Fair Classification with Variational Fair Autoencoders

Author
Zuzanna Dubanowska
University of Copenhagen
vpz558@alumni.ku.dk

Abstract

In representation learning a system learns features reflecting variations in the data informative for machine learning tasks. In deep learning, neural networks create representations to optimise prediction performance. Following the remarkable successes of these systems, they are progressively being introduced to make decisions in high-risk areas such as loans or criminal intelligence. These algorithms are trained on historical data, which can contain artefacts of previous discrimination against certain subgroups, in which case they will propagate the bias in future decisions. This phenomenon affects increasingly many people. A possible venue to counteract it is by creating "fair" data representations. This paper is a reproduction and extension of the Variational Fair Autoencoder (VFAE) paper (Louizos et al., 2015). We implement the algorithm and test it on new datasets. Results show that the algorithm generalises well to new data, and successfully removes known sensitive data variations. The fair latent representations produced by the VFAE retain enough information to be used in prediction tasks and produce meaningful results.

1 Introduction

Representation learning is a set of machine learning techniques in which a system aims to discover a latent representation of data informative for feature detection or classification tasks. This allows the machine to learn the features and apply them to the given task, reducing the requirement for manual feature engineering. The objective of representation learning is to encapsulate different explanatory, otherwise "latent" or "hidden", factors of variation while removing uninformative factors of variation, otherwise "noise". How well the representations reflect variations in data is detrimental to the success of the modelling activity (Bengio et al., 2012). Many machine learning algorithms can be considered in these terms. Principal component analysis or manifold learning aim to extract informative factors affecting variation in data, often with the goal of visualising the data. On the other hand, algorithms such as deep neural networks produce latent representations useful for e.g. classification tasks.

Deep learning is a class of representation learning models, where neural networks are used to create representations of data meaningful to the model. Deep neural networks, since their advent, have been achieving remarkable successes both in academia and in the industry (Bengio et al., 2012), making them the preferred alternative to machine learning systems where feature engineering is used. Machine learning is increasingly present in all areas of life. Algorithms tailor advertising, make movie recommendations, are used to secure our phones. Algorithmic decision-making is also progressively introduced to high-risk areas such as loans (Mukerjee et al., 2002), hiring (Bogen and Rieke, 2018) or criminal intelligence (Perrot, 2017). The benefits of automated decision-making range from reducing human error, to increased effectiveness as machines do not become bored to increased magnitude of factors that can be taken into considerations compared to humans (Mehrabi et al., 2021).

One of the aspects which both people and algorithms are vulnerable to are biases. Humans can make decisions rendered unfair or discriminatory based on prejudices, while algorithms can propagate whatever bias is present in the data by creating "unfair" latent representations. The historical data used in modelling can contain artefacts of previous discrimination against certain groups, like race, ethnicity, gender, sexual orientation, or age. In this case the algorithm would create latent features that enunciate the bias and perpetuate discriminatory practices.

This unintended consequence of machine learning based automated-decision systems is already affecting and continues to affect progressively more people. In recent years, several such examples have

been observed, in areas from predictive policing (Lum and Isaac, 2016; Perrot, 2017) to face recognition (Phillips et al., 2011). As we cannot change the historical data, there is a clear need for machine learning models to account for the bias to avoid propagating it when assisting or executing decisions.

1.1 Fairness in machine learning systems

Machine learning is used to support or make decisions increasingly across different areas. Reported incidents of unfair or biased decisions made by these systems (Barocas and Selbst, 2016; O'neil, 2017; Romei and Ruggieri, 2014) raise a question of their legitimacy altogether (Barocas et al., 2019). This in turn led to interest in fair representation learning amongst researchers (Barocas et al., 2019).

In machine learning and deep learning, fairness is divided into different subdomains. Fair classification aims to make decisions (e.g. whether to grant a loan) based on input data (e.g. financial and demographic information), while at the same time avoiding discrimination based on one's group membership (e.g. age, race, gender) encoded in a sensitive variable (McNamara et al., 2019). On the other hand, learning fair representations involves encoding the input data to retain as much information as possible, while simultaneously concealing any information about group membership i.e. sensitive factors of variation (Zemel et al., 2013).

Early work in fairness classification research can be divided into two general approaches. One was to modify the labels, so that the proportion of positive labels (in binary classification) was equal for protected and unprotected groups. This approach assumed that that the "equal-opportunity" balance will be captured in the representations and propagate to the test set (Kamiran and Calders, 2009). The second one was to add a regularisation term to the classification training objective function quantifying the degree of bias. The algorithm's goal is then to minimise loss and degree of bias.

Zemel et al. were the first ones to propose an algorithm capable of learning fair representations. Authors formulated fairness as an optimisation task with two opposing goals: to create the most informative representation of the data while minimally informative about the (known) sensitive variable. The algorithm proposed, Learning Fair Representations (LFR), was a semi-supervised neural network clustering model. The idea is to cluster data points in a way that each cluster has equal proportions of members of each protected group in $\mathbf{s}$. The algorithm was successfully applied to a few datasets and outperformed other fairness algorithms available at the time. The formulation as a clustering algorithm, however, restricted the possibility to leverage the representational power of the model's distributed representations. Moreover, Louizos et al. argue that some information about the sensitive variable could persist in the latent representation leading to an information leak.

Louizos et al. built upon Zemel et al. and addressed both these issued in their proposed novel algorithm addressing learning invariant fair representations. Authors introduced a Variational Autoencoder (VAE) based semi-supervised model called the Variational Fair Autoencoder (VFAE) which is explicitly invariant to known sensitive variables in data. They introduced an additional Maximum Mean Discrepancy penalty to combat information leaks to the posterior. The model has been applied to three tasks; fair classification i.e. prediction accuracy of VFAE on $\mathbf{y}$, domain-adaptation i.e. quantifying how "invariant" to specific domain the latent representations are, and the general task of learning invariant representations i.e. whether or not all information about s was removed from the representation while retaining predictive power. The VFAE was tested against the LFR algorithm proposed by Zemel et al. and outperformed it, as well as performed competitively against recent (at the time of publishing) adversarial approaches.

1.2 Aim

This work is a reproduction and extension of the Variational Fair Autoencoder paper by Louizos et al.. We implement the VFAE algorithm proposed in the paper, capable of producing fair representations invariant to known sensitive variables. We aim for a minimal implementation which will be available publicly online to enable more to use it. We then extend the study by Louizos et al. by testing the algorithm on some data not considered previously. Specifically, we choose one dataset used in the study as a reference, but also explore other variables as sensitive, and a previously unseen large and sparse dataset.

The rest of this paper is organised as follows. In Section 2 we introduce the reader to the methodology, we briefly discuss the Variational Autoencoders and then the Variational Fair Autoencoder.

In Section 3 we present the experimental setting. In Section 4 we discuss the results, specifically; classification on $\mathbf{y}$, producing invariant representations and fair representations. We conclude our findings and suggest future directions for research in Section 5.

2 Methodology

2.1 Variational Autoencoder

Autoencoders (Rumelhart et al., 1985) are a family of models which learn efficient lower dimensional data manifolds in an unsupervised manner. The comprise of two networks: the encoder and the decoder. The encoder projects the data from a higher to a lower dimensional space (otherwise latent space) and the decoder reconstructs the original input from these latent representations. The aim is to find an optimal representation to minimise the reconstruction loss, i.e. to make the input and the reconstructed input as similar as possible.

Variational Autoencoders (VAE) (Kingma and Welling, 2013) follow the same encoder-decoder architecture, but instead of fixed latent representations, the encoder outputs parameters (mean and variation) of the latent distribution of the data. An example is sampled from the distribution and fed to the decoder, which aims to reconstruct it. Since the VAE operates on distributions, not fixed vectors in latent space, it can generate new examples by sampling from the learned distribution. The VAE learns by minimising the reconstruction loss and enforcing the latent distribution to be normal using Kullback-Leibler Divergence [1] as a regularisation term (also called similarity loss). The objective function of the VAE is defined as follows (Kingma and Welling, 2013):

$$Loss = \mathcal{L}(x, \hat{x}) + \mathcal{D}_{KL}(\mathcal{N}(\mu_x, \sigma_x) \| \mathcal{N}(0, \mathbf{I})) \tag{1}$$

2.2 Variational Fair Autoencoder

The Variational Fair Autoencoder (VFAE) (Louizos et al., 2015) is based on a VAE architecture, as VAE can naturally encourage separation between latent variables $\mathbf{z}$ and sensitive variables $\mathbf{s}$ by using factorised priors $p(\mathbf{s})p(\mathbf{z})$. In the base form the VAE produces latent representations of $\mathbf{x}$ of one data source $\mathbf{z}$. Here, we are dealing with two independent sources of data $\mathbf{s}$, which accounts for variations we want to remove and $\mathbf{z}$, a continuous latent variable encoding the informative variations we want to keep. This generative process can be formally defined as (Louizos et al., 2015):

$$\mathbf{z} \sim p(\mathbf{z}), \mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{s}) \tag{2}$$

This generative process, however, can result in removing informative factors of variation and produce degenerate representations if $\mathbf{s}$ is correlated to the label $\mathbf{y}$ i.e. if the data is biased. Therefore authors propose a process with two layers of latent variables, where information about the label is "injected" into the latent representation. Then there are two distinct sources for $\mathbf{z_1}$: the label $\mathbf{y}$ of the data point $\mathbf{x}$ and a continuous latent variable $\mathbf{z_2}$ encoding the variation on $\mathbf{z_1}$ not explained by the label $\mathbf{y}$:

$$\begin{aligned} \mathrm{y}, \mathrm{z}_2 &\sim Cat(\mathbf{y})p(\mathbf{z_2}) \\ \mathrm{z}_1 &\sim p_\theta(\mathbf{z_1}|\mathbf{z_2}, \mathbf{y}) \\ \mathrm{x} &\sim p_\theta(\mathbf{x}|\mathbf{z_1}, \mathbf{s}) \end{aligned}$$

Authors, following Kingma and Welling add an ability for the model to handle data even if the label is missing. Thus we arrive at the loss function of a 'stacked' semi-supervised VAE following Kingma and Welling, where $\mathcal{L}_s$ denotes the supervised loss and $\mathcal{L}_u$ denotes unsupervised loss i.e. when labels are missing from data. [2]

$$\begin{aligned} &\mathrm{F}_{\mathrm{VAE}}(\phi, \theta; \mathbf{x}_n, \mathbf{x}_m, \mathbf{s}_n, \mathbf{s}_m, \mathbf{y}_n) = \\ &\textstyle\sum_{n=1}^{N} \mathcal{L}_s(\phi, \theta; \mathbf{x}_n, \mathbf{s}_n, \mathbf{y}_n) + \\ &\textstyle\sum_{m=1}^{M} \mathcal{L}_u(\phi, \theta; \mathbf{x}_m, \mathbf{s}_m) + \\ &\alpha \textstyle\sum_{n=1}^{N} E_{q(\mathbf{z}_n|\mathbf{x}_n, \mathbf{s}_n)}[-\log q_\phi(\mathbf{y}_n \mid \mathbf{z}_{1n})] \end{aligned}$$

Where the last term is introduced to assure the predictive posterior learns both labeled and unlabeled data.

2.3 Maximum Mean Discrepancy

Despite Variational Autoencoders naturally encouraging statistical independence of $\mathbf{s}$ and $\mathbf{z_1}$ a priori, some information about the sensitive variable $\mathbf{s}$ might be retained in the marginal posterior $q_\theta(\mathbf{z_1}\|\mathbf{s})$. In particular, if the label $\mathbf{y}$ is correlated to the sensitive variable $\mathbf{s}$ there can be an information leakage about $\mathbf{s}$ to the posterior.

To counteract this issue, and assure further invariance of data to the sensitive variable, authors introduce an additional regularisation term Maximum Mean Discrepancy (Gretton et al., 2006). Let X $\sim P_0$ and X' $\sim P_1$, we want to determine

[1] Kullback-Leibler Divergence (KL divergence) is a measure of divergence between two distributions.

[2] See Louizos et al. for full mathematical derivation.

whether are drawn from the same distribution, i.e., P0 = P1. We can achieve that by measuring the distance between empirical statistics $\psi(.)$ of the two datasets X and X' (Gretton et al., 2006):

$$\ell_{\text{MMD}} = \| \frac{1}{N_0} \sum_{i=1}^{N_0} \psi(\mathbf{x_i}) - \frac{1}{N_1} \sum_{i=1}^{N_1} \psi(\mathbf{x'_i}) \|^2 \quad (3)$$

Minimizing this equation can be viewed as matching all of the moments of P0 and P1. Therefore, in our case, adding MMD to the loss function as an extra regularises will result in the model trying to match the moments between the marginal distrubutions $q(\mathbf{z_1} \| \mathbf{s} = 0)$ and $q(\mathbf{z_1} \| \mathbf{s} = 1)$(in the case of binary sensitive variable $\mathbf{s}$).

Computing MMD has a large space complexity (Zhao and Meng, 2015) and in order to reduce it, authors use random kitchen sinks algorithm (Rahimi and Recht, 2008) to approximate the full MMD. Let K be the dimension of $\mathbf{x}$, D number of random features. We draw a random K × D matrix $\mathbf{W}$, where each entry of W is drawn from a standard isotropic Gaussian distribution, and $\mathbf{b}$ is D-dimensional vector drawn from uniform random distribution in range $[0, 2\pi]$. The feature expansion is then given as:

$$\psi_{\mathbf{W}}(\mathbf{x}) = \sqrt{\frac{2}{D}} \cos\left(\sqrt{\frac{2}{\gamma}} \mathbf{x}\mathbf{W} + \mathbf{b} \right) \quad (4)$$

Lastly, we add the MMD regulariser to the loss function $\mathcal{L}_{VAE}$ as defined previously to obtain the objective for the Variational Fair Autoencoder (Louizos et al., 2015):

$$\begin{gathered} \mathrm{F}_{\text{VFAE}}\left(\phi, \theta; \mathbf{x}_n, \mathbf{x}_m, \mathbf{s}_n, \mathbf{s}_m, \mathbf{y}_n\right) = \\ \mathrm{F}_{\text{VAE}}\left(\phi, \theta; \mathbf{x}_n, \mathbf{x}_m, \mathbf{s}_n, \mathbf{s}_m, \mathbf{y}_n\right) - \\ \beta \ell_{\text{MMD}}\left(\mathbf{Z}_{1\mathbf{s}=0}, \mathbf{Z}_{1\mathbf{s}=1}\right) \end{gathered}$$

3 Experimental Setting

3.1 Datasets

We experimented with two datasets. UCI Adult Dataset which contains socio-economical data labeled with whether an individual makes more than [$]50K per year. We chose age (binarised) and sex as sensitive variables. Both are known to be correlated to the label. We also experiment with a subset of Stanford Open Policing Dataset (Pierson et al., 2020), collected in the city of San Francisco, CA, USA, which is a repository of traffic stops in roads. We chose this dataset as it is sparse and highly imbalanced - most stops are conducted rutinely and are not a result of a violation, and only approx. 1% of individuals get arrested overall. We chose arrest as the label and race (binarised i.e. we conduct multiple one-vs-others experiments) as the sensitive variable. Dataset statistics are summarised in Table 1. In Tables 2 and 3 we summarise statistics about the distributions of individuals from protected groups (e.g. males and females or race) in the datasets.

Dataset	Size	Sensitive variable	Characteristics
Stanford Open Policing Project (SF[1]) (Pierson et al., 2020)	31,778,515	Race (binarised)	sparse, highly imbalanced
UCI Adult Dataset a	45,222	Age, Sex (binary)	used in original study

Table 1: Description of datasets used for experimentation. [1] *SF stands for San Francisco. We chose the subset of data collected in San Francisco only.*

3.2 Experimental setup

For the Adult dataset we used the same setup as in (Louizos et al., 2015). Both encoders $\mathbf{z_1}$ and $\mathbf{z_2}$ and both decoders $\mathbf{z_1}$ and $\mathbf{x}$ had one hidden layer with 100 units, and there were 50 latent dimension for $\mathbf{z_1}$ and $\mathbf{z_2}$. We used α = 1, β = 0.1 and γ = 1 and 500 MMD dimensions as suggested by Louizos et al. whose MMD implementation we closely followed. For the Stanford Open Policing Dataset, both encoders $\mathbf{z_1}$ and $\mathbf{z_2}$ and both decoders $\mathbf{z_1}$ and $\mathbf{x}$ had one hidden layer with 150 units, and 50 latent dimensions for $\mathbf{z_1}$ and $\mathbf{z_2}$. We used We used α = 1, β = 0.1 and γ = 1 and 500 MMD dimensions. For the predictive posterior $q_\theta(\mathbf{z_1} \| \mathbf{s})$ we chose a Logistic regression classifier like Louizos et al. We chose Adam as the objective function with learning rate of 0.001.

Our evaluation had two goals in mind; (1) classification accuracy for $\mathbf{y}$, (2) removing the variance due to sensitive variable $\mathbf{s}$ from the data and producing fair representations, i.e. removing bias from data. We assess the Adult dataset in terms of accuracy, while the SF dataset in terms of F1 since it is highly imbalanced. We evaluate our classification results against a Random Forest (RF) and a Logistic Regression (LR) classifier. To measure the information about $\mathbf{s}$ in the new representation, we train a Logistic Regression classifier to predict $\mathbf{s}$ from $\mathbf{z_1}$.

Our implementation of the algorithm in Pytorch (Paszke et al., 2019) is publicly made available on GitHub.

% of individuals in protected groups earning less than USD 50k p.a	
Males vs. Females	
Males	69.4%
Females	89.0%
Age below vs. Age above mean	
Age below mean	85.5%
Age above mean	64.8%

Table 2: Distribution of individuals in protected groups earning less than USD 50k p.a. in Adult dataset before any data transformations.

% of individuals in protected group (race) arrested	
Asian / Pacific	0.77%
Black	2.32%
Hispanic	1.97%
White	1.01%

Table 3: Distribution of individuals in protected groups arrested (race) in the SF dataset. Order of racial origins alphabetical. 'Other' omitted as inconclusive.

4 Results

On the Adult dataset, a LR classifier trained on the VFAE produced representations obtains 75.9% accuracy with age being the sensitive variable. A LR trained on $\mathbf{x}$ (unchanged) obtains 85.2%, and an RF classifier trained on $\mathbf{x}$ obtains 85.0%. When taking sex as the sensitive variable an LR trained on VFAE latent representations obtains 75.7% accuracy, an LR trained on $\mathbf{x}$ obtains 85.1% and an RF trained on $\mathbf{x}$ obtains 84.9%.

To establish whether all sensitive information has been removed we trained an LR classifier to predict $\mathbf{s}$ from the latent representation $\mathbf{z_1}$. The accuracy was 59.6% for $\mathbf{s}$ = age, which was slightly above chance (53.7%). The prediction accuracy of correct sex from the latent representation $\mathbf{z_1}$ was 65.8% which is below chance (66.9%).

As Louizos et al. remark in the original paper, more important than accuracy on $\mathbf{y}$ is the trade-off between removing sensitive information and retaining predictive information. The given results showcase the model is indeed capable of performing classification while also (somewhat in the case of $\mathbf{s}$ = age) capable of removing s from latent representations.

On the Stanford Policing dataset, we observe similar behaviour to Louizos et al. on the "Health" dataset with similar sparseness used in their study. Namely, the model predicts the majority class (not arrested) as a label for every data point. The F1 score is around 99% for all sensitive variables for the majority class, and 0% for the minority class

	Logistic Regression	
Protected group	*F1 (not arrested / arrested)*	*% predicted arrested*
Asian / Pacific	99.2% / 55.1%	1.29%
Black	99.5% / 59.7%	1.28%
Hispanic	99.6% / 58.1%	1.27%
White	99.5% / 59.6%	1.33%

Table 4: Metrics of predictions performed by the Logistic Regression classifier on unchanged data.

	Random Forest	
Protected group	*F1 (not arrested / arrested)*	*% predicted arrested*
Asian / Pacific	99.5% / 63.8%	1.34%
Black	99.6% / 63.6%	1.33%
Hispanic	99.6% / 65.3%	1.27%
White	99.5% / 65.5%	1.35%

Table 5: Metrics of predictions performed by the Random Forest classifier on unchanged data.

(arrested). This still yields an approx. 98-99% F1 weighted score, but results in the label of interest disappearing. The results of classification on original data is summarised in Tables 4 and 5. Similarly to VFAE, the F1 score for 'not arrested' label is around 99% for both the LR and RF classifier. The classifiers show better recall of the arrested label than the VFAE, with scores above 55% for the LR and above 60% for the RF. The percentage of individuals who got assigned the arrested predicted label in each protected group appears to be balanced compared to the distribution of the data before. With both LR and RF classifiers predicting around 1.3% people arrested regardless their racial origin.

We also sought to establish whether all sensitive information was removed, the results are summarised in Table 6. For Black and Hispanic protected groups, the sensitive information was removed from the latent representations, as the prediction accuracy is below chance. For the Asian / Pacific group, the accuracy is equal to chance which denotes at least partially successful removal. The only group for which for information was not fully removed was White, where the accuracy is 2% above chance.

5 Conclusion

We implement the Variational Fair Autoencoder, originally introduced by (Louizos et al., 2015). We aim to provide a simple implementation of the algorithm, made available publicly. We extend the original study to new data aiming to test the algorithm's generalisation capabilities. We chose a dataset appearing in (Louizos et al., 2015), and test

Protected group (binarised)	Accuracy	Chance
Asian / Pacific	82.5%	82.5%
Black	82.6%	83.2%
Hispanic	86.0%	87.2%
White	60.8%	58.8%

Table 6: Prediction accuracy of the binarised sensitive variable from VFAE latent representations with chance levels for reference.

model's performance trying to remove a different sensitive variable than the authors used. We also choose a sparse dataset where the labels balance is 99% to 1%.

Our results offer compelling evidence that the Variational Fair Autoencoder generalises well to new datasets, and is capable of removing sensitive data variations with considerable success. Moreover, the latent representations the VFAE produces retain enough information to successfully perform classification tasks with acceptable accuracy.

While the Variational Fair Autoencoder does not outperform the Logistic Regression and Random Forest classifiers, as Louizos et al. point out in the original paper, removing the sensitive information and retaining predictive latent representations is more important than the accuracy on $\mathbf{y}$. This is especially true of the Adult dataset. On the Stanford Policing SF Dataset, we observe that conventional machine learning methods (Logistic Regression and Random Forest classifier in our case) have better recall of the arrested label. Moreover, they balance the arrests between protected groups, so that there's an even amount (%-wise) of members of each race arrested. Further investigations could be directed to examine whether these are in fact fair representations.

Future research could focus on improving upon the algorithm to achieve better accuracy; generating synthetic data using the VFAE and examining it's levels of bias and usability; and investigating whether the algorithm does not produce some undiscovered artefacts propagating bias.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California law review*, pages 671–732.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2012. Representation learning: A review and new perspectives.

Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Kristian Lum and William Isaac. 2016. To predict and serve? *Significance*, 13(5):14–19.

Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. 2019. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 263–270.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research*, 9(5):583–597.

Cathy O'neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Patrick Perrot. 2017. What about ai in criminal intelligence: From predictive policing to ai perspectives. *Eur. Police Sci. & Res. Bull.*, 16:65.

P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. 2011. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, 4(7):736–745.

Ali Rahimi and Benjamin Recht. 2008. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21.

Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.

Ji Zhao and Deyu Meng. 2015. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372.