Understanding the Impossibility of Machine Learning Fairness with Data Examples

Keshav Pillutla
Mountain View High School, Mountain View, California

Abstract

This review paper examines how the three common criteria of machine learning fairness, despite their common sense and moral appeal, are often mutually exclusive. This frequently presents a challenge and the need to prioritize one criterion above the others. Specifically, the paper highlights the general overview of a machine learning model and the historical aspect that can play a role in the existing biases in models today. The paper then dives into the three criteria of machine learning fairness: independence, separation, and sufficiency. It explains the conflict between them and how it is impossible to satisfy all three conditions simultaneously, creating an imperfect machine-learning model. This paper illustrates how this comes up and plays out with real problems with real data examples and code.

Keywords: Machine learning fairness, mutually exclusive, independence, separation, sufficiency, data examples

1) Introduction

A machine learning model takes information and data from the past and uses it to predict trends and future outcomes of a certain event. This concept has a historical aspect, as discrimination and demographic disparities have influenced the trends created by these models. The societal aspect of machine learning raises the question of the impossibility of fairness, due to the inevitable bias of humans associated with artificial intelligence (AI). Machine learning fairness refers to the idea of a perfect machine learning model, one that can predict trends without bias. To explain machine learning fairness, we take a deep dive into the three most common criteria, which allow for a fair model if satisfied. These criteria are independence, separation, and sufficiency (Wikipedia contributors, 2024). Independence is the idea that the probability that the model will predict something happens is independent of a sensitive characteristic changing. A sensitive characteristic is a defining characteristic in a human being that could cause bias in a model. This means the probabilities (or error rates) will be the same even if one changes that sensitive characteristic, for example, whether the gender is male or female. Separation is similar, where an event's predictive probability must be the same even if the sensitive characteristic changes, given that the event has already happened. Sufficiency is the notion that the likelihood of something happening is the same when the sensitive characteristic changes, given that the model predicted the event would happen regardless of the characteristic. Prior research has both mathematically and theoretically proven that satisfying these three criteria simultaneously is impossible (*Inherent Trade-Offs in the Fair Determination of Risk Scores*, 2016), meaning that a perfect machine-learning model is something that we cannot achieve.

This issue emerged through the controversy with the “COMPAS” AI model, initially discovered by ProPublica (Mattu, 2023). COMPAS is a model that predicts the likelihood of

recommitting a crime based on many factors associated with the criminal being judged. Race, among many other factors, was a sensitive characteristic causing bias in the model. For example, colored people were more often predicted to commit a crime again compared to white criminals, even if their criminal records said otherwise. This started to raise doubts about the future of artificial intelligence in our world. For instance, how could artificial intelligence adapt to society's biases against certain groups of people? This paper's purpose is to explore a model's fairness criteria and demonstrate its mutual exclusivity using real-world data, so it is proven that a fair machine-learning model is often impossible to achieve.

1.1) Background on Fairness Criteria

When talking about machine learning fairness, the concept of sensitive variables, or characteristics, generally comes up. It may be considered unfair when a computer decision is based on these types of variables. Gender, ethnicity, sexual orientation, and disability are examples of what constitutes a sensitive variable. This matters in our world because when computers use these sensitive characteristics in machine learning, they introduce societal inequalities into the system, causing one race or gender, for example, to be judged more harshly than another. Rather than using human history and all our discriminatory actions and implementing it into a machine-learning model, a fair model would address these historical biases and ensure that its decisions align with the fundamental human values of equality. This type of model is only hypothetical, however, as achieving this goal is often mathematically impossible.

The basis of machine learning fairness can be narrowed down to three criteria, which are common when discussing what a machine learning model should have. These are independence, separation, and sufficiency. Independence is the idea that the probability that something will be predicted to happen should stay the same even if a sensitive characteristic changes. This can be written as follows:

$$P(R = r \mid A = a) = P(R = r \mid A = b)$$

where R is set to the predicted outcome of r and A is a sensitive characteristic, which is the changing variable, being either a or b. For example, if the model were trying to predict whether or not someone would get into UC Berkeley, the probability that the model predicts they would get in should not change if this person's race changes. Suppose two different people had the same qualifications except one person was Asian and one was European. In that case, independence states that the probability that they get admitted to Berkeley should be the same.

Separation is similar to the idea of independence. It is the idea that given the true target variable Y and sensitive characteristic A, the probability that R is predicted to happen should stay the same if the sensitive characteristic changes. This can be written as follows:

$$P(R = r \mid Y = q, A = a) = P(R = r \mid Y = q, A = b)$$

In this case, Y is set to the true outcome of q. If, for example, we were to take a male and a female who were both diagnosed with COVID-19, separation is the concept that the probability that they are predicted to have that disease must be the same, and the probability they are predicted to not have that disease must be the same. Because of this, we can say that the false

positive rates for males and females must be equal, along with the true positive rates. We can also say that the false negative rates must be equal and that the true positive rates must be equal

Sufficiency, unlike separation, is conditioned on the prediction, rather than the result. It is the concept that given the prediction R and sensitive characteristic A, the probability of Y happening should stay the same if the sensitive characteristic changes. This can be written as follows:

$$P(Y = q \mid R = r, A = a) = P(Y = q \mid R = r, A = b)$$

If we go back to the UC Berkeley example and use disability as our sensitive characteristic, the notion of sufficiency states that two people, one without disability and one with disability must have the same probability of getting into UC Berkeley, given that they were predicted to be admitted. Unlike separation, this equalizes the positive predictive value and negative predictive value and also equalizes the false omission and the false discovery rate.

It is a commonly held belief that these three criteria being satisfied pave the way for machine learning fairness. However, as stated before, these three criteria are incompatible and cannot fulfill this requirement. The mathematical aspect of this idea is beyond the scope of this paper, but another way we can prove this is through using a real-world dataset and calibrating our model to try and simultaneously satisfy all three fairness criteria considering only binary outcomes. We demonstrate this using a case study in the later sections.

1.2) Literature Review: COMPAS Scandal

As machine learning continues to grow and improve, and its concept begins to spread worldwide, doubts will naturally arise. One particular doubt was the question of whether a machine learning model can make fair judgments in certain situations without bias. The public made this doubt clear when ProPublica released an article on COMPAS AI (Mattu, 2023). COMPAS was a machine learning model created to decide the likelihood of a person, who had already committed a crime, committing another crime in the future. This was so that it could influence decisions about criminals, and when they can be set free at every stage of the criminal justice system (bond amounts, sentence, etc.) It was supposed to revolutionize the concept of prosecution, as many in the US believed human bias played too much of a role in the criminal justice system. This algorithm predicted the likelihood of recommitting another crime using several variables, including a handful of sensitive characteristics, such as race. The ProPublica article covers many instances of colored people being charged with the label that they were at high risk of committing another crime in the future, simply because race was a factor in the machine learning model.

One story the article follows is the story of Brisha Borden and Vernon Prater. Brisha Borden was charged with burglary and petty theft, and Vernon Prater was also charged on similar accounts. However, Prater had committed crimes in the past, while Borden had not. Borden, being a colored individual, was still labeled as high risk to commit another crime, while Prater, who is white, was labeled as low risk. Two years later, it had been discovered that Borden had not committed another crime, while Prater was serving an 8-year prison sentence, showing how COMPAS was completely wrong. Along with other examples, this article also goes into the overall analysis when it comes to black and white individuals charged with crimes, and

their risk scores for how likely it is they will commit another crime. It described how black defendants were twice as likely to be wrongfully labeled as high-risk than white defendants. Even when race was isolated as a factor from criminal history, age, and gender, black defendants were 77% more likely to be labeled with a high risk of committing a future violent crime and 45% with a high risk of committing a future crime overall. All of this goes to show the bias in the COMPAS predictive model, and how it could not accurately and equitably predict the likelihood of a defendant recommitting a crime.

This connects back to the fairness criteria described in this paper, as they are in clear violation since the decisions made by the model are not independent of the sensitive characteristic of race. This issue brought the uncertainty centered around machine learning fairness to the public eye, and it raised the question of whether it was possible to truly create a perfect and fair machine learning model. This question stirred up controversy because some believed that it was impossible, while others believed it was possible. One example of people thinking it is possible is a paper trying to disprove the theory of the mutual exclusivity of the fairness criteria (Flores et al., n.d.). This paper predominantly looks at other studies done on this subject and gives us reasoning as to why these studies either had limitations or had false information. It looks at the ProPublica article above and gives us insight into why it does not give the full story. The basis of this paper is that ProPublica is holding information from us. It takes a look at the different cases listed in this article from another, from a much broader perspective, to give us an understanding of how COMPAS did not have bias and was fairly accurate when judging both black and white defendants. Although all this may be true, COMPAS is one specific case of a much larger problem, meaning the question of machine learning fairness and whether it is possible to achieve still stands. That question will be answered using a dataset in the real world in later sections.

2) Methods

The questions surrounding machine learning fairness have quickly risen through the popularity of multiple artificial intelligence platforms. Can a machine learning model make decisions without bias playing a role? That question is explored through the student success dataset used in this experiment, which is publicly accessible. The dataset being used is from the country of Portugal, and it is of 3630 students who are studying a diverse set of undergraduate degrees. This dataset focuses on the students who dropped out or graduated, based on several factors, or sensitive characteristics, with this notebook focusing on gender. This experiment is centered around satisfying the three fairness criteria of independence, separation, and sufficiency at the same time when it comes to predicting the status of a student. First, the main dataset must be split into two different datasets, one with the male students and one with the females. A logistic regression is run on each dataset. This means that based on the independent variables in the dataset, the probability of the situation, in this case graduating, is estimated through machine learning. Using these probability values, the three fairness criteria can be computed. Table 1 demonstrates the results obtained from the confusion matrix created by the code in the notebook.

To calculate independence, we must find the threshold where a certain level of people drop out. A threshold, in this case, means the minimum predictive probability where the algorithm decides to label a student as a graduate. To find that certain amount, we simply saw how many people dropped out in the training set of our data, which was around 39%. Then, we

found the quantile of the male and female validation predictive probability dataset, meaning we found the value where 39% of the students' predictive probability falls below for males and females.

To calculate separation, we must take a look at the ROC curve. A ROC curve takes the relationship between false positive and true positive rates of each threshold and plots them on a graph. In this case, we must find where the female ROC curve and male ROC curve intersect, and which threshold this happens at. We can do this by interpolating the true positive and false positive rates, along with the thresholds, and using the functions being outputted to calculate where the true and false positive rates would be equal.

To calculate sufficiency, we must figure out the thresholds where a certain positive predictive value is equal. To find that certain amount, we looked at the dataset containing male and female students and saw the positive predictive value of the model run on that set. The result of this was a value of around 71.6%, and from this, we can calculate the thresholds where this is satisfied. The results of all of these allowed us to find the true and false positive rates at the thresholds given and plot these points on the ROC curve. The threshold at which all three fairness criteria are satisfied is the threshold that must be used to create this hypothetical fair machine learning model.

Table 1: *Describes the thresholds, true positive, and false positive rates where each fairness criterion is satisfied, which is all plotted on the ROC curve*

Criteria being satisfied	Threshold(s) Where Criteria is Satisfied	True Positive Rates (males, females)	False Positive Rates (males, females)
Independence	49.5%, 74.9%	77.1%, 67.8%	46.6%, 47.9%
Separation	2.6%	96.6%, 96.6%	69.4%, 69.4%
Sufficiency	61.4%, 80.2%	12.7%, 32.5%	4.6%, 30.6%

3) Results

As the notebook displays, there is no particular point on the ROC curve where independence, separation, and sufficiency are satisfied. This means that there is no threshold that we can use for this machine learning model that allows the model to be fair and unbiased. Table 2 shows that independence is satisfied at the thresholds of 74.9% and 49.5% when trying to predict that 40% of the students will drop out. At these thresholds, the false positive rates are fairly close, at around 1.3% apart, but the true positive rates differ greatly, with a difference of around 9%, showing an error in separation. Sufficiency at this threshold has an error of 16.7%, which is the difference in positive predictive values between males and females. Separation is satisfied at around the threshold of 2%. When satisfied, there is an 18% difference in the proportion of students dropping out, along with a 13.9% difference in the positive predictive values, demonstrating how independence and sufficiency are not satisfied. The sufficiency criterion is met at the thresholds of 61.4% and 80.2%. When sufficiency is satisfied, the false and true positive rates differ greatly, with a difference of 26% and 19.8%, respectively, showing

a great error in separation. The proportion of students dropping out also differs by 23.5%, which shows a large error in independence.

Table 2: *Error chart displaying how much error is in the other fairness criteria when one criterion is satisfied (what threshold or thresholds independence, separation, or sufficiency are taking place)*

Threshold	Independence (Difference in proportion of students dropping out)	Separation (Difference in false positive rates and true positive rates, respectively, for males and females)	Sufficiency (Difference in positive predictive value for males and females)
49.5%, 74.9%	Satisfied	1.3%, 9.3%	16.7%
2.6%	18%	Satisfied	13.9%
61.4%, 80.2%	23.5%	26%, 19.8%	Satisfied

These percentage errors all demonstrate the trade-offs that are presented when creating a machine learning model. In this case, satisfying independence may be correct because the percentage errors are not too bad, but it depends on the perspective of the person creating the model. All of these results are shown on the ROC curve, shown in Figure 1, with the “x” marking separation satisfaction, the dot marking independence, and the square marking sufficiency. The model, along with the ROC curve shows that it is impossible to audit our model to make it fair amongst all genders using solely these criteria.

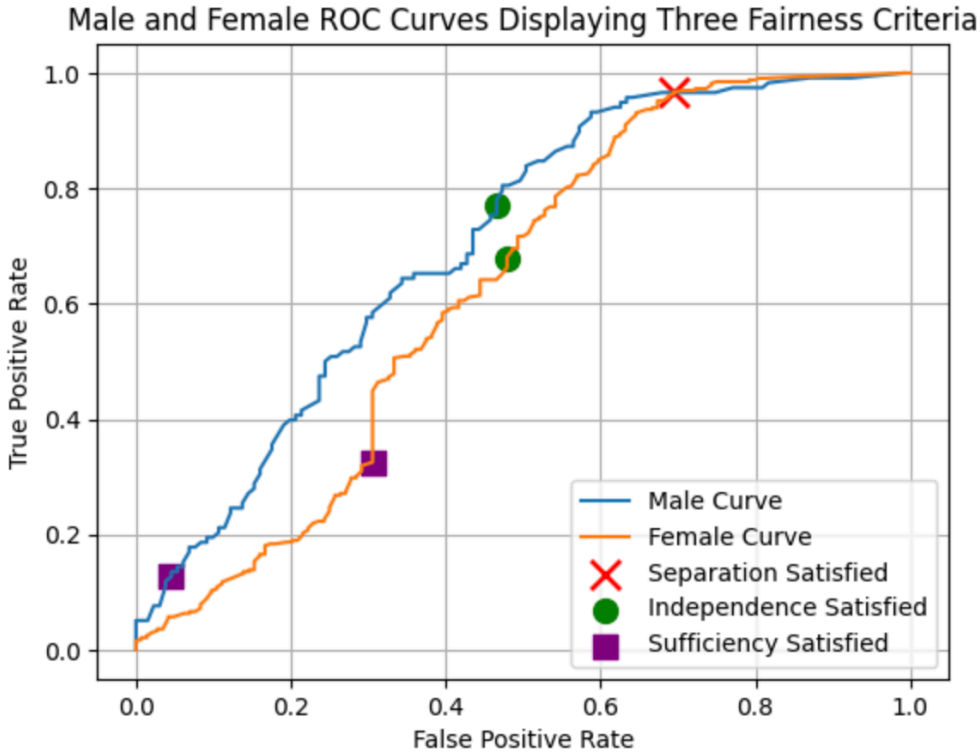

Figure 1: *Male and female ROC curves for the model fit onto the student performance data, along with the points at which independence, separation, and sufficiency are satisfied. Green dots are where independence is satisfied, red 'x' is where separation is satisfied, and purple squares are where sufficiency is satisfied.*

4) Discussion

The calibration of the machine learning model used in this experiment did not produce the desired results, demonstrating the tradeoff that occurs when creating one of these models. At certain thresholds, we can only satisfy one of these fairness criteria, which implies that when fitting a model to a dataset, one must consider the factor of which fairness criteria one wants to satisfy, whether it be independence, separation, or sufficiency. The experiment that was conducted gives a glimpse into the general problem of machine learning fairness in our world. The implications of these results are extensive because they can impact the future of our world, as it transitions to a state of artificial intelligence, machine learning, and text generation. It connects back to past prejudices and human values against groups of people because these have impacted how artificial intelligence perceives our world, and how its perception impacts its decisions. These doubts about bias in models relate to multiple studies based on the recurring issue of machine learning fairness in this world.

One example of this is the study done on these three fairness criteria, and how their mutual exclusivity can be disproved by mathematics (*Inherent Trade-Offs in the Fair Determination of Risk Scores*, 2016). This paper gives us a detailed analysis of each criterion and then proves the theory that these cannot be satisfied simultaneously, using mathematics and logic. It gives us the tradeoffs that are a result of the mutual exclusivity of these criteria, along with finding constrained, special, hypothetical cases where they could be satisfied at the

same time. The book “Fairness and Machine Learning: Limitations and Opportunities” (Barocas et al., 2023) also details the fairness criteria from a logical standpoint. After giving us an introduction to machine learning, it presents us with ideas about the historical aspects of society, like discrimination laws, along with multiple case studies to comment on machine learning fairness. The paper and book mentioned above, along with this paper give us multiple methods of proving the mutual exclusivity of the fairness criteria, which further solidifies the claim of machine learning fairness being impossible to achieve.

5) Conclusion

This paper has detailed the three common fairness criteria considered when exploring machine learning fairness. These three are independence, separation, and sufficiency. The experiment that was conducted throughout this paper demonstrates the mutual exclusivity of these three fairness criteria. Through the auditing of this model, by changing the threshold that allows for the model to output the graduating status, we discovered that different thresholds satisfy different fairness criteria. However, the ROC curve demonstrates how there is no overlap between these thresholds, implying that there is no one threshold that can satisfy all three fairness criteria. This experiment is simply an empirical example of the recurring issue in the real world. As artificial intelligence continues to grow in popularity, and as we continue to integrate it with so many aspects of our lives, the questions surrounding it regarding fairness in models become more prevalent. The controversy revolving around machine learning fairness has caused a major uproar in society because of the largely impactful repercussions it can have, especially when it comes to discrimination. This is because biased models are influenced by existing prejudices, resulting in societal inequalities. One person’s race, gender, or sexual orientation should not define the outcome of a certain situation, which is what these three common fairness criteria are centered around. However, satisfying all three of these can only happen in extremely rare cases, which is why it is generally considered impossible to do so. This paper and many other studies preceding it demonstrated the phenomenon centered around the impossibility of machine learning fairness and the mutual exclusivity of independence, separation, and sufficiency.

6) Data Availability:

The script for the code used in this manuscript can be found in this link:
https://github.com/githubkshav/StudentPerformance

7) References:

Barocas, S., Hardt, M., & Narayanan, A. (2023). *FAIRNESS AND MACHINE LEARNING*. https://fairmlbook.org/pdf/fairmlbook.pdf

Flores, Anthony, et al. *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks." the Authors Wish to Thank*. http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf

Kleinberg, Jon. "Inherent Trade-Offs in Algorithmic Fairness." *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 1, 17 Jan. 2019, pp. 40–40, https://doi.org/10.1145/3308809.3308832.

Wikipedia Contributors. "Fairness (Machine Learning)." *Wikipedia*, Wikimedia Foundation, 23 July 2024, en.wikipedia.org/wiki/Fairness_(machine_learning)#Group_fairness_criteria. Accessed 11 Sept. 2024.

Angwin, Julia, et al. "Machine Bias." *ProPublica*, 23 May 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.