Technical Solutions to Evaluating Fairness in Algorithms

Kristy Guo, Xuanqi Liang, Jingjing Lu, Qiqi Tang, Chiebuka Onwuzurike

Introduction

Topics on fairness and biases of algorithms and machine learning models have become more recognized by the public. Where there has been an increasing effort made to ensure fairness and biases, until October 2020, "only 53% of organizations have a leader who is responsible for the ethics of AI systems." The enforcement of ethics in algorithms is vital in increasing customer's trust on Machine Learning and AI, as there are no more than half of the population trusting current AI-based enabled interactions with organizations.

It is also us as business analysts' responsibility to create a safe environment for people when building these machine learning models. One of the actions we can practice is to validate the fairness of models through existing technical techniques, which is discussed below.

Example illustrations

We'll construct a model based on the US Homicide Reports 1980-2014 to predict the perpetrator's sex and race(skin color). Further analysis would be focusing on checking if the model is biased on the victim description and how to evaluate fairness using five metrics.

1. Statistical Parity Difference (Chiebuka Onwuzurike)

Formula

Variables

$X = Sample$
$S = Subset\ of\ Sample\ (S \supset X)$
$D = Distribution\ of\ X$
$Y = (possibly\ randomized)\ classifier.\ Can\ Either\ be\ \{1, 0\}$
$\varepsilon = fairness\ threshold$

Equation

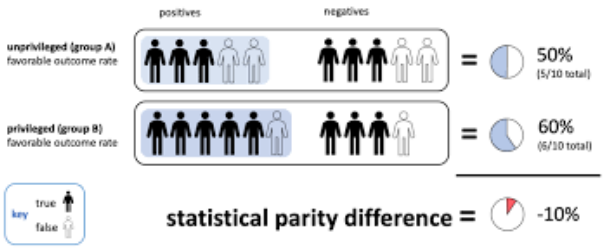
$$bias_h(X, S, D) = Pr(Y = 1 | D = unprivileged) - Pr(Y = 1 | D = privileged)$$

Fairness

$$\left|bias_h(X, S, D)\right| < \varepsilon$$

Definition

Statistical parity difference is the difference in probabilities of one subset sample's likelihood for a particular outcome and the rest of the sample set. The smaller the statistical parity difference the more fair for a particular outcome because it is statistically saying the probability of one subset sample is not more or less likely than the probability of the sample. The absolute value of of statistical parity difference of bias ($\left|bias_h(X, S, D)\right|$) needs to be lower than some discretionary fairness range (ε). This measure of fairness is also seen in the legal system under Title VII and Title VI of the Civil rights act of 1964 under the legal theory of adverse impact



Factors that have an Effect

Two things that can determine if statistical parity difference will be useful are the effect size of the disparity and underlying differences. When the effect size is small or there isn't enough statistical power, bias can not reach the threshold. Additionally there are disparities that may be affected by underlying variables, called confounders, which would imply that the disparity is due to underlying differences that are not predicated on group membership.

Situation when it is a good metric

Situation when it is a bad metric

An example where statistical parity difference is a bad metric can be seen in fire departments. Fire departments have a 100 lb minimum requirement for firefighters. This minimum is to ensure firefighters can carry people out of burning buildings if need be. However this affects the gender makeup of fire departments because men and women tend to have different average weights. Because men tend to be taller and have more muscle, they tend to weigh. Consequently there tends to be less female firefighters than male.

2. Equal Opportunity Difference (Xuanqi Liang)

Formula

$$TPR_{D=unprivileged} - TPR_{D=privileged}$$

Definition

Equal Opportunity Difference(EOD) is the difference in true positive rates between unprivileged and privileged groups. Here, the true positive is that you estimate the positive sample correctly, both the prediction and truth are positive. True positive rate is the true positive divide by the total actual positive. In order to get the equal opportunity difference, we need to use the true positive rate for unprivileged group minus the true positive rate for privileged group.

Situation when it is a good metric

The ideal value for the equal opportunity difference is 0. When the difference is 0 or close to 0 means there is no or less difference between unprivileged and privileged groups. We usually consider it fair when the metric is between -0.1 and 0.1.

Situation when it is a bad metric

The larger number of the equal opportunity difference means there are more differences between unprivileged and privileged groups. There are two two situations in this case. If the EOD value is less than 0, which means the privileged group have a higher benefit. And if the EOD result is greater than 0, which means higher benefit for the unprivileged group.

3. Average Absolute Odds Difference (Qiqi Tang)

Formula

$$\frac{1}{2}\left[\left|FPR_{D=unprivileged} - FPR_{D=privileged}\right| + \left|TPR_{D=unprivileged} - TPR_{D=privileged}\right|\right]$$

Definition

Average Odds difference is one of the fairness metrics to check for bias in datasets and models. We need both the original dataset containing true labels and the predicted dataset with predicted labels to check accuracy and fairness with average odds difference matric on our model.

	Predicted: positive (privileged)	Predicted: negative (unprivileged)
Actual: positive (privileged)	True Positive (TPR privileged)	False negative (FPR privileged)
Actual: negative (unprivileged)	False positive (FPR unprivileged)	True Negative (TPR unprivileged)

This metric is concerned with the whole confusion matrix. To do the calculation, we should split our interested column into privileged and unprivileged groups. For instance, in our lectures, we have an example dataset that shows that people have been arrested and have not been arrested. In this dataset, we can set the arrested people to the privileged group and people not being arrested to the unprivileged group. In the formula, TPR privileged is when both prediction and actual are positive and TPR unprivileged is when both prediction and actual are negative. At the same time, FPR privileged is when the actual is positive and the predicted is negative and FPR unprivileged is when the actual is negative and the predicted is positive. Applying two groups into the formula above, we can get the result which is the average odds difference. When the result is close to zero, our prediction model is fair. It is easy to be understanded that if the difference of FPR and TPR or FNR and FNR between predictions and true samples are 0, it means that we estimate samples 100% correctly, indicating our model perfectly has no bias.

Situation when it is a good metric

The dataset I mentioned above, which we were exposed to in class, is a good example of how Average Odds difference is good to be used.From the standpoint of equity as well as humanitarianism, arresting someone can have a significant impact on that person's life, so it is important to detect the difference between someone who is low risk, but we wrongly arrest and

someone who really should be arrested. However, from the perspective of the general public, for their own safety, they are more concerned about those who are at high risk but are not arrested. In this case, it is very important to consider the fairness of the algorithm in a comprehensive manner so that the right of both the suspect and the public is protected. This is where Average Odds difference, a method that takes into account the entire confusion matrix, is ideally suited for testing the fairness of our model.

Situation when it is a bad metric

There are two sides to everything, The Average Odds difference, while being very comprehensive, is inevitably more cumbersome than the other four methods. We need to slowly improve our algorithm, in order to make the result of this method appear close to 0.

4. Disparate Impact (Jingjing Lu)

Formula

$$\frac{Pr(Y=1|D=unprivileged)}{Pr(Y=1|D=privileged)}$$

Definition

Like the first metric we use both probabilities of a random individual drawn from unprivileged or privileged with a label of 1 but here it's a ratio.
It changes the objective, for the disparate impact it's 1 that we need.
We can use disparate impact to check discrimination that is unintentional. The algorithm, procedures or policies appear to deliver neutral results, but people in a protected class are negatively affected.

Situation when it is a good metric

For instance, while job applicants are applying for a certain job, they get tested on their reaction times; and only those with a high score will be hired. This kind of test will be biased against older workers who are less likely to have fast reaction times. So, after the results of the test are stored into an algorithm, those older workers will more likely be ruled out of the application pool.

Situation when it is a bad metric

On the other hand, if the employer shows that it has legitimate business reasons for this reaction test as it needs to run business on workers with sharp reactions, the reaction test can be justified. Also, it could cost the employer higher if the testing on all applicants is prepared in a differential

way for regular class and protected class. Standing on a consequentialist's point, the employer would weigh between the potential disparate impact of its testing given and the cost of resolving the issue.

5. Theil Index (Kristy Guo)

Formula

$$\frac{1}{n}\sum_{i-0}^{n}\frac{b_i}{\mu}ln\frac{b_i}{\mu}$$

Where $bi = y\hat{}i - yi + 1$

where N is the number of cases (e.g., households or families), yi is the income for case i and a is a parameter which regulates the weight given to distances between incomes at different parts of the income distribution. For large a the index is especially sensitive to the existence of large incomes, whereas for small a the index is especially sensitive to the existence of small incomes.

Definition

The Theil index measures an entropic "distance" the population is away from the "ideal" egalitarian state of everyone having some defined standard. A higher numerical result indicates more order and that is further away from the "ideal" of maximum disorder, which in our case is equality. A result of 0 reflects perfect equality.

Situation when it is a good metric

In the case of the US Homicide model, inequality of subgroups can be observed. For example, to measure if there is any gender inequality within the model, it can be observed by Theil index with predicted gender.

In addition, Theil index is widely used to measure income inequality. A high Theil index indicates the total income is not distributed evenly among individuals or observed groups. For example, the US Census Bureau uses Theil Index as one of the measurements to evaluate income inequality between zip code, gender, race etc. We can evaluate ineuqliaty in models by comparing theil indices calculated from predicted value between male and female.

Conclusion (Jingjing Lu)

We used a python package and 5 different fairness metrics to evaluate that our model or dataset are biased. Rather than simply using confusion metrics as one estimator of potential bias, we

introduced 5 proposed definitions for fairness. We do consider that there should be more measures like these shown in our presentation as every analyst would have his or her own standard of fairness. For our next step, we would emphasize on how to mitigate bias if there is any measure. Based on our research, it would be difficult to remove bias once our machine learning model has been implemented. So, we've suggested ways to deal with bias at the stages: data collection, pre-processing and training pipeline. For instance, at the pre-processing stage, we can modify training labels in such a way that the disadvantaged group are more likely to get a better outcome and retrain the model.

Post-Processing	In-Processing	Pre-Processing	Data Collection
• Change thresholds • Trade off accuracy for fairness	• Adversarial training • Regularize for fairness • Constrain to be fair	• Modify labels • Modify input data • Modify label/data pairs • Weight label/data pairs	• Identify lack of examples or variates and collect

Presentation slide & Organizations (Xuanqi Liang)