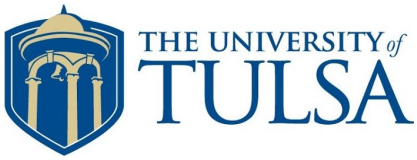

Machine Learning
CS-7333

Fairness and Bias in Machine Learning

Instructor:
Prof. Sandip Sen

Author:
Selim Karaoglu

May 5, 2022

1 Introduction

This report is the final project report for Machine Learning class. In this project we focus on Fairness and Bias in Machine Learning (ML). In the first part, we review some researches that influenced the search of fairness in ML systems. We provide some insights to explain these concepts and how to avoid the bias and ensure fairness in ML system designs.

2 Fairness and Bias in Machine Learning

Machine Learning algorithms are widely used in numerous industries. As the impact of the ML algorithms increases, our lives became more dependent on these technologies on many different aspects. We use the advances of ML in education, security, entertainment, economy etc. Since ML affects our lives so much, we expect ML to be fair. But what it is mean for a machine to fair, or how can we know if it's fair? Researchers are focused on questions like these to provide insight about the fairness and bias in ML. Here we present a report that presents a literature review to explain fairness and bias concerns with existing ML applications. We begin our report with the introduction of the concepts fairness and bias, after we explain these concepts in detail, we provide information about how to ensure fairness and avoid bias in ML and we talk about how to detect the possible bias and unfairness.

2.1 Fairness in ML

Integration of the ML systems in many different areas of our lives made us rely on these technologies. As a result of this, people using these technologies are expecting ML systems to meet both the society norms and legal standards to make decisions that can be considered fair. In literature, researchers emphasised the unfairness in ML systems as a significant problem. Several researchers showed that ML systems' quality drastically decreases against for a protected group comparing to the whole population. This issue took a lot of attention and lot of researchers focused either on the definition of fairness or developing methods to avoid the bias in data to provide fairness to the system[1].

As the attention to the fairness increased, researchers first illuminated the concept of fairness by presenting their definitions. However, our literature search showed that the fairness is still a trendy topic, and the approach of the researchers on fairness evolves as they discover different aspects that affect the fairness of the system. In one of the pioneering researches in the area, Chouldechova et al. defined the fairness in two main families of definitions; statistical and individual definition. For the statistical definition, they stated; "Most of the literature on fair classification focuses on statistical definitions of fairness"[2]. Mehrabi et al., additionally considers subgroup fairness definition in their research[3]. This family of definitions fixes a small number of protected demographic groups such as racial groups, and then ask for approximate parity of some statistical measure across all of these groups. Furthermore, they noted that individual fairness, on the other hand; ask for constraints that bind on specific pairs of individuals, rather than on a quantity that is averaged over groups[2]. A good example for this definition is provided by Dwork et al., they defined the fairness constraint in their work that as "similar individuals should be treated similarly", where similarity is defined with respect to a task-specific metric that must be determined on a case by case basis[4]. This example shows that Dwork et al., supposes that there exists a similarity metric agreed upon, which by definition requires to solve a non-trivial problem in fairness since there's no such metric. As this example suggests, there is not a single and agreed fairness definition on ML systems, many researchers keep explaining how we can be more fair on ML systems. Although, by focusing on the fairness definitions provided in different researches, we can summarize the main concerns on fairness as simply as; achieving equal results for different people unless a real meaningful distinction can be drawn between the individuals. The duality of this point of view on fairness concept leaded two different approaches in researchers, while some of them support the fairness for statistical approach, some of them put emphasis on the individual fairness. Dwork et al., pointed fingers to this issue by stating that; statistical parity speaks to group fairness rather than individual fairness, and appears desirable, as it equalizes outcomes across protected and non-protected groups. However, we demonstrate its inadequacy as a notion of fairness through several examples in which statistical parity is maintained, but from the point of view of an individual, the outcome is blatantly unfair[4]. As the information provided from these research suggest; there is no single solution that can magically fixes unfairness of every ML system, therefore understanding the reasons behind the unfairness helps researchers to solve the issues.

2.2 Understanding Unfairness

Several different researchers showed that unfairness of an ML system is based on different aspects of the environment and therefore to provide fairness to the ML we should understand and eliminate the causes of unfairness. First we present some previous work that explores the unfairness of different ML systems to provide insight about the causality behind unfairness. In theory, ML systems are not designed to be unfair, however; when in real life applications they might prove to be unfair and even can cause damage to public. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a fairly known example for this since it received a lot of attention in the literature[5, 6]. COMPAS is a tool used by U.S. to manage cases and support decision making process of courts to measure the likelihood of a defendant would rather repeat the offence or not in the future. The unfairness of this ML system discovered as the system inaccurately predicted that black defendants carried a higher risk of repeating the offence than they actually were. In another example, Goodman uncovered; Amazon's ML system that automates the applicant resume review process has unfairness against women. Amazon decided to shut down the ML system after it was found to be discriminating against women[7]. Similar to Amazon, Apple is also suffering unfairness in their ML system. The investigation is still continuing about the Apple Card's ML system that's responsible for credit limits are whether discriminating women or not. Researchers explained the unfairness of the ML system are not the algorithms fault, it is related to data that's utilized for the project. Boyd-Graber explains this claim with the pneumonia example; where the task is to classify patients into two different risk groups, low and high risk. The most accurate ML system for this task was the multilayer neural network with logistic regression, however; the rule learned by the ML system inaccurately classified asthma patients as low risk, even though asthma and pneumonia together poses a higher death risk. The algorithm resulted unfair due to the asthmatic pneumonia patients receive aggressive treatment and they go to healthcare more often, and treatments decrease the risk of death[8]. To prevent this type of unfairness, we need to understand the data first. With the big data into consideration, most of the times it's too hard to fully understand the data. One way to solve this issue is introducing humans in the loop. Experts of the area can work on the data before the ML system design to overcome possible unfairness. Similar to the pneumonia example; race, gender, socioeconomic etc. bias causes the unfairness in the ML systems. Zhao et al. stated that bias on gender, race, sociculturel etc. factors, inaccurate generalizations and malicious input are some of the factors that caused unfairness in ML systems[9]. This example supports the underlying cause of the unfairness can be caused by the bias in dataset.

2.3 Bias

Bias is the flaw of an unfair ML system. ML algorithms train and learn from provided data, but that data might contain the reflections of the biases of humans that gather them. This suggests that, every data collected by humans might contain the traces of the collectors biases and this brings the risk of having an unfair ML system. Therefore, understanding the bias is the most important task to avoid unfairness. Essentially, bias refers to the systematically distorted predictions caused by wrong assumptions. An ML system with bias produces large errors on training and make distorted predictions, this results with unfairness on the application of the ML system. As Ntoutsi et al. suggested to ensure fairness in an ML system, we need to understand the bias, alleviate the bias and be accountable for bias[10]. They suggest that, bias is a result of several different factors, therefore understanding the bias should be the first approach to avoid bias. Debiasing the data can only be applicable after we have an understanding about the bias in the dataset[11]. By understanding the data, we can reduce the bias with several different methods and with all this information in our hands, we can account for bias.

2.4 Understanding Bias

As Mitchell stated; bias is an old ML concept that refers to the assumptions made by a specific model[12]. Following this definition, Ntoutsi et al. said they consider bias as the prejudice or inclination of a decision made by an ML system which is for or against one person or group, especially in a way considered to be unfair[10]. To understand the bias in the ML systems, researchers explored different aspects of the biases in datasets.

Karimi et al. emphasized the human factor as the main cause of the bias[13]. Tufekci et al. stated that the ease of access to the distorted data sources like Twitter causes unfairness[14]. Just like these examples, there are several other works focus on the socio-technical causes of bias. In addition to socio-technical causes, Ntoutsi suggest that sensitive features and their influence on the dataset might result with over or under-representation of certain groups[10]. Foster and Grimes et al. also stated that understanding influences of the sensitive features is fundamental to avoid bias in the legal circles[15] and in medical research[16]. Furthermore, researchers also

claim that the representativeness of the data is another important aspect behind bias, they state that the representativeness of the data is caused by different biases, namely; selection, self-selection, reverse exclusion, reporting and detection bias[8, 10, 14]. Bolukbasi et al. claims another cause of the bias is data modalities. In their research, they exposed a large number of offensive associations related to gender and race on publicly available word embeddings causes unfairness in the ML system[11].

Researchers put their efforts in order to understand the bias in the data. To alleviate the bias in the data, we need to know what is the cause of the bias is. Our literature review showed that researchers discovered important information about the bias in data. The main causes of the bias are; biased labels, biased sampling, protected attributes and insufficient training. Biased labels are mainly the mistakes of humans when labeling the dataset. Biased sampling refers to the degradation of bias through initial biased labels. Protected attributes are presented as discriminations on CORPUS, Amazon and Apple cases[5, 6, 7]. Insufficient data also causes bias, when the data is not sufficient to represent small groups[1], this might result with distorted predictions.

2.5 Alleviating Bias

There are numerous researchers focused to reduce the bias in the ML system and they implemented different methods to solve different bias problems. Just like there are several reasons to cause bias, there are also several methods can applied to alleviate the bias in the data. These approaches can be splitted into three by considering at which step they are applied in the ML system. Some researchers brought pre-processing approaches to mitigate the bias, while others focused on in-processing or post-processing approaches.

Pre-processing approaches focuses on alleviating the bias before training the ML algorithm. This approach assumes fairer training data provides less distorted predictions, hence results with less bias. Researchers modified the original data by handpicking the examples close to the decision boundaries or local neighborhoods, reassigning the weights based on groups or sampling from each group. Calmon et al. presented a new method adjust to data to increase fairness while controlling the per-instance distortion and by preserving data utility for learning[17].

In-processing approaches reduces the bias by reformulating the classification problem through constraints or regularizations in the ML system. Researchers applied different methods to modify their ML algorithms to be fairer, Dwork et al. redesigned their classifier by minimizing an arbitrary loss function[4], Agarwal et al. and Zafar et al. presented constraint-based methods for logistic regression and support vector machines[6, 18], Kamiran et al. modified the splitting criterion of decision trees[19]. While most of the in-processing approaches focus on the classification process, some researchers also provided information on alleviating bias in unsupervised networks which can be considered as an emerging problem in the area[10].

Post-processing approaches mitigate the bias after the ML system is trained and learned from the data. There are two different applications of post-processing approaches; modifying the ML algorithm or the predictions of the ML algorithm. Researchers altered the algorithm in several ways such as; correcting the classification rules, changing the probability distributions in Naive Bayes models or modifying the labels of decision tree's leaves. A group of researchers conducted experiments on altering the predictions. Examples of this approach is applied by promoting or demoting predictions close the decision boundary[13], differentiating the decision boundary over groups[20] or adding an additional classifier.

Considering that understanding the bias and alleviating the bias are explained, we can focus on how we can account for bias. In this report's perspective, accountability refers to the responsibility for the design of the algorithm and its outcomes. These outcomes can have an impact on society, therefore shows importance. ML systems' accountability has different aspects; hence researchers approached to accountability in two angles, they either collect bias-aware data or explain the ML system in-depth to make sense of predictions.

2.6 Avoid Bias - Ensure Fairness

Machine learning applications are parts of the daily lives now and they affect our lives more and more everyday. Many different industries adopted ML systems to increase their productivity and different aspects of our lives became dependant to it. These ML systems make critical predictions that can have impact in our lives. This brings a question, how fair is the ML system? To answer this question, we need to understand the fairness and bias in ML systems.

Machine learning is a complicated process, therefore the problems come with it are also complicated. Fairness is one of these concepts that there is no single solution that works for every unfair ML system. To improve the fairness of the system, the underlying reason behind the distorted predictions should be examined and understood. In addition to that, there is no single definition or metric to explain fairness, hence; fairness might have different meanings with different considerations[4, 20]. Furthermore, our literature review showed that, every solution brought by researchers in order to achieve fairer ML systems approaches the issue as a process rather that an action.

There are different approaches on how an ML system is considered fair due to different fairness definitions[2]. Understanding the unfairness of the ML system can help the designers of the system to ensure fairness. A large amount of researches on unfairness of an ML system suggest that the main reason for unfairness is caused by the bias in the data rather than the ML algorithm. If the bias is the source of the problem, it should be corrected to increase fairness. Numerous researchers focused to achieve this goal, their findings agree upon the importance of understanding the bias is the first task to do. Just like different people have different assumptions, bias can be a result of different wrong assumptions. One of the reasons that there is bias on datasets is the human factor since people's assumptions might affect the data gathering or data labeling process. Understanding the bias in the data allows to alleviate it by applying the necessary processes. Researchers tried to mitigate the bias with several different methods. A group of researchers corrected the dataset by debiasing it before the training of ML algorithm. Another group of researchers modified the ML algorithm to reduce the bias, while some researchers put their efforts on altering the predictions of the ML system.

3 Conclusion

To conclude it; understanding the unfairness of the ML system should be the first step. Following the concept of unfairness, understanding the bias in the data should be the center of the focus owing to the fact that without understanding the bias it would not possible to alleviate it. After understanding the underlying reasons behind unfairness, we can eliminate those reasons with several methods. The concerns of fairness and bias should be taken into consideration on every step of the ML system's design; since collecting data, labeling data, ML algorithm training, predictions and evaluation can all be affected by wrong assumptions - bias, the fairness of the ML system might be affected by it.

References

[1] Beutel A., et al., "Data decisions and theoretical implications when adversarially learning fair representations". arXiv:1707.00075, 2017.

[2] Chouldechova A., Roth A., "The frontiers of fairness in machine learning". arXiv:1810.08810, 2018.

[3] Mehrabi N., et al., "A Survey on Bias and Fairness in Machine Learning". ACM Comput. Surv. 54, 6, Article 115 (July 2022), 35 pages. https://doi.org/10.1145/3457607

[4] Dwork C., et al. "Fairness through awareness". In Proceedings of the 3rd innovations in theoretical computer science conference, pages 214–226. ACM, 2012.

[5] Angwin J., et al., "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks", 2016.

[6] Agarwal, A., et al., "A Reductions Approach to Fair Classification", Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research. 80:60-69 Available from https://proceedings.mlr.press/v80/agarwal18a.html, 2018

[7] Goodman R., "Why Amazon's Automated Hiring Tool Discriminated Against Women". url = https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against, 2018.

[8] Boyd-Graber J., "Fairness, Accountability, and Transparency in Machine Learning" url = https://home.cs.colorado.edu/ jbg/teaching/CMSC_726/22a.pdf

[9] Zhao J., et al., "Men also like shopping: Reducing gender bias amplification using corpus-level constraints". In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017.

[10] Ntoutsi E., et al., "Bias in data-driven artificial intelligence systems—An introductory survey". WIREs Data Mining and Knowledge Discovery. 10. 10.1002/widm.1356, 2020

[11] Bolukbasi T., et al., "Man is to computer programmer as woman is to homemaker?", Debiasing word embeddings. In NIPS (pp. 4349–4357), 2016.

[12] Mitchell T. M., "Machine learning (1st ed.)". New York, NY: McGraw-Hill, 1997

[13] Karimi F., et al., "Homophily influences ranking of minorities in social networks." Scientific Reports, 8, 2018.

[14] Tufekci Z., "Big questions for social media big data: Representativeness, validity and other methodological pitfalls". In ICWSM. The AAAI Press, 2014.

[15] Foster S. R., "Causation in antidiscrimination law: Beyond intent versus impact". Houston Law Review, 41(5), 1469, 2004.

[16] Grimes D. A., Schulz K. F., "Bias and causal associations in observational research." Lancet, 359, 248–252, 2002.

[17] Calmon F., et al., "Optimized pre-processing for discrimination prevention". In NIPS (pp. 3992–4001), 2017.

[18] Zafar M. B., et al., :Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment". In WWW (pp. 1171–1180). ACM, 2017.

[19] Kamiran F., et al., "Discrimination aware decision tree learning". In ICDM (pp. 869–874). IEEE Computer Society, 2010.

[20] Hardt M., et al., "Equality of opportunity in supervised learning". In NIPS (pp. 3315–3323), 2016.