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 [1], Heidari et al [3], and Gajane et al[2], 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.

Our project is currently focused on both theoretically and empirically answering how changes to model parameters affect fairness, and if possible, we will analyze how fairness optimizations affect accuracy as well.

1.1 Overview

We currently consider three datasets: simulated data, a German creditability dataset which matches features of individuals to whether they are "creditable", and the Washington State HDMA dataset which matches features of individuals to whether they can get mortgages. While the first dataset is clearly fairly clean (as it has been synthesized), we've had to clean and preprocess both the German and Washington datasets using Pandas (removing missing entries, encoding categorical data properly, etc.) before experimentally testing the lemmas we give on real datasets via fitting logistic regressions, changing the weights slightly, and observing the disparity between groups.

After running preliminary analyses on both simulated datasets and the German dataset, we've tested an intuitive approach to improving fairness, and show that certain metrics are less sensitive to changes than other, providing insight on which metrics to focus on. We then focus on thorough explanations/definitions for both simulated and real datasets, and we test whether or not the experimental findings generalize.

2 Simulated Data

2.1 Model

We are interested in a supervised learning setting, where an agent 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 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 $X \sim \text{Bern}(p_1, \ldots, p_n)$. 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 $w^{\natural}$.

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 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; i.e., $P(\hat{Y}_1 = 1|X_1 = x) = P(\hat{Y}_2 = 1|x_2 = x)$. This has the application in settings where the decision making agent 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*.

2.2.1 Example

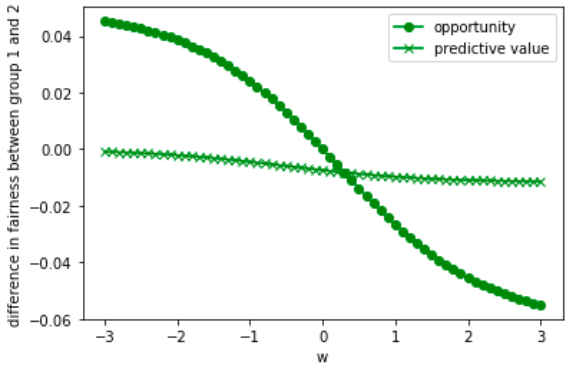
Suppose $X_1 \sim \text{Bern}(0.2, 0.2, 0.2, 0.2)$ and $X_2 \sim \text{Bern}(0.7, 0.2, 0.2, 0.2)$, and $w^\natural = [0.3, 0.2, -0.1, 0.4]$, $b^\natural = -0.3$. Suppose $\hat{W} = w^\natural$, $b^\natural = -0.3$, so the analyst is using the Bayes Optimal Classifier. The statistics are as follows:

$$P(Y_1 = 1) - P(Y_2 = 1) = -0.037 \quad (1)$$

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

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

We see that group 2 has an inherent advantage over group 1 from (1). Intuitively, to improve fairness, the analyst can lower the value of $\hat{w}_1$, since group 2 has a higher probability of having feature 1 than group 1. Fig 1 is a plot of how the difference in opportunity and predictive value changes as $\hat{w}_1$ is varied. In this example, the difference in opportunity increases monotonically as $\hat{w}_1$ increases. Note that at $\hat{w}_1 = -10$, group 1 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.



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

Proof. Omitted due to the 3 page limit, but found in the "Probability Calcualtions Bernouilli" jupyter notebook. □

Lemma 3.1 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.

In addition, Lemma 3.2 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 $\hat{w}$. The results that demonstrate these findings are omitted due to the page limit, but can be found in the notebook.

3 German Credit Data

The German Creditability Dataset contains datasets where the label is a person's creditability. The dataset contains one protected attribute that would ideally not influence someone's predicted label: gender. The dataset contains coninuous, Boolean, and categorical features, but most of the features of Boolean or categorical.

First, we will fit a logistic regression model to the data. Because we don't know the distribution of the data, we can't calculate the each group's opportunity and predictive value exactly. However, we can estimate each quantity by counting the number of points classified positively, and 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:

$$P(Y_1 = 1) - P(Y_2 = 1) = 0.08 \quad (4)$$

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

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

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

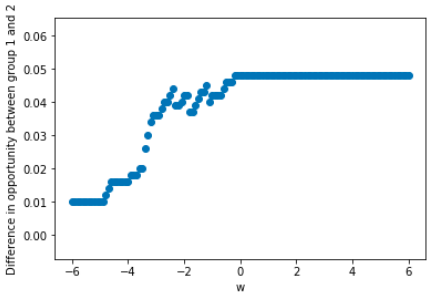
Thus, we see that group 2 is at an empirical disadvantage, because the probability of their true label being positive is less than the probability of group 1's true label being positive. The classifier also is not fair in terms of any of the metrics.

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 where (1) the advantaged group has a higher probability of having than the disadvantaged group, and (2) that particular feature 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. By varying the logistic model's weight vector on it on $[-8, 8]$, we obtain the following graph, where the y-axis : Due to the noise associated with



estimating the probabilities, we find that the difference does not increase monotonically, but the trend is there. More results can be found in the "German Credit" notebook.

4 Next Steps

Ideally, we would be able to prove facts about the monotonicity/number of critical points of the difference in opportunity as a function of a component of the weight vector, without the assumption that X is Bernouli distributed and the features are independent.

We will perform more analysis on real data sets and potentially show how to optimally improve fairness with respect to accuracy.

5 Datasets

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] 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.

[2] 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.

[3] 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.