The Societal Challenge: Legal Perspectives on Discrimination

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

Abstract

This template provides guidance on the structure for your 4 page report. Please feel free to deviate from the proposed structure if you feel that it is useful; but try to follow the spirit of the guidelines. In the abstract, summarize the topic of your report in about 5-8 lines of text. Do not cite your assigned papers here, but instead give a very concise overview over the insights you report on in this text.

1 Introduction

Description of Fairness in ML (target variable/class labels, training data, feature selection, Proxies, Masking)

Defining any kind of target variable or class labels is always a very subjective process, where the presented problem could unintentionally be parsed in a way which systematically disadvantages certain classes [1]. In addition to that, 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 [1]. If this data gets used, it would lead to a discriminatory model.

As the statistical framework to describe the process of machine learning [3] proposes the idea of a population with a limitless number of I.I.D observations that are sampled from a single joint probability distribution $P(Y, L, S)$. 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.

2 Relevant Work

2.1 Fairness and machine learning: Limitations and Opportunities

Discuss your first assigned paper [2]. Outline the main idea and key results. If suitable, reproduce key mathematical insights. Ideally, also provide critical comments of your own were suitable. But make sure to clearly delineate the ideas and experiments in the assigned paper from your personal opinion or analysis.

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

In the paper "Fairness in Criminal Justice Risk Assessments" [3] 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 $\hat{Y}$.

Following these definitions is 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."[3]
The authors also explore the conflict between conditional use and procedure accuracy equality by citing the 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."[5] [4]. 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."[3]

2.3 Big Data's Disparate Impact

The whitepaper "Big Data's Disparate Impact" [1] 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 discriminitate 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) [6] 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) [1]. All of these topics have already been clarified and described regarding their extent in section 1 and therefore won't need special attention here.

[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"[3]

In the american civil rights act, especially in title VII, there are two presented cases for discrimination, 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 particular 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.

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 [1]. 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 [1].

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" [1] [6].[4] In response, the defendant-employer is then allowed to justify the challanged practice by showing the job relation and business necessity.[5] The plantiff then still has the chance to show that an alternative, less discriminatory employment practice could have been used instead [1]. For the case of data mining this means that liability regarding disparate impact can be caused by using a non job related target variable [1]. 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 [1]. 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 [1].

The presented legal reform possibilities can be separated into *internal* data mining issues as well as politicial and constitutional *external* constraints [1].

2.4 Machine Bias

Discuss your forth assigned paper. Outline the main idea and key results. If suitable, reproduce key mathematical insights. Ideally, also provide critical comments of your own were suitable. But make sure to clearly delineate the ideas and experiments in the assigned paper from your personal opinion or analysis.

3 Discussion

In this section you can summarize and link your assigned reading. Try to distill an overall insight from the papers, not to make a laundry list of individual results. Did you come across open questions that were not answered in the papers? Are there hidden pitfalls or problems that, in your opinion, the papers do not solve or marginalize? Provide a critical but constructive reading without being dismissive. Ideally, try to do some literature research of your own to find follow-on papers or related works.

4 Summary

Provide a concise summary of your findings, in about 3-10 lines of text.

5 Appendix: Possible Additions

Here we can provide all relevant additional information.

[4] 42 U.S.C.§2000e-2(k)(1)(A)

[5] *Id.*

	$\hat{Y}_f$ 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] S. Barocas and A. D. Selbst. Big data's disparate impact. *SSRN Electronic Journal*, 2016. doi: 10.2139/ssrn.2477899.

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

[3] 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.

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

[5] 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`.

[6] 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.