A Design for Algorithmic Bias Mitigation

Theo McArn
CSCI 1952B Socially Responsible Computing in Practice

Introduction

For this project, I have designed and developed a tool kit that acts as a prototype for a system that could be used to analyze machine learning models for fairness and could help developers mitigate bias. The tool kit measures fairness using multiple metrics and displays these metrics to the user in graphical plots. Additionally, the tool kit contains simple mitigation approaches which developers can test out to see how they affect the fairness metrics. This product has a lot of limitations but acts more as a concept for what a ML auditing and testing system could look like.

In order to test this tool out, I analyzed a CDC Dataset on Heart Disease (Pytlak) which contains patient information and whether or not they are experiencing heart conditions. This dataset contains information about the patients sex and this is the protected group that I used for analysis.

In this report I will explain the motivations behind a project like this as well as some of the challenges and design decisions that I have made. I will finish this report with the specifications of the project as well as explain some of the limitations.

Motivation

As Machine Learning is becoming more popular and sophisticated, its applications are becoming more and more prevalent. Following this extreme growth in machine learning, there has been a strong push for more attention on algorithmic fairness, the notion that these machine learning models have inherent bias which needs to be addressed. However, addressing this algorithmic bias and making these models fair is not straightforward. Because there is no single way to define fairness, it follows that there is no absolute way to measure and counteract unfairness as well. What fairness means in the context of a machine learning model depends on the context of the problem the model is trying to solve. There is no objective view on fairness and the stakeholders involved as well as the consequences of misclassifications all play a role in how fairness is portrayed.

However, while there may be no perfect solution, fairness is crucial and something that still has to be strived for. There have been countless companies and institutions which have been “caught” using poorly tested biased algorithms. These algorithms are everywhere and effect users in a variety of ways. Bias exists in smart-home speakers (Harwell) where some accents are understood better than others. Bias also exists in algorithms which hold more consequences, such

as the COMPAS algorithm (Larson et al.), used by judicial systems to influence whether or not a criminal defendant can post bail. This algorithm has been proven to score Black defendants as higher risk of recidivism as compared to White defendants with the same background. These are only the algorithms that have been externally audited. Many more biased algorithms like this operate behind closed doors and continue to be undetected as well.

As a result of these findings, among all of the other exposed algorithms, there has been a growing protest pushing for regulation which would make these companies responsible for developing and ensuring fairness in their products. This movement has gone so far as it has been a strong component of the Biden Administration, as shown by the Executive Order enacted on this topic (Biden) as well as the Blueprint for an AI Bill of Rights (OSTP), which outlines a general ideal for rights and regulations.

However, in order for a real regulatory system to be enacted, it is important to have a specific set of guidelines that can be understood by developers and enforced by auditors. For this project, I attempted to design a toolkit that could be used by auditors and developers alike for assessing the fairness/unfairness of a model and mitigating bias.

Conflicting Definitions and Metrics

The main issue as I have previously mentioned is developing a formal way to measure a concept which has no formal definition. Depending on the individual interacting with the model, there are very different interpretations of what fairness might be. In general there are a few schools of thought around fairness.

One common metric is statistical parity which means that all groups have the same probability of a positive prediction. In the context of the example dataset, as long as men and women have the same probability of being diagnosed, the model is fair. In other words, as long as the **selection rates are equal**, the model is fair.However, this framing of fairness does not take into account the ground truth classifications. Even if female patients were classified randomly, while men were classified normally, as long as the same proportion were given positive results, statistically parity is still met, even without fairness.

Another common metric for fairness is sufficiency. Sufficiency states that out of all predictions, the ratio of correct outcomes should be equal for each group. In our example, that means that out of all of the positive diagnoses, the proportion of these patients who truly have heart disease should be the same for both men and women. This is akin to the **precision being equal** between groups.

The third common metric for fairness is separation. Separation states that for each true outcome, the ratio of correctly predicted examples should be equal between groups. In our example, this means that everyone who truly has heart disease, the ratio of people predicted to have heart disease is the same. In other words, sufficiency is met if the **false negative and positive rates are equal**. Therefore it also follows that the true positive and negative rates should be equal as well, this is known as the Equalized Odds ratio.

It is also important to note, as stated by the Impossibility Theorem of Machine Learning (Saravanakumar), it is impossible to satisfy both sufficiency and separation. Therefore, it is always necessary to prioritize one sense of fairness over another when designing and evaluating machine learning algorithms. Having a good understanding of the context and the stakeholders is important in deciding which fairness definitions and metrics to select.

Navigating These Decisions

The purpose of this project is design a way for developers and auditors to analyze the fairness of a model by not only looking at any one of these metrics in isolation but by being able to look at all of the metrics together as a whole to understand the ways that they interact with each other and the inherent tradeoffs that exist between them. In doing so, the goal is that it gives the developer or auditor a more holistic view of the model and gives them the knowledge and ability to either select the best model, or in the case of the auditor, assess the current model's fairness. Therefore, in order to allow for the most informed decision, I design the interface in a way which communicates all of the metrics in a clear graphical manner so that it is easier to visualize the relationships and changes. Once presented with this information, it is up to the developers or auditors to deem which models are fair and which are unfair.

In addition to providing metrics, I also included common training methods used to combat and mitigate bias. I included this as a way for developers to see how they might improve their model and provided initial steps and suggestions on how to do so. I intentionally chose simple mitigation techniques as I just wanted to demonstrate their effectiveness/ineffectiveness, and did not want it to seem like a quick solution, more as inspiration. In the case of the Heart Disease dataset, a lot of these mitigations made very minor changes while others made much more significant improvements. However, just like the metrics, these mitigation solutions are context dependent and work differently on different models and datasets, therefore, I included a variety in order to give developers a variety of suggestions.

For the sake of this project, I also took on the role of the developer of the machine learning model, and wanted to use the platform I have developed to assess the fairness of models and mitigate bias. I wanted to see how this toolkit might be applied to real world applications if it was released as a product. I put myself in the position of someone trying to develop a fair heart disease detector. Given this context, I was able to narrow down the metrics provided to ones I think that are most relevant indicators of my definition of fairness. In order to gather a sense of what my method of fairness might be, I decided to talk to an expert, one of my friends, who is entering into medical school next year and knows a lot about the priorities and ethics of medicine. She mentioned that the missed diagnoses are a lot more harmful then false positive diagnoses. If a patient receives a false negative, they will be unaware of their true condition and will miss out on the valuable chance of preventative measures and treatments. On the other hand, if a patient is falsely diagnosed, while still harmful, the consequences are much lower, as follow up testing can confirm false diagnoses. Therefore, she explained that reducing false negatives is much more important than reducing false positives, in the context of medical diagnosis.

Therefore, when looking at the different models produced, each designed to reduce bias in different ways, I had an easier time deciding which models and mitigation methods were better because I knew which metrics to look for. However, it is important to consider other stakeholders too, apart from the patient, who would have other motives for fairness. For example health insurance companies, who possibly fund this machine learning model, might want to make sure that false positive rates are fair across groups instead. A possible scenario could be that the follow up tests that come after this preliminary screening are extremely expensive and so therefore, the insurance company wants to reduce the amount of unnecessary follow ups. Because of unanticipated other stakeholders like this, I wanted this tool kit to contain all of the available metrics, for decision making.

Toolkit Specifications

So far I have explained the motivations, challenges, and design considerations for this project and now I will explain the design I came up with given all of this information. This toolkit takes in a dataset and a sensitive feature that the user wants to be analyzed. The sensitive feature typically corresponds to demographic datum such as Sex, Race, Age, Nationality etc but could be anything that the user believes could result in disparate predictions. Given this input, the user is able to generate five different logistic regression models[1] to fit this data. One regression model where the data is unmodified, and another four models where mitigation is attempted.

Types of Mitigation

1. Balancing Training Samples
    a. A typical source of error is when a model is trained on one group of samples much more than another. This is common in facial recognition models where it sees many more white faces than black faces and therefore, recognizes white faces much more easily
    b. This is counteracted by making sure that there is the same number of samples from each class.
2. Excluding Sensitive Features
    a. A strategy that is less popular now but still prominent is excluding sensitive features all together during training. This is also known as the “Unaware Approach” .
    b. Often this is not effective because even though the sensitive features are removed, they are still often encoded into other features which remain.
3. Adversarial Networks[2]

[1] The Logistic Regression Models were built and trained using the sklearn library
[2] Trained using fairlearn library

    a. This is the most advanced form of mitigation but it is extremely effective. This is often called the “Aware Approach”, as the model is “aware” of the bias and trains an adversarial network which attempts to mitigate the bias that the predictor model creates.
4. A combination of balancing and adversarial networks

After applying each type of mitigation, the toolkit then analyzes the fairness of each model and plots 7 different metrics[3], each split for the given sensitive feature:

1. Accuracies
2. Precisions (Sufficiency)
3. False Positive Rates
4. False Negative Rates
5. Selection Rates
6. Statistical/Demographic Parity Ratio
7. Equalized Odds Ratio

Additionally, since the models are trained using logistic regression, it is possible to analyze and interpret the weights of the model to understand which features of the data affect the classification the most. I created a function which returns the ranking of top 10 features that influenced classification the most. The reason that I added this was because it can be very telling of a model's fairness. If sensitive features have very high weights in the models, then it is clear that the model is unfair, without even looking at the metrics. When I looked at the features for the Heart Disease data, after training on the raw data, sex was the 6th most important feature for prediction out of the 37 total number of features. This function clearly shows the biases and unfairness of the model. However, it cannot easily be applied to other machine learning models such as neural networks, which cannot be easily interpreted like this.

To use the toolkit and understand the inner workings of it, you can view the repository here. The toolkit can be run by uncommenting the main function calls in `main.py`.

Limitations

As this is merely a proof of concept, there are quite a few limitations to this toolkit. Firstly, this was only designed to support binary classification with logistic regression. In reality, models being used in industry are much larger, using more complex data, and making much more complex predictions. In these cases fairness is even more difficult to quantify and analyze. This toolkit was developed to demonstrate that even in the most ideal and simplistic case, there is no

[3] Calculated and visualized using the fairlearn library

solution to this issue of fairness. Therefore, while this product is not applicable to larger models, it accurately demonstrates the challenges of algorithmic machine learning.

Conclusion

Algorithmic fairness is an important issue and something that will only become more and more relevant as Machine Learning is further integrated into our lives. Fairness in machine learning is a very challenging problem to solve and is entirely context dependent. Therefore, in order to provide developers and auditors with the tools needed to detect bias, it is necessary to provide them with an arsenal of metrics and mitigation strategies since there is no one solution or quick fix. The toolkit I have created demonstrates a proof of concept for this sort of arsenal. A variety of ways to assess and mitigate fairness, which allows auditors or developers to set their own standards of fairness as they see fit, based on the context and the stakeholders that only they can understand. While this still puts a lot of pressure on the developers and auditors to understand how they define fairness, this tool kit allows them to optimize their model no matter which decision they come to.

Work Cited

Biden, Joseph. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." *The White House*, 30 October 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. Accessed 5 May 2024.

Harwell, Drew. "Why some accents don't work on Alexa or Google Home." *Washington Post*, 19 July 2018, https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/. Accessed 5 May 2024.

Larson, Jeff, et al. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, 23 May 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 5 May 2024.

OSTP. "Blueprint for an AI Bill of Rights." *The White House*, October 2022, https://www.whitehouse.gov/ostp/ai-bill-of-rights/. Accessed 5 May 2024.

Pytlak, Kamil. *Indicators of Heart Disease*. 2020 annual CDC survey data of 400k+ adults related to their health status. 2020. *Kaggle*, Kaggle,

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. Accessed May 2024.

Saravanakumar, Karthik. “[2007.06024] The Impossibility Theorem of Machine Fairness -- A Causal Perspective.” *arXiv*, 12 July 2020, https://arxiv.org/abs/2007.06024. Accessed 5 May 2024.