Fairness constraints

Zahra Khatti
Lehigh University, zak223@lehigh.edu

This article focuses on the importance of ensuring fairness in machine learning to address and reduce biases in predictive models. The aim is to engage the audience's attention to read further on this topic. It introduces key fairness measures such as Disparate Impact, Equalized Odds, and Equal Opportunity, explaining how these measures contribute to treating diverse demographic groups fairly. The article delves into fairness constraints, exploring strategies discussed in the literature review to enforce fairness without sacrificing model accuracy. The literature review highlights three research papers, emphasizing their contributions to extending fairness frameworks and keeping the balance between fairness and accuracy.

Key words: Fair machine learning, Fairness constraints

---

1. Introduction Artificial intelligence, despite its revolutionary impact on various industries, often creates biases that lead to unfair decisions. This bias shows itself when outcomes disproportionately favor one particular group over others [8, 3]. To address this issue, fairness has emerged as a significant research area within machine learning, aimed at mitigating bias and helping in fair decision-making. As automated decision-making systems increasingly spread across diverse application domains, ranging from credit approval and criminal risk assessment [7] to hiring [9, 6] and education [10], the need to ensure fairness in their outcomes becomes crucial.

There are several studies showing the unfairness, arising in different areas. In a study by Charles et al. [2], it was shown that black individuals receive higher interest rates for auto loans, while in another research by Alesina et al. [1], it was shown that small business loans have higher interest rates for women. Female candidates were discriminated against by Amazon's machine learning hiring system, particularly for technical and software development positions [4]. Similarly, according to another study [5, 11], Google's ad-targeting algorithm suggested higher-paying executive jobs more to men than to women. By exploring fairness in machine learning in different areas, it can be ensured that AI technologies are not only powerful but also unbiased in their impact on society.

In light of the growing interest in fairness constraints, there are several open questions to be answered.

1. How to strike a balance between accuracy and fairness?
2. What ethical concepts should be incorporated in the model to reach the highest fairness?
3. If some individuals belong to several protected groups, how can fairness be satisfied?

In this paper, the goal is to engage the audience. It aims to encourage you to explore it further and find it engaging enough to read by giving examples, introducing some of the fundamental concepts, and exploring some of the most recent papers in this area.

2. Background In this section, we introduce concepts in fairness constraints for machine learning, including some notion and methodologies. In machine learning, the aim is to train a prediction function that, given a feature vector for a data point, predicts an appropriate label for the data point. Throughout this section, we denote Y as a true label, $\hat{Y}$ as a prediction, and Z as a label for the group type (sensitive features). For the purpose of this article, we assume that Y, $\hat{Y}$, and Z take the values of zero and one.

2.1. Fairness Measures In this section, we introduce some of the metrics to address the fairness of algorithms and models.

TABLE 1. Common classification criteria

Event	Condition	(P(event | condition))
$\hat{Y}=1$	$Y=1$	True Positive Rate
$\hat{Y}=0$	$Y=1$	False Negative Rate
$\hat{Y}=1$	$Y=0$	False Positive Rate
$\hat{Y}=0$	$Y=0$	True Negative Rate

Disparate Impact Disparate impact refers to discriminatory predictions for different groups based on their sensitive attributes. The aim to avoid disparate impact is to enforce the equation

$$P(\hat{Y}=1|Z=0)=P(\hat{Y}=1|Z=1). \tag{2.1}$$

This equation ensures that the positive prediction is independent of the sensitive attribute.

Equalized Odds Equalized Odds dictate that the probability of a positive prediction for a true label and the probability of a positive prediction for a negative label should be the same for different groups, expressed as

$$P(\hat{Y}=1|Z=0,Y=y)=P(\hat{Y}=1|Z=1,Y=y) \quad \text{for all } y\in\{0,1\}.$$

Equal Opportunity (EO) or disparate mistreatment Equal Opportunity states that different sensitive groups should have equal true positive rates, expressed as

$$P(\hat{Y}=1|Z=0,Y=1)=P(\hat{Y}=1|Z=1,Y=1).$$

This equation ensures that the model provides an equal probability of a positive prediction for individuals from different groups, when they truly deserve it, regardless of their sensitive feature.

2.2. Fairness Constraints In this section, we delve into fairness constraints. This takes into account both the objectives and constraints to strike a balance between accuracy and fairness. In the literature review section, we take a look at the contributions regarding these definitions.

Business Necessity Clause In scenarios where the correlation between labels and sensitive attributes is substantially high, low accuracy may occur, conflicting with business objectives. To address this, the business necessity clause is introduced, aiming to minimize disparate impact while maintaining an acceptable threshold for loss. Compared to the conditional probability given by Eq. (2.1) which makes the problem nonconvex, the paper presented by Zafar et al. [12] uses the relaxation of the no disparate impact probability constraint using a novel covariance measure of decision boundary unfairness. This paper presents disparate impact as the covariance between the model's decision boundary (θ) and sensitive attributes (z_i) over the training dataset (x_i). The equation is expressed as

$$\mathrm{Cov}(z,d_\theta(x))=\mathbb{E}\left[(z-\bar{z})d_\theta(x)\right]-\mathbb{E}(z-\bar{z})\bar{d}_\theta(x)\approx\frac{1}{N}\sum_{x,z}(z_i-\bar{z})d_\theta(x_i),$$

and the optimization problem is formulated as follows

$$\begin{aligned}&\underset{\theta}{\text{minimize}} \quad \left|\frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})d_\theta(x_i)\right| \\ &\text{subject to } L(\theta)\leq(1+\gamma)L(\theta^*).\end{aligned}$$

N : The number of training examples.
z_i : The value associated with the sensitive feature of the i-th training example.
$\bar{z}$: The mean or average value of the sensitive features over all training examples.
$d_\theta(x_i)$: The decision function associated with the i-th training example, parameterized by θ.
θ : Decision boundary parameters that are optimized to minimize the given expression.
$L(\theta)$: The loss associated with the decision boundary parameters θ.
$L(\theta^*)$: The optimal loss over the training set.
γ : A non-negative parameter specifying the maximum additional loss allowed compared to its optimal loss L(θ^*).

The model's loss should not be more than a little bit worse than the loss of an unconstrained (unfair) model ($L(\theta^*)$). The parameter γ represents how much worse the model's loss can be compared to the unfair model. This constraint shows business necessity because it ensures that the model's fairness improvements (reducing disparate impact) do not lead to a significant drop in the model's accuracy [12].

Fair Empirical Risk Minimization (FERM) In this framework [6], the optimization problem tries to minimize empirical risk with fairness constraints. We have a dataset D consisting of n samples drawn from an unknown probability distribution over $X \times Z \times Y$. Each input x in X may or may not include the sensitive feature z from Z. The goal is to create a model that minimizes the risk or loss. The error (risk) of f is defined as

$$\text{minimize} \quad \mathbb{E}[l(f(x), y)]$$

where f(x) is defined as a prediction. Since the distribution over $X \times Z \times Y$ is unknown, the risk cannot be computed. But, we can compute the empirical risk $\hat{L}(f) = \hat{E}[l(f(x), y)]$, where $\hat{E}$ denotes the empirical expectation. Then, Empirical Risk Minimization (ERM), tries to minimize the empirical risk within a set of functions.

The Equal Opportunity constraint ensures that the model provides an equal probability of a positive prediction for individuals from different groups, when they truly deserve it, regardless of their sensitive feature. The constraint will be calculated as follows

$$|P\{f(x) > 0 \,|\, y = 1, z = 1\} - P\{f(x) > 0 \,|\, y = 1, z = 0\}| \leq \epsilon.$$

Stochastic Sequential Quadratic Optimization Algorithm This is an optimization algorithm for solving complex optimization problems that have deterministic constraints and a stochastic (random) objective function. In this framework, which has been proposed by Curtis et al. [3], objective functions are defined over large datasets, and constraints are defined over relatively small datasets. We desire to find the optimal solution to the following problem

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}[l(f(x), y)] \\ \text{s.t.} \quad & c(x) + s = 0 \\ & s \geq 0, \quad x \in \mathbb{R}^n, s \in \mathbb{R}^m \end{aligned}$$

where $f(x)$ represents the predicted outcome, and $c(x)$ represents deterministic fairness constraints. The ϵ-constraint method integrates fairness constraints as an additional objective in a multiobjective optimization framework, formulating a problem to simultaneously minimize the loss and the introduced fairness objective. To address scalability, a stochastic optimization algorithm is proposed for efficient solutions to the constrained optimization problems.

3. Literature Review In this literature review, we summarize the contribution of three research papers that explore fairness constraints, a technique for adding constraints to create a fair model.

The authors of [6] contribute to the field of fair machine learning by introducing a framework that extends the traditional concept of Equal Opportunity (EO) fairness. This framework focuses on constraining the conditional risk of classifiers for positive samples within different groups, ensuring stability in group membership. They achieve this by using a specified loss function. The authors introduce Fair Empirical Risk Minimization (FERM) for optimizing models within this fairness framework and provide statistical consistency bounds. Knowing the computational challenges associated with nonconvex fairness constraints, they propose a convex FERM problem instead, accompanied by an empirical verification method. Furthermore, the paper shows the applicability of this framework to kernel methods, particularly support vector machines (SVMs), showing how it naturally leads to fairness.

Zafar et al. [12] tackle the issue of fairness through a comprehensive approach that addresses potential gaps in prior works. It emphasizes that ignoring sensitive features, such as gender or race, can inadvertently result in fairness through unawareness. To bridge this gap, the paper presents a method that covers all sensitive features while addressing various fairness criteria, including disparate treatment and disparate impact. Furthermore, unlike the study by Donini et al. [6] the paper extends its coverage to a wide range of classification models. This framework aims to remove the obstacles of prior studies, ensuring that fairness is not compromised. Notably, it offers a dual formulation that maximizes fairness under accuracy constraints, thus aligning with the business necessity clause of anti-discrimination methods, which has not been discussed in previous studies [6].

By using constrained stochastic optimization, Curtis et al. [3] try to mitigate bias in models. The method involves minimizing loss while constraining unfairness on a relatively small dataset defined by sensitive attributes, essentially engaging in multiobjective optimization. This approach focuses on a subset of data characterized by sensitive attributes, making it both effective and scalable, particularly in the context of large-scale supervised learning. This method overcomes the gaps of prior studies. They relied on using all available data, thereby having data privacy concerns or having substantial computational expenses. They even overlooked some sensitive features, resulting in fairness through unawareness. The paper's approach is more applicable to large-scale problems compared to the Zafar et al.(2019) [12].

References

[1] Alesina AF, Lotti F, Mistrulli PE (2013) Do women pay more for credit? evidence from italy. *Journal of the European Economic Association* 11(suppl_1):45–66.

[2] Charles KK, Hurst E, Stephens Jr M (2008) Rates for vehicle loans: race and loan source. *American Economic Review* 98(2):315–320.

[3] Curtis FE, Liu S, Robinson DP (2023) Fair machine learning through constrained stochastic optimization and an ϵ-constraint method. *Optimization Letters* 1–17.

[4] Dastin J (2022) Amazon scraps secret ai recruiting tool that showed bias against women 296–299.

[5] Datta A, Tschantz MC, Datta A (2015) Automated experiments on ad privacy settings. vol. 2015, issue 1. *Proceedings on Privacy Enhancing Technologies* 92–112.

[6] Donini M, Oneto L, Ben-David S, Shawe-Taylor JS, Pontil M (2018) Empirical risk minimization under fairness constraints. *Advances in neural information processing systems* 31.

[7] Perry WL (2013) *Predictive policing: The role of crime forecasting in law enforcement operations* (Rand Corporation).

[8] Pessach D, Shmueli E (2022) A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55(3):1–44.

[9] Posse C (2016) Cloud jobs api: machine learning goes to work on job search and discovery.

[10] Romero C, Ventura S (2011) Preface to the special issue on data mining for personalised educational systems. *User Modeling and User Adapted Interaction* 21(1):1.

[11] Simonite T (2015) Probing the dark side of google's ad-targeting system. *MIT Technology Review* .

[12] Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP (2019) Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 2737–2778.