Project Extention: Adversarially Reweighted Learning

Guy Shapira
guy.shapira@campus.technion.ac.il

Shai Feldman
shai.feldman@campus.technion.ac.il

Abstract

As of today, most previous machine learning (ML) fairness literature assumes that protected features are present in the dataset and thus can be used in order to mitigate fairness concerns. However, in practice factors like privacy and regulation often preclude the collection of protected features, severely limiting the applicability of traditional fairness research. In this work we build upon [1] weho were the first to present an ML model that tries to improve fairness when protected group memberships is unknown. In this work we address this problem by proposing Adversarial Orthogonal Learning(AOL). We believe that non-protected features and task labels are valuable for identifying fairness issues, and can be used to co-train in an adversarial approach for improving fairness. Our results show that AOL improves Rawlsian Max-Min fairness, with notable AUC improvements for worst-case protected groups in multiple datasets, outperforming state-of-the-art alternatives as well as the Adversarially Reweighted Learning(ARL) it is based upon.

1 Introduction

Learning algorithms are increasingly prevalent within consequential real-world systems, where fairness is an essential consideration: deploying learning algorithms to be fair across all protected features (e.g., race and gender) requires more than high prediction accuracy in controlled testbeds [2–4]. Consider, for example, estimating the effects of a drug for a specific person given their demographic information and medical measurements. In such a high-stakes setting, the decision-maker must be fair. If it is not fair for a certain group, the algorithm will be responsible for health damage resulting from racist motives. Recent research show that this happens in practice, where there are significant accuracy disparities across demographic groups in face detection [5], health-care systems [6], and more.

The approaches available today focus on proposing formal notions of fairness, and propose "de-biasing" methods to satisfy the fairness criteria [3, 7, 8]. However, these algorithms suffer from a major limitation—they assume that the protected features are known to the model [8–11]. In many practical situations, it is not feasible to collect or use the protected features, and thus they are unknown to the model. A recent work proposed an approach to deal with this limitation, achieving state of the art results [12]. However, this method suffers from major limitations which we tackle in this work. We propose a novel method to achieve fairness across all sub-populations of the data, which is more flexible and not over-conservative, as opposed to the existing method. Technically, we achieve this by augmenting the cross-entropy loss function with an additional term that promotes appropriately balanced accuracy across the feature space.

Formally, consider a classification problem where we are given n training samples $\{(X_i, Y_i)\}_{i=1}^n$, where $X \in \mathbb{R}^p$ is a feature vector, and $Y \in \{0, 1\}$ is a response variable. At test time, we observe a feature vector X_{n+1} and our goal is to predict the unknown value of Y_{n+1} and—importantly—to succeed independently on the unknown protected features. The predictor is denoted as h_θ, were θ are the parameters of the model. In this work, we seek to produce predictions that achieve high accuracy while being independent to the protected features:

$$(h_\theta(X_{n+1}) = Y_{n+1}) \perp\!\!\!\perp s_{n+1}, \tag{1}$$

where s_{n+1} are the protected features of X_{n+1}. Notice that such predictions are both correct in that they achieve high accuracy, and are also fair, as the protected features are independent to the probability of predicting correctly.

As an example, consider predicting presence of cancer from demographic variables: age, gender, health status, and so on. The independence requirement in (1) asks that predictions are correct for any age, gender, and health status combination. That is, no matter what an individual's value of the features s_{n+1}, the model must be valid, and thus, be fair.

In this work, we propose a novel regularization scheme to push classification algorithms towards solutions that better satisfies the orthogonal requirement (1). The core idea is to force the known features and the model's loss to be approximately independent, since this independence must hold for an optimal oracle predictor. A method whose predictions' correctness depend on the input, i.e., has lower accuracy for certain inputs, is unfair, since there is a sub-population that suffers from a discrimination.

1.1 Problem Formulation

The setup of this work is the problem of binary classification, given a training dataset consisting of n individuals $D = \{(x_i, y_i)\}_{i=1}^n$, where the samples are drawn from the distribution P_{XY}. x_i is a m-dimensional vector of non-protected feature, and y_i is a binary label. We assume that there exist K protected groups where for each example x_i there exists an unobserved $s_i \sim S$, Where S is a random variable over $\{k\}_{i=1}^K$. The set of examples with membership in group s is given by $\mathcal{D}_s = \{(x_i, y_i) : s_i = s\}$. Our goal is to construct a model that is fair across all groups $s \in S$, where we use Definition 1 as the definition of a fair model.

Definition 1 (Rawlsian Max-Min Fairness)**.** Suppose H is a set of hypotheses, and $U_{D_s}(h)$ is the expected utility of the hypothesis h for the individuals in group s, then a hypothesis h^* is said to satisfy Rawlsian Max-Min fairness principle [12] if it maximizes the utility of the worst-off group, i.e., the group with the lowest utility.

$$h^* = \arg\max_{h \in H} \min_{s \in S} U_{\mathcal{D}_s}(h) \tag{2}$$

The utility metric we use to assess the fairness of a model is AUC (area under the ROC curve).

2 Adversarially Reweighted Learning

Similarly to [1] we also turn Rawlsian Max-Min Fairness in 2 into a learning objective, by replacing the expected utility with a loss function $L_{\mathcal{D}_s}(h)$ over the set of individuals in group s:

$$h^*_{\text{max}} = \arg\max_{h \in H} \min_{s \in S} L_{\mathcal{D}_s}(h), \tag{3}$$

where $L_{\mathcal{D}_s}(h) = \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}_s}[\ell(h(x_i), y_i]$ is the expected loss for the points in group s.

Since the sub-populations in the data (defined as the group S) are unknown, we can't solve equation 3. Therefore, we turn to the following alternative.

Minimax Problem Similar to Agnostic Federal Learning [13], we formulate the objective in 3as a zero-sum game between two players θ and λ. The process is comprises of T game rounds, where in round t, player θ learns the best parameters θ that optimizes his expected loss. In round $t + 1$, the adversary λ learns the best parameters that maximizes his loss.

$$J(\theta, \lambda) := \min_\theta \max_\lambda L(\theta, \lambda) = \min_\theta \max_\lambda \sum_{s \in S} \lambda_s L_{\mathcal{D}_s}(h) = \min_\theta \max_\lambda \sum_{i=0}^{n} \lambda_{s_i} \ell(h(x_i), y_i) \tag{4}$$

A concrete algorithm that is derived from equation 4 is defined as follows. The θ player uses an iterative learning method for classification, to optimize its parameters. For player λ we use the concept of *computationally-identifiable subgroups* [14].

Definition 2 (Computationally-Identifiable Subgroup)**.** Given a family of binary functions $\mathcal{F}$, we say that a subgroup G is computationally-identifiable if there is a function $f : X \times Y \rightarrow \{0, 1\}$ in $\mathcal{F}$ such that $f(x, y) = 1$ if and only if $(x, y) \in G$.

Building on this definition, we define $f_\phi : X \times Y \rightarrow [0, 1]$ to be an adversarial neural network parameters by ϕ whose task, implicitly, is to identify sub-populations in the data in which the learner makes significant errors. The adversarial examples weights $\lambda_\phi : f_\phi \rightarrow \mathbb{R}$ is defined by rescaling f_ϕ to put a high weight on regions which are likely to have larger errors. This way, we encourage h_θ to improve the predictions in these regions. The deriving of λ_ϕ from $f\phi$ is done as:

$$\lambda_\phi(x_i, y_i) = 1 + n\frac{f_\phi(x_i, y_i)}{\sum_{i=1}^{n} f_\phi(x_i, y_i)}$$

ARL Objective We formalize the intuition above and develop the Adversarially Reweighted Learning (ARL) scheme, which performs a minimax game between a *learner* and an *adversary*. Both players are trained alternatively, and implemented as a neural network that optimizes its own objective. The *learner* optimizes the main classification task, and aims to learn the optimal parameters θ that minimize the expected loss. The *adversary* aims to find a sub-population in the data that the *learner* is likely to have high loss over them. He does that by learning a mapping function $f_\phi : X \times Y \rightarrow [0, 1]$, that maps to *computationally-identifiable* regions with high loss. The *learner* minimizes the reweighted loss, which is adjusted by the adversarial:

$$J(\theta, \lambda) = \min_{\theta} \max_{\lambda} \sum_{i=0}^{n} \lambda_\phi(x_i, y_i)\ell(h_\theta(x_i), y_i) \tag{5}$$

In practice, we use cross-entropy loss for ℓ. The adversarial highlights the problematic regions with high loss, so the learner is more aware of these regions and train more carefully over them. This approach aims to achieve similar loss for every sample, so the model will achieve the same accuracy over every sup group in the data. It is worth noting that a complex adversary model is likely to overfit to outliers, and in experiments, a linear adversary performed the best.

Limitations While working on this project, we found the ARL model to be particularly vulnerable to outliers in data. The model tends to overfit to outliers, what harms it's performance on out of sample data. Moreover, we believe that while the motivation of ARL is correct, the demand of achieving the exact same loss on every sample (thus of course receiving the same average loss on every sub-group) is too conservative, in some cases might even be wrong (e.g., outliers). In the following section we present our approach (AOL) that deals with these problems.

3 Adversarial Orthogonal Learning

Motivation Since we find the ARL objective to be too conservative, we investigated similar ideas in order to achieve fairness over all sub-groups. We decided to focus on the demand that the result of the loss function will be independent of the input's feature vector, as can be seen in 6.

$$\forall i \in [0, n-1] : x_i \perp \ell(h_\theta(x_i), y_i) \tag{6}$$

Formal Definition Building on that motivation we further formalize our orthogonal loss function in 7. We denote by ρ(Z,W) the Pearson's correlation between Z and W, and denote x_i^j as the j-th feature of the i-th sample in the data set.

$$\sum_{j=1}^{m} \rho(x_i^j, \ell(h_\theta(x_i), y_i) \tag{7}$$

Using this formula we can define the "adversarial-orthogonal" $\mathcal{A} : x_i \rightarrow \gamma_i \in \mathbb{R}$ to be the following min-max problem where the adversary tries to maximize‘ (8) w.r.t. to Θ_2 and the "learner" tries to minimize the same equation w.r.t. Θ_1

$$\mathcal{J}(\Theta_1, \Theta_2) := \min_{\Theta_1} \max_{\Theta_2} \sum_{i=0}^{n} \sum_{j=1}^{m} \rho(\gamma_{\Theta_2 j}, \ell(h_{\Theta_1}(x_i), y_i) \tag{8}$$

4 Experimental results

Armed with the motivation and ideas described in Section 3, we now head to demonstrate the effectiveness of our proposed AOL approach through experiments over two real datasets: Compas [15] and Adult [16]. In order to compare our results to those of [1] we chose AUC as our utility metric as it robust to class imbalance. We will present the minimum AUC over all protected groups, the average AUC over all protected groups and the AUC on the smallest protected group. Software implementing the proposed method and reproducing our experiments can be found at [`https://github.com/Shai128/236802/tree/master/236802`].

4.1 Results

Table 1: Adult

	Accuracy	min subgroup AUC	average subgroup AUC	minority subgroup AUC
Adult_AOL	0.841	**0.737**	**0.888**	**0.975**
Adult_ARL	0.841	0.700	0.882	0.963
Adult_baseline	0.841	0.709	0.886	0.956

Table 2: Compas

	Accuracy	min subgroup AUC	average subgroup AUC	minority subgroup AUC
Compas_AOL	0.678	**0.653**	**0.717**	**0.770**
Compas_ARL	0.676	0.653	0.717	0.758
Compas_baseline	0.675	0.652	0.717	0.760

Tables 1 and 2 summarizes the accuracy, min subgroup AUC, average subgroup AUC and smallest subgroup AUC over Compas and Adult dataset.

Main Results: Our main comparison is of course with ARL which AOL is based upon. Additionally, in order to provide a baseline result we report report results for the vanilla group-agnostic method (as in [1]). Both tables reports results based on average performance across runs (15 different seeds), with the best average performance highlighted in bold.

From both table it can be concluded that our AOL method improves or equals worst-case performance in both datasets, while also outperforming ARL in the smallest subgroup AUC metric. Our method manages to gain an edge over AUC in all metrics without compromising the overall accuracy of prediction over the whole dataset.

5 Conclusion

As the authors of [1] worded it perfectly "improving model fairness without directly observing protected features is a difficult and under-studied challenge". Although we managed to push the limit a bit further in this work, sadly the same can still be said today. While having this statement in mind it, we still believe that our AOL method is better the the state of the art ARL at improving both the AUC for worst-case protected groups and the AUC for the minority protected groups across multiple dataset. We believe that both AOL and ARL should be better improved and build upon in order to further advance the field, thus keeping on the pursue for fairness without access to demographics.

References

[1] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning, 2020.

[2] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[3] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126, 2016.

[4] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE, 2010.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[6] Romana Hasnain-Wynia, David W Baker, David Nerenz, Joe Feinglass, Anne C Beal, Mary Beth Landrum, Raj Behal, and Joel S Weissman. Disparities in health care are driven by where minority patients seek care: examination of the hospital quality alliance measures. *Archives of internal medicine*, 167(12):1233–1239, 2007.

[7] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. *arXiv preprint arXiv:1707.00010*, 2017.

[8] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[9] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[11] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[12] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc., 2020.

[13] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 09–15 Jun 2019.

[14] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 Jul 2018.

[15] Matias Barenstein. Propublica's compas data revisited, 2019.

[16] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.