Fairness in Machine Learning Using a Logistic Decision Rule

Eric Landgrebe, Vineet Parikh, Christopher Qian

November 2019

1 Introduction

Fairness in machine learning is a topic that has seen an explosion in interest as machine learning becomes more widespread. The use of machine learning in important decisions, such as the COMPAS model for determining risk of criminal recidivism, has come under scrutiny lately because of the discrepancy in predictions across groups. However, fairness metrics can be flawed and fundamentally incompatible, as seen in Kleinberg et al [4], Heidari et al [6], and Gajane et al[5], and the extent of how flawed individual metrics can be, and how well the differences in metrics between groups can be minimized, hasn't been fully explored.

In our project, we will consider the perspective of a decision making analyst, who has to label an individual (this could be, for example, determining whether to hire them, or give them a credit card). We will assume that due to federal law, the analyst cannot take into consideration the person's group when making a decision (the classifier is unaware). Let's say that the analyst sees in an audit that the company's current decisions are unfair across groups. We will explore a natural attempt at improving fairness: by looking at the features for which the disadvantaged group has a lower probability of having, and decreasing how much they weigh those features. How does this intuitive approach affect fairness?

1.1 Related Work

Current research has shown the limitations of unaware classifiers. In 2008, Pedreshi et al. showed that taking away the protected attributes in a classifier does not result in a fair classifier, and propose a new classification scheme [3]. However, as we have seen in class, there are also disadvantages of classifiers that do take into account protected attributes. Fryer et al. demonstrate the economic issues of unaware classifiers[2]; meanwhile Goldin and Rouse demonstrate that it can actually work quite well in certain situations by analyzing a case study on musical audtions [1]. Most current research is focused on aware classifiers, because of the limitations of unaware classifiers, but since we currently have laws that require unawareness, it is important to analyze the performance of unaware classifiers. Chen et al. present recent theoretical analysis of differences in *demographic disparity* in unaware classifiers [7]. However, little work has been done on the theoretical results of changing classifier weights on measures of fairness like equal opportunity and predictive value parity. Throughout this paper, we use the notation and terminology of Hediari et al [6].

We currently consider three datasets: simulated data, a German creditability dataset which matches features of individuals to whether they are "credit-able", and the Washington State HDMA dataset which matches features of individuals to whether they can get mortgages.

2 Simulated Data

2.1 Model

We are interested in a supervised learning setting, where an analyst is to classify an individual's label. In particular, we consider a loan repayment analyst who wants to determine whether or not an individual will default on their loan, and we assume that the

analyst will classify based off of a logistic decision rule:

$$P(\hat{Y} = \hat{y}|X = x) = \frac{1}{1 + \exp(-y(\hat{w}^T x + \hat{b})}$$

where $\hat{y} \in \{-1, +1\}$. In particular, we assume an unaware classifier, so

$$P(\hat{Y}_1 = 1|X_1 = x) = P(\hat{Y}_2 = 1|X_2 = x)$$

for groups 1 and 2, as a logistic decision rule intuitively fits how reviewers might grade applications for loans, credit, etc. In our simulated data, we assume that $X = [X_1, \ldots, X_k]$, where each $X_i \sim \text{Bern}(p_i)$ and are mutually independent. As a shorthand, we will say $X \sim \text{Bern}(p_1, \ldots, p_n)$. That is, we assume that the analyst has a dataset such that all of the features are Bernoulli. We believe this assumption is not overly simplistic because a typical dataset consists of mostly binary and categorical variables (as seen in the datasets we will analyze). The analysis we do when assuming Bernoulli random variables can be extended to Categorical random variables in the future. We also assume that an individuals true label is also determined by a logistic decision rule:

$$P(Y = y|X = x) = \frac{1}{1 + \exp(-y(w^{\natural T} x + b^{\natural})}$$

so the analyst's weight vector $\hat{w}$ may be different than the true weight vector $w^{\natural}$. In addition, the true label does not use the person's group:

$$P(Y_1 = 1|X_1 = x) = P(Y_2 = 1|X_2 = x)$$

2.2 Definitions

Suppose that we have two groups, whose features have potentially different distributions: $X_1 \sim \text{Bern}(p_{11}, \ldots, p_{1k})$, $X_2 \sim \text{Bern}(p_{21}, \ldots, p_{2k})$. We define group i's *opportunity* to be $P(\hat{Y}_i = 1|Y_i = 1)$, and *predictive value* to be $P(Y_i = 1|\hat{Y}_i = 1)$. We want the difference in opportunity:

$$P(\hat{Y}_1 = 1|Y_1 = 1) - P(\hat{Y}_2 = 1|Y_2 = 1)$$

and the difference in predictive value:

$$P(Y_1 = 1|\hat{Y}_1 = 1) - P(Y_2 = 1|\hat{Y}_2 = 1)$$

to both be as small as possible, given that the analyst uses the same $\hat{w}$ to classify both groups (the classifier is unaware). This has the application in settings where the decision making analyst is unable to use the applicant's group in the classification process, but they still have access to it (perhaps for auditing purposes).

In addition, we consider two other important quantities. The difference in probability of the true label:

$$P(Y_1 = 1) - P(Y_2 = 1)$$

And the difference in probability of the predicted label:

$$P(\hat{Y}_1 = 1) - P(\hat{Y}_2 = 1)$$

When this difference is 0, we say that the classifier satisfies *statistical parity*.

2.3 Calculating Probabilities

In this setting, we are able to easily calculate each of the fairness metrics because we have a closed form solution for each:

$$\begin{aligned} P(\hat{Y} = \hat{y}|Y = y) &= \sum_{x \in \mathcal{X}} P(\hat{Y} = \hat{y}|Y = y, X = x) P(X = x|Y = y) \\ &= \frac{\sum_{x \in \mathcal{X}} P(\hat{Y} = \hat{y}|X = x) P(Y = y|X = x) P(X = x)}{P(Y = y)} \end{aligned}$$

and

$$\begin{aligned} P(Y = y|\hat{Y} = \hat{y}) &= \sum_{x \in \mathcal{X}} P(Y = y|\hat{Y} = \hat{y}, X = x) P(X = x|\hat{Y} = \hat{y}) \\ &= \frac{\sum_{x \in \mathcal{X}} P(Y = y|X = x) P(\hat{Y} = \hat{y}|X = x) P(X = x)}{P(\hat{Y} = \hat{y})} \end{aligned}$$

since we assume X is Bernoulli distributed, we can simply iterate over all k length vectors of $0's$ and $1's$ to compute these quantities.

2.3.1 Example

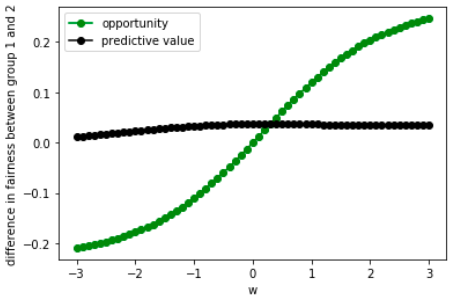
Suppose $X_1 \sim \text{Bern}(0.7, 0.2, 0.2, 0.2)$ and $X_2 \sim \text{Bern}(0.2, 0.2, 0.2, 0.2)$, and $w^\natural = [0.3, 0.2, -0.1, 0.4]$, $b^\natural = -0.3$. Suppose $\hat{w} = w^\natural$, $b^\natural = -0.3$, so the analyst is using the Bayes Optimal Classifier. The statistics are as follows:

$$P(Y_1 = 1) - P(Y_2 = 1) = 0.037 \tag{1}$$

$$P(\hat{Y}_1 = 1 | Y_1 = 1) - P(\hat{Y}_2 = 1 | Y_2 = 1) = 0.037 \tag{2}$$

$$P(Y_1 = 1 | \hat{Y}_1 = 1) - P(Y_2 = 1 | \hat{Y}_2 = 1) = 0.036 \tag{3}$$

We see that group 1 has an inherent advantage over group 2 from (1). Intuitively, to improve fairness, the analyst can lower the value of $\hat{w}_1$, since group 1 has a higher probability of having feature 1 than group 2. Let $\hat{w}_i = w_i^\natural$ for $i = 1, 2, 3$. The following is a plot where we vary $\hat{w}_1$ from -3 to 3.



In this example, the difference in opportunity between group 1 and group 2 increases monotonically as $\hat{w}_1$ increases. Note that at $\hat{w}_1 = -3$, group 2 has an advantage over group 2, because it has a lower probability of having the first feature. Also, the difference in predictive value does not change very much. We also see that although the curve is always decreasing, the function tapers off with large values of $\hat{w}_1$, so the maximum difference in opportunity is bounded. We see that to achieve the lowest difference in opportunity, we should set $\hat{w}_1$ to be around 0.1.

This intuition leads into the following lemmas, which assume independent feature distributions:

Lemma 2.1. *Suppose $X_1 \sim Bern(p_1, p_2, \ldots, p_k)$ and $X_2 \sim Bern(\bar{p}_1, p_2, \ldots, p_k)$, with $p_1 > \bar{p}_1$ and $w_1^\natural > 0$. Then increasing $\hat{w}_1$ increases the difference in opportunity between group 1 and group 2 monotonically.*

Lemma 2.2. *Let $m_i = \max\{P(Y_i = 1 | X = x) : x \in \mathcal{X}\}$ and $l_i = \min\{P(Y_i = 1 | X = x) : x \in \mathcal{X}\}$ for $i = 1, 2$. Then the maximum difference in predictive value is $\max(m_1 - l_2, m_2 - l_1)$.*

Proof. The proofs can be found in the Appendix. □

Lemma 2.1 makes strong assumptions on the distributions of the features for both groups, but experimental results still show that varying an individual w_i causes the difference in opportunity to change monotonically, provided that the differences in probability for the feature for the two groups is sufficiently large. This has a practical application for an analyst wanting to improve fairness on his classifier: decrease the weight on the feature for which the advantaged group has a higher probability in to decrease the difference in opportunity between the groups, until the difference is 0, or the difference tapers off. Also, note that although Lemma 2.1 assumes the feature with the different probabilities across groups is feature 1, this assumption is not important, and the fact holds if the probability for feature k is different across groups, as long as all the other probabilities are the same.

In addition, Lemma 2.2 provides intuition as to why optimizing for the difference in opportunity is better than predictive value: the difference is bounded by a value that does not depend on $\hat{w}$.

2.3.2 Example

Suppose $X_1 \sim \text{Bern}(0.1, 0.2, 0.2, 0.2)$ and $X_2 \sim \text{Bern}(0.1, 0.2, 0.2, 0.6)$. Using Lemma 2.2, the maximum difference in predictive value is:

$$P(Y_1 = 1|X_1 = [1, 1, 1, 1, 1]) - P(Y_2 = 1|X_2 = [0, 0, 0, 0, 0]) = 0.475 - 0.332 = 0.143$$

Note that this difference does not depend on $w^\natural$. Now suppose $w^\natural = [0.2, 0.1, 0.2, 0.1, -0.7]$. How unfair can an analyst make the classifier? Suppose the analyst decides the weight the one feature that group 2 has an advantage over group 1 in: feature 5, and uses the weight vector $\hat{w} = [0.2, 0.1, 0.2, 3, -0.7]$. Then we have:

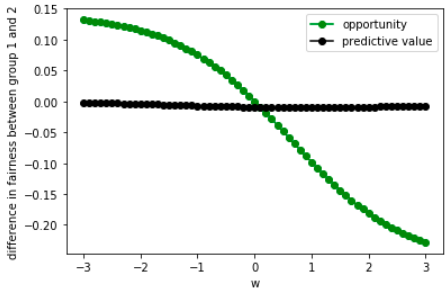
$$P(\hat{Y}_2 = 1|Y_2 = 1) - P(\hat{Y}_1 = 1|Y_1 = 1) = 0.699 - 0.470 = 0.288$$

So we see that the maximum difference in opportunity easily exceeds the maximum difference in predictive value, even though the bound in Lemma 2.2 greatly overestimates the potential difference in predictive value. Indeed, the true difference is

$$P(Y_2 = 1|\hat{Y}_2 = 1) - P(Y_1 = 1|\hat{Y}_1 = 1) = 0.370 - 0.361 = 0.009$$

Intuitively, we see that the difference in predictive value is more robust to changes in $\hat{w}$ because $P(\hat{Y})$, which is the quantity that the analyst has control over, occurs in the denominator and numerator, so it is more difficult to change a group's predictive value. For the difference in opportunity, $P(\hat{Y})$ only occurs in the numerator.

When we let $\hat{w}_i = w_i^\natural$ for $i = 1, 2, 3$ and vary $\hat{w}_4$ from -3 to 3, it produces the following graph:



Since the difference in predictive value is more robust to changes in the classifier weight vector, we argue that the analyst wanting to make their classifier more fair should focus on decreasing the difference in opportunity, which can be changed more.

2.4 When Group Probabilities Differ Significantly

So far, our simulated data involved two groups such that the probability of each group having each feature is the same, except for one. Realistically, we would not expect this to hold. Therefore, we conducted an experiment where we randomly generated 4 probabilities for X_1, and 4 probabilities for X_2 using numpy.random.randomsample with a fixed $w^\natural = [0.1, 0.3, -0.1, 0.1, -0.5]$. Then, we pick $\hat{w}$ so that $\hat{w}_2, \hat{w}_3, \hat{w}_4$ are random and vary $\hat{w}_1$ from $[-3, 3]$ and plot the difference in opportunity.

We find that in accordance with the results in the previous section, increasing $\hat{w}_1$ causes the difference in opportunity between the group with the higher probability in feature 1, and the group with the lower probability in feature 1, to increase. In addition, the increase in the difference of opportunity tends to increase monotonically for the vast majority of the trials.

When the two groups have very similar probabilities of having one feature, we find that changing $\hat{w}_1$ may not cause a monotonic change in the difference in opportunity. Thus, it may be beneficial to only change adjust the weight on features that you know have somewhat different probabilities between groups.

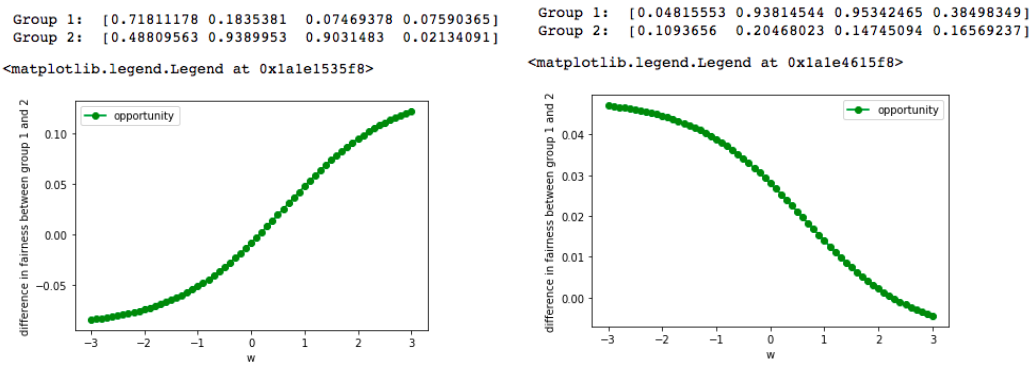


Figure 1: Increasing $\hat{w}_1$ causes the difference in opportunity to increase between the group with the higher probability in feature 1, and the group with the lower probability in feature 1. Note that what's plotted is the difference in opportunity between group 1 and group 2.

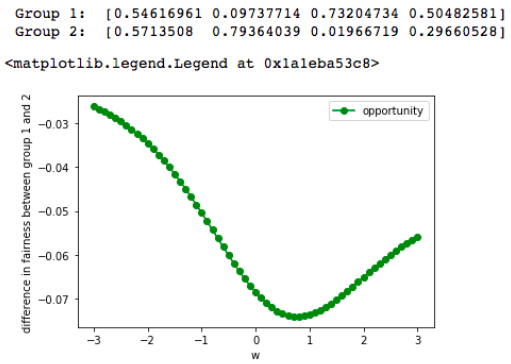


Figure 2: When the probabilities for the first feature are similar, may cause non-monotonicity

For future work, we would like to prove the following hypothesis:

Conjecture. *Suppose* $X_1 \sim Bern(p_1, p_2, \ldots, p_k)$ *and* $X_2 \sim Bern(q_1, q_2, \ldots, q_k)$*, with* $p_k > q_k$ *and* $w_k^{\natural} > 0$*. There exists* $\delta > 0$ *such that if* $p_k - q_k > \delta$*, then increasing* $\hat{w}_k$ *increases the difference in opportunity between group 1 and group 2 monotonically.*

This is Lemma 2.1 with fewer assumptions. It essentially states that if the natural weight on a feature is positive, then increasing the classifier weight on that feature benefits the group that has a higher probability of having that feature.

We would also like to investigate the conditions for which monotonicity is guaranteed. In addition, we hypothesize that there should be a bound on the number of times the function of the difference in opportunity changes convexity as a function of $\hat{w}_1$, and we hope to prove this in the future.

3 German Credit Data

The German Creditability Dataset contains where the binary label indicating whether a person is "credit-able". The dataset contains one protected attribute that would ideally not influence someone's predicted label: gender and marital status (which isn't a binary classification). The dataset contains continuous, Boolean, and categorical features, but most of the features are Boolean or categorical.

First, we will fit a logistic regression model to the data. To do so, we use a one-hot encoding on the categorical features. Because we

don't know the distribution of the data, we can't calculate each groups opportunity and predictive value exactly. However, we can estimate each quantity by counting the number of points classified positively, and the number of points whose true label is positive.

Using the base model obtained from the logistic regression, we obtain the following statistics about fairness, where group 1 is male, and group 2 is female:

$$P(Y_1 = 1) - P(Y_2 = 1) = 0.08 \tag{4}$$

$$P(\hat{Y}_1 = 1) - P(\hat{Y}_2 = 1) = 0.098 \tag{5}$$

$$P(\hat{Y}_1 = 1|Y_1 = 1) - P(\hat{Y}_2 = 1|Y_2 = 1) = 0.05 \tag{6}$$

$$P(Y_1 = 1|\hat{Y} = 1) - P(\hat{Y}_2 = 1|Y_2 = 1) = 0.03 \tag{7}$$

Thus, we see that group 2 is at an empirical disadvantage, because the probability of their true label being positive is less than the probability of group 1's true label being positive. The classifier also is not fair in terms of any of the metrics.

3.1 Making our Real Classifier Fairer

We attempt to adopt a similar methodology as in our simulated data. The idea will be to look at features that (1) the advantaged group has a higher probability of having than the disadvantaged group, and (2) is weighted positively in the true classifier (if we assume that the true classifier has a logistic form, then this means that $w^\natural$ is positive for that feature).

To do so, we will estimate the distribution of the features for both groups to estimate which features the advantaged group has a higher probability of having than the disadvantaged group, and we will look at the coefficients returned by the logistic model to estimate which features are truly beneficial to have.

We find that group 2 is less likely to have the "Not Foreign Worker" attribute, which has a positive coefficient, at 0.66. The difference in probabilities of having it between the advantaged group and disadvantaged group was 0.03. By varying the logistic model's weight vector on it on $[-8, 8]$, we show that we can change the difference in opportunity as from the simulated data.

We do the same with the "Housing" feature, which tells whether the individual owns their home, rents, or has free housing (this is a categorical feature). The coefficient for having free housing is positive, at 0.28, and we estimate that the disadvantaged group has a lower probability of having free housing (the difference is 0.11). Varying the weight vector for it on $[-8, 8]$, we show the same effect applies, even for categorical, rather than binary features.

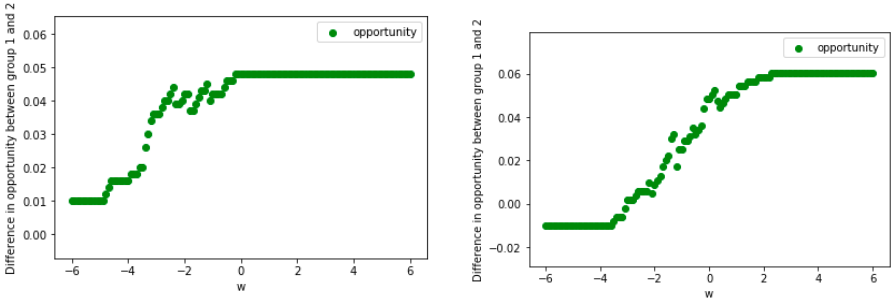


Figure 3: Left: Not Foreign Worker Feature; Right: Free Housing Feature

We see that the strategy from the simulated data generalizes to real data. The curves are noisy due to the estimation of probabilities from the discrete results. We are able to make the difference in opportunity 0 by setting the weight for "Free Housing" to be around -3. We're also able to decreasing the difference in opportunity to around 0.01 by setting the weight for "Not Foreign Worker" to be -6.

Also as noted in the simulated data, when the probability of each group having the feature is very similar, we may observe a graph where increasing the weight on that feature does not change the difference in opportunity monotonically. We can see this when we vary the weight on the feature "No Guarantor". We estimated the difference in the probability of both groups having the feature to be .0086, and varying the weight on it from -8 to 8 produces the following graph:

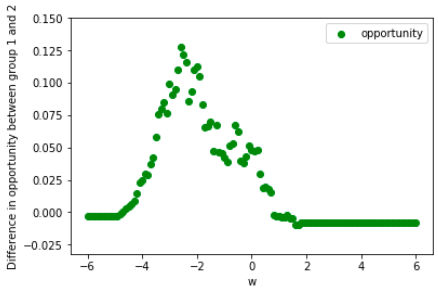


Figure 4: Non-Monotonicity when probabilities are close

Thus, our experimental results on the German Credit dataset verify the results from our simulated data.

4 Washington Home Loans, 2016

The Washington State Home Loans dataset contains data on the statuses of over 450 thousand home loans. It contains many features including demographic information including gender, as well information on property type, loan type, and loan purpose.

As with the German credit dataset, we fit a logistic model to the data, using 13 features, including the county, data on income, and data on the demographic composition of the area. Many of the features we use are categorical, so we represent these features using a one-hot encoding. The dataset is also unbalanced, with over 80 percent of the loans being approved, so we remove data from the positive class until the dataset is balanced, and randomly sample to get an 80-20 train-test split. We further divide the test set by gender to obtain male and female test sets

4.1 Experiment 1

In our first experiment we train a logistic classifier with access to the gender data. One possible source of unfairness in classifiers which are blind to protected attributes is the presence of a feature which is highly correlated with the protected attribute. Of course, including the protected group itself is a trivial case of such a "correlated feature." We therefore examine the effect of moving this weight over 10 small steps in the direction that would favor approval of females in the loan process. We are not suggesting that gender should be considered in practice, but we use gender as a clear example of what happens when adjusting the weight on a highly correlated feature. We see that the accuracy goes essentially unaffected (increasing by under 1 percent) and the difference in opportunity falls much faster than the difference in predictive value increases.

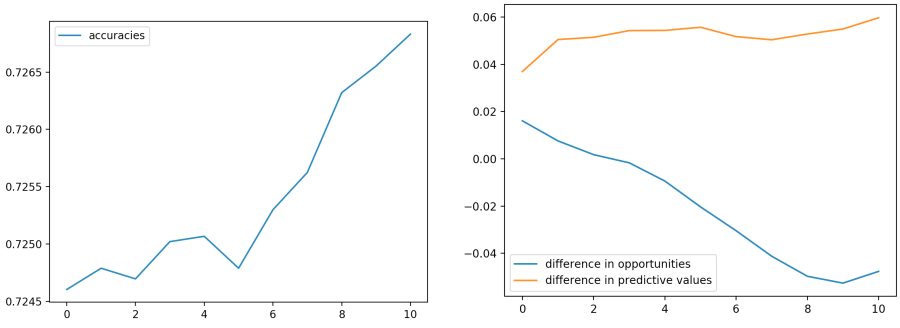


Figure 5: Weight change on gender (Left) Change in Accuracy (Right) Change in Fairness

4.2 Experiment 2

In our second experiment we train a logistic classifier without access to the gender data, and vary the weight corresponding to income. In our test set the average male earns 118,000 while the average female earns only 91,000, so we adjust the weight to reduce the relative importance of income. Here the result is very different than the previous. While we still see that the difference in opportunity drops faster than the corresponding increase in the difference in predictive value, the accuracy also drops considerably as we move the weight vector. One possible explanation for the difference in behavior between this experiment and the previous is that income could be a strong predictive feature for home loan approval, while gender likely is not. It thus seems intuitive that the accuracy could drop due to perturbing the weight on a gender-correlated predictive feature, as opposed to perturbing the weight on the non-predictive feature of gender itself, although we leave a thorough analysis to future work. These results suggest that it may be possible to improve the fairness of classifiers with a logistic decision rule without incurring a large cost to accuracy of predictions by moving the weight on a feature highly correlated with the belonging to a protected group given that that feature is not very predictive.

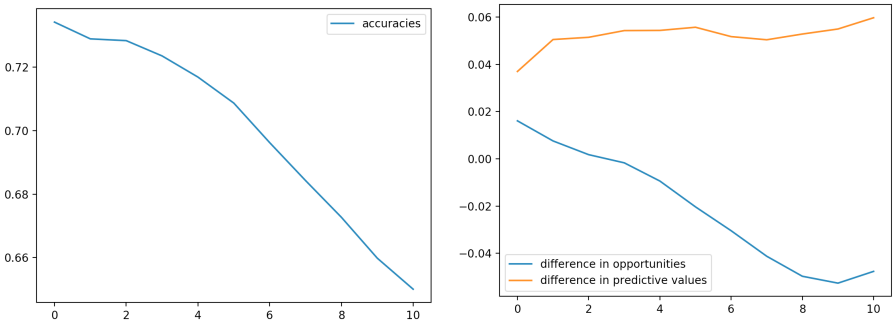


Figure 6: Weight change on income (Left) Change in Accuracy (Right) Change in Fairness

5 Conclusion

We have shown empirically that for an unaware classifier, increasing the weight on a feature tends to result in a monotonic change in the difference in opportunity, where the change favors the group with the higher probability of having the feature, if the feature has a positive weight in determining the true label. We have proved this for when the two groups have the same probability for each feature, except for one, in the case of binary features. In addition, we find that it may be preferable to optimize for opportunity rather than predictive value, because the difference in predictive value appears to not change as much as the weight vector is altered. We run experiments on simulated data and real data that corroborate our result. We also do preliminary analysis on the trade-off between accuracy and fairness; in the case of the Washington dataset, we also find that it may be possible to improve fairness without a large drop in accuracy.

Thus, we have provided theoretical groundwork for the intuitive idea that changing how much a classifier weights certain features can improve fairness in an unaware classifier. This is especially useful when considering settings where classifiers, by regulation, must be unaware.

For future work, we would like to develop more theoretical results on how the difference in opportunity changes when a feature of the weight vector is changed, and prove the conjecture we proposed. In particular, we would like to also consider continuous features more thoroughly as well. In addition, we would like to do more analysis on how the difference in predictive value changes and obtain a tighter maximum bound (that is more consistent with experimental results). Finally, we would like to perform further analysis on the impact that changing the classifier weight vector has on accuracy.

6 Appendix

6.1 Proof of Lemma 2.1:

Proof. We have

$$\begin{aligned} P(\hat{Y}=\hat{y}|Y=y) &= \sum_{x\in\mathcal{X}} P(\hat{Y}=\hat{y}|Y=y, X=x)P(X=x|Y=y) \\ &= \sum_{x\in\mathcal{X}} P(\hat{Y}=\hat{y}|Y=y)P(X=x|Y=y) \qquad (1) \\ &= \frac{\sum_{x\in\mathcal{X}} P(\hat{Y}=\hat{y}|X=x)P(Y=y|X=x)P(X=x)}{P(Y=y)} \end{aligned}$$

where (1) follows because the conditional probability of $\hat{Y}$ is determined completely by Y (that is, $\hat{Y}$ is conditionally independent of X, given Y). In addition, using the Law of Total Probability, we have

$$P(Y=y) = \sum_{x\in\mathcal{X}} P(Y=y|X=x)P(X=x)$$

Then we have:

$$\begin{aligned} \frac{dP(\hat{Y}=1|Y=1)}{d\hat{w}_1} &= \frac{\sum_{\substack{x\in\mathcal{X}\\x_1=1}} \frac{\exp(-x^T\hat{w})}{(1+\exp(-x^T\hat{w}))^2} P(Y=1|X=x)P(X=x)}{P(Y=1)} \\ &= \frac{\sum_{\substack{x\in\mathcal{X}\\x_1=1}} \frac{\exp(-x^T\hat{w})}{(1+\exp(-x^T\hat{w}))^2} P(Y=1|X=x)P(X=x)}{\sum_{\substack{x\in\mathcal{X}\\x_1=1}} P(Y=1|X=x)P(X=x) + \sum_{\substack{x\in\mathcal{X}\\x_1=0}} P(Y=1|X=x)P(X=x)} \\ &= \frac{\sum_{\substack{x\in\mathcal{X}\\x_1=1}} \frac{\exp(-x^T\hat{w})}{(1+\exp(-x^T\hat{w}))^2} P(Y=1|X=x)p_1\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i}}{d_1} \\ &= \frac{p_1\sum_{\substack{x\in\mathcal{X}\\x_1=1}} \frac{\exp(-x^T\hat{w})}{(1+\exp(-x^T\hat{w}))^2} P(Y=1|X=x)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i}}{d_2} \\ &= \frac{p_1\sum_{\substack{x\in\mathcal{X}\\x_1=1}} \frac{\exp(-x^T\hat{w})}{(1+\exp(-x^T\hat{w}))^2} P(Y=1|X=x)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i}}{d_3} \end{aligned}$$

where

$$d_1 = \sum_{\substack{x\in\mathcal{X}\\x_1=1}} P(Y=1|X=x)p_1\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i} + \sum_{\substack{x\in\mathcal{X}\\x_1=0}} P(Y=1|X=x)(1-p_1)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i}$$

$$\begin{aligned} d_2 = p_1 &\sum_{\substack{x\in\mathcal{X}\\x_1=1}} P(Y=1|X=x)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i} - p_1\sum_{\substack{x\in\mathcal{X}\\x_1=0}} P(Y=1|X=x)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i} \\ &+ \sum_{\substack{x\in\mathcal{X}\\x_1=0}} P(Y=1|X=x)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i} \end{aligned}$$

$$\begin{aligned} d_3 = p_1&\left(\sum_{\substack{x\in\mathcal{X}\\x_1=1}} P(Y=1|X=x)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i} - \sum_{\substack{x\in\mathcal{X}\\x_1=0}} P(Y=1|X=x)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i}\right) \\ &+ \sum_{\substack{x\in\mathcal{X}\\x_1=0}} P(Y=1|X=x)\prod_{i=2}^k p_i^{x_i}(1-p_i)^{1-x_i} \end{aligned}$$

Let

$$k = \sum_{\substack{x \in \mathcal{X} \\ x_1 = 1}} \frac{\exp(-x^T \hat{w})}{(1 + \exp(-x^T \hat{w}))^2} P(Y = 1 | X = x) \prod_{i=2}^{k} p_i^{x_i} (1 - p_i)^{1 - x_i}$$

$$r = \left(\sum_{\substack{x \in \mathcal{X} \\ x_1 = 1}} P(Y = 1 | X = x) \prod_{i=2}^{k} p_i^{x_i} (1 - p_i)^{1 - x_i} - \sum_{\substack{x \in \mathcal{X} \\ x_1 = 0}} P(Y = 1 | X = x) \prod_{i=2}^{k} p_i^{x_i} (1 - p_i)^{1 - x_i} \right)$$

$$q = \sum_{\substack{x \in \mathcal{X} \\ x_1 = 0}} P(Y = 1 | X = x) \prod_{i=2}^{k} p_i^{x_i} (1 - p_i)^{1 - x_i}$$

Then we have

$$\frac{d}{dw_1} P(\hat{Y}_1 = 1 | Y_1 = 1) = \frac{p_1 k}{p_1 r + q}$$

and

$$\frac{d}{dw_1} P(\hat{Y}_2 = 1 | Y_2 = 1) = \frac{\hat{p_1} k}{\hat{p_1} r + q}$$

Thus,

$$\begin{aligned}
\frac{d}{dw_1} \Big(P(\hat{Y}_1 = 1 | Y_1 = 1) - P(\hat{Y}_2 = 1 | Y_2 = 1) \Big) &= \frac{d}{dw_1} P(\hat{Y}_1 = 1 | Y_1 = 1) - \frac{d}{dw_1} P(\hat{Y}_2 = 1 | Y_2 = 1) \\
&= \frac{p_1 k (\hat{p_1} r + q) - \hat{p_1} k (p_1 r + q)}{(p_1 r + q)(\hat{p_1} r + q)} \\
&= \frac{q(p_1 - \hat{p_1})}{(p_1 r + q)(\hat{p_1} r + q)} \\
&< 0 \qquad (1)
\end{aligned}$$

where (1) follows because when $w_1^\natural \geq 0$, $r > 0$ because given two arbitrary vectors of the form $[1, x_2, \ldots, x_k], [0, x_2, \ldots, x_k]$, we have:

$$\begin{aligned}
P(Y = 1 | X = [1, x_2, \ldots, x_k]) &= \frac{1}{1 + \exp(-(w_1 + x_2 w_2 + \cdots + x_k w^k))} \\
&\geq \frac{1}{1 + \exp(-(x_2 w_2 + \cdots + x_k w^k))} \\
&= P(Y = 1 | X = [0, x_2, \ldots, x_k])
\end{aligned}$$

This shows that as w_1 increases, the difference in opportunity increases. □

Remark. First, we note that although we assumed that the first feature was the one that differed in probability between the two groups, but the same proof is easily generalized to the two groups differing in two probabilities.

Third, we note the assumption that $w_1^\natural \geq 0$. This means that it applies for $w_1^\natural = 0$. This means that even if one particular feature has no effect on the true probability, the analyst can decrease the difference in opportunity by decreasing how much he cares about that feature.

If $w_1^\natural < 0$, the difference in opportunity can increase or decrease monotonically. (In the proof, if $w_1^\natural < 0$, then $r < 0$, so $(p_1 r + q)(\hat{p_1} r + q)$ can be positive or negative. However, we note that the difference is still either positive or negative, which means that if the analyst increases the weight on a feature and observes that the difference in opportunity goes up or down, he knows that it will continue going up or down.

6.2 Proof of Lemma 2.2:

Proof. A group's (positive) predictive value is given by:

$$P(Y=1|\hat{Y}=1) = \frac{\sum_{x\in\mathcal{X}} P(Y=1|X=x)P(\hat{Y}=1|X=x)P(X=x)}{P(\hat{Y}=1)}$$
$$= \frac{\sum_{x\in\mathcal{X}} P(Y=1|X=x)P(\hat{Y}=1|X=x)P(X=x)}{\sum_{x\in\mathcal{X}} P(\hat{Y}=1|X=x)P(X=x)}$$

Let $m_i = \max\{P(Y_i=1|X=x) : x\in\mathcal{X}\}$. Then we have:

$$P(Y_i=1|\hat{Y}=1) = \frac{\sum_{x\in\mathcal{X}} P(Y_i=1|X=x)P(\hat{Y}_i=1|X=x)P(X=x)}{\sum_{x\in\mathcal{X}} P(\hat{Y}_i=1|X=x)P(X=x)}$$
$$\leq \frac{\sum_{x\in\mathcal{X}} m\cdot P(\hat{Y}_i=1|X=x)P(X=x)}{\sum_{x\in\mathcal{X}} P(\hat{Y}_i=1|X=x)P(X=x)}$$
$$\leq m_i$$

Likewise, let $l_i = \min\{P(Y_i=1|X=x) : x\in\mathcal{X}\}$. Then we have:

$$P(Y_i=1|\hat{Y}_i=1) = \frac{\sum_{x\in\mathcal{X}} P(Y_i=1|X=x)P(\hat{Y}_i=1|X=x)P(X=x)}{\sum_{x\in\mathcal{X}} P(\hat{Y}_i=1|X=x)P(X=x)}$$
$$\geq \frac{\sum_{x\in\mathcal{X}} l_i\cdot P(\hat{Y}_i=1|X=x)P(X=x)}{\sum_{x\in\mathcal{X}} P(\hat{Y}_i=1|X=x)P(X=x)}$$
$$\geq l_i$$

Thus, we have $P(Y_1=1|\hat{Y}_1=1) - P(Y_2=1|\hat{Y}_2=1) \leq m_1 - l_2$ and $P(Y_2=1|\hat{Y}_2=1) - P(Y_1=1|\hat{Y}_1=1) \leq m_2 - l_1$. Thus, $|P(Y_1=1|\hat{Y}_1 1) - P(Y_2=1|\hat{Y}_2=1)| \leq \max(m_1-l_2, m_2-l_1)$. □

6.3 Datasets and Techniques Used

Links to both datasets are provided below. Within the German Creditability Dataset, we found that the "Sex Marital Status" feature was categorical, and we cluster this column into simply "Male" and "Female" values. Within the Washington HDMA dataset, we create a train-test split for the individual classifier, and create a one-hot encoding for features such as county name, loan occupancy name, etc. As mentioned above, we fit a logistic regression-based classifier to both.

1. German Creditability Dataset: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
2. Washington HDMA Dataset https://www.kaggle.com/miker400/washington-state-home-mortgage-hdma2016

References

[1] Claudia Goldin and Cecilia Rouse. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians". In: *American Economic Review* 90.4 (Sept. 2000), pp. 715–741. DOI: 10.1257/aer.90.4.715. URL: http://www.aeaweb.org/articles?id=10.1257/aer.90.4.715.

[2] Roland Jr, Glenn Loury, and Tolga Yuret. "An Economic Analysis of Color-Blind Affirmative Action". In: *Journal of Law Economics and Organization* 24 (Oct. 2008). DOI: 10.1093/jleo/ewm053.

[3] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. "Discrimination-aware Data Mining". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA: ACM, 2008, pp. 560–568. ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401959. URL: http://doi.acm.org/10.1145/1401890.1401959.

[4] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *CoRR* abs/1609.05807 (2016). arXiv: 1609.05807. URL: http://arxiv.org/abs/1609.05807.

[5] Pratik Gajane. "On formalizing fairness in prediction with machine learning". In: *CoRR* abs/1710.03184 (2017). arXiv: 1710.03184. URL: http://arxiv.org/abs/1710.03184.

[6] Hoda Heidari et al. "A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity". In: *CoRR* abs/1809.03400 (2018). arXiv: 1809.03400. URL: http://arxiv.org/abs/1809.03400.

[7] Jiahao Chen et al. "Fairness Under Unawareness". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (2019). DOI: 10.1145/3287560.3287594. URL: http://dx.doi.org/10.1145/3287560.3287594.