Seminar Thesis

FairML and the SQF dataset

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, January, 15th 2025

Submitted in fulfillment of the requirements for the degree of B. Sc.
Supervised by FairML and the SQF dataset

Abstract

In this study we provide an introduction to the most common fairness definitions and subtleties that come with them. We advocate for tackling fairness in a wholeistic way, taking into account how the data was generated and how it will be used. This will be illustrated by a case study on the Stop, Question, and Frisk data (SQF) from the New York Police Department (NYPD).

Acknowledgement

Contents

List of Figures

List of Tables

1 Introduction

Here will be an introduction of common fairness metrics (group, individual, causal) similar to my presentaiton. Inspired by Verma and Rubin 2018 Caton and Haas 2024 Castelnovo et al. 2022
Problem of Inframarginality Corbett-Davies et al. n.d. "In this example, the incompatibility between threshold policies and classification parity stems from the fact that the risk distributions differ across groups. This general phenomenon is known as the problem of inframarginality in the economics and statistics literature, and has long been known to plague tests of discrimination in human decisions" For our case this would mean the risk of risk of the target (of being arrested, of being searched, of having a weapon) is really not the same for all groups in the true populaiton. This is realistic and is to be assumed. THen Corbett-Davies et al. n.d. argues that group fairness will no lead to individual fairness, and the optimal classifier from a utility maximazation perspective for the individual will not be fair for the group.
In case of SQF, the observed crime rate among african americans is higher (according to official statistics from NYPD). So it would make sense that more african americans are stopped because they have higher risk scores in general. But the higher risk scores for african americams should be questioned in the first place. They are of course not due to the fact that african americans are truly more likely to committ crime in the first place, but they have developed over many centuries of racial discrimination and targeted policing (lower socio-economic status and more repoerted crime rates because more polcie in these regions). So in this dataset we basically have the problem that - risk scores in the ture population are really different (african americans higher crime rate than white people) $\rightarrow$ due to historical bias, no objective truth process - do not know yet whether risk scores (a.k.a. crime rate) is higher for african americans in my sample - could be that the crime rate in sample is distributed in the same way as in the true population (african americans have higher crime rate than white people) - could be that the crime rate in sample is distributed in a different way than in the true population (african americans have lower crime rate than white people) $\rightarrow$this would be the extreme strict effect described in Kallus and Zhou n.d. where the stop decision is so biased that we explicitly target innocent african americans (this is likely not the case)
This ties into the comment in Castelnovo et al. 2022 that Separation is appropriate when the true lable Y is an objective truth. Here at first sight we would say, whether someone has commited a crime or not is an objective truth. But in reality, the fact that someone committed a crime is influenced by historic bias. Then enfrocing statistical parity here would be good ?? No because then this would e.g. lead to many innocent white people being wrongly accused of crime because after all at the present white people commit less crimes than african americans.
The story I actually want to tell in the end is that hey this is what our results show. It is not a super clear picture and maybe not what you expect, but to see the situation here clearly we have to take into account historical bias ("infected" Y) and sampling bias (PoC more easily stopped). Historical bias currently no specific method or idea to show it, sampling bias is addressed in the literature e.g. simulating the target population with weighing method.

Definitions of Fairness in Machine Learning

When one starts to get into the topic of fairness in machine learning, it is easy to get overwhelmed by the sheer amount of definitions and metrics that are out there. In this chapter we try to group them in an intuitive way and motivate them in the hope to bring some clarity to readers. It is helpful to make one or both of the two classifications. We can distinguish (a) between observational vs. causality-based criteria; or (b) group vs. individual criteria Castelnovo et al. 2022. In this paper we will combine these two, as a fairness metric is always in one of (a) and (b) e.g. a group metric is also an observational metric. This will become clearer in the following.
Broadly speaking, group fairness aims to create equality between groups and individual fairness aims to create equality between two individuals within a group. Observational fairness metrics act descriptive and use the observed distribution of the data to assess fairness while causality-based criteria make assumptions about the causal structure of the data and base their notion of fairness on this (so basically observational says, fairness is when I can measure equality from my distribution and causality says fairness is when the cause for my decision is not discriminatory against someone or a group). On the basis of these fundamental ideas, a plethora of formalisations have emerged. Most of them concern themselves with defining fairness for a binary classification task and one binary protected attribute (PA). The extension to a multiclass PA is the easiest. Multiple sensitive attributes bring challenges with them that are not as straightforward to address. Also, the extension from binary classification to other tasks, such as neural networks, LLMs is subject to ongoing research. As this work is meant to help you start thinking about fairness in machine learning, we will limit ourselves to the binary classification case. Specifically, we want to use the following running example, inspired by our case study on real data in chapter x. The crime rates in NYC should be decreased with the help of a new AI tool. Specifically, the administration orders a team of machine learning experts to design an automated decision-making system that should predict criminal activity of a person. It should be employed by police officers to decide whether to stop a person or investigate them further. Past police stops serve as training data. Given the history of racial profiling in the United States, it is reasonable to raise concerns about racial decision patterns the algorithm could learn from. First, we approach this from a group fairness perspective.

Group fairness

The notion of fairness underlying group metrics is that discrimination of certain groups of the population defined via the protected attribute should be prevented. Group fairness can be grouped into three main categories, independence, separation, and sufficiency.
Independence is in a sense the simplest group fairness metric. It requires that the prediction $\hat{Y}$ is independent of the protected attribute A, so $\hat{Y} \perp A$. In other words, the positive prediction ratio (ppr) should be the same for all values of A. For a binary classification task with binary sensitive attribute this can be formalised as demographic parity/statistical parity $P(\hat{Y}|A = a) = P(\hat{Y}|A = b)$. The other two groups of group fairness metrics, Separation and Sufficiency can both be derived from the error matrix. **Separation** requires independence between $\hat{Y}$ and A conditioned on the true label Y, so

$\hat{Y} \perp A|Y$. This means that the focus is on equal error rates between groups, which gives rise to the following list of fairness metrics:

- Equal opportunity/ False negative error rate balance: $P(\hat{Y} = 0|Y = 1, A = a) = P(\hat{Y} = 0|Y = 1, A = b)$ or $P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$ `mlr3`: `fairness.fnr`, `fairness.tpr`
- Predictive equality/ False positive error rate balance: $P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b)$ or $P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$ `mlr3`: `fairness.fpr`, `fairness.tnr`
- Equalized odds: $P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = b) \forall y \in \{0, 1\}$ `mlr3`: `fairness.equalized.odds`
- Overall accuracy equality: $P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = b)$ `mlr3`: `fairness.acc`
- Treatment equality: $\frac{\text{FN}}{\text{FP}}\big|_{A=a} = \frac{\text{FN}}{\text{FP}}\big|_{A=b}$

Equal opportunity requires the false negative rates, the ratio of actual positive people that were wrongly predicted as negative, to be equal between groups. Therefore, it is also called false negative error rate balance. Since equal true positive rates between groups are simultaneously fulfilled, one could also define equal opportunity via the true positive rate. Predictive equality follows the same principle as equal opportunity but instead of focusing on the false negatives, it focuses on the false positives. Again, if a classifier has equal false positive rates between groups, it also has equal true negative rates. With its focus on the false positive rates, predictive equality is also presented in the context of punitive tasks. Since people could experience potential harm on the basis of a positive prediction, the proportion of truly innocent people that do not deserve punishment should be kept at a minimum. For assistive tasks, such as deciding who receives some kind of welfare, a focus on minimising the false negative rate could be more relevant. Equalized odds combines equal opportunity and predictive equality. It requires that the false positive and true positive rates are equal between groups, and is in this sense stricter than either of them alone. Treatment Equality is another variation that forms the error ratio for each group and requires it to be equal. Finally, overall accuracy equality simply requires equal accuracy between groups, meaning equal proportion of correctly classified individuals in each group. **Sufficiency** requires independence between Y and A conditioned on $\hat{Y}$, so $Y \perp A|\hat{Y}$. Intuitively this means that we want a prediction to be equally credible between groups. Instead of conditioning vertically on the true labels, we now condition horizontally on the predictions. This leads to the following fairness metrics:

- Predictive parity/ outcome test: $P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b)$ `mlr3`: `fairness.ppv`
- Equal true negative rate: $P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = b)$ `mlr3`: `fairness.npv`

- Equal false omission rate: $P(Y = 1|\hat{Y} = 0, A = a) = P(Y = 1|\hat{Y} = 0, A = b)$ `mlr3: fairness.fomr`
- Equal false discovery rate: $P(Y = 0|\hat{Y} = 1, A = a) = P(Y = 0|\hat{Y} = 1, A = b)$
- Conditional use accuracy equality: $P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b) \wedge P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = b)$

Predictive parity requires that the probability of actually being positive, given a positive prediction is the same between groups. Following the same principle, we can require that the probability of actually being negative, given a negative prediction is the same between groups. If we instead look at errors again, we can require equal false omission rates between groups or equal false discovery rates between groups. False omission describes the case in which an actual positive person is predicted as negative and can be highly relevant in assistive settings, such as description of a medical treatment. False discovery rate describes the case in which an actual negative person is predicted as positive. This should be taken into account in punitive settings, in which we do not want to convict innocent people. By not only requiring one of these criteria but two simultaneously, we can build a stronger metric, like conditional use accuracy equality that requires same positive predictive values between groups and same negative predictive values between groups. Hopefully, the pattern becomes clear now. While it is easy to get overwhelmed by the amount of definitions at first, taking a closer look, it becomes clear that they are constructed in a structured way. In fact, equal false omission rate and equal false discovery rate were not introduced in the paper Verma and Rubin 2018 but it is clear that they follow the same pattern as the other metrics.
Most (binary) classifiers work with predictions scores and a hard label classifier is applied only afterwards in form of a treshold criterion. It should therefore come as no surprise that instead of formulating fairness with $\hat{Y}$ there exist fairness metrics that use the score S, which typically represents the probability of belonging to the positive class.

- Calibration: $P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b)$
- Well-calibration: $P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b) = s$
- Balance for positive class: $E(S \mid Y = 1, A = a) = E(S \mid Y = 1, A = b)$
- Balance for negative class: $E(S \mid Y = 0, A = a) = E(S \mid Y = 0, A = b)$

Calibration requires that the probability for actually being positive, given a score s is the same between groups. As the score can usually take values from the whole real number line, this can in practive be implmented by binning the scores Verma and Rubin 2018. Well-calibration is a stronger version of this, requiring that the probability for actually being positive, given a score s is the same between groups and equal to the score itself. This means, when for a set of suspects the classifier predicts a certain probability s of crime, then the proportion of people that actually committed crime should be s. Balance for the positive class takes the expectation over the predictions scores of the people that are actually positive and wants them to be equal across groups. We do not want that one the positive people of one group get on average a higher score than the positive people

of another group. The same holds for the negative class, formalized as balance for the negative class. To constrast the group fairness criteria, sufficiency takes the perspective of the decision-making instance, as usually only the prediction is known to them in the moment of decision. For example, the police, who do not yet know the true label at the time when they are supposed to decide whether someone would become a criminal. As separation criteria condition on the true label Y it is suitable when we can be sure that Y is free from any bias, so to say when Y was generated via an objectively true process (this will become clearer in the chapter on bias). Independence is best, when we want to enfore a form of equality between groups, regardless of context or any potential personal merit. While this seems be useful in cases in which the data contains complex bias, it is unclear whether this enforcments has the intended benefits, especially over the long term. reference?

Individual fairness

By creating equality between groups, it can happen that individual people are being treated unfairly. If we want to equalise e.g. the false positive rates between two groups and currently group a has a higher false positive rate than group b, this would lead us to lowering the prediction threshold for b, such that more actual negative people would get classified as positive. Or it we would need to set a higher thresholkd for group a, such that it becomes harder for them to be classified as positive. Depending on the context, either option can seem unfair. Individual metrics therefore shift the focus. The underlying idea of fairness is that similar individuals should be treated in a similar way. Different individuals should be treated in a different way. It is an intuitive idea that was already formulate by greek philosopher bothmann citation. **Fairness through awareness (FTA)** formalizes this idea as Lipschitz criterion.

$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$

This simply puts an upper bound to the distance between predictions of two individuals, which depends on the features of them. In other words, if two people are similar in their features, they should also get similar predictions from the algorithm. The challenge of FTA is the definition of precisely this equality in the feature space. Defining when two individuals are similar is not much different from defining fairness in the first place Castelnovo et al. 2022. Threre is no clear solution to this. In any case, the choice of d_X should take context-specific information into account. We want to find a distance metric, that suits the target and represents an ethical formalisation of similarity in the features.
Fairness through unawareness (FTU) or blinding This is primarly formulated as procedural rule. Blinding tells us to not use the protected attribute explicitly in the decision-making process. So at first this would simply mean to discard the protected attribute from the data during training. After training, FTU can be tested by simulating a doppelganger for each person in the dataset. This doppelganger has the exact same features with the excpetion of the protected attribute, which is flipped (easy in binary PA case). If both these instances have the same prediction, the algorithm would satisfie FTU Verma and Rubin 2018. This is actually also a from of FTA, in which we chose the distance metric to measure a distance of zero only if two people are the same on all

their features except for the protected attribute. Blinding is however to be seen critically. FTU has the problem of proxies. These are variables that are strongly correlated with the protected attribute. Therefore, it's not enough to simply mask the information of the sensitive attribute druing training because discrimination can persist via these proxis. This becomes clearer, when imagine that we remove information, such that this feature is simply not available to the classifier during training. The place of residence, however, is strongly correlated with the person's ethnicity. Thus, indirect discrimination based on ethnicity remains, even though the information was not directly available during training. Suppression therefore extends the idea of blinding and the goal is to develope a model that is blind to the sensitive attribute and the proxis. The drawback is, that it is unclear when a feature is sufficiently correlated with the sensitive attribute to be counted as proxi. Additionally, we could loose important information by removing too many these features Castelnovo et al. 2022.

Causality-based notions

The previously discussed notions of fairness can be counted a observational. In contrast to them, causality-based notions ask whether the sensitive attribute was the *reason* for the decision. If a certain (harmful) decision was made *because of* the value of the sensitive attribute of a person, we deem the algorithm as unfair. There are causality-based concepts that focus on group-level fairness and also some that focus on individual-level fairness. We want to give an intoduction to all of them, but since this category requires a new theory we will not get into great detail.
How do we make it fairer now? In principle, there are three approaches, which can be categorised as preprocessing, inprocessing or postprocessing. Depending on when they take place in the machine learning pipeline. Preprocessing methods have the idea that the data should be processed before training, so that our algorithm learns, for example, practically on corrected data and thus no discrimination can take place. This can be achieved by sampling or transformations, for example. In Processing, we have methods for which we need to have access to the algorithm, because the approach here is that we really modify the optimisation problem itself by actually taking the loss function and, for example, appending a regularisation term and thus optimising our algorithm not only for high accuracy, but also for one of the fairness metrics mentioned above. Postprocessing now works again with black box algorithms, just like preprocessing, because here we simply take the finished labels that our algorithm has predicted and also set the scores and, for example, individual group thresholds in order to then do justice to predictive equality or whatever. So here again, there really are a wide variety of methods, some adapted to specific algorithms, some more universal. With all these methods and all these definitions of fairness, it's really easy to get lost, especially when you're applying the whole thing. It is therefore very important. But when it comes to fairness, it's very important that you don't get bogged down in the details, that you always take a step back and look at your data and your algorithm in a wider context.
You can think of the whole thing as a kind of cycle, also known as a feedback loop, which consists of personal data and an algorithm. Because the fact is that our algorithm learns from data, but of course this data practically reflects our reality, i.e. our society to a

certain extent, and we then make decisions based on this trained algorithm. And so we practically have a kind of feedback loop, a kind of cycle, at each of these three points bias can be introduced into the process and, above all, bias can also be reinforced in the course of this process. So, for example, let's take a closer look at the whole thing. So let's look at the whole thing again. Let's illustrate the whole thing again using our example from before. So it's like this now, so we can imagine that in the past the police have experienced a lot of discrimination, a lot of discrimination towards the police, towards people of colour, and therefore, for example, people of colour are stopped much more often by the police, much more police presence simply takes place in certain regions where people of colour live and we therefore record much higher crime rates among people of colour. And that is then reflected. And so we, as people, as a society, make practically discriminatory decisions, whereby it can also be that bias is introduced into this cycle through the data, for example, in which our data is sampled through a distorted process and does not actually reflect our basic population, our true population, i.e. we have a disproportionate number of white people in our data set and therefore the faces of white people are then later better recognised than those of black people, simply because our sample does not actually reflect the world population.
Or the algorithm simply uses biased estimates and therefore introduces discrimination into the cycle even with correct data and no historical bias. In principle, we have now illustrated the topic of fair machine learning using the simplest case of a binary classification and a protected attribute. Of course, in reality it is often more complex. People don't just belong. We have several protected attributes People belong to different groups and we naturally have more diverse tasks such as regressions and supervised learning, where we also want to ensure fairness. So there are many more in research, so there are many more extensions of this. And in research, there are many more extensions to the current definitions of fairness. And there is also the fact that we always have to take our context and the data into account. And that is why there is not yet a single definition of fairness. And it is always important to keep the specific context, the task and the data in mind. I hope that I have been able to arouse your curiosity about the topic and thank you for your attention.

Residual Unfairness

Preposition 2: For group a the scores of the target population are always strictly higher than of the training population. This means that we will learn a comparatevily low threshold for group a. When we employ the algorithm in the target population, group a member will receive the positive outcome more easily (receive benefit of the doubt) because the thresholds is so low. For gorup b the opposite is true. The scores in the training data are really high compared to the overall population. This means we learn a high threshold for group b. When the system is applied on the whole population it will be harder for a random person from group b to receive the advantage because their threshold is so high. Applied on the SQF data this could translate as follows. First of all, the interpretation shifts. $\hat{Y} = 1$ is no longer desirable and we can interpreate scores as riskscores G_g^E. This means a high thresholds for being classified as $\hat{Y} = 1$ is desirable, a low threshold is undesirable. We assume that officers were more lenient to stop

black individuals, which means that the scores (probability of actually having committed crime) in the training population of black people are lower than the scores of the target population of black people. $G_b^{Z=1} \preceq G_b^{T=1}$. This means we will learn a lower threshold for black people(???) [1] When we apply the algorithm to the target populaiton we will be more likely to classify black people as $\hat{Y} = 1$ because the threshold is so low. White people, on the other hand, were selected more strictly. This means that the scores of white people in the training population are higher than the scores of white people in the target population. $G_w^{Z=1} \succeq G_w^{T=1}$. This means we will learn a high threshold for white people. When we apply the algorithm to the target population we will be less likely to classify white people as $\hat{Y} = 1$ because the threshold is so high. – Still unsrue if this makes sense, if a transfered it correctly.
For the other group we have many truly guilty and less truly innocent. When now 80% of truly guilty are classified as guilty in the advantaged group then we would want want 80% of the truly guilty to be correctly labelled as guilty in the disadvantaged group. This would only results in lowering the threshold for the disadvantaged group (so making it easier to predict them as guilty) if we predicted low risk scores for truly guilty people in the disadvantaged group Because for equal opportunity we are only looking at the people who were really guilty. So we are basically saying that the large proportion of truly innocent people in our sample of the disadvantaged leads to lower risk scores even in the truly guilty group of the disadvantaged (like a spill over effect). Only then it would make sense to say that a fairness intervention would compensate by setting lower thresholds for the disadvantaged group. Is this happening?
Chaper 6: Case study on SQF data Their main message is always, bias in, bias out. fairness interventions, done on the trainign data are not enough, if your sample is biased, your model will be biased (even after fairness interventions). They show this in the following way. The goal is to predict innocence of an individual. Such an ADM could help officers decide who to stop in the first place. The SQF data serves as training data and is naturally censored. The censoring process is that we only observe innocence of a person if they were stopped. But the decision to stop someone could be based on a biased decision policy. So we have our censored training data (SQF data). We know that this training data is not representative of the population of NYC in general defined via location specific variables. Kallus and Zhou use train a logistic regression classifier on the SQF data as is and use post-processing proposed by Hardt et al. to ensure Equal Opportunity or Equalized Odds. They use their a weighing technique (proposed by them and inspired by propensity score matching) to simulate the target population. The fairness intervention in the training population produces group-specific thresholds that are then applied to the target population. They use these fairness-adjusted threshold for the target population and still observe unfairness.
But of course they observe unfairness because the fairness intervention they do is a post-processing step and doesn't modifiy the classifier. What am i not getting here?

[1] Why do we learn a lower threshold for black people. Maybe something like this happens: So when a group is super leniently stopped we will have many truly innocent and few truly guilty.

Bias in, bias out - an alternative perspective

Rambachan and Roth n.d. take a different perspective on the problem of biased training data than Kallus and Zhou n.d. The mechanism they describe works as follows I think!!?): Black people are more leniently stopped, leading to higher stopping rates in for black people in the training data, meaning more training data for this group. Because we stop black peopel more leniently, we record nany innocent black people in our data. In Kallus and Zhou n.d. this would lead to a lower learned threshold [2] for black individuals. Applied on the target population this would mean that we would predict too many false positive. The threshold estimated from the training data is so low that we classify to many people as guilty because in the target populations the scores are actually higher and meet the threshold easily. In Rambachan and Roth n.d. they say that by stopping (searching, they actually talk about searching, not stopping) black people so leniently, our sample for black people comes actually pretty close to the target population. In other words, the training data for black people is pretty close to the target data for black people, which means that our classifier will work well on the target population for black people.
To summarise, in Kallus and Zhou n.d. bias against a group results in a less representative sample. In Rambachan and Roth n.d. bias against a group results in a more representative sample.
Theorem 1
The prediction for african americans is weakly decreasing in tau. This means, as tau increases (so racial bias increases), the expected value for Y gets actually lower, so closer to zero, so less often predicted to have a contraband. What is happening? Higher tau means lower searching threshold for african americans. So the data for african americans becomes "more noisy", more and more innocent people come into our sample, so we predict lower risk for african americans. In Rambachan and Roth n.d. paper this translates to a more representative training data for african americans and thus also better performance on the general population of african americans. In Kallus and Zhou n.d. paper the mechanisms is the same, we also estimate lower risks cores for african americans, but then sth else happens. I think in Kallus we then do a fairness intervention that leads us to setting a LOWER threshold for african americans, meaning we predict them as guilty more easily to achieve the same FPR as in the other group. I think in kallus they first formulate it in the strict way, where the police is so biased against african americans that the stopped african americans are LESS likely to actually have a weapon than the general population. But they relax this setting afterwards.
My big questions is is these two papers are actually contradicting each other. I think they do not. What both are essentially saying is that if the distribution of the target in training and target population is different, then there will be a problem. Kallus and Zhou n.d. looks at the situation in which training and target datahave different distributions in both groups. In the stricter scenario the difference in target and train exists for both gorups and is going in opposite directions (e.g. train of a is underestimation and train of b is overestimation). In the relaxed scenario the difference in target and train exists for both groups but goes in the same direction, I think is is just more severe for the

[2] first this leads to lower risk scores for black individuals. And then via fairness adjustments (e.g. for equalized odds) this leads to lower thresholds for black individuals.

disadvantaged group. In Rambachan and Roth n.d. we say that our limited sample is not biased necesariliy in itself in the sense that the distribution of the target in our sample is different from the distribution of the target in the target population per se. But what happens is that the sample is limited and therefore only cuts out a piece of the target population that is not representative. Therefore, when we collect more data we come closer to the target population and our classifier will work better on the target population for the group with more data.
What happens if we train the logistic classifier (to predict weapon yes no) on the SQF as is (Kallus), don't do a post processing fairness intervention (NO Hardt et. al) and test the classifier on the target population (that is created via the weighing method of Kallus and Zhou)? I think according to Rambachan and Roth n.d. we should observe bias reversal.

A Electronic Appendix

Data, code and illustrations are available in electronic form.

References

Castelnovo, Alessandro et al. (Mar. 2022). "A Clarification of the Nuances in the Fairness Metrics Landscape". In: *Scientific Reports* 12.1, p. 4209. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07939-1. (Visited on 12/23/2024).

Caton, Simon and Christian Haas (July 2024). "Fairness in Machine Learning: A Survey". In: *ACM Computing Surveys* 56.7, pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3616865. (Visited on 12/23/2024).

Corbett-Davies, Sam et al. (n.d.). "The Measure and Mismeasure of Fairness". In: ().

Kallus, Nathan and Angela Zhou (n.d.). "Residual Unfairness in Fair Machine Learning from Prejudiced Data". In: ().

Rambachan, Ashesh and Jonathan Roth (n.d.). "Bias In, Bias Out? Evaluating the Folk Wisdom". In: ().

Verma, Sahil and Julia Rubin (May 2018). "Fairness Definitions Explained". In: *Proceedings of the International Workshop on Software Fairness*. Gothenburg Sweden: ACM, pp. 1–7. ISBN: 978-1-4503-5746-3. DOI: 10.1145/3194770.3194776. (Visited on 11/16/2024).

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, January, 15th 2025

Juliet Fleischer