2p_0 \mathbb{E}(W_0 I(0 \leq W_0 \leq t_0)), \end{aligned}$$

where F_1 and F_0 are the CDF of $m_1(X)|A=1$ and $m_0(X)|A=0$ and W_1 is the random variable with the same distribution of $m_1(X) - \frac{1}{2}|A=1$ and W_0 is the random variable with the same distribution of $\frac{1}{2} - m_0(X)|A=0$.

Under the optimal choice with fairness constraint in equation (10), we replace t_0 by $t_0^*(t_1)$, which leads to the following expression of the excess risk

$$\begin{aligned} E(t_1) &= R(c_{t_0^*(t_1),t_1}) - R(c^*) \\ &= 2p_1 \mathbb{E}(W_1 I(0 \leq W_1 \leq t_1)) + 2p_0 \mathbb{E}(W_0 I(0 \leq W_0 \leq t_0^*(t_1))). \end{aligned} \tag{11}$$

The optimal t_1^* should be chosen as

$$t_1^* = \mathsf{argmin}_{t_1 > 0} 2p_1 \mathbb{E}(W_1 I(0 \leq W_1 \leq t_1)) + 2p_0 \mathbb{E}(W_0 I(0 \leq W_0 \leq t_0^*(t_1))) \tag{12}$$

and the fair Bayes classifier will be

$$c^*_{t_0^*(t_1^*),t_1^*}(x,a).$$

Note that this is for the case of $D^* > 0$, by the symmetric argument we can construct a unified fair Bayes classifier for any D^*. In fact, equations (11) and (12) remain the same and the only difference is the random variables W_0 and W_1:

$$\begin{aligned} W_1 &\overset{d}{=} \frac{D^*}{|D^*|}\left(m_1(X) - \frac{1}{2}\right)\Bigg|A=1 &&\overset{d}{=} \begin{cases} m_1(X) - \frac{1}{2}|A=1, & \text{if } D^* > 0 \\ \frac{1}{2} - m_1(X)|A=1, & \text{if } D^* < 0 \end{cases} \\ W_0 &\overset{d}{=} \frac{-D^*}{|D^*|}\left(m_0(X) - \frac{1}{2}\right)\Bigg|A=0 &&\overset{d}{=} \begin{cases} \frac{1}{2} - m_1(X)|A=0, & \text{if } D^* > 0 \\ m_0(X) - \frac{1}{2}|A=0, & \text{if } D^* < 0 \end{cases} \end{aligned}. \tag{13}$$

The technical challenge of obtaining a closed-form of t_1^* is that its closed-form depends on the distribution of both $m_1(X)|A=1$ and $m_0(X)|A=0$, it is unclear if we can derive a simple form of the fair Bayes classifier.