Towards Diagnosing Accuracy Loss in Discrimination-Aware Classification: An Application to Predictive Policing

Zubin Jelveh ZJ292@NYU.EDU
New York University, Brooklyn, NY, USA

Michael Luca MLUCA@HBS.EDU
Harvard Business School, Boston, MA, USA

Abstract

Prediction algorithms are increasingly used to forecast outcomes of processes that are societally sensitive. In response, algorithms have been developed to produce fair classifications but at the potential cost of accuracy. In this work, we present a framework for modeling the pathways by which sensitive variables influence – and are influenced by – nonsensitive variables. These pathways allow us to discern between two types of accuracy loss: justified reduction due to underlying discrimination in the data, and over-adjustment due to the removal of nonsensitive predictive information. We also present a framework for adjusting input data to remove the association between sensitive and nonsensitive predictors and assess its ability to produce fair classifications. Finally, we apply our methodology to a new dataset in the criminal justice domain.

1. Introduction

The promise of predictive policing, or the use of data analysis to forecast future criminal activity (Perry, 2013) has gained in popularity in recent years. In 2013, the Chicago Police Department used administrative data on previous criminal justice-related activity to construct a list of people in the city predicted most likely to be involved in a violent act (Stroud, 2014). Individuals at the top of the list received visits by the police department. A concern with this and other individual-level predictive tools in the realm of criminal justice is that algorithmic results will be driven by the historical association between ethnicity and criminal activity. As has been recently documented (Calders & Verwer, 2010), if the underlying data generating process shows an imbalance in outcomes involving a protected group, a classifier built on this imbalanced data is likely to make predictions that, if acted upon, would increase the imbalance.

A nascent literature has developed in recent years on the construction of discrimination-aware algorithms, i.e. predictive tools which protect against results driven by a sensitive variable such as race, gender, or neighborhood (Pedreshi et al., 2008; Kamiran & Calders, 2009; Zemel et al., 2013; Friedler et al., 2014). A common concern across this work is that prediction accuracy tends to fall when a dataset or algorithm is adjusted to account for discrimination.

There are two chief causes of accuracy loss. First, if the underlying data generating process has indeed been discriminatory, then a discrimination-aware algorithm by its nature will produce predictions targeted for an environment where this disparate impact does not exist. Measuring the accuracy of such a classifier on data from a discriminatory regime is highly likely to produce weaker results when compared to a classifier that does not take discrimination into account. Second, accuracy loss could be due to over-adjusting (Berk, 2009) for sensitive variables. This occurs when, in the process of achieving fair predictions, a discrimination aware algorithm removes nonsensitive predictive information in the modeling process.

In this work, we introduce a simple framework for adjusting input data to produce fair results. An advantage of our method is it allows us to reason about the situations in which we should expect accuracy loss to result from actual bias, and hence be acceptable, and the situations where accuracy loss is due to over-adjustment for sensitive variables. Another strength of our approach is that the transformed predictors retain the same interpretation as the original variables. Additionally, we argue that our method should be a baseline against which other algorithms should be compared. Finally, we introduce a new publicly available dataset for discrimination-aware classification in the criminal justice realm and apply our approach to it.

2. Related Literature

A key aspect of the recent work in discrimination-aware algorithms has been determining the proper criterion for certifying fair predictions. Let C be the binary class to be predicted, $\widehat{C}$ the class predictions, S the binary sensitive attribute, and X the set of other nonsensitive attributes. To make things concrete and aligned with the criminal justice setting in the United States, if $C = 1$ then an individual is predicted as high-risk for committing a crime in the future. If $S = 1$ then an individual is of a minority ethnicity such as black or Hispanic and white otherwise.

Various papers (Kamiran et al., 2013; Zemel et al., 2013; Kamishima et al., 2011) have proposed some combination of the following four metrics for determining whether the output of a classifier is fair or discrimination-free: 1) $P(\widehat{C} = 1|S = 1) - P(\widehat{C} = 1|S = 0) \approx 0$, the probability of being predicted high-risk is the same between the sensitive and nonsensitive groups. 2) $P(S = 1|\widehat{C} = 1) - P(S = 1) \approx 0$, the proportion of high-risk individuals who are in the sensitive group is the same as for the entire population. 3) The sensitive attribute should not be predictable from the variables entering the discrimination-aware classifier. 4) The association as measured by between $\widehat{C}$ and S should be very small.[1][2]

A number of the proposed approaches for discrimination-aware classifiers directly optimize one of the above the criteria. For example Zemel et al. (2013) minimizes 1) and Kamishima et al. (2011) include a regularization term to reduce 4). Our approach is distinct in that it is agnostic to the notions of fairness above, but we show that it is still able to substantially reduce discrimination.

With respect to empirical data, our paper is most similar to Kamiran et al. (2012) who used Dutch data to predict whether a person is a crime suspect. Our dataset includes an order of magnitude more predictors. In future work, we will show that accuracy loss can be attenuated by the inclusion of additional predictors.

3. Concept

Our goal is to predict an outcome C using predictors x_1 to x_n. There also exists a collection of sensitive variables s_1 to s_k. The idea is to transform each predictor x_i such that it so no longer correlated with some function of the sensitive variables, $f(s_1, \ldots, s_n)$. In the simplest case, $f(s_1, \ldots, s_n) = \sum_{i=1}^{n} w_i s_i$, where w_i are variable-specific weights. For the following, we assume there is only one sensitive variable, however it should be noted that any function of sensitive variables can be taken into account. To produce the transformation, we model each x_i as a linear function of s:

$$x_i = \alpha + \beta s_i + e_i \tag{1}$$

where α and β are parameters and e is the empirical error. When α and β are estimated via least squares, e will be uncorrelated from s. The interpretation is that e is the part of x that can't be explained by s. It's straightforward to show the zero correlation between e and s. The simple linearity assumption provides a baseline against which other discrimination-aware classifiers can be compared. In particular, algorithms that target both linear and nonlinear associations between sensitive and nonsensitive predictors should do at least as well as our approach to justify potentially extra complexity.

A valid critique of our approach is that it's too indiscriminate in removing predictive information from the model, i.e. it over-adjusts for sensitive variables. While our primary goal is prediction, it's insightful to form a causal perspective of how the sensitive variable affects, or is affected by, other nonsensitive variables. The causal process we refer to here is that which generates the discriminatory data. In other words, the reason an outcome such as earnings is "caused" by gender is due to discrimination in the pathway from gender to earnings, and not a statement about the direct effect of gender *only* on earnings. By decoupling the gender-earnings pathway we aim to break the discriminatory link. This framework allows us to reason about the scenarios under which accuracy reduction from altering the input variables is due purely to discrimination or both discrimination and over-adjustment.

Imagine the simple linear causal model on the left-hand side of Figure 1 where x_1 affects both x_2 and x_3 which then affect the outcome of interest C. As per the statistical discrimination literature in economics (Phelps, 1972), one reason to be shy about removing the information contained in a sensitive variable is that it stands in for unobserved nonsensitive characteristics (such as productivity or intelligence). Where the sensitive and unobserved variables are located in this causal model inform the extent to which we should care about indiscriminate information removal.

First, let x_1 be the sensitive variable and x_3 be unobserved. We may be tempted to include x_1 as a predictor since x_1 could act as a good stand-in when predicting C. Of course, if our goal is fairness, then the only part of x_3 we can capture is the part affected by levels of the sensitive variable. Hence, x_1 should not be included as a predictor. Additionally, de-correlating x_2 from x_1 leaves us the part of x_2 that is not driven by the sensitive variable. This causal structure is appropriate if our concern is limiting the influence

[1] See Friedler et al. for a formal discussion of 3) and Kamishima et al. for definitions of 4)

[2] When both C and S are binary, then 1) and 2) are equivalent and imply independence between $\widehat{C}$ and S.

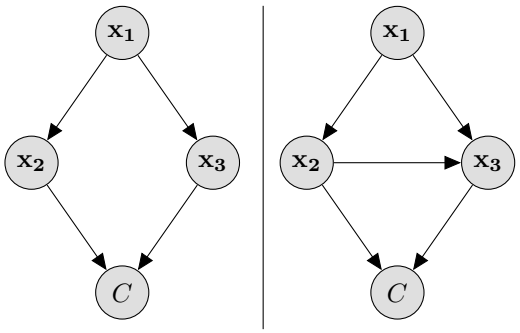

Figure 1. Two causal models for how a sensitive variable affects, or is affected by nonsensitive variables.

of variables that are either set at birth (race, ethnicity, gender, sexual orientation) or unlikely to be manipulated (age, disability status). The end result would be that indiscriminate information removal is not a concern and any accuracy loss would be due to discrimination-purging only.

Second, let x_2 be the sensitive variable (such as quality of college attended) and x_1 be unobserved (and represent a nonsensitive variable such as productivity). This structure implies that x_2 is a manipulable sensitive variable and we would not want to de-correlate x_3 from it since we don't believe there is a causal pathway. To get fair predictions, we would just not include x_2 in the model. But if we let x_2 affect x_3 as in right-hand side of Figure 1, then decoupling the two would result in a loss of information in x_3 about their common cause x_1. While we can still make progress in this situation since we have another measure of productivity via x_3, accuracy loss would be due to both discrimination-reduction *and* over-adjustment.

Finally, let x_2 be the sensitive variable and x_3 be unobserved.[3] Again, this implies that x_2 is manipulable. The appropriate approach here would not be to de-correlate x_1 from x_2 since there is predictive information in x_1 about C that is independent of x_2. To avoid over-adustment, x_2 should not be included as a predictor. It's worth noting that the sensitive variable, x_2, could still be predictable.

The end result of this thought experiment is that in a domain where the sensitive variable is a plausible root cause, then any accuracy loss in a prediction algorithm with inputs de-correlated from the sensitive variable(s) is a reflection of the underlying discrimination in the data. If a sensitive variable is not located at x_1, then a loss in accuracy additionally implies over-adjustment.

[3]We note one final scenario to consider which is when x_1 is both the sensitive variable and also unobserved. This is beyond the scope of this work.

4. Data and Experimental Setup

The National Longitudinal Survey of Youth 1997 (NLSY97) is a survey of men and women born between 1980 and 1984. The initial interviews occurred in 1997 with annual follow-ups. The survey consists of close to nine thousand individuals and contains detailed information on demographics, education, employment, household characteristics, sexual activity, health, and substance use. Importantly, NLSY also includes questions on arrest, incarceration, and criminal histories.[4]

We split the NLSY data into pre- and post-periods where in the pre-period – which includes all information collected through 2003 – we create predictors from the captured variables. Our outcome variable indicates whether a self-reported arrest was reported between 2004 to 2011. Our results below are insensitive to differing pre- and post-period. We treat ethnicity as the sensitive variable. We only keep ethnicities that are well represented in the data, i.e. contain more than 100 individuals, which leaves us with four ethnic groups: White ($N = 4,413$), Black ($N = 2,335$), Hispanic ($N = 946$), and White-Hispanic ($N = 879$). For simplicity, we concentrate on the difference between black and white results below. We convert all categorical variables into one-of-k encoding, resulting in 389 predictors.

The NLSY does not contain a representative sample of arrest rates. Data from the Bureau of Justice Statistics (BJS) indicates black adults were arrested 2.68 times more than white adults in 2003.[5] In the NLSY data, black respondents were 1.52 times more likely than white respondents to indicate an arrest in 2003. In results presented below we show how bias changes when we shift the arrest distribution in the NSLY to be comparable with BJS. We perform this shift by dropping 1,075 black survey respondents who did not report an arrest in the post-period. We also examine an extreme class imbalance by dropping 1,500 black survey respondents not reporting an arrest.

We run a ridge regression classifier with 5-fold cross-validation to predict future arrest. We refer to a model which uses the actual distribution of ethnicities in the data as **NLSY**, a model which alters the black arrest rate to mirror the BJS as **BJS**, and the extreme skew model as **SKEW**. Similarly, we refer to a model which removes the correlation between the sensitive variable and all other predictors as **uncorrelated** and **correlated** otherwise. We report results for six models: **NLSY-correlated**, **NLSY-uncorrelated**, **BJS-correlated**, **BJS-uncorrelated**, **SKEW-correlated**, **SKEW-uncorrelated**. We do not include the sensitive variable as a predictor in

[4]Data used in this work can be found here: https://github.com/zjelveh/fairness_data

[5]http://www.bjs.gov/

Table 1. The proportion of each ethnicity with at least one self-reported arrest in the post-period.

MODEL	BLACK	HISP.	WHITE	WH.-HISP.
NLSY*	0.1799	0.1501	0.1244	0.1354
BJS*	0.2927	0.1501	0.1244	0.1354
SKEW*	0.5030	0.1501	0.1244	0.1354

* $p < 0.01$

uncorrelated models. Table 1 shows the proportion of each ethnicity with at least one self-reported arrest in the post-period for the **NLSY** ($N = 8,573$), **BJS** ($N = 7,498$), and **SKEW** ($N = 7,073$) models. A proportion test rejects the hypothesis of equality across the four ethnicities.

5. Results

We assess the ability of our method to reduce bias and compare its accuracy to non-adjusted models. Given predicted scores, a threshold is chosen above which individuals are predicted risky. For simplicity, the threshold is set so the percentage of individuals predicted risky is equal to the percentage of individuals with self-reported arrests (14.3%).

In the context of predictive policing, a law enforcement agency will target those with highest predicted risk. Hence, it's important to ensure that criteria 1) and 2) from above are maintained. Table 2 shows the probability of being predicted risky for each ethnicity where it's evident that the uncorrelated models show greater balance across ethnicities than do the correlated models. Similarly, Table 3 shows the proportion of individuals of each ethnicity predicted to be risky. For comparison, we also include the overall sample distribution for ethnicities (suffix NAT). The uncorrelated models reduce the percentage of black ethnicity individuals predicted to be risky, particularly as class imbalance grows. It should also be noted as class skew increases, the uncorrelated models appear to retain a small portion of skewness in their predictions via fairness criteria 1) and 2). This may be a result of not explicitly targeting these criteria.

To assess whether the sensitive variable can be predicted by our altered variables, we run a ridge regression where the outcome variable is whether an individual is of black ethnicity. Not surprisingly, given how predictors were constructed, AUC is .514 indicating no ability to predict. Additionally, the correlation between the predicted scores from the uncorrelated models have very low correlation and mutual information with binary indicators of ethnicity. Finally, we examine how well the various models perform in out-of-sample prediction metrics via precision and AUC in Table 4. As we introduce class skew we notice that the correlated models improve in accuracy while the uncorrelated ones perform worse. This is to be expected since by introducing skew we are strengthening the relationship between the sensitive variable and the outcome.

Table 2. Probability of predicted risky given an ethnicity.

MODEL	BLACK	HISP.	WHITE	WH.-HISP.
NLSY-CORR.	0.173	0.142	0.129	0.138
NLSY-UNCORR.	0.147	0.128	0.144	0.148
BJS-CORR.	0.261	0.153	0.131	0.150
BJS-UNCORR.	0.178	0.149	0.159	0.152
SKEW-CORR	0.408	0.165	0.134	0.159
SKEW-UNCORR	0.217	0.156	0.172	0.163

Table 3. Probability of ethnicity given predicted risky.

MODEL	BLACK	HISP.	WHITE	WH.-HISP.
NLSY-CORR.	0.326	0.105	0.463	0.106
NLSY-UNCORR.	0.280	0.098	0.514	0.109
NLSY-NAT.	0.272	0.110	0.515	0.103
BJS-CORR.	0.307	0.131	0.456	0.107
BJS-UNCORR.	0.185	0.119	0.581	0.115
BJS-NAT.	0.168	0.126	0.589	0.117
SKEW-CORR	0.277	0.127	0.482	0.114
SKEW-UNCORR	0.147	0.120	0.616	0.116
SKEW-NAT	0.118	0.134	0.624	0.124

6. Conclusion

In this work, we highlight the unique use case of predictive models in the criminal justice setting. Second, we propose a flexible and straightforward method for altering input variables such that they are no longer correlated with sensitive variables. Third, we introduce a causal framework for reasoning about scenarios under which accuracy loss is due to over-adjustment. Fourth, we show our approach reduces discrimination using a dataset on individual criminal activity. In future work, we will compare our model to other approaches and other datasets. In particular, we will explore data containing manipulable *and* non-manipulable sensitive variables. Additionally, we will investigate why some of the imbalance remains in predictions from uncorrelated models. Finally, since correlation is a linear relationship, we intend to investigate the feasibility of using the altered predictors as inputs into nonlinear classifiers.

Table 4. Precision and AUC

DATA SET	PRECISION	AUC
NLSY-CORR.	0.4243	0.7958
NLSY-UNCORR.	0.4187	0.7910
BJS-CORR.	0.4634	0.8073
BJS-UNCORR.	0.4407	0.7842
SKEW-CORR.	0.4853	0.8213
SKEW-UNCORR.	0.4309	0.7568

References

Berk, Richard. The role of race in forecasts of violent crime. 1(4):231–242, 2009.

Calders, Toon and Verwer, Sicco. Three naive bayes approaches for discrimination-free classification. 21(2): 277–292, 2010.

Friedler, Sorelle, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Certifying and removing disparate impact. 2014.

Kamiran, Faisal and Calders, Toon. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pp. 1–6. IEEE, 2009.

Kamiran, Faisal, Karim, Asim, Verwer, Sicco, and Goudriaan, Heike. Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. pp. 370–377. IEEE, 2012.

Kamiran, Faisal, liobait, Indr, and Calders, Toon. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. 35(3):613–644, 2013.

Kamishima, Toshihiro, Akaho, Shotaro, and Sakuma, Jun. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 643–650. IEEE, 2011.

Pedreshi, Dino, Ruggieri, Salvatore, and Turini, Franco. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 560–568. ACM, 2008.

Perry, Walt L. *Predictive policing: the role of crime forecasting in law enforcement operations*. RAND, 2013.

Phelps, Edmund S. The statistical theory of racism and sexism. *The american economic review*, pp. 659–661, 1972.

Stroud, Matt. The minority report: Chicago's new police computer predicts crimes, but is it racist?, 2014.

Zemel, Rich, Wu, Yu, Swersky, Kevin, Pitassi, Toni, and Dwork, Cynthia. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 325–333, 2013.