IPP report for
Investigating Fair Classification

Adrien Morisot

Supervised by John Pate and Gianluca Corrado

April 2018

1 Introduction

Machine learning tools increasingly shape human decision-making, and as such deserve close ethical scrutiny. This scrutiny ought to be particularly intense when the decisions made by these machine learning tools profoundly affect the lives of other individuals, for example when they involve hiring someone, firing someone, giving emergency medical care to one person over another, etc. And yet, machine learning tools are not immune to prejudice with respect to age, race, sex, gender, nationality, or religion. This has been demonstrated multiple times [2, 1, 9, 4].
For example, Bolukbasi et al. [2] found that word embeddings trained on Google News articles were eerily sexist, with men being more closely associated with words such as "maestro", "skipper" and "protégé", while women were more closely associated with words such as "homemaker", "nurse" and "receptionist". In addition, analysis by ProPublica [1], a nonprofit newsroom, found that a tool used by US judges when trying to determine whether or not to release a defendant was racially biased in its recommendations. Indeed, ProPublica's analysis found that out of all defendants that were labelled high risk but did not re-offend, 23.5% of them were white, whereas 44.9% of them were African-American. Conversely, they found that out of all defendants labelled low risk but that did re-offend, 47.7% were white and 28% were African-American.

In addition, these algorithms are for the most part incapable of justifying themselves, or of explaining why they came to the conclusions they did. For example, neural networks and other non-linear models are composed merely of a series of innumerable weights and biases, whose meanings are entirely opaque to humans, and they output a single inscrutable number (e.g. a risk or recidivism score). Their inscrutability means that it's important to get the fairness element right from the get go, as human operators cannot detect the biased reasoning of the

algorithm without a lot of analysis.

Furthermore, contrary to classical hard-coded algorithms (that have well-defined commands written by humans), the fact that machine learning tools often function as black boxes means that prejudice can go undetected, and might even create self-reinforcing and self-fulfilling prophecies. For example, a model could extend the prison sentences of a group of inmates with characteristic X because it classifies them as likely to commit crimes. It could be that extending prison sentences leads to individuals committing more crimes (for example, a long prison sentence on a person's résumé might make them significantly less employable). This would reinforce the model's belief that inmates with characteristic X are more likely to commit crimes, creating a feedback loop where unfair decisions lead to unfair data which lead back again to unfair decisions.

However, despite all these issues, designing classifiers that are more fair than humans should be possible. Machine learning tools already outperform humans on a variety of tasks, from image recognition [17] to Go playing [12, 11], and they can conceivably be designed in such a way that they can be free of the biases and prejudices that subconsciously permeate most humans. Not to mention the fact that the human baseline for fairness is quite weak. For example, Goldin and Rouse [5] show that orchestras are significantly more likely to hire women if auditions are held such that the judges cannot distinguish the sex of the applicant.

For all these reasons, there is a need for fair classification systems capable of making decisions that are both more accurate than humans, and non discriminatory with respect to race, gender, age, etc.
The purpose of this MSc project is twofold. I will first evaluate and compare two different fair classification techniques, with respect to both fairness and accuracy, and then investigate more deeply the negative feedback loop issue described above.

2 Background and Purpose

There are currently, as far as I can tell, two broad axes in the literature surrounding fair classification: de-correlating data from sensitive attributes, and adding fairness as part of a classifier's loss function.

2.1 Preprocessing

The first involves preprocessing data in such a way that de-correlates it from the sensitive attributes. This usually involves finding the plane that is equidistant between the feature vectors representing two elements of the protected classes, and then projecting down all the other feature vectors onto that plane, which de-biases the feature vectors. For example, in the word embeddings paper [2], bias is removed by projecting all word-vectors onto the plane separating the vectors "he" and "she".
In another paper, Zemel et al. [16] define a mapping between original feature vectors and a lower dimensional subspace that minimises a loss function with three main parameters. The first is the final prediction accuracy achieved by the classifier they used, in order for the mapping to not strip important information from the feature vector. The second is the measure of fairness, specifically, whether or not the false-positive and false-negative rates are the same for the protected and unprotected classes (what we refer to as equalised odds in section 3). The third is the reconstruction accuracy, in order to minimise the amount of information lost by the mapping.
Calmon et al. [3] employ a similar strategy, using probability distributions instead of projecting feature vectors into lower dimensional subspaces. They use the sensitive features as part of the training process, in order to learn a transformation of the non-sensitive features and the outputs, such that when that transformation is applied to them, the output is independent of the sensitive features. In probabilistic terms, if X is the input data of the feature vectors, D the sensitive attributes, and Y the labels that we want to classify the feature vectors into, then Calmon et al.'s method turns X into $\tilde{X}$ and Y into $\tilde{Y}$, such that: $\mathbb{P}[\tilde{Y} \mid \tilde{X}, D] = \mathbb{P}[\tilde{Y} \mid \tilde{X}]$, i.e. $\tilde{Y} \perp D$. Calmon et al. impose similar constraints on their transformation as Zemel et al. [16] do on theirs: the accuracy of the prediction must be maximised, while the distortion of the original features needs to be minimised.

After this preprocessing of the vectors is done, a vanilla classifier can be used to classify each vector, and the resulting classification should be fair, since the feature vectors were de-correlated from the sensitive attributes.
Alternatively, one could skip the preprocessing step, and enforce fairness during the classification process.

2.2 Direct fair classification

A classifier, in short, attempts to draw boundaries in a high-dimensional space, in order to separate certain regions in that space from others. It then bins points in each delineated region of space into separate categories. The better it separates the space, the more points it will correctly bin, i.e. the more accurate

it will be. Classifiers are usually designed to delineate the space in order to maximise classification accuracy. However, fair classification will often involve taking into account a fairness penalty term that the classifier needs to minimise in order to obtain the best score. For example, Zafar et al. [13], during training, attempt to maximise accuracy while also minimising the covariance between the feature vector's sensitive features, and the distance between it and the decision boundary drawn by the classifier. In doing so, they are trying to make the decision boundary drawn by the classifier independent of sensitive features in the data.
Alternatively, Hardt et al. [6] propose first training a classifier without any fairness penalties, and then adjusting the decision boundaries of the classifier in order to satisfy a fairness metric (in this case, equalised odds: cf section 3). They do this by averaging their unfair classifier with a random classifier, in order to equalise the false-positive and false-negative of protected and non-protected classes, much like the strategy described in Pleiss et al. [10] in order to achieve equalised odds.

2.3 The limits of fair classification

Finally, it is worth pointing out that there are limits to the fair classification techniques described above. Classifiers are only as good as the data they are trained on. If the data is collected in a biased or unfair manner, then one should expect the classifiers to output unfair classifications. Thus, while building fair classifiers is an important component of fair classification, the data that is used to train these classifiers, and the way in which the data is collected, are also worth paying attention to. Lum et al. [9] illustrate this phenomenon through the case study of predictive policing methods employed by the Oakland, California police department. As illustrated in Figure 1 below, they found that although a large proportion of the population around the city of Oakland uses drugs, drug-related arrests were strongly concentrated in one specific part of Oakland. They speculate that this is due to the fact that that particular neighbourhood had been historically over-policed, and that past arrest data, when fed into the predictive policing algorithm, had made the algorithm predict a disproportionate amount of crime in that area. As a consequence, the police department sent more policemen to that location, and thus more crimes were reported, and subsequently fed into the predictive policing algorithm, thus reinforcing its belief that that region was crime-infested. Subsequent research by Ensign et al. [4] supports Lum et al.'s unfair feedback loop hypothesis.

2.4 Purpose

The purpose of this MSc project is twofold.

Figure 1: TOP: Police arrests due to drug-related crime around Oakland, California. BOTTOM: Estimated drug use around Oakland. Note that the police disproportionately target a specific area, which has historically been an over-policed minority district [9], despite the fact that actual drug use is spread out over a much larger area in the city. Figure obtained from Lum et al. [9].

First, I will evaluate and compare two different fair classification techniques on the same data sets (for example, that of ProPublica), by:
- Preprocessing feature vectors in the dataset by de-correlating features from sensitive attributes, and then running an ordinary classifier on that dataset, like [3, 16, 2].
- Enforcing fairness during the classification process, like [15, 14, 6].
To my knowledge, this is not a comparison that anyone has done yet.

Second, I will study the unfair feedback loop issue identified above, by simulating a population of agents whose behaviour is affected by the behaviour of a classifier. My approach will be similar to that of Ensign et al. [4].

3 Evaluation

This project involves juggling two principal metrics: fairness and performance. The latter is defined in most machine learning scenarios as the classifier's accuracy, i.e. its incorrectly-classified to correctly-classified ratio. Fairness, however, is a fuzzy ethical, philosophical and legal notion, which is hard to quantify and turn into numbers that can be easily manipulated by a computer. There is no clear winning standard for the notion. Instead, there are several different definitions, each with pros and cons. Two of these fairness definitions are:
Calibration
A classifier is calibrated when it assigns a probability p to some object belonging to some class and that is the true probability of that object belonging to that class. For example, if a machine learning system gives a set of 100 inmates a 60% chance of re-offending within a year of their release, and each of those inmates is released, and after a year we find that 60 of those inmates truly have re-offended (irrespective of their race), then that machine learning system is calibrated. The ProPublica classifier, for example, is calibrated: for inmates given a recidivism score of 6/10, 60% of white inmates and 61% of African-American inmates re-offended [8].
Equalised odds
A classifier respects equalised odds if no race is specifically disadvantaged by one type of error. More specifically, it respects equalised odds if different groups (defined with respect to the sensitive attributes) get false positive and false negative classifications at equal rates. For example, suppose that a system allowed the release of 100 inmates, 50 of them African-American and 50 of them white, and after a year none of them committed a crime. The system respected equalised odds if it assigned on average the same recidivism score to the white and African-American inmates. This is not the case for the ProPublica classifier, where in this situation, the score for African-American inmates was significantly higher than that of white inmates [1].
Throughout my project, I will study the fairness of different classification tech-

niques through the lens of these two notions, as they together embody important elements of fairness. There are other, more sophisticated definitions of fairness (e.g. Zafar et al.'s preferred impact and preferred parity [14]), but I focus on the former because they are both older and more established, and other papers [15, 7, 10] also use them as a proxy for fairness, making comparison easier.
It's important to note that as discussed by Pleiss et al. [10], these two notions of fairness are mutually incompatible if the base rates for the protected and unprotected classes are different.

4 Methods

Throughout my project, I intend on following good machine learning practice. Since I will be studying large data sets (e.g. the ProPublica data set [1], which has around 7,000 data points), I will split up each of them into a 70% training set, a 15% validation set and a 15% test set. The training set will be used to train models, and the validation set will be used to tune hyperparameters. The test set will be held out until the end, and will only be used to compare the most promising methods and models that I come up with.
In addition, I will make an effort to make my code agnostic to the type of data that is fed into it, in order to make my work more generalisable, and easily applicable to different data sets.
Finally, since the field of fairness in machine learning is quite new, it has not settled on a standard for notation. For example, there is no universally agreed upon symbol for feature vectors, or protected classes, or calibration. As such, in order to keep my report clear, I will include a table at the beginning of the dissertation summarising the notation that I will use throughout the work.

5 Outputs

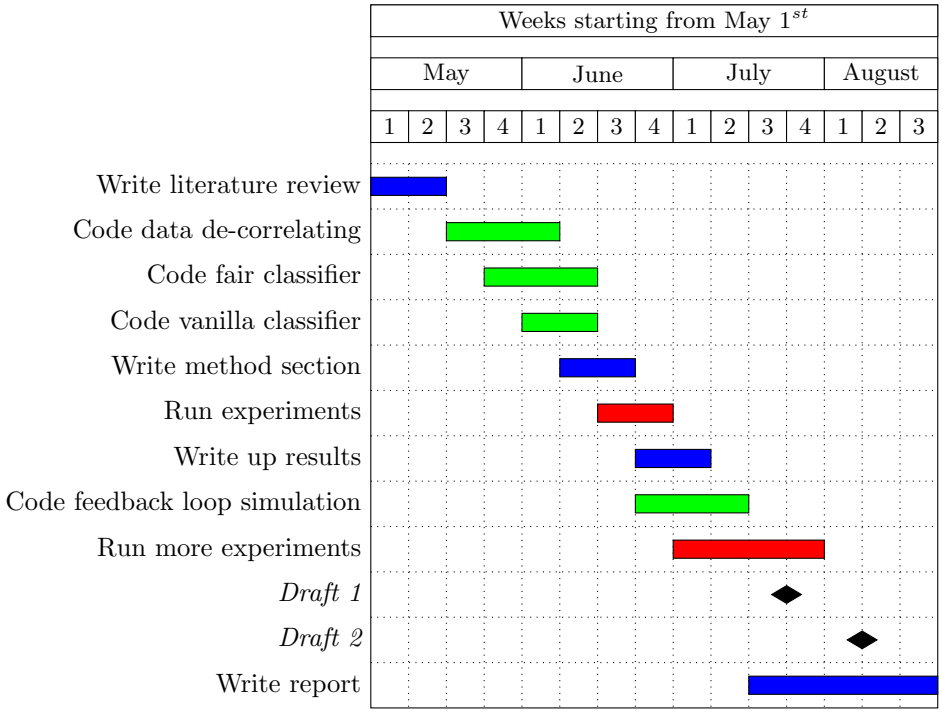
The aim of my project is twofold:
- First, I will aim to answer the question: does preprocessing/debiasing the data that is then fed into an ordinary classifier produce better results, both in terms of fairness and prediction accuracy, than using a classifier that is optimised for producing fair results (i.e. that tries to maximise calibration or equalised odds).
- Second, I will produce a simulation illustrating the effects of feedback loops in classification. This will be accompanied by a list of potential fixes, which will each be evaluated with respect to both accuracy and fairness.

As a stretch goal, by the end of my thesis I also hope to publish on the internet a step by step guide to making fair classifiers, in order to inform machine learning

practitioners of fairness best practices.

6 Work Plan

Below is a Gantt chart of how I intend on using my time. My last exam is on May 22^{nd}, so from now until then I will spend most (maybe 70%) of my time studying for that, and the rest of my time thinking about the MSc project. The chart is colour coded: blue represents periods of time where I will be mostly reading and writing, green represents periods of time where I will write code, and finally red represents periods of time where I will run experiments. This is not a definitive plan, and will probably change, as I am likely to encounter more difficulties with some parts of the project than others. However, I am committed to start the writing process early, and to have a complete draft of the thesis at least 2 weeks before the deadline.



References

[1] Julia Angwin et al. *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.* Ed. by ProPublica. [Online; posted 23-May-2016]. May 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[2] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4349–4357.

[3] Flavio Calmon et al. "Optimized Pre-Processing for Discrimination Prevention". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3995–4004.

[4] Danielle Ensign et al. "Runaway feedback loops in predictive policing". In: *arXiv preprint arXiv:1706.09847* (2017).

[5] Claudia Goldin and Cecilia Rouse. "Orchestrating impartiality: The impact of "blind" auditions on female musicians". In: *American Economic Review* 90.4 (2000), pp. 715–741.

[6] Moritz Hardt, Eric Price, Nati Srebro, et al. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems*. 2016, pp. 3315–3323.

[7] Niki Kilbertus et al. "Avoiding discrimination through causal reasoning". In: *Advances in Neural Information Processing Systems*. 2017, pp. 656–666.

[8] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". In: *arXiv preprint arXiv:1609.05807* (2016).

[9] Kristian Lum and William Isaac. "To predict and serve?" In: *Significance* 13.5 (2016), pp. 14–19.

[10] Geoff Pleiss et al. "On fairness and calibration". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5684–5693.

[11] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587 (2016), pp. 484–489.

[12] David Silver et al. "Mastering the game of Go without human knowledge". In: *Nature* 550.7676 (2017), p. 354.

[13] Muhammad Bilal Zafar et al. "Fairness constraints: Mechanisms for fair classification". In: *arXiv preprint arXiv:1507.05259* (2017).

[14] Muhammad Bilal Zafar et al. "From parity to preference-based notions of fairness in classification". In: *Advances in Neural Information Processing Systems*. 2017, pp. 228–238.

[15] Muhammad Bilal Zafar et al. “Learning fair classifiers”. In: *arXiv preprint arXiv:1507.05259* (2015).

[16] Rich Zemel et al. “Learning fair representations”. In: *International Conference on Machine Learning*. 2013, pp. 325–333.

[17] Barret Zoph et al. “Learning transferable architectures for scalable image recognition”. In: *arXiv preprint arXiv:1707.07012* (2017).