Degree of Fairness in Machine Learning Algorithms

Bodhankar Komal
Illinois Institute of Technology
Chicago, Illinois
kbodhankar@hawk.iit.edu

Sarkar Mousam
Illinois Institute of Technology
Chicago, Illinois
msarkar3@hawk.iit.edu

December 2, 2022

Abstract

Decision making is getting affected my fairness. A wide variety of fairness metrics have been proposed in the literature to identify the extent to which the algorithm derives fair results. These fair variables are a set of features chosen by users. Effect of these fair variables is irrelevant in assessing how fair a decision making algorithm is. This report demonstrates the traditional notions of demographic parity and equalized odds as special cases of the conditional fairness. Moreover, the model utilizes the Derivable Conditional Fairness Regularizer (DCFR) and is integrated with a decision-making model, to track the trade-off between precision and fairness of decision-making.

1 Introduction

Machine Learning (ML) is being increasingly applied to decision-making in sensitive domains such as healthcare, crime, justice management, and finance. Although it has the potential to overcome undesirable aspects of human decision-making biases, concern continuous to mount that its opacity is getting reflected in several datasets. Machine Learning field, even now, is in an infant stage while deriving predictions from this unreliable data, resulting in unaccountable predictions proposed by different algorithms. In recent years, attention has been focused on how predictive models may be biased - a now overloaded word that, in popular media, has come to mean that the model's performance (however defined) unjustifiably differs along social axes such as race, gender, and class. Therefore impacting various human lives. To address this issue, the measure of fairness and various fairness notions of an algorithm are taken into consideration. The existing research considers Derivable Conditional Fairness Regularizer (DCFR) as a metric in the in-processing framework to track the trade-off between precision and fairness of major

decision-making algorithms. Additionally, it addresses the paramount importance of fairness in algorithms for both academic and practical research applications.

2 Previous Works

There are several common types of fairness notions including individual fairness, group fairness, and causality-based fairness. The most commonly used individual fairness notion is fairness through awareness. This notion requires that similar individuals should be treated similarly. Since it is difficult to define the similarity function between different individuals, individual fairness still lacks further research today. In group fairness notions, the algorithm treats different groups of individuals equally. The most commonly used group fairness notions are demographic parity, equal opportunity, and equalized odds. These fairness notions only use sensitive attributes and outcomes as measuring features. As a result, these notions may fail to distinguish between fair and unfair parts in the problem. For an in-depth understanding of fairness, causality-based fairness notions were proposed, wherein it first define the causal graph among the features and afterward, the model distinguishes the unfair causal effect caused due to sensitive attributes toward the outcome. However, the aforementioned fairness notions require strong assumptions and are not scalable.

Later in [2013] Kamiran et al. proposed a similar fairness notion. Instead of defining features as sensitive attributes, the author defined variables as explanatory variables and proposed algorithms to mitigate the illegal discrimination they define. However, this method is limited as it may do great harm to accuracy and cannot be applied in practice.

Methods that mitigate biases in the algorithms fall under three categories: pre-processing, in-processing, and post-processing algorithms. Representation learning is a common in-processing method which is first proposed by Zemel et al. [2013]. In this, the author tries to mitigate individual unfairness and demographic discrimination simultaneously. Later Edward and Storkey [2015] first proposed the adversary learning representation method and provided a framework that mitigated demographic discrimination. Several works followed this framework, in particular, Madras et al. [2018] proposed to use of different adversarial loss functions when faced with different fair notions. Zhao et al. [2020] redesigned the loss functions to mitigate the gap of demographic parity and equalized odds simultaneously.

However, these works all focus on the most commonly used group fairness notions. Therefore they cannot be applied to the general conditional fairness target. Conditionally independent tests have been popularly used in causal structure discovery problems. However, these methods cannot be mixed with gradient-based machine learning algorithms, since they usually calculate a statistic first and estimate a p-value with random methods. Therefore, the DCFR method is based on an equivalent relation of condi-

tional independence and is traceable in common machine learning algorithms.

3 Conventions

3.1 Notations

We suppose the dataset consists of a tuple D = (S, X, Y), where S represents sensitive attributes such as gender and race, X represents features, and Y represents the outcome. Furthermore, we divide features X into two parts X = (F, O), where F represents fair variables and O represents other features. We use m_X, m_F, m_O to denote the dimension of the features and we have $m_X = m_F + m_O$. We use calligraphic fonts to represent the range of corresponding random variables. For example, $\mathcal{X}$ represents the space of X and $\mathcal{X} \subset \mathbb{R}^{m_X}$. Similarly, we have $\mathcal{F} \subset \mathbb{R}^{m_F}$. To simplify, we suppose the sensitive attribute and the outcome are binary, which means $\mathcal{Y}$,$\mathcal{S}$ = 0,1. We set $\mathcal{S} = 1$ as the privileged group and $Y = 1$ as the favored outcome. We suppose there are N samples in total and we use S_i, X_i, Y_i, F_i, and O_i to represent the features of the i^{th} sample. In addition, for condition E, we use D(E) to represent the samples that satisfy the condition and $|D(E)|$ to represent the number of these samples. For example, D(Y = 1) means the samples that satisfy $Y_i = 1$, and $|D(Y = 1)|$ is the total number of such samples. A fair machine learning problem is to design a fair predictor Yˆ with parameters θ: $\mathcal{X}$ x $\mathcal{S} \rightarrow \mathcal{Y}$, which maximizes the likelihood P (Y, X, S|θ) while satisfying some specific fair constraints, which we will introduce in the next section.

3.2 Fairness Notions

We first introduce some well-known fair notions in machine learning problems.

Definition 1 (Demographic parity (DP)). Given the joint distribution D, the classifier $\hat{Y}$ satisfies demographic parity with respect to sensitive attribute S if $\hat{Y}$ is independent of S, i.e.

$$\hat{Y} \perp S \tag{1}$$

The definition of DP is clear and concise, representing that S has no predictive power to $\hat{Y}$, but in practice, we are also interested in some evaluation metrics to reveal how fair the system is. Thus the following equivalent form ΔDP is proposed to measure the degree of fairness. Easy to show that $\hat{Y} \perp S$ if and only if $\Delta DP = 0$.

$$\Delta DP \triangleq |P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S = 0)| \tag{2}$$

One of the drawbacks of ΔDP is that when the base rate differs significantly among two groups, i.e., $P(Y = 1|S = 0) \neq P(Y = 1|S = 1)$ the utility could be limited. Hardt et al. [2016] further proposed another notion Equalized Odds to avoid this problem.

Definition 2 (Equalized odds (EO)). Given the joint distribution D, the classifier $\hat{Y}$ satisfies equalized odds with respect to sensitive attribute S if $\hat{Y}$ is independent of S con-

ditional on Y, i.e.

$$\hat{Y} \perp S|Y \tag{3}$$

Similarly, the metric ΔEO is defined as the expectation of the absolute difference between the true positive rate and false positive rate across two groups.

$$\Delta EO \triangleq \mathbb{E}_y[|P(\hat{Y}=1|S=1,Y=y) - P(\hat{Y}=1|S=0,Y=y)|] \tag{4}$$

It is also easy to show that $\hat{Y} \perp S|$ Y if and only if $\Delta EO = 0$.

Definition 3 (Conditional fairness (CF)). Given the joint distribution D, the classifier $\hat{Y}$ satisfies conditional fairness with respect to sensitive attribute S and fair variables F if $\hat{Y}$ is independent of S conditional on F, i.e.

$$\hat{Y} \perp S|F \tag{5}$$

In addition, similar to ΔEO, we define a metric ΔCF as:

$$\Delta CF \triangleq \mathbb{E}_f[|P(\hat{Y}=1|S=1,F=f) - P(\hat{Y}=1|S=0,F=f)|] \tag{6}$$

Specifically, when fair variables are continuous, Equation 6 becomes:

$$\Delta CF = \int_{f\in\mathcal{F}} \mathbb{E}_f[|P(\hat{Y}=1|S=1,F=f) - P(\hat{Y}=1|S=0,F=f)|]d\mathbb{P}(f) \tag{7}$$

and when fair variables are categorical, ΔCF becomes

$$\Delta CF = \Sigma_{f\in\mathcal{F}} \mathbb{E}_f[|P(\hat{Y}=1|S=1,F=f) - P(\hat{Y}=1|S=0,F=f)|]P(F=f) \tag{8}$$

ΔCF aims to calculate the mean of the absolute difference between two groups among all potential values of the fair variables. Similarly, we have $\hat{Y} \perp S|F$ if and only if $\Delta CF = 0$.

Compare CF with DP and EO: On the one hand, conditional fairness can take more complex situations into account. On the other hand, conditional fairness is more general and it can be easily reduced to DP and EO. Consider the data-generating graph for a toy example of a college admission case in Figure 1. Because qualification requirements usually differ among various departments, it is fair to determine outcomes according to department choices and qualifications. Hence any predictors with the form $\hat{Y} = f(Q, D)$ can be considered fair in practice. It is easy to show that when setting department choice D as the fair variable, $\hat{Y}$ is conditional fair. However, DP and EO fail to judge the fairness of $\hat{Y}$ as Equation 1 and Equation 3 may not be satisfied.

In addition, conditional fairness is a more flexible fairness notion as:

– If we believe none of the features X is fair, which means F = , the conditional independence target is reduced to the independence condition as shown in Equation 1 and conditional fairness is reduced to DP.

– If we set F as Y, the conditional independence target is reduced to the conditional independence as shown in Equation 3 and conditional fairness is reduced to EO.

Compare CF with causality-based fairness notions Generally speaking, conditional fairness requires much fewer assumptions than causality-based fairness notions, which makes CF practical in real problems. Under some circumstances, a conditional fair decision-making system can sat-

isfy causality-based fairness notions. Consider path-specific fairness [Chiappa, 2019] in the example shown in Figure 1. The directed path S $\rightarrow$ Y can be viewed as an unfair path while S $\rightarrow$ D $\rightarrow$ Y and Q $\rightarrow$ Y are fair paths. Hence, the historical decision Y is not path-specific fair for the existence of unfair path S $\rightarrow$ Y. However, the conditional fair decision-making system $\hat{Y}$ = f (Q, D) successfully satisfies the requirement as the unfair path S $\rightarrow$ $\hat{Y}$ does not exist. As for deeper connections between conditional fairness and causality-based fairness notions, we remain as future works.

3.3 Problem Formulation

Next, we will apply our definition of conditional fairness to real fair problems. In general, the goal of a fairness problem is to achieve a balance between fairness and algorithm performance. Formally, we need to design a loss function on prediction $L_{pred}(Y, \hat{Y})$ and another loss function on fairness $L_{fair}(Y, S, F)$. The optimization goal of a fairness problem can be formulated as:

$$\theta = \arg\min_{\theta} L(\hat{Y}) = \arg\min_{\theta} L_{pred}(\hat{Y}) + \lambda \cdot L_{fair}(\hat{Y}, S, F) \tag{9}$$

where the hyper-parameter λ provides a trade-off between fairness and performance. When λ is large, the target tends to make L_{fair} small which can ensure fairness while doing harm to performance, and the result is the opposite when λ is small.

As for the prediction loss, any form of traditional loss functions are suitable such as cross-entropy or L1 loss. While the fairness loss targeted for conditional fairness is difficult to design relatively. When fair variables are categorical, we can use the ΔCF metric as a loss function. However, in practice, the fair variables may contain many different values or they may be continuous. Under this circumstance, the metric can no longer be a suitable loss function for optimization. Inspired by this issue, we will propose a new derivable loss function that can deal with these situations in the next section.

4 Proposed Method

Our solution to this problem is to learn a latent representation Z, which satisfies the condition (Equation 5). Suppose the representation has m_Z dimensions, g: $\mathbb{R}^{m_X} \times 0{,}1 \rightarrow \mathbb{R}^{m_Z}$ is the function from the space of X and S to representation space. The prediction function k: $\mathbb{R}^{m_Z} \rightarrow [0{,}1]$ yields the probability of the sample in the positive class. The framework of our model is shown in Figure 2. We now rewrite Equation 9 under this representation learning framework as:

$$\theta = \arg\min_{\theta} L_{pred}(k(g(X, S)), Y) + \lambda \cdot L_{fair}(g(X, S), F, S) \tag{10}$$

4.1 Adversarial Learning

Now we combine the total loss function as shown in Equation 10 and the conditional independence and get:

$$\begin{aligned} \theta &= \arg\min_{g,k} L_{pred}(k(g(X, S)), Y) + \lambda \cdot \sup_{h} Q(h) \\ &= \arg\min_{g,k} L_{pred}(k(g(X, S)), Y) + \lambda \cdot Q(h) \end{aligned} \tag{11}$$

As Q(h) is actually a weighted L1 loss, the loss function above can be optimized with the method of adversarial learning by setting the Q(h) as the adversarial loss. There are several works that use adversarial learning to solve fairness notions. While the frameworks among these works are similar, the main difference lies in the design of loss functions. Our method is most close to LAFTR [Madras et al., 2018]. And actually, when F = ϕ, which means the conditional independence constraint $\hat{Y} \perp S|F$ is reduced to $\hat{Y} \perp S$, our method is exactly the same as theirs.

Algorithm 1 *Derivable Conditional Fairness Regularizer*

Input: *(Dataset D = X, Y , S, X = (F, O), Epoch, batch-size, steps)*
Output : *g, k, h as in Equation 11*
Step 1

for epoch$_i \leftarrow$ 1 to EPOCH **do**
    Random mini-batch D' = (X'=(F', O'), Y', S') from D
    Freeze h. Un-freeze g, k
    Optimize g, k with gradient descent according to D'
    Freeze g, k. Un-freeze h
    for Step $\leftarrow$ 1 to STEPS **do**
        Optimize h with gradient descent according to D'
    end for
end for

Step 2

Freeze h. Un-freeze g, k
for epoch$_i \leftarrow$ 1 to EPOCH **do**
    Random mini-batch D' = (X'=(F', O'), Y', S') from D
    Optimize k with gradient descent according to D'
    if accuracy on validation set does not increase for continuous 20 epochs **then**
        break
    end if
end for
return g,h,k

5 Experiments

5.1 Datasets

For the purpose of this project, we have implemented the model on the Adult dataset. The aim of the adult dataset is to predict whether a person makes more than $50k per year or not. There are 112 attributes including sex, gender, education, level, etc. Among these, we gender as the sensitive attribute and consider occupation along with 14 other variables as the fair variable.

5.2 Results

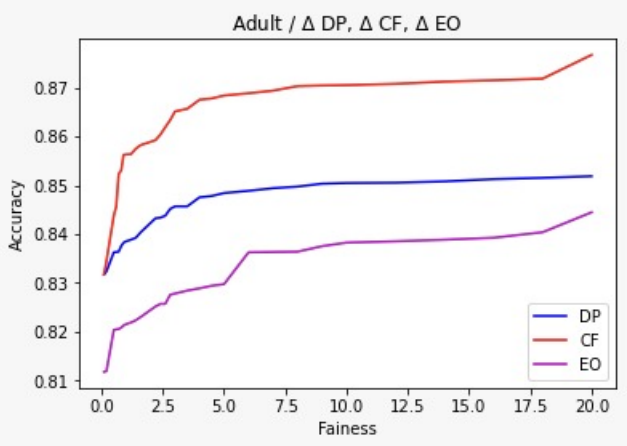

Figure 1: Model DCFR trained and tested with fairness coefficients ranging from 0.1 to 20.0 for three fairness notations, namely demographic parity (DP), Equalized odds (EO) and Conditional Fairness (CF).

6 Conclusion

From this project, we can conclude that conditional fairness concerning fair variables is difficult to optimize directly as it cannot be written directly as a derivable loss function especially when fair variables are continuous or contain many categorical values.

There is ample scope for future work on this project, especially the application of this model to different settings and measuring the performance between different models.Also the application of this method to the unsupervised setting.

References

[1] Renzhe Xu, Peng Cui, Kun Kuang. 2020. Algorithmic Decision Making With Conditional Fairness. In proceedings of KDD 2020, Virtual Event, USA.

[2] Junyi Chai, Xiaoqian Wang. 2022. Fairness with Adaptive Weights. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[3] Mariya I. Vasileva, 2020. The Dark Side of Machine Learning Algorithm: How and why they can leverage bias, What can be done to pursue algorithm fairness. In Proceedings of 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2020.

[4] Shira Mitchell, Eric Potash, Solon Barocas. 2021. Algorithm Fairness: Choices, Assumptions, and Definitions. In proceedings of the Annual Review of Statistics and its Applications, Virtual Event, USA, November 2020.

[5] https://github.com/komalbodhankar/MLFinal.git