The Societal Challenge: Legal Perspectives on Discrimination

Robin Schmidt & Floyd Kretschmar
MSc Informatik
University of Tübingen
Matriculation number 4255055 and 4205979
[rob.schmidt|floyd.kretschmar]@student.uni-tuebingen.de

Abstract

This report gives a basic introduction to fairness with regards to machine learning. We give a short overview over a multitude of existing works in the field and discuss their findings as well as their implications. The discussed papers and articles describe why the issue of fairness is a fundamental challenge for the field of machine learning and how it relates to the idea of learning from data. Furthermore, the impact of machine learning in the specific area of criminal risk assessment and automated decision making is evaluated. We put special attention on the relationship between machine learning systems and existing laws and judicial processes and also present a set of short and long term challenges to be addressed by both the machine learning community as well as overall society with regards to the fair use of machine learning.

1 Introduction

The rise of machine learning in today's society through automated, data-powered decision making systems in various application fields with high impact raised increasing attention regarding the possible inherited discrimination for people with sensitive characteristics (e.g. gender, race or religion) through the commonly used machine learning techniques.

For the remainder of this paper we will use the statistical framework introduced in [4] to describe the process of machine learning. Here Berk et al. propose the idea of a population with a limitless number of i.i.d observations that are sampled from a single joint probability distribution $P(Y, L, S)$. The target variable Y is the outcome of interest, L represent the legitimate predictors and S are the protected predictors like race or gender. In this population there exists a function $f(L, S)$ linking the predictors L, S to the expectation of Y. When a fitting procedure $h(L, S)$ is applied to the data, it produces a so called hypothesis $\hat{f}(L, S)$ which is the source of the predictions $\hat{Y}$. The procedure $h(L, S)$ can either be seen as approximating the true response surface, resulting in a $\hat{f}(L, S)$ that will be a biased estimator for $f(L, S)$. If the estimation target of $h(L, S)$ is instead acknowledged to be an approximation of the true response surface, this approximation can be estimated by $\hat{f}(L, S)$ in an asymptotically unbiased manner.

In this context, defining any kind of target variable (outcome of interest) or multiple class labels is always a very subjective process, where the presented problem could unintentionally be parsed in a way which systematically disadvantages certain classes [2, 3]. Moreover, the training data could be biased by either considering cases in which prejudice has played a role or simply over- or underrepresenting a certain class [2, 3]. If such data gets used, it would lead to a discriminatory model, which would have a disadvantageous impact on a certain subgroup of people. This process can get enhanced through disadvantageous feature selection, proxies (non sensitive features correlate

to class membership) or masking, which allows decision maskers to mask their prejudicial views by using any of the previously mentioned approaches [2, 3]. All of these problems introduce a lot of legal issues which are yet to be solved.

2 Relevant Work

In this section we want to individually discuss some of the related literature [1, 2, 4] by outlining the main idea and key results.

2.1 Fairness in Criminal Justice Risk Assessments: The State of the Art

In the paper "Fairness in Criminal Justice Risk Assessments" [4] the authors explore the concept of fairness in the technical and mathematical context. More specifically, they are interested to find a way, a notion of fairness can be operationalized in the context of machine learning and how it applies to the specific case of criminal justice risk assessment. The following discussions of fairness are based on the statistical framework presented in section 1. For ease of exposition the paper limits itself to the case where Y and $\hat{Y}$ are binary.

In the first parts of the paper, the underlying statistical framework is defined. The fairness definitions proposed are based on the accuracy measurements defined by the confusion matrix[1]. More specifically, the authors define fairness as an equality of accuracy across all protected group categories. Imposing this equality constraint for each kind of accuracy defined by the confusion matrix, leads to the following definitions of fairness:

1. **Overall accuracy equality:** equal probability of correct classification $\frac{t_p+t_n}{N}$
2. **Statistical parity:** equal probability of predicting failure/success $\frac{t_p+f_p}{N}$ or $\frac{t_n+f_n}{N}$
3. **Conditional procedure accuracy equality:** equal probability of correct classification, given the actual outcome: $\frac{t_p}{t_p+f_n}$ or $\frac{t_n}{t_n+f_p}$
4. **Conditional use accuracy equality:** equal probability of an actual outcome, given the prediction: $\frac{t_p}{t_p+f_p}$ or $\frac{t_n}{t_n+f_n}$
5. **Treatment equality:** equal ratio between false negatives and positives: $\frac{f_p}{f_n}$ and $\frac{f_n}{f_p}$
6. **Total fairness:** All previously notions of fairness are achieved simultaneously

For the discussion of fairness it is assumed that $f(L, S) = \hat{f}(L, S)$ as to no conflate discussions about fairness and accuracy. But the notions about accuracy discussed in section 1 also hold true for the probabilities defined by the confusion matrix of Y and $\hat{Y}$.

In the next part follows a discussion of necessary tradeoffs between accuracy and fairness as well as between different kind of fairness. The paper states that " [...] excluding S will reduce accuracy. Any procedure that even just discounts the role of S will lead to less accuracy." [4]. The authors also explore the conflict between conditional use and procedure accuracy equality by citing the following impossibility theorem: "When the base rates[2] differ by protected group and when there is not separation[3], one cannot have both conditional use accuracy and equality in the false negative and false positive rates." [6, 10]. The authors suggest, that "the key tradeoff will be between the false positive and false negative rates on the one hand and the conditional use accuracy on the other." [4].

In the final parts of the paper multiple approaches for solving the issue of fairness in machine learning are introduced and briefly discussed. The paper explores three different main strategies, which can be combined.

1. **Pre-Processing** is the elimination of sources of unfairness in the data before formulating $h(L, S)$. Examples include the removal of linear dependencies between L and S, the

[1] A full explanation of the confusion matrix can be found in section 5

[2] proportion of actual failures/successes $\frac{t_p+f_n}{N}$ or $\frac{t_n+f_p}{N}$

[3] separation = "perfectly accurate classification is possible" [4]

rebalancing of base rates or the random transformation of predictors, such that $P(Y, L, S)$ is less dependent on S.

2. **In-Processing** means including the adjustments for fairness in the process of constructing $h(L, S)$. One example of this, is enforcing more fair results according to the defined notions of fairness, if the initial prediction $\hat{Y}$ had substantial uncertainty.
3. In **Post-Processing** $h(L, S)$ is applied first, and its results are adjusted afterwards to account for fairness. A possible approach for Post-Processing is the reassignment of class labels after classification with the goal of minimizing classification errors subject to a particular fairness constraint.

The authors note, that all the corrections for fairness presented are themselves agnostic about "what the target outcome for fairness should be." [4]. They argue that a discussion about the benchmark according to which equality is achieved, is just as important, as it makes the determination of tradeoffs more complicated.

2.2 Big Data's Disparate Impact

The whitepaper "Big Data's Disparate Impact" [2] by Solon Barocas and Andrew D. Selbst is separated into three main parts, which deal with slightly different topics regarding fairness in machine learning, in particular data mining. The first part focuses on the various ways through which data mining can discriminate certain classes, while the second and third part discuss the liability issue of discrimination in data mining for the american title VII (equal employment opportunity) [12] of the civil rights act and the difficulty for future legal reforms. According to their studies, there are five main structures in data mining which can cause discrimination for certain classes. In particular, these are the "definition of the target variable and class labels" (I), "training data" (II), "feature selection" (III), "proxies" (IV) and "masking" (V) [2]. All of these topics have already been clarified and described regarding their extent in section 1 and therefore won't need special attention here.

In the american civil rights act, especially in title VII, there are two presented notions for unfairness, namely "disparate treatment" and "disparate impact", which also find usage in the presented whitepaper. While disparate treatment describes an uneven behavior towards a certain person due to a sensitive characteristic (e.g. gender, race or religion), disparate impact represents a neutral rule which treats everyone equally in form, but has a damaging effect on a subset of people with such a certain characteristic [13, 14].

In their whitepaper Barocas and Selbst argue that formal liability in disparate treatment doesn't correspond to any special step within data mining and that using a protected class as an input for any classificatory model should be a legal harm, because this process corresponds to the employer classifying and differentiating potential hires according to exactly this protected class [2]. They also show that the disparate treatment either occurs at the decision to apply a biased predictive model or when the biased result gets used for the ultimate hiring decision and draw the conclusion that the disparate treatment doctrine doesn't regulate discriminatory data mining to a satisfying extent [2].

While considering the disparate impact doctrine, the authors state that in such a case the plaintiff must prove that "a particular facially neutral employment practice causes a disparate impact with respect to a protected class" [2, 12]. In response, the defendant-employer is then allowed to justify the challenged practice by showing the job relation and business necessity. The plaintiff then still has the chance to show that an alternative, less discriminatory employment practice could have been used instead [2]. For the case of data mining this means that liability regarding disparate impact can be caused by using a non job related target variable [2]. As soon as the target variable is shown to be job related, there are two questions which need to be answered. First, whether or not the model is predictive of the trait and secondly if the model with statistical significance predicts what it is supposed to predict [2]. Barocas and Selbst also explain that it is hard to know which features would make an existing model more or less discriminatory and therefore proving that a less discriminatory alternative would exist becomes a very hard task to solve [2].

The presented *internal* issues with data mining are fundamental questions that need to be addressed or can't be solved properly. For example, the target variable will always inherent certain kinds of biases, because a target variable must contain judgments about what is really important in the presented problem [2]. Additionally, a solution to the issues with training data labeling need to

compromise between forbidding employers from using past discrimination and allowing them to use historical data of good employees [2]. For skewed data sets the employer needs to recognize the type of bias, have access to the underlying data and needs the possibility to collect more data [2]. Otherwise, oversampling underrepresented communities can clear up some of the bias [2]. Statistical discrimination in the area of feature selection is avoidable if there is the possibility to gather additional or more granular data [2]. Otherwise minimizing the error rate between groups can help to improve this aspect [2]. Lastly, for proxies there needs to be a threshold which defines when a correlation between an attribute and class membership becomes alarming, as well as when it is sufficiently relevant to use it despite being highly correlated to class membership [2].

2.3 Machine Bias

The article "Machine Bias" by Julia Angwin and Jeff Larson describes the findings of ProPublica with regards to the risk assessment tool COMPAS. Their findings seem to confirm a lot of the problems described by [2], [3] and [4]. First of all, they found problems with regards to overall accuracy of predicting future crimes, which was "only 61 percent [...] for [committing] any subsequent crimes withing two years." [1]. Moreover they found that false positive rates for black defendants was "almost twice the rate as white defendants" [1] and "White defendants were mislabled as low risk more often than black defendants" [1].

The authors go on to explain history of criminal risk assessment in the US overall and explain the motivation for deploying algorithmic risk assessment: "If computers could accurately predict [...] new crimes, the criminal justice system could be fairer and more selective about who is incarcerated and for how long." [1]. The article then explores in more detail, how modern algorithmic tools have performed in making risk predictions and how these predictions differ among protected groups.

In the final part of the article the authors describe how algorithmically generated risk scores generated are being used today. They explain how these systems are deployed in various US states and jurisdictions and which degree of impact they have had on decision making processes of judges, prosecutes, defenders and other parts of the criminal justice system.

3 Discussion

From the presented papers as well as additional literature research for the missing EU law structures we conclude that currently the discrimination law structures aren't well prepared for the challenges which are brought up by automated decision making systems [2, 3, 5]. This applies for article 14 of the European Convention on Human Rights [7], the article 21 of the Charter of Fundamental rights of the EU [8], as well as title VII [12] of the american civil rights act [2, 5]. Moreover, it is going to be interesting how the European General Data Protection Regulation (GDPR) will influence this subject in the upcoming years [5, 11].

All of the presented papers describe issues within the machine learning process which are hard if not impossible to solve for future liability improvements and the authors of [2, 4] argue that discussion about fairness in machine learning is fundamentally a discussion about tradeoffs. Berk et al. state that the problem at the heart of fairness in machine learning is the difference in base rate across protected groups which "can cascade through fairness assessments and lead to difficult tradoffs." [4]. In terms of discussing these tradeoffs, the authors make multiple suggestions. First, the tradeoffs need to be explicitly represented and available as tuning parameters. Secondly, future measures of fairness should be formulated in such a way, that trade-offs can be made with them. And thirdly, the determination of tradeoffs should ultimately fall in the hands of the stakeholders.

In this field there have also been various works like [9, 13, 14] and many more for detecting and removing discrimination in automated decision making systems.

4 Summary

In conclusion, all of the papers present the issue of fairness in machine learning as a fairly complex one. Recognizing and exploiting patterns in data is at the very core of machine learning. If the underlying structures contain discriminatory or biased patterns, this will be reflected in the resulting algorithms. The goal has to be to find ways to quantify and operationalize fairness in a way that

makes these biases obvious and allows to account for them. This becomes especially important, as these systems are used more and more frequently to make decisions about peoples lives and that have long lasting effects, like automated decision making systems used in the criminal justice system.

5 Appendix

	$\hat{Y}_f$ Failure predicted	$\hat{Y}_s$ Success Predicted	Conditional Procedure Accuracy
Y_f Failure - A Positive	t_p true positive	f_n false negative	$\frac{t_p}{t_p+f_n}$ True Positive Rate
Y_s Success - A Negative	f_p: false positive	t_n: true negative	$\frac{t_n}{t_n+f_p}$ True Negative Rate
Conditional Use Accuracy	$\frac{t_p}{t_p+f_p}$	$\frac{t_n}{t_n+f_n}$	$\frac{t_p+t_n}{t_p+f_p+t_n+f_n}$

- **Sample Size:** $N = t_p + f_p + t_n + f_n$
- **Base Rate:** proportion of actual failures/successes $\frac{t_p+f_n}{N}$ or $\frac{t_n+f_p}{N}$
- **Prediction Distribution:** proportion of predicted failures/successes $\frac{t_p+f_p}{N}$ or $\frac{t_n+f_n}{N}$
- **Cost Ratio:** ratio between false negatives and positives: $\frac{f_p}{f_n}$ or $\frac{f_n}{f_p}$

References

[1] J. Angwin and J. Larson. Machine bias, May 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`. [Accessed: 2019-05-26].

[2] S. Barocas and A. D. Selbst. Big data's disparate impact. *SSRN Electronic Journal*, 2016. doi: 10.2139/ssrn.2477899.

[3] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.

[4] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments. *Sociological Methods & Research*, 104(6), 2018. ISSN 0049-1241. doi: 10.1177/0049124118782533.

[5] L. Carmichael, S. Stalla-Bourdillon, and S. Staab. Data mining and automated discrimination: A mixed legal/technical perspective. *IEEE Intelligent Systems*, 31(6):51–55, Nov 2016. ISSN 1541-1672. doi: 10.1109/MIS.2016.96.

[6] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163, 2017.

[7] Council of European Union. European convention for the protection of human rights and fundamental freedoms, as amended by protocols nos. 11 and 14, 1950. `https://www.refworld.org/docid/3ae6b3b04.html` [Accessed: 2019-05-26].

[8] Council of European Union. Charter of fundamental rights of the european union, 2012. `https://www.refworld.org/docid/3ae6b3b70.html` [Accessed: 2019-05-26].

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams, editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268. ACM, 2015. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783311. URL `https://doi.org/10.1145/2783258.2783311`.

[10] J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016. URL `http://arxiv.org/abs/1609.05807`.

[11] J. Singh. The tech-legal aspects of machine learning: Considerations for moving forward, 2016. URL `http://www.mlandthelaw.org/papers/singh.pdf`.

[12] U.S. Equal Employment Opportunity Commission. Title VII of the Civil Rights Act of 1964, 1964. URL `https://www.eeoc.gov/laws/statutes/titlevii.cfm`. [Accessed: 2019-05-19].

[13] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1171–1180. ACM, 2017. ISBN 978-1-4503-4913-0. doi: 10.1145/3038912.3052660. URL https://doi.org/10.1145/3038912.3052660.

[14] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. URL http://jmlr.org/papers/v20/18-262.html.