Trustworthy Machine Learning: Addressing Fairness in Classification with a Disparate Impact-Informed Loss Function

Vivek Parthasarathy

Department of Computer Science

University of Chicago

vivek@parthasarathy.tv

Abstract

We study loss–based mitigation of group-level bias in binary classification. Building on the disparate-impact criterion of Feldman *et al.* (2015), we introduce a differentiable penalty that is added to the standard binary cross-entropy loss and optimised jointly with model parameters. The penalty drives the ratio of positive-prediction rates between protected and non-protected cohorts toward parity, and is agnostic to the underlying model architecture. Using a feed-forward neural network and the ADULT INCOME benchmark, we show that the proposed objective substantially narrows the demographic-parity gap with only a marginal decrease in overall accuracy. Notably, improvements extend to *race*, a sensitive attribute excluded from the loss, suggesting that penalising one source of imbalance can propagate to correlated sub-populations. We discuss conditions under which this behaviour is desirable and outline limitations when protected attributes are latent or partially observed.

I. INTRODUCTION

In this project, we delve into the complex topic of fairness in classification tasks, a realm of machine learning that simultaneously presents a significant challenge and an undeniable necessity. More specifically, we shift our focus to the inherent biases that can be embedded in predictive models—biases that become particularly prominent when such models are

deployed in sensitive domains. These biases often can reflect or amplify existing undesirable societal disparities. To combat these negative externalities, researchers have introduced quantitative distributional notions of fairness. In this project, we encode the desired fairness metrics into the loss function, enabling the trained model to better achieve our desired distributional notion while also optimizing for accuracy. In particular, we adopt the work of Feldman et al. from their paper titled "Certifying and Removing Disparate Impact," implementing a disparate impact measure term into a custom "fairness-aware" loss function. In this project, our aspirations are as follows:

- Explore the potential weaknesses or pitfalls of the fairness-aware loss function approach in the protection of certain classes (in particular we discuss the ramifications of the methodology on hidden protected classes).
- Experiment with the methodology on data to gain insight.
- Better understand the realities of reasoning about fairness in machine learning.

II. Theoretical Basis: Certifying and Removing Disparate Impact

To navigate the domain of fairness in machine learning, as stated, we anchor our approach on the theoretical framework provided by Feldman et al., whose work presents a measurable definition of fairness. Drawing from the legal doctrine of disparate impact, they propose a method to measure and subsequently alleviate this impact in classification tasks.

According to Feldman et al., disparate impact arises when seemingly neutral decision-making processes inadvertently lead to substantial disparities across different demographic groups. They advocate for a fairness measure that quantifies this disparity, proposing that a truly fair classifier should yield outcomes devoid of disproportionate disadvantages or benefits to any group. This conception of group fairness is compelling because it strives to rectify biases at a societal level, unlike individual fairness measures that focus on single entities. It bears particular relevance in scenarios where decision-making carries broad demographic implications, such as in credit scoring or job recruitment.

In this project, we deploy the disparate impact measure within a custom loss function to guide the training of a Neural Network classifier. This methodology not only allows us

to assess the fairness of our model, but also to actively work towards reducing unfairness during the training process itself. Our custom loss function is defined as follows:

$$L_{\text{final}}(y, \hat{y}, z) = \text{BCE}(y, \hat{y}) + \alpha \cdot L_{\text{DI}}(y, \hat{y}, z)$$

where $\text{BCE}(y, \hat{y})$ is the binary cross entropy between the true label y and the predicted label $\hat{y}$, and $L_{\text{DI}}(y, \hat{y}, z)$ is the disparate impact loss defined as:

$$L_{\text{DI}}(y, \hat{y}, z) = 1 - \min\left(\frac{\frac{1}{N_{z=1}}\sum_{i:z_i=1}\hat{y}_i}{\frac{1}{N_{z=0}}\sum_{i:z_i=0}\hat{y}_i}, \frac{\frac{1}{N_{z=0}}\sum_{i:z_i=0}\hat{y}_i}{\frac{1}{N_{z=1}}\sum_{i:z_i=1}\hat{y}_i}\right)$$

Here, z denotes the protected attribute (such as sex), $N_{z=1}$ and $N_{z=0}$ represents the number of instances with $z = 1$ and $z = 0$, respectively, and α is a hyperparameter determining the trade-off between binary cross-entropy loss (accuracy) and disparate impact loss (fairness). Note that $\sum_{i:z_i=1}\hat{y}_i$ denotes the sum of predicted labels for instances where the protected attribute $z = 1$, and similarly for $\sum_{i:z_i=0}\hat{y}_i$ where $z = 0$.

Disparate impact (DI) is typically framed as a ratio to be maximized for fairness, with a perfect DI of 1 indicating no disparity. However, in a loss function for training a neural network, we need a quantity to minimize. Hence, we take $1-\text{DI}$ to formulate a minimization problem. The closer this term is to 0, the smaller the disparity, leading to fairer predictions. Consequently, our custom loss function minimizes both prediction error (binary cross-entropy) and unfairness (disparate impact loss) to ensure accurate and fair model outcomes.

III. Hidden Protected Classes

While it is well-known that optimizing one distributional notion of fairness exhibits a trade-off in terms of other distributional notions, we also explore a more foundational issue: the issue of hidden protected attributes. In real-world applications, there will always be attributes that ethically should be treated with some notion of fairness, but for which data was not collected. While we can change the loss function to accommodate known protected attributes, no such methodology can be used for those protected attributes which are unknown.

There are two hypotheses:

1) That downweighting the features which are correlated with inclusion in a protected class results in greater reliance on other features that could be separately correlated with another protected class. In this case, well-intentioned decision makers or fairness researchers could inadvertently do more harm than good for unknown protected classes.
2) That predictive features spuriously correlated with one discriminated protected class are more likely to be correlated with another protected class.

IV. Methodology

In this section we discuss the specifics of our dataset, task, model, and experimental design.

We employ the Adult Income dataset, which encompasses a wide range of socioeconomic and personal characteristics such as age, education level, marital status, occupation, and, most importantly, gender and race. The goal of this dataset is to predict whether an individual's income surpasses $\$50$K per year. Recognizing the potential for bias, sex and race are identified as the sensitive attributes. However, we choose to implement the fairness-aware loss function with respect to sex only. Our objectives are twofold: First, we wish to observe the extent to which the fairness-aware model performs better on various fairness metrics with respect to sex. Second, we wish to observe the consequences of training a model with respect to a loss function that does not consider a different sensitive attribute (race in our case), in terms of fairness with respect to that attribute. In particular, note that a methodology which seeks to optimize fairness for multiple classes simultaneously would not solve the issue of hidden attributes.

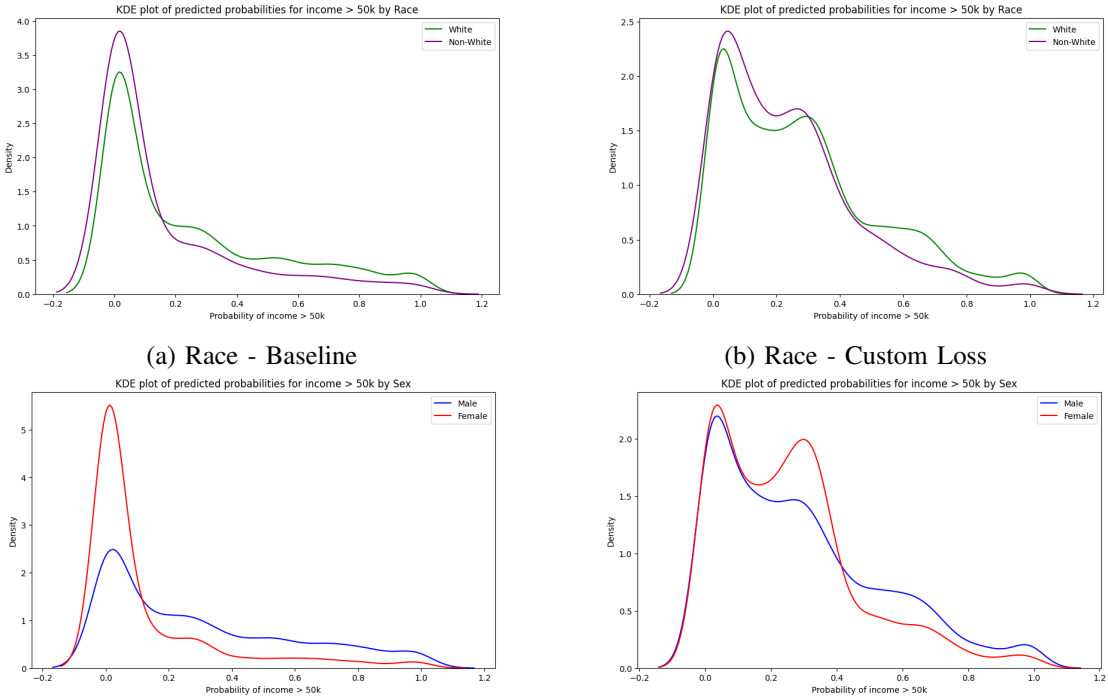
Our prediction model of choice is a Neural Network. This selection stems from the model's capability to encapsulate complex non-linear relationships within the data, its adeptness in managing both numerical and categorical variables, and, most importantly, its appropriateness for the implementation of a custom loss function that includes a fairness criterion.

It is important to note that there is a trade-off between disparate impact (or demographic parity) and equal opportunity metrics. Our focus gravitates towards disparate impact in

the context of this dataset primarily because, among other reasons, predicting the positive outcome of high income could enable banks to extend additional credit. Thus, the same amount of loans could be granted.

We first train a baseline model using standard binary cross entropy loss and compute performance metrics and then train the modified model. We compute the same metrics for the modified model and compare accordingly.

V. Results



(a) Race - Baseline

(b) Race - Custom Loss

(c) Sex - Baseline

(d) Sex - Custom Loss

Fig. 1: Kernel Density Estimation Plots by Sensitive Attribute and Model

	Male	Female	White	Other Races	Overall
ROC-AUC	-	-	-	-	0.91
Accuracy	81.7%	92.8%	84.6%	89.9%	85.4%
Precision	73.8%	72.1%	73.7%	72.1%	73.6%
Recall	62.3%	56.2%	62.0%	55.6%	61.8%
False Positive Rate	9.7%	2.7%	7.6%	3.9%	-
True Positive Rate	62.3%	56.2%	62.0%	55.6%	-
Positive rate	25.8%	8.5%	21.5%	11.7%	-

TABLE I: Baseline Model: Performance Metrics by Demographic Group

	Sex	Race
Demographic Parity	0.173	0.098
Equalized Odds (TPR Diff)	0.062	0.065
Equalized Odds (FPR Diff)	0.071	0.037

TABLE II: Baseline Model: Fairness Metrics by Sensitive Attribute

	Male	Female	White	Other Races	Overall
ROC-AUC	-	-	-	-	0.88
Accuracy	81.2%	92.2%	84.0%	89.3%	84.8%
Precision	78.9%	63.9%	76.4%	69.4%	75.7%
Recall	52.3%	65.5%	54.4%	53.5%	54.3%
False Positive Rate	6.1%	11.2%	5.8%	4.2%	-
True Positive Rate	52.3%	65.5%	54.4% %	53.5%	-
Positive rate	20.2%	11.2%	18.2%	11.7%	-

TABLE III: Modified Model: Performance Metrics by Demographic Group

	Sex	Race
Demographic Parity	0.090	0.064
Equalized Odds (TPR Diff)	0.132	0.009
Equalized Odds (FPR Diff)	0.016	0.015

TABLE IV: Modified Model: Fairness Metrics by Sensitive Attribute

VI. Discussion

To begin, we note that the baseline model exhibits a commendable predictive performance as evidenced by an ROC-AUC of 0.91 and an overall accuracy of 85.4%. These figures highlight the model's ability to effectively discriminate between positive and negative

outcomes. However, shifting our attention to fairness measures, we find several indications of potential bias within the model's predictions. Examining the positive rates across gender reveals a disparity, with 25.8% for males and a much lower 8.5% for females. This disparity suggests a difference in how the model's predictions are distributed across gender groups, perhaps indicating that the model's predictions may favor males over females.

However, when we deploy the modified model using a custom loss function, we note several changes in performance and fairness metrics. Inspecting the results of the modified model reveals a decrease in predictive performance, with the ROC-AUC falling to 0.88 and overall accuracy to 84.8%. However, it is also apparent the modified model performs significantly better on fairness metrics. One notable difference is the reduced disparity in positive rates across gender groups, with 20.2% for males and 11.2% for females; which indicates the modified model has mitigated some of the bias in the baseline model. Such observations exemplify the widely known trade-off that exists between accuracy and fairness.

Moreover, the Demographic Parity fairness metric for sex is decreased in the modified model from 0.173 to 0.090. This reduction signifies a narrowing of the gap in the probability of receiving positive outcomes between males and females, thereby indicating improved fairness with respect to sex. In addition, in the above KDE plots, albeit slight, we qualitatively note a reduction in the gap between the prediction distributions for males and females as the curves for the two sexes have increased overlap, indicating less bias in the modified model.

Turning our attention to race, we note that, perhaps surprisingly, training to ensure fairness with respect to sex does not sacrifice fairness metrics for race. In fact, these metrics are slightly improved across the board with respect to the baseline model, as the positive rates have also grown closer together and the demographic parity, the true positive rate difference, and the false positive rate difference are all closer to zero, indicating better achievement of both demographic parity and equalized odds. Turning once more to the above KDE plots, this is further reinforced by the fact that the curves for the two racial demographics observed have slightly more overlap and a slightly smaller gap between them in the plot corresponding to the modified model compared to that of the baseline model.

VII. Concluding Remarks

Reflecting upon the aspirations laid out at the beginning of this project, our deep dive into fairness in classification tasks has yielded intriguing findings. Though perhaps surprising, we note that our results are indicative of success in implementing the disparate impact measure within a custom loss function inspired by the work of Feldman et al.

A striking result of this project is that our custom loss function, which intended only to improve disparate impact only for the sex attribute, actually also improved our metrics for fairness with respect to the race attribute. This outcome highlights a potential intersectionality in the biases that predictive models may inadvertently amplify, implying that similar discriminatory obstacles might indeed influence both race and sex. Indeed, perhaps the predictive features spuriously correlated with race are also more likely to be correlated with sex.

An interesting avenue for future study could be to construct datasets with a very large set of protected classes which should be sensitive to fairness. This would enable further research to handle the question of whether hidden protected classes pose significant fundamental ethical dilemmas and pitfalls for trustworthy machine learning methodologies going forward.

Upon comparing our model implementing the fairness-aware loss function with the baseline model, we observed a slight decrease in accuracy while improving fairness metrics. In real-world applications, assessing the optimal trade-off between fairness and performance is heavily dependent on the situation and may also be contingent on potentially hidden protected classes and their correlates. Looking forward, the insights gained from this project serve as a stepping stone, underscoring the complexity and intricacies of achieving fairness in machine learning tasks while also looking to maximize accuracy.

VIII. References

[1] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). ACM.

Code Availability

The code used for the analysis in this paper can be found at the following link: https://colab.research.google.com/drive/1twwAc_datpf6frQ0pSjbaztFoYUD3u4H#scrollTo=O9ZOVromfTAx.