Non-Discriminatory Machine Learning through Convex Fairness Criteria

Naman Goel and **Mohammad Yaghini** and **Boi Faltings**
Artificial Intelligence Laboratory,
École Polytechnique Fédérale de Lausanne,
Lausanne, Switzerland, 1015
{naman.goel, mohammad.yaghini, boi.faltings}@epfl.ch

Abstract

Biased decision making by machine learning systems is increasingly recognized as an important issue. Recently, techniques have been proposed to learn non-discriminatory classifiers by enforcing constraints in the training phase. Such constraints are either non-convex in nature (posing computational difficulties) or don't have a clear probabilistic interpretation. Moreover, the techniques offer little understanding of the more subjective notion of fairness.

In this paper, we introduce a novel technique to achieve non-discrimination without sacrificing convexity and probabilistic interpretation. Our experimental analysis demonstrates the success of the method on popular real datasets including ProPublica's *COMPAS* dataset. We also propose a new notion of fairness for machine learning and show that our technique satisfies this subjective fairness criterion.

1 Introduction

Algorithmic decision making powered by state of the art machine learning techniques is quickly gaining popularity in many diverse application areas including the ones which have larger social impact. Some examples of these applications are calculating risk scores of criminal recidivism, stop-and-frisk programs, predictive policing, university admissions, bank loan approvals and jobs(salary) screening/recommendation (Northpointe 2012),(Hvistendahl 2016),(Miller 2015). There have been strong evidences that the trained systems often show significant bias against certain social groups for e.g. against a certain race while calculating recidivism risk score (ProPublica 2017) or against a certain gender while recommending jobs and salaries (Datta, Tschantz, and Datta 2015). This has led to a widespread concern that machine learning systems not only appear discriminatory (which may have legal/financial consequences) but can potentially create (or enhance) imbalance in the society (ACM 2017),(WhiteHouse 2016). Consequently, there has been a lot of recent research towards making machine learning fair, accountable and transparent. In this paper, we focus mainly on the issues of non-discrimination and fairness.

The first objective, in this paper, is to design classifiers that are non-discriminatory. An ideal non-discriminatory classifier is perhaps the one in which prediction is statistically independent of the sensitive attribute (for e.g. race, gender etc). However, as we discuss in the paper, this objective is not always practical as it may cause considerable loss in utility. Another less strict non-discrimination criterion that finds its roots in legal literature and has recently gained popularity in machine learning literature is the *p-rule* criterion. The *p-rule* criterion requires that the ratio of the probabilities of favoring two different social groups is at least p. A solution that looks promising is to omit the sensitive attribute from both training and decision making phases. However, this solution doesn't work (Calders, Kamiran, and Pechenizkiy 2009). The reason being that other non-sensitive attributes are often correlated with the sensitive attributes (because of historical bias) and algorithms can learn this bias even without access to sensitive attributes. The two broad ideas in advanced approaches (Kamishima et al. 2012), (Zafar et al. 2017b), (Zafar et al. 2017c) are : i) using a regularizer term that penalizes discrimination; and ii) enforcing non-discrimination constraints on the learning objective. A general problem in these approaches is that they are non-convex in nature or achieve convexity at the cost of probabilistic interpretation.

The second objective, in this paper, is to offer a fairness interpretation of the trained classifiers. Fairness being a subjective issue, it may not be acceptable to say that a classifier which is non-discriminatory is also fair. The designer of the systems should, however, be able to offer an explanation that in what sense her classifier is fair. Fairness in machine learning is a far less studied issue, with most works using the terms fairness and non-discrimination interchangeably. Unlike *p-rule* criterion for non-discrimination, there is no popular notion of fairness in machine learning. Very recently, there has been some work towards defining the notions of envy-freeness (Zafar et al. 2017a), Rawlsian and meritocratic fairness (Joseph et al. 2016a),(Joseph et al. 2016b), (Jabbari et al. 2017) in machine learning.

To meet both these objectives, we propose a novel technique called the *weighted sum of logs* technique. Instead of aiming to directly minimize the difference in the probabilities of favoring individuals of different groups to enforce non-discrimination, we assign different weights to the favoring probabilities and minimize the negative of the weighted sum of their logs (subject to accuracy constraints). By using the logarithm and avoiding to use the difference between probabilities distributions, we manage to keep our problem convex. Unlike (Zafar et al. 2017b), we use the analytical expressions for these probabilities as a function of model parameters and don't use any proxy during the training phase. Thus, our algorithm has an easy to understand probabilistic interpretation. The solution is not limited to binary valued sensitive attributes and naturally extends to non-binary discrete values. We show through experiments on two real datasets that our technique can efficiently achieve given *p-rule* criterion with very little drop in accuracy. The technique is applicable to any learning algorithm for probabilistic discriminative models. The idea of weighted sum of logs can also be extended to the slightly different settings of minimizing mistreatment (difference in false positive or false negative rates for different social groups), which otherwise needs a convex-concave formulation (Zafar et al. 2017c). We next define a notion of fairness that we call as the 'weighted proportional fairness' in machine learning. We show that our weighted sum of logs technique guarantees weighted proportional fairness for a given set of weights.

The summary of our novel contributions in this paper are as follows :

1. We compare the two popular definitions of non-discrimination : demographic parity (Calders, Kamiran, and Pechenizkiy 2009) and equalized odds (Hardt et al. 2016) by analyzing the trade off between non-discrimination and accuracy of respective classifiers. We show that the structure of the real world (data) poses restrictions on this trade off.
2. We propose a technique called the "weighted sum of logs" for learning non-discriminatory and fair classifiers.
3. We define a new notion of fairness for machine learning, called the "weighted proportional fairness" and show that our technique satisfies the weighted proportional fairness criterion.
4. We discuss heuristics that can be used to tune our weighted sum of logs idea to satisfy a given non-discrimination requirement.
5. We show through experiments that our technique gives a performance competitive to the state-of-the art work by (Zafar et al. 2017b). While Zafar et al. have to sacrifice probabilistic interpretation to achieve convexity and similar performance, our classifier does this while retaining both these desired properties. Unlike Zafar et al., our classifier also comes with a guarantee on subjective fairness.

2 Related Work

In very recent past, there has been a large amount of research in making machine learning fair and non-discriminatory. The works that are most relevant to our work are (Calders and Verwer 2010),(Kamishima et al. 2012), (Zafar et al. 2017b) and (Zafar et al. 2017c). (Calders and Verwer 2010) propose three approaches for making naïve Bayes classifier discrimination free. In first approach, they modify the conditional probabilities of the classifier. In the second approach, they train two different models for two different social groups (for e.g. male and females). In the third approach, they introduce a latent variable for the actual class that would have been present in the training set if the set was discrimination-free. The extension of these approaches to classifiers other than naïve Bayes is not discussed. (Kamishima et al. 2012) also consider settings similar to ours. They propose a regularization technique based on mutual information of the sensitive attribute and the prediction class for probabilistic discriminative classifiers. The idea in this technique is to learn different parameter sets for different groups, while using a regularizer that penalizes the increase in probability of favoring conditional on the group membership. The regularizer term is non-convex but has a clear probabilistic interpretation. The overall training process is slow also because of multiple models being learned. (Zafar et al. 2017b) propose an interesting solution by modeling the problem as a constrained optimization problem of minimizing the loss of the classifier (as usual, for e.g. negative likelihood) but also subject it to the constraints that the absolute difference in probabilities of favoring different social groups doesn't exceed a given threshold. As this would lead to a non-convex problem, they use the covariance between the sensitive attribute and the dot product of attribute values and the parameter vector as a proxy. They minimize the loss subject to the constraints that the absolute value of the covariance doesn't exceed a certain threshold. While the technique works well, the use of proxy term strips the model of a meaningful probabilistic interpretation (which is a highly desired property in machine learning). (Zafar et al. 2017c) target a slightly different setting in which the goal is to limit *mistreatment*. They define mistreatment as the difference in false positive (or false negative) rates for different groups. The resulting constrained problem is non-convex in nature and they use a special DCCP (convex-concave) package to solve the problem.

Another line of research is on pre-processing of the training data (Kamiran and Calders 2009), (Kamiran and Calders 2012), (Dwork et al. 2012) to enable fair learning. Most notable among these is the concept of ϵ-predictability (and fairness) (Feldman et al. 2015) to measure and correct bias in the training data. This line of research is significantly different from ours because we are interested in developing algorithms that can be used to learn fair classifiers irrespective of the data being fed and can provide direct control on accuracy-fairness trade-off during training phase.

Finally, theoretical work of (Dwork et al. 2012) discusses a notion of individual fairness to guarantee that people who are similar in non-sensitive characteristics are treated in same way, (Hardt et al. 2016) discusses an approach to modify a learned classifier (post training) to make its decision look non-discriminatory and (Kleinberg, Mullainathan, and Raghavan 2017) studies the theoretical trade off and incompatibility of different notions of non-discrimination.

There has also been a significant amount of work in computational social science towards achieving fairness in data collection, social networks etc (Olteanu et al. 2016),(Tufekci 2014),(Morstatter and Liu 2017),(Hajian, Bonchi, and Castillo 2016).

3 Preliminaries

We consider the settings of learning to predict a discrete target variable y_m, for a sample m with a set of features $X_m \in \mathbb{R}^d$. Every sample has a discrete valued sensitive attribute $z_m \notin X_m$. The examples of such sensitive attributes include race, color, religion, gender etc.

Let's, for easier understanding, assume that the protected attribute is binary valued. For e.g., if the protected attribute is gender, then it may take values 'male' (0) or 'female' (1). Further, assume that the prediction space is binary i.e. $y_m \in \{+1, -1\}$. We will be using $+$ for $+1$ and $-$ for -1 in the rest of the paper.

World Bias Matrix The world bias matrix is a 2×2 matrix W, where $W_{ij} = W[i][j]$ is the probability of the true class of a data sample being $i \in \{+, -\}$, given that the value of the protected attribute is $j \in \{0, 1\}$.

The world bias matrix represents the inherent real world bias in the true class of a *test*[1] sample for different values of the protected attribute. We will be using notations $W[i][j]$ and W_{ij} interchangeably.

Classifier Confusion Matrix The confusion matrix of a classifier is a 2×2 matrix C, where $C_{ij} = C[i][j]$ is the probability of the predicted class $\hat{y}$ of a test data sample being $i \in \{+, -\}$, given that the true class y is $j \in \{+, -\}$.

It remains to define what makes a classifier non-discriminatory. There are various ways to measure the difference in the treatment received by different groups, leading to different ways of defining non-discrimination. The strongest definition of non-discrimination is the demographic parity.

Demographic Parity (Calders, Kamiran, and Pechenizkiy 2009) Demographic parity requires that the prediction of a classifier is independent of the sensitive attribute i.e.

$$P(\hat{y} = +|z = 1) = P(\hat{y} = +|z = 0)$$

A weaker definition of non-discrimination is equalized odds.

Equalized Odds (Hardt et al. 2016) Equalized odds requires that the prediction of a classifier is independent of the sensitive attribute conditional on the true class i.e.

$$P(\hat{y} = +|y, z = 1) = P(\hat{y} = +|y, z = 0), y \in \{0, 1\}$$

There are several other (more restrictive) ways of defining non-discrimination such as equal false negative rate for different values of sensitive attribute (Zafar et al. 2017c), which are important in specific application scenarios. In this paper, we limit ourselves to the above two common notions of non-discrimination namely demographic parity and equalized odds.

Discussion Judging which of these two definitions (demographic parity or equalized odds) should be used to measure and enforce non-discrimination, is a policy-making issue. For various reasons, demographic parity is an important definition but difficult to satisfy without causing significant loss in utility. Demographic parity is specially important when the decisions involved are the ones that can have a potentially big impact on the society in the long term or can even create a feedback cycle of injustice by affecting the training data samples to be seen in future. In this paper, we strive to develop techniques that can meet stronger requirements of non-discrimination and not settle for weaker ones. In the rest of the paper, *we call a classifier non-discriminatory if and only if it satisfies demographic parity*. Our goal is to design classifiers that are as accurate and as non-discriminatory as possible on the test samples. Before discussing how to achieve this, we look at the loss that an approach satisfying demographic parity must have to face as compared to an approach that only satisfies equalized odds.

4 Accuracy-Discrimination Trade-off

Classifier Bias Matrix The bias matrix of a classifier is a 2×2 matrix D, where $D_{ij} = D[i][j]$ is the probability of the classifier predicting the class $\hat{y}$ of a data sample as $i \in \{+, -\}$, given that the value of the protected attribute is $j \in \{0, 1\}$.

Lemma 1 *The bias matrix of a classifier satisfying equalized odds is :*

$$D = CW$$

where C is the confusion matrix of the classifier and W is the world bias matrix.

Proposition 1 *If the world is not biased, any classifier with arbitrary confusion matrix and satisfying equalized odds, is also non-discriminatory.*

The above proposition follows by setting $W[+][1] = W[+][0]$, which gives equal values for $D[+][1]$ and $D[+][0]$.

Proposition 2 *If the world is biased, a classifier with identity confusion matrix and satisfying only equalized odds, can not be non-discriminatory.*

Clearly, if the classifier is always accurate i.e. if the confusion matrix C is identity, then the bias matrix of the classifier is equal to the world bias matrix. Thus, unless the world is non-discriminatory, a perfectly accurate classifier satisfying equalized odds will surely be discriminatory.

Proposition 3 *If the world is biased, a classifier satisfying equalized odds is non-discriminatory if and only if*

$$C[+][+] = C[+][-]$$

This subsumes the naïve ways of achieving non-discrimination, for e.g., always classifying as $+$, always classifying as $-$ and classifying randomly i.e. $C[+][+] = C[-][-] = 0.5$.

Impossibility of Non-Discrimination Proposition 3 effectively renders a non-discriminatory classifier useless for practical purpose by requiring the predicted class to be independent of the true class. Thus, we arrive at a negative conclusion that *any practically useful classifier satisfying equalized odds in a biased world can't be non-discriminatory*. Note that while we discussed only the settings with binary sensitive attribute and binary classification, the conclusions can be generalized to non-binary setting by simply increasing the dimensionality of the bias and confusion matrices.

[1] We use the word 'test' samples for the data instances on which classifier is actually used for prediction.

Approximately Non-Discriminatory Classifiers Proposition 3 only establishes the impossibility of achieving perfect non-discrimination in a biased world. We next explore the possibility of existence of classifiers that are approximately non-discriminatory in biased world and still useful in practice.

p-rule A popular way to measure and limit discrimination, called the *p-rule* (Biddle 2006), is given as follows.

$$min\Big(\frac{P(\hat{y}=+|z=1)}{P(\hat{y}=+|z=0)}, \frac{P(\hat{y}=+|z=0)}{P(\hat{y}=+|z=1)}\Big) \geq p$$

where $P(\hat{y}=+|z=1)$ is the probability of the classifier predicting the class of a test sample as $+$, given that the value of the protected attribute is 1 and $P(\hat{y}=+|z=0)$ is the probability of the classifier predicting the class of a test sample as $+$, given that the value of the protected attribute is 0. p is set to a positive value less than 1. A *p-rule* value[2] equal to 1 implies non-discriminatory classifier satisfying demographic parity. The general requirement under law is to have *p-rule* value above 0.8 for the classifier to be considered reasonably non-discriminatory.

Theorem 1 *If the world satisfies p-rule, then any classifier with arbitrary confusion matrix and satisfying equalized odds satisfies the p-rule.*

Theorem 2 *If the world doesn't satisfy p-rule, then the following holds for a classifier satisfying equalized odds :*
If $\frac{W_{+1}}{W_{+0}} < p$*, then the classifier satisfies the p-rule if and only if*

$$\frac{-(1-p)}{W_{+0}-pW_{+1}} \leq \frac{C[+][+]}{C[+][-]} - 1 \leq \frac{-(1-p)}{W_{+1}-pW_{+0}}$$

If $\frac{W_{+0}}{W_{+1}} < p$ *, then the classifier satisfies the p-rule if and only if*

$$\frac{-(1-p)}{W_{+1}-pW_{+0}} \leq \frac{C[+][+]}{C[+][-]} - 1 \leq \frac{-(1-p)}{W_{+0}-pW_{+1}}$$

$\frac{C[+][+]}{C[+][-]} - 1$ is the relative difference between probabilities of predicting a $+$ class for a true $+$ class sample and predicting a $+$ class for a true $-$ class sample. Thus, this quantity can be seen as measuring the level of 'correlation' in the classifier's predicted class and the true class. Theorem 2 provides bounds on this correlation. This correlation directly affects the accuracy of the classifier. Clearly, depending on the bias in the world (inversely proportional), the theorems require some accuracy to be sacrificed but also provide hope for getting around the impossibility result and building useful and approximately non-discriminatory classifiers. These results will help us justify the drop in accuracy that we will observe later in our experimental analysis. It may be noted that the condition in Theorem 2 is never satisfied by a perfectly accurate classifier and always satisfied by a useless classifier, making Propositions 2 and 3 special cases.

There is another (less common) way to measure and limit discrimination and is called the *CV-score* (Calders and Verwer 2010). *CV-score* measures the absolute difference in probabilities of the classifier favoring different groups instead of the ratio. Similar conclusions about trade off between accuracy and *CV-score* can also be drawn (see extended version of the paper).

5 Weighted Sum of Logs Technique

Our technique, called the *weighted sum of logs* technique, can be used to lower discrimination in classifiers. This technique is applicable to probabilistic discriminative models. More specifically, we will be discussing the logistic regression classifier in the paper. Let θ be the variable (non-discriminatory parameter weight vector) for logistic regression that we are interested in learning and θ^v be the parameter weight vector (discriminatory) learned with vanilla logistic regression. Here, vanilla logistic regression may also involve the regularized (for e.g. L2) version of logistic regression classifier. We limit our discussion to the case of only a single sensitive attribute z, which takes discrete values. For e.g., if z is race, then we may be interested in lowering discrimination w.r.t. black, white and asian groups.

To learn the optimal non-discriminatory θ, we solve the following optimization problem :

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & \sum_{i=1}^{N} w_{g(i)} \cdot \log \hat{P}_i^+(\theta) \\ \text{subject to} \quad & \mathcal{L}(\theta) \leq (1+\delta)\mathcal{L}(\theta^v) \end{aligned} \tag{1}$$

where $w_{g(i)}$ is the empirical estimate of historical bias in the training data (of size N) against the group (for e.g. race) of the i^{th} individual. More precisely,

$$w_{g(i)} = \frac{\sum_{m=1}^{N} \mathbb{1}_{z_m=g(i)} \cdot \mathbb{1}_{y_m=-}}{\sum_{m=1}^{N} \mathbb{1}_{z_m=g(i)}} \cdot \frac{1}{\sum_{m=1}^{N} \mathbb{1}_{z_m=g(i)}} \tag{2}$$

$\log \hat{P}_i^+(\theta)$ is the log of the empirical estimate of favorable bias of the classifier towards i, i.e.,

$$\log \hat{P}_i^+(\theta) = -\log(1+e^{-X_i^T\theta}) \tag{3}$$

$\mathcal{L}(\theta)$ is the negative log likelihood given by :

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log(1+e^{-y_i(X_i^T\theta)})$$

It is important to note the difference between $w_{g(i)}$ and $\hat{P}_i^+(\theta)$. $w_{g(i)}$ is only the historical estimate of unfavorable bias against a group and hence depends only on the true class present in training data. On the other hand, $\hat{P}_i^+(\theta)$ is the empirical estimate of the favorable bias of the classifier being learned and hence depends on the predicted class. $w_{g(i)}$ is a constant, independent of the optimization variable θ but $\hat{P}_i^+(\theta)$ depends directly on θ.

[2] We refer to the left hand side of the inequality in *p-rule* as the *p-rule* value.

Also note that $\underset{\theta}{\text{maximize}} \sum_{i=1}^{N} w_{g(i)} \cdot \log \hat{P}_i^+(\theta)$ is equivalent to $\underset{\theta}{\text{minimize}} \sum_{i=1}^{N} -w_{g(i)} \cdot \log \hat{P}_i^+(\theta)$. Thus, we have a convex optimization problem in θ.

The constraints restrict the optimization space to the set of classifier that don't suffer a loss more than $(1+\delta)$ times the loss the vanilla classifier which doesn't enforce non-discrimination. δ is a hyper-parameter which allows to set the desired level of drop in accuracy that a decision maker is willing to suffer to maximize the non-discrimination objective. We will show through experiments on real datasets that one can use this formulation to tune the classifier for satisfying a given *p-rule*. The reason why this technique works in practice can be intuitively understood in the following manner. If the training data shows historical bias i.e. weights $w_{g(i)}$ are different for different groups, then the optimizer gets different utility gains from increasing the $\log \hat{P}_i^+(\theta)$ for individuals of different groups. In other words, the group with higher weight ($w_{g(i)}$) is compensated more by increasing $\log \hat{P}_i^+(\theta)$ of its individuals more than others. The weight $w_{g(i)}$ will be higher for historically disfavored and minority group. The first term $\frac{\sum_{m=1}^{N} \mathbb{1}_{z_m=g(i)} \cdot \mathbb{1}_{y_m=-}}{\sum_{m=1}^{N} \mathbb{1}_{z_m=g(i)}}$ in the weight measures disfavoring bias against the group. The second term $\frac{1}{\sum_{m=1}^{N} \mathbb{1}_{z_m=g(i)}}$ can be seen as a normalization factor for the whole group since there are equal number of log terms in the numerator for that group. This factor will make the weight higher for the group which has less samples in the training data.

Before moving to the experiments results, we look at an interesting *fairness* interpretation of our technique.

6 Fairness Interpretation

So far, we have used demographic parity and equalized odds as definitions of only non-discrimination and not discussed the issue of "fairness". The definitions of non-discrimination in machine learning often find their motivation (Barocas and Selbst 2016),(Zarsky 2016),(Roemer 2009) in the need for preventing the automatic decision making systems from disfavoring (or appearing to disfavor) particular social groups because this has social, legal and even financial consequences. However, the definition of fairness in machine learning remains largely unclear. Fairness is a very subjective issue and using the terms fairness and non-discrimination interchangeably perhaps over simplifies the issue. But both non-discrimination and fairness are important issues. A classifier that enforces non-discrimination must be able to offer a fairness interpretation of its training process and its decisions (even if the explanation is subjective). In this paper, we draw inspiration from literature on rate control in communication networks (Kelly 1997) to offer fairness interpretation of a machine learning classifier.

A machine learning classifier f is said to be ***proportionally fair*** if for any other *allowed* classifier u,

$$\sum_{i=1}^{N} \frac{(P_u)_i^+ - (P_f)_i^+}{(P_f)_i^+} \leq 0 \tag{4}$$

where $(P_c)_i^+$ is the probability of classifier c favoring i as discussed in Section 5 (Equation 3). The inequality can also be seen as the empirical average of the proportional changes in favoring probabilities being non-positive. Thus, in expectation no other *allowed* classifier can cause a positive proportional change in probability of favoring an individual.

In simpler words, if out of two allowed actions, an action causes small relative disadvantage to the one individual as compared to the other action but much more relative advantage to another individual, then the action is proportionally fair. There are two interesting points that should be noted about this definition. The first is that it takes the summation of relative changes and prefers the classifier that gives best *global* (across all individuals) impact. The second is that it calculates this global impact based on relative changes in probabilities and not absolute changes. It can be shown that such a classifier is *Pareto optimal* and has many other interesting game theoretic properties (which are outside the scope of this paper : see (Bertsimas, Farias, and Trichakis 2011)).

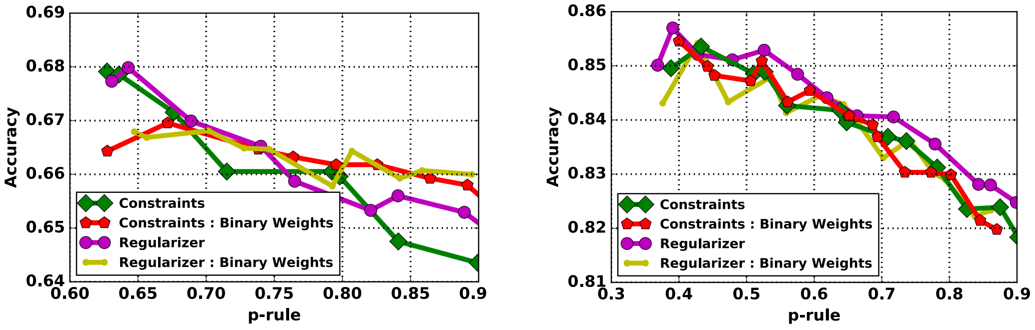
This definition of proportional fairness can be extended to weighted proportional fairness as follows. A classifier f is called to be ***weighted proportionally fair*** if for any other *allowed* classifier u,

$$\sum_{i=1}^{N} w_i \frac{(P_u)_i^+ - (P_f)_i^+}{(P_f)_i^+} \leq 0 \tag{5}$$

Here w_i's are interpreted as *costs* paid by different individuals in history. Thus, if an individual belongs to a social group that has been more unfairly treated in history (i.e. has paid more cost), gets more weight while calculating the sum. As cost is a subjective term (i.e. hard to measure), our definition of weight proportional fairness also remains subjective as expected.

Weighted Sum of logs : The *weighted sum of logs* technique guarantees a classifier that is weighted proportionally fair among all classifiers that suffer a loss not more than $(1+\delta)$ times the loss of the vanilla classifier. This implies that if a decision maker is confined to a set of machine learning classifiers that are no worse than a certain threshold of the vanilla classifier (because of utility concerns), then the classifier learned by our technique is guaranteed to be weighted proportionally fair among all such *allowed* classifiers. The weights are the historical bias in the training data against different social groups to which individuals belong.

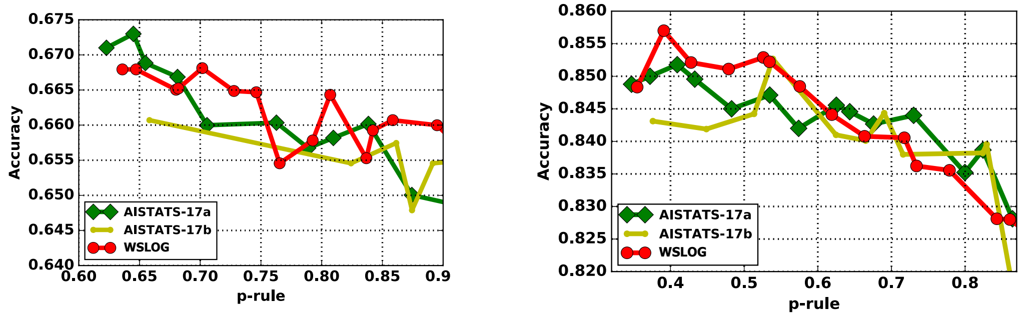
Having discussed the fairness interpretation of the weighted sum of logs technique, we now discuss the experimental results to evaluate the practical suitability of the technique in lowering discrimination.



(a) COMPAS Data

(b) Adult Data

Figure 1: Comparison of different weighted sum of logs heuristics



(a) COMPAS Data

(b) Adult Data

Figure 2: Comparison with baseline

7 Experimental Evaluation

Heuristics

As mentioned earlier, the weighted sum of logs technique needs to be tuned for meeting a given *p-rule* requirement while ensuring as little drop in accuracy as possible. We evaluate three additional techniques that can used for this.

Negative Weighted Sum of Logs as Regularizer - In this heuristic, instead of solving a constrained problem, we directly minimize the following objective :

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}(\theta) + \gamma \cdot \sum_{i=1}^{N} -w_{g(i)} \cdot \log \hat{P}_i^+(\theta)$$

The parameter γ is analogous to the parameter δ in Equation 1 and controls the trade-off between accuracy and non-discrimination.

Exponentially scaled weights - Recall that it is because of the difference in the historical weights $w_{g(i)}$ that the optimizer gets different utility gains by increasing favoring bias towards different groups. However, this difference may not be sufficient to meet a given *p-rule* requirement on the trained classifier. In this heuristic, we introduce another hyper-parameter k, which is used to make the difference in weights $w_{g(i)}$ of different groups more pronounced. Specifically, we replace the weight $w_{g(i)}$ by $(w_{g(i)})^k$.

For an example, if we consider the above mentioned regularizer version, then our optimization problem becomes :

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}(\theta) + \gamma \cdot \sum_{i=1}^{N} -(w_{g(i)})^k \cdot \log \hat{P}_i^+(\theta)$$

Note that as k increases, the overall magnitude of weighted sum becomes small. Hence, γ needs to be scaled up appropriately to balance the effect.

Binarized weights - Since tuning hyper-parameter k causes additional overhead, we also evaluate another heuristic which is an extreme case of exponentially scaled weights. It is applicable to the case when sensitive attribute is binary valued i.e. $\text{domain}(g(i)) = \text{domain}(z) = \{0, 1\}$. In this heuristic, we set the weight for one of the groups as 1 and the other as 0. For e.g. if $w_{female} > w_{male}$, then w_{female} is set to 1 and w_{male} to 0. The optimization problem is thus :

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}(\theta) + \gamma \cdot \sum_{i=1}^{N} -\mathbb{1}_{w_{g(i)} > w_{1-g(i)}} \cdot \log \hat{P}_i^+(\theta)$$

Note that in all of these heuristics, a common hyper-parameter γ continues to provide direct control on accuracy similar to the hyper-parameter δ in the original technique.

Datasets

We work with two datasets in our experimental evaluation. The first dataset is the ProPublica's *COMPAS* dataset (Larson et al. 2016). It contains data from Broward County, Florida, compiled by ProPublica. The goal is to predict whether a convicted individual would commit a crime again in the following 2 years or not. The dataset has personal information about the offenders such as gender, race, age and criminal history etc. The prediction class is whether or not these individuals actually recidivated within 2 years after the screening. We call recidivating as negative class and not recidivating as positive class. We consider offenders whose race was either black or white. The dataset was filtered for quality reasons such as missing values etc. In this sample of 6150 points, 61% while people belonged to the positive class and 49% black people belonged to the positive class. Total number of black examples is 3696 and number of white examples is 2454. A vanilla logistic regression classifier trained on this data classifies 74% of while examples and 48% of black examples as belonging to positive class.

The second dataset is the popular *Adult Income* dataset (UCI 1996). This dataset has $45,222$ instances and 14 attributes such as age, occupation, working hours per week and work class etc. The prediction classes are whether a subjects incomes is above or below 50K USD. We call income above 50K USD as positive class and below 50K USD as negative class. Gender is treated as a sensitive attribute in this dataset. We use a random sample of size $10,000$ in our experiments. In this sample, 12% of females belonged to the positive class and 31% of males belonged to the positive class. Total number of female examples are 3303 and total number of male examples is 6397. A vanilla logistic regression classifier trained on this data classifies 24% of males and 8% of females as belonging to positive class.

Results

Results of our experiments are presented in Figures 1 and 2. All the results presented in this paper are k-fold cross-validated. In Figure 1a, we compare the performance of the different techniques we discussed on the COMPAS dataset. 'Constraints' refer to the original optimization problem discussed in Section 5. The regularized version of the problem is called 'Regularizer'. In both these techniques, we used exponentially scaled weights as discussed earlier. Finally, the corresponding versions with binarized weights are called 'Constraints : Binary Weights' and 'Regularizer : Binary Weights' respectively. As one can see in the figure that all four techniques give competitive results with the binarized weights technique being at marginal advantage over the rest. We manage to bring up the *p-rule* value from about 64% to 90% while bringing the accuracy down marginally from around 67% to 66%. A similar trend can be observed on the Adult dataset as well. The results are given in Figure 1b. On this dataset also, all four techniques given competitive results. We can bring up the *p-rule* value from about 34% to 90% at the cost of a small drop in accuracy (from around 85% to 82%). It should be noted that a drop in accuracy was indeed expected as discussed earlier in Theorem 2 because world in these datasets is biased.

We also compare our approach with the techniques proposed in (Zafar et al. 2017b), which have been briefly discussed in Section 2. There are two heuristics proposed in that paper : minimizing loss subject to constraints on covariance and minimizing covariance subject to constraints on loss. We refer to these two techniques as *AISTATS-17a* and *AISTATS-17b* respectively. We used the publicly available implementation by the authors for comparison. Our implementation uses the same set of libraries and code for optimization and cross validation etc. Thus, the experimental conditions are identical. As we can see in Figure 2a, both techniques are competitive on COMPAS data with our method being marginally better than the baseline. The same is true about the Adult dataset as clear from Figure 2b. It should be noted that (Zafar et al. 2017b) had also compared their method with the non-convex regularization approach proposed in (Kamishima et al. 2012) and had observed a similar trend about the two approaches giving identical results. By transitivity, we can compare our approach with (Kamishima et al. 2012).

Hence, the experimental results confirm that our method can achieve non-discrimination without significant loss in accuracy, with performance being comparable to the other methods. However, our method achieves this without sacrificing convexity and probabilistic interpretation, while guaranteeing weighted proportional fairness.

It may be worth noting that the weights used in the optimization problem are constants, independent of the optimization variable and depend only on the training data. These weights need not be recomputed during every iteration. Thus, the weights cause no computational overhead. Finally, binarized version not only gives competitive performance but also has computational advantages. First, it has only one parameter, making hyper-parameter tuning through cross-validation fast. Second, it sets one of the terms in weighted sum of logs to 0 and thus, avoids computing it.

8 Conclusions

In this paper, we addressed an important problem of gender, race, religion based discrimination by machine learning systems. With many incidents of discrimination being noticed frequently, the problem is getting attention of not only computer science researchers and government organizations but also general public and users of the machine learning powered services. In this paper, we discussed the definitions of non-discrimination and compared their inherent accuracy-(non) discrimination trade-off. We defined a notion of fairness in machine learning called the proportional fairness. We further discussed the idea of weighted proportional fairness that gives more weight to the group that has been treated unfairly in the past. We proposed a novel technique called the weighted sum of logs that uses a convex fairness criterion to enforce non-discrimination. Our formulation has a clear probabilistic interpretation and results in a convex optimization problem, thus avoiding the issues in previous approaches. We showed that this technique also guarantees a weighted proportionally fair classifier. We demonstrated through experiments on very relevant datasets that our technique achieves non-discrimination without much loss in accuracy.

References

ACM. 2017. Statement on algorithmic transparency and accountability.

Barocas, S., and Selbst, A. D. 2016. Big data's disparate impact. 104 california law review 671.

Bertsimas, D.; Farias, V. F.; and Trichakis, N. 2011. The price of fairness. *Operations research* 59(1):17–31.

Biddle, D. 2006. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd.

Calders, T., and Verwer, S. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21(2):277–292.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, 13–18. IEEE.

Datta, A.; Tschantz, M. C.; and Datta, A. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015(1):92–112.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ACM.

FATML. 2017. http://www.fatml.org.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

Hajian, S.; Bonchi, F.; and Castillo, C. 2016. 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*, 2125–2126. ACM.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

Hvistendahl, M. 2016. Can 'predictive policing' prevent crime before it happens? *AAAS Science Magazine*.

Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2017. Fairness in reinforcement learning. In *International Conference on Machine Learning*, 1617–1626.

Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2016a. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016b. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, 325–333.

Kamiran, F., and Calders, T. 2009. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, 1–6. IEEE.

Kamiran, F., and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1):1–33.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases*.

Kelly, F. 1997. Charging and rate control for elastic traffic. *Transactions on Emerging Telecommunications Technologies* 8(1):33–37.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent trade-offs in the fair determination of risk scores. *8th Innovations in Theoretical Computer Science Conference (ITCS)*.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. https://github.com/propublica/compas-analysis, 2016.

Miller, C. C. 2015. Can an algorithm hire better than a human? *The New York Times*.

Morstatter, F., and Liu, H. 2017. Tutorial : Social data bias in machine learning: Impact, evaluation, and correction. *AAAI*.

Northpointe. 2012. Compas risk and need assessment system, url : http://www.northpointeinc.com/files/downloads/faq document.pdf.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kiciman, E. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries.

ProPublica. 2017. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

Roemer, J. E. 2009. *Equality of opportunity*. Harvard University Press.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *ICWSM* 14:505–514.

UCI. 1996. Adult income data set. url: https://archive.ics.uci.edu/ml/datasets/adult.

WhiteHouse. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; Gummadi, K. P.; and Weller, A. 2017a. From parity to preference-based notions of fairness in classification. *Neural information processing systems*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017b. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *26th International World Wide Wed Conference (WWW)*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017c. Fairness constraints: Mechanisms for fair classification. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Zarsky, T. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values* 41(1):118–132.