Fairness-aware Naive Bayes Classifier for Data with Multiple Sensitive Features

Stelios Boulitsakis-Logothetis

University of Durham
Durham, United Kingdom
stelios.b.logothetis@gmail.com

Abstract

Fairness-aware machine learning seeks to maximise utility in generating predictions while avoiding unfair discrimination based on sensitive attributes such as race, sex, religion, etc. An important line of work in this field is enforcing fairness during the training step of a classifier. A simple yet effective binary classification algorithm that follows this strategy is two-naive-Bayes (2NB), which enforces statistical parity - requiring that the groups comprising the dataset receive positive labels with the same likelihood. In this paper, we generalise this algorithm into N-naive-Bayes (NNB) to eliminate the simplification of assuming only two sensitive groups in the data and instead apply it to an arbitrary number of groups.

We propose an extension of the original algorithm's statistical parity constraint and the post-processing routine that enforces statistical independence of the label and the single sensitive attribute. Then, we investigate its application on data with multiple sensitive features and propose a new constraint and post-processing routine to enforce *differential fairness*, an extension of established group-fairness constraints focused on intersectionalities. We empirically demonstrate the effectiveness of the NNB algorithm on US Census datasets and compare its accuracy and debiasing performance, as measured by disparate impact and DF-ϵ score, with similar group-fairness algorithms. Finally, we lay out important considerations users should be aware of before incorporating this algorithm into their application, and direct them to further reading on the pros, cons, and ethical implications of using statistical parity as a fairness criterion.

1 Introduction

Today, countless machine learning-based systems are in use that autonomously make decisions or aid human decision-makers in applications that significantly impact individuals' lives. This has made it vital to develop ways of ensuring these models are trustworthy, ethical, and fair. The field of fairness-aware machine learning is centered on enhancing the fairness, explainability, and auditability of ML models. A goal many research works in this field share is to maximise utility in generating predictions while avoiding discrimination against people based on specific sensitive attributes, such as race, sex, religion, nationality, etc.

Researchers have devised many formalisations to try and capture intuitive notions of fairness, each with different priorities and limitations. We summarise the ones we will mention here in Table 1. Traditionally, the proposed notions have been classified into two categories. The simplest and most well-studied, group fairness, is based on defining distinct protected groups in the given data. Then, for each of these groups, a user-selected statistical constraint must be satisfied. This has notable disadvantages: It requires groups to be treated fairly in aggregate, but this guarantee does not necessarily extend to individuals (Awasthi et al. 2020). Further, different statistical constraints prioritise different aspects of fairness. Many of them have also been shown to be incompatible with each other, making the choice even more difficult for users. Finally, the choice of the protected groups that should be considered is an open question (Blum et al. 2018; Kleinberg, Mullainathan, and Raghavan 2017).

An orthogonal notion to group fairness is individual fairness. Put simply, this notion requires that ”similar individuals be treated similarly” (Dwork et al. 2012). This approach addresses the previous lack of any individual-level guarantees. However, it requires strong functional assumptions and still requires the step of choosing an underlying metric over the dataset features (Awasthi et al. 2020).

Alternative models of fairness have been proposed to address the disadvantages of the two traditional definitions. One model is causal fairness, which examines the unfair causal effect the sensitive attribute value may have on the prediction made by an algorithm (Mhasawade and Chunara 2021). Another, which is explored in this paper, is differential fairness (DF). This is an extension of the established group fairness concepts that applies them to the case of intersectionalities, meaning groups that are defined by multiple overlapping sensitive attributes (Foulds et al. 2020; Morina et al. 2019).

A similar model is statistical parity subgroup fairness (SF), which focuses on mitigating intersectional bias by applying group fairness to the case of infinitely many, very small subgroups (Kearns et al. 2018). SF and DF are notable because they both enable a more nuanced understanding of unfairness than when a single sensitive attribute and broad, coarse groups are considered. A key difference between them, however, is DF's focus on minority groups. The SF measure of subgroup parity weighs larger groups more heavily than very small ones, while DF-parity considers all groups equally. This means DF can provide greater protection to very small minority groups since, in SF, their impact

---

In T. Kido, K. Takadama (Eds.), Proceedings of the AAAI 2022 Spring Symposium “How Fair is Fair? Achieving Wellbeing AI”, Stanford University, Palo Alto, California, USA, March 21–23, 2022.

on the overall score is reduced (Foulds et al. 2020).

Despite the lack of consensus on any universal notion of fairness, research has proceeded using the existing models. A major line of work in the development of fair learning algorithms is enforcing fairness during the training step of a classifier (Donini et al. 2018). A simple yet effective algorithm that follows this strategy is Calders and Verwer's two-naive-Bayes algorithm (Calders and Verwer 2010) (2NB). This algorithm was originally proposed as one of three ways of pursuing fairness in naive Bayes classification. It received further attention in the 2013 publication (Kamishima et al. 2013) which asserted its effectiveness in enforcing group fairness in binary classification and explored its underlying statistics. It works by training separate naive Bayes classifiers for each of the two (by assumption) groups comprise the dataset, the privileged and the non-privileged group. Then, the algorithm iteratively assesses the fairness of the combined model and makes small changes to the observed probabilities in the direction of making them more fair (Friedler et al. 2019).

A recent publication exploring the arguments for and against statistical parity (Räz 2021) has served as motivation to re-visit algorithms based around it. Statistical parity (also referred to as demographic parity or independence) is a group fairness notion which requires that the groups comprising the dataset receive positive labels with the same likelihood. An assumption that is at the core of 2NB and many other research works, however, is that of a single, binary sensitive feature (Oneto, Donini, and Pontil 2020). This assumption has been noted to rarely hold in the real world, and eliminating it is one of the essential goals of the previously introduced notions of differential fairness and subgroup parity fairness (Foulds et al. 2020; Kearns et al. 2018).

This opens the question of how 2NB can be applied to data with multiple, overlapping sensitive attributes while avoiding oversimplification. The 2NB algorithm is applicable to a wide range of tasks and its effectiveness, even in comparison to more complex algorithms, has been demonstrated (Kamishima et al. 2013; Friedler et al. 2019). At the same time, its' design is sufficiently elegant and intuitive to be approachable to practitioners across many disciplines - an important advantage. Thus, extending the algorithm to cover more use cases will be the focus of this work.

Contributions

This paper seeks to build upon Calders and Verwer's work by exploring the following:

- We adapt the original 2NB structure and balancing routine to support multiple, polyvalent (categorical) sensitive features.
- We use this new property of the algorithm to apply it to differential fairness.
- To support the above, we examine the extended algorithm's performance on real-world US Census data.
- Finally, we lay out important considerations users should be aware of before using this algorithm. We draw upon the literature to lay out the pros, cons, and ethical implications of using statistical parity as a fairness criterion.

Name	Definition
Statistical Parity	Likelihood of positive prediction given group membership should be equal for all groups.
Disparate Impact	Mean ratio of positive predictions for each pair of groups should be 1 or greater than $p\%$.
Subgroup Fairness	Group fairness applied to infinite number of very small groups.
Differential Fairness	Group fairness applied to groups defined by multiple overlapping sensitive attributes.
Individual Fairness	Distance between the likelihood of outcomes between any two individuals should be no greater than similarity distance between them.
Causal Fairness	Use of causal modelling to find effect of sensitive attributes on predictions.

Table 1: Some notable formalisations of fairness.

Related Work

Naive Bayes Naive Bayes is a probabilistic data mining and classification algorithm. In spite of its relative simplicity, it has been shown to be very competent in real-world applications that require classification or class probability estimation and ranking[1]. Various strategies have been explored for improving the algorithm's performance by weakening its conditional independence assumption. These include structure extension, attribute weighting, etc. These techniques focus on maximising accuracy or averaged conditional log likelihood (Jiang 2011). Calders and Verwer's proposal of composing multiple naive Bayes models instead aims to enforce independence of predictions with respect to a binary sensitive feature, thus satisfying the statistical parity constraint between the two groups (Calders and Verwer 2010).

Fair Classification There is a large body of research into designing learning methods that do not use sensitive information in discriminatory ways (Oneto, Donini, and Pontil 2020). As mentioned, various formalisations of fairness exist but the most well-studied one is group fairness (Blum et al. 2018). Many algorithms designed around this notion are introduced as part of the comparative experiment in Section 3.

A more recent proposal, differential fairness (DF), extends existing group fairness concepts to protect subgroups defined by intersections of and by individual sensitive attributes. The original papers by (Foulds et al. 2020) and (Morina et al. 2019) explore the context of intersectionality, and provide comparisons of DF with established concepts. The first paper asserts DF's distinction from subgroup parity and demonstrates its usefulness in protecting small minority groups. The latter paper gives methods to robustly estimate the DF metrics and proposes a post-processing technique to enforce DF on classifiers.

---

[1] Recent, novel applications include (Valdiviezo-Diaz et al. 2019; Feng et al. 2018; Niazi et al. 2019) among others.

Humanistic Analysis A line of work that is parallel to fair algorithm development focuses on analysing these proposals from an ethical, philosophical, and moral standpoint. A recent such publication, which examines statistical parity among other notions, and which motivated and influenced this paper, is by Hertweck, Heitz, and Loi (Hertweck, Heitz, and Loi 2021). They propose philosophically-grounded criteria for justifying the enforcement of independence/statistical parity in a given task. They include scenarios where enforcing statistical parity is ethical and justified, as well as counter-examples where the criteria are met but independence should not be enforced. As with many similar works, they conclude by directing the reader to strike a balance between fairness and utilitarian concerns (such as accuracy) in their task. (Heidari et al. 2019) do similar work, laying out the moral assumptions underlying several popular notions of fairness. In (Räz 2021), Räz critically examines the advantages and shortcomings of statistical parity as a fairness criterion and makes an overall positive case for it.

(Friedler, Scheidegger, and Venkatasubramanian 2016) introduce the concept of distinct *worldviews* which influence how we pursue fairness. One of them is that We're All Equal (WAE) i.e. there is no association between the *construct* (the latent feature that is truly relevant for the prediction) and the sensitive attribute. The orthogonal worldview is that *What You See Is What You Get*, wherein the observed labels are accurate reflections of the construct. In (Yeom and Tschantz 2021), Yeom and Tschantz give a measure of *disparity amplification* and dissect the popular group fairness models of statistical parity, equalised odds, calibration, and predictive parity through the lens of worldviews. They argue that under WAE, statistical parity is required to eliminate disparity amplification. However, deviating from this worldviews introduces inaccuracy when we enforce parity.

2 N-Naive-Bayes Algorithm

The proposed N-naive-Bayes algorithm is a supervised binary classifier that allows the enforcement of a statistical fairness constraint in its predictions. Given an (ideally large) training set of labelled instances, the algorithm partitions the data based on sensitive attribute value and trains a separate naive Bayes sub-estimator on each of the sub-sets. This is an extension of the original two-naive-Bayes structure, where exactly two sub-estimators are trained. The next step of the training stage is for the conditional probabilities $P(Y|S)$ to be empirically estimated from the training set. Where N_s is the number of instances that belong to group s, and $N_{y,s}$ the number of instances of that group that have label y, the empirical conditional probability[2] is given as:

$$P(y|s) = \frac{N_{y,s} + \alpha}{N_s + 2 * \alpha} \tag{1}$$

Finally, the algorithm modifies the joint distribution $P(Y, S)$ to enforce the given fairness constraint. Then, the final predicted class probabilities, for a sample xs (where x is the feature vector excluding the sensitive feature s), is:

$$\begin{aligned} P(y|xs) &= P(x|y) * P(s|y) * P(y) && (2) \\ &= C_s(x) * P(s|y) * P(y) && (3) \\ &= C_s(x) * P(s \cap y) && (4) \end{aligned}$$

Where C_s is the the sub-estimator for sensitive group $s \in S$.

Enforcing Statistical Parity

To satisfy the statistical parity constraint, the original 2NB algorithm runs a heuristic post-processing routine that iteratively adjusts the conditional probabilities $P(Y|S)$ of the groups in the direction of making them equal. During its execution, this probability-balancing routine alternates between reducing $N(Y = 1, S = 1)$ and increasing $N(Y = 1, S = 0)$ depending on the number of positive labels outputted by the model at each iteration. This is to try and keep the resultant marginal distribution of Y stable. Once balancing is complete, the value of $P(S|Y)$ can be induced from $N_{y,s}$ similar to (1). The first contribution of this paper is to extend this routine to suit the polyvalent definition of statistical parity we will use:

Definition 1. *Statistical (Conditional) Parity for Polyvalent S (Ritov, Sun, and Zhao 2017):*

For predicted binary labels $\hat{y}$ and polyvalent sensitive feature S, statistical (conditional) parity requires [3]*:*

$$P(\hat{y} = 1|s) = P(\hat{y} = 1|s') \; \forall \; s, s' \in S \tag{5}$$

We modify the probability-balancing routine to subtract and add probability to the group with the highest (max) and lowest (min) current $P(Y = 1|s)$ respectively. These probabilities are re-computed with each iteration, and the max and min groups re-selected. Further, we introduce the constraint that only groups designated by the user as privileged can receive a reduction in their likelihood of getting a positive label [4]. This is to avoid making any assumptions about which groups it would be appropriate to demote positive instances of. It allows the balancing routine to terminate immediately if it over-corrects, or if the data is such that $P(\hat{y} = 1|s_{np}) > P(\hat{y} = 1|s_p)$ to begin with, as is the case in the well-known UCI Adult dataset, for example. This gives us the final form of our statistical parity criterion:

Definition 2. *Statistical Parity Criterion for NNB:*

For predicted binary labels $\hat{y}$ and sensitive feature S:

$$P(\hat{y} = 1|s_p) = P(\hat{y} = 1|s_{np}) \; \forall \; (s_p, s_{np}) \in S_p \times S_{np} \tag{6}$$

Where S_p and S_{np} are the sub-sets of all privileged and non-privileged sub-groups of S respectively.

We adapt the above definition into a score that the algorithm can minimise:

$$disc = \max P(\hat{y} = 1|s_p) - \min P(\hat{y} = 1|s_{np}) \tag{7}$$

---

[2] Equation (1) gives a smoothed empirical probability, where the constant α is the parameter of a symmetric Dirichlet prior with concentration parameter $2 * \alpha$, since a binary label is assumed.

[3] The cited definition requires this to hold for all values of $\hat{y}$, however for a binary label it is sufficient to check $\hat{y} = 1$.

[4] A similar constraint is explored by (Zafar et al. 2017).

Algorithm 1: Pseudocode for a probability-balancing routine to enforce statistical parity

```
Calculate the parity score, $disc$, of the predicted classes by the current model and store $s_{max}$, $s_{min}$
while $disc > disc_0$ do
    Let $numpos$ be the number of positive samples by the current model
    if $numpos <$ the number of positive samples in the training set then
        $N(y=1, s_{min}) += \Delta * N(y=0, s_{min})$
        $N(y=0, s_{min}) -= \Delta * N(y=0, s_{min})$
    else
        $N(y=1, s_{max}) += \Delta * N(y=1, s_{max})$
        $N(y=0, s_{max}) -= \Delta * N(y=1, s_{max})$
    end if
    If any $N(y, s)$ is now negative, rollback the changes and terminate
    Recalculate $P(Y|S)$, $disc$, $s_{max}$, $s_{min}$
end while
```

Note that the above criterion can easily be relaxed to apply the four-fifths rule for removing disparate impact (or its more general form, the $p\%$ rule (Zafar et al. 2017)) instead of perfect statistical parity. For the purposes of this paper, however, we explore the effect of statistical parity in its base form.

We also note the definition of disparate impact we use in the evaluation stage:

Definition 3. *Disparate Impact (Mean) for Polyvalent S:*

$$\frac{1}{|S_p \times S_{np}|} \sum_{(s_p, s_{np})} \frac{P(\hat{y}=1|s_{np})}{P(\hat{y}=1|s_p)}$$

Algorithm 1 describes the extended probability balancing heuristic for enforcing parity. The values of s_p, s_{np} in the parity criterion (Equation 7) are referred to as s_{max} and s_{min} respectively. At each iteration, the routine determines these groups and adjusts their conditional probabilities. A further modification from the original is that the proportion by which the probabilities are adjusted with each iteration is now proportional to the size of the group itself, instead of the size of the opposite group. In experiments, this yields a great performance improvement, especially where the distribution of samples over S is very imbalanced.

Enforcing Differential Fairness

An alternative measure of fairness we explore is differential fairness, as given in (Foulds et al. 2020).

Definition 4. *A classifier is ϵ-differentially fair if:*

$$e^{-\epsilon} \leq \frac{P(\hat{y}|s)}{P(\hat{y}|s')} \leq e^{\epsilon} \; \forall \; s, s' \in S, \hat{y} \in Y \tag{8}$$

The (smoothed) empirical differential fairness score, from the empirical counts in the data, assuming a binary label, is:

$$e^{-\epsilon} \leq \frac{N(\hat{y}, s) + \alpha}{N(s) + \beta} \frac{N(s') + \beta}{N(\hat{y}, s') + \alpha} \leq e^{\epsilon} \; \forall \; s, s' \in S, \hat{y} \in Y \tag{9}$$

This is used in experiments to estimate the value of ϵ (the ϵ-score) from the predicted labels on the dataset[5]. In experiments we set $\beta = 2 * \alpha$ and substitute with the observed conditional probability estimates from the dataset. An additional measure given in (Foulds et al. 2020) to assess fairness from the standpoint of intersectionality is *differential fairness bias amplification*. This measure gives an indication of how much a black-box classifier increases the unfairness over the original data (Foulds et al. 2020; Zhao et al. 2017).

Definition 5. *Differential Fairness Bias Amplification*
A classifier C satisfies ($\epsilon_2 - \epsilon_1$)-DF bias amplification w.r.t. dataset D if C is ϵ_2-DF fair and D is ϵ_1-DF fair.

To adjust the joint distribution $P(Y, S)$ to minimise satisfy DF-fairness and minimise the ϵ-score, we propose a new heuristic probability-balancing routine and associated discrimination score. The distinction from the balancing routine given in Algorithm 1 is that this focuses on outputting a narrower range of probabilities, while still avoiding negatively impacting groups that are designated as non-privileged. To form the new discrimination score, we apply the principle of separating privileged and non-privileged sub-groups of S from the previous section to the ϵ-score definition:

$$e^{-\epsilon} \leq \frac{P(\hat{y}=1|s_{np})}{P(\hat{y}=1|s_p)} \leq e^{\epsilon} \; \forall \; (s_p, s_{np}) \in S_p \times S_{np} \tag{10}$$

We then express this restricted ϵ-score as the maximum of two ratios: $e^{\epsilon} = max(\rho_d, \rho_u)$, where for $(s_p, s_{np}) \in S_p \times S_{np}$:

$$\rho_d = \max \frac{P(\hat{y}=1|s_{np})}{P(\hat{y}=1|s_p)}, \; \rho_u = \max \frac{P(\hat{y}=1|s_p)}{P(\hat{y}=1|s_{np})} \tag{11}$$

The execution of the proposed balancing routine is determined by these ratios. If ρ_d is greater, then the non-privileged sub-group with smallest probability at that iteration receives an increase in probability. If ρ_u is greater, then the privileged group with highest probability receives a decrease in probability. These conditions can be expected to alternate as the conditional probabilities $P(Y|S)$ converge. Iteration continues until ρ_d is close to zero. The s_{max} and s_{min} groups are determined as in the previous section.

This routine disregards the number of positive labels the model produces, while Algorithm 1 attempts to keep that number close to the number of positive labels in the training data. This allows it to avoid situations where a single, non-privileged sub-group with small probability would require the probabilities of the privileged groups to be reduced significantly. In such cases, other non-privileged sub-groups might maintain much higher probabilities, therefore giving a poor ϵ-score. An further difference is the proportion by

---

[5] Note that this definition produces noisier estimates for sub-groups with fewer members. (Morina et al. 2019) shows that as the dataset grows, the given estimate converges to the true value, and that this happens regardless of the chosen smoothing parameters. However, for small or imbalanced datasets, more robust estimation methods should be used.

Algorithm 2: Pseudocode for a probability-balancing routine to enforce DF parity

```
Calculate the ratios $\rho_d, \rho_u$ empirically from the predicted classes by the current model, store $s_{max}$, $s_{min}$
while $\rho_d > disc_0$ do
    if $\rho_u \leq \rho_d$ then
        $N(y = 0, s_{min}) -= \Delta * N(y = 0, s_{min})$
        $N(y = 1, s_{min}) += \Delta * N(y = 1, s_{min})$
    else
        $N(y = 0, s_{max}) += \Delta * N(y = 0, s_{max})$
        $N(y = 1, s_{max}) -= \Delta * N(y = 1, s_{max})$
    end if
    Recalculate $P(Y|S)$, $\rho_d$, $\rho_u$, $s_{max}$, $s_{min}$
end while
```

which each $N_{y,s}$ is modified grows/decreases exponentially. In experiments, this allows the routine escape local minima that occur during the adjustment of $P(Y|S)$ and lead to inefficiency. This routine does, however, offer a theoretical accuracy trade-off compared to Algorithm 1, which we investigate in the following section.

Finally, note that all the above probability-balancing routines (including Calders and Verwer's original one) are based around the assumption that the distribution of labels over the sensitive feature(s) in the training set is reflective of the test setting. This assumption is not unique to this model (see (Agarwal et al. 2018; Hardt, Price, and Srebro 2016)), and under it, we can conclude that minimising the given fairness measure on the training set generalises to the test data (Singh et al. 2021).

3 Experimental Results

Setup

We implement NNB in Python within the scikit-Learn framework, using Gaussian naive Bayes as the sub-estimator. We then evaluate its performance in two experiments.

For both experiments, we use real-world data from the US Census Bureau[6]. (Ding et al. 2021) define several classification tasks on this data, each involving a sub-set of the total features available. We consider two:

- `Income`: Predict whether an individual's income is above \$50,000. The data for this problem is filtered so that it serves as a comparable replacement to the well-known UCI Adult dataset.
- `Employment`: Predict whether an individual is employed

The details of which features are included in each task and what filtering takes place can be found in the paper (Ding et al. 2021) and the associated page on GitHub[7]. To evaluate NNB we use data from the 2018 census in the state of California. The sensitive feature(s) used in each task are indicated after its name, e.g. `Income-Race-Sex` is the `Income` task using race and sex as the sensitive features. To best capture intersectional fairness when using multiple sensitive features, we follow the approach from (Foulds et al. 2020) and define each group s as a tuple of the sub-groups of each sensitive feature that each sample belongs to.

First Experiment This experiment compares NNB's performance with other algorithms. The comparison includes "vanilla" models as baselines for performance, and several group-fairness-aware algorithms that have a similar focus to NNB - ensuring non-discrimination across protected groups by optimising metrics such as statistical parity or disparate impact. Specifically, we consider the following:

- `GaussianNB`, `DecisionTree`, `LR`, `SVM`: scikit-Learn's Gaussian naive Bayes, Decision Trees, Logistic Regression, and SVM.
- `Feldman-DT`, `Feldman-NB`: A pre-processing algorithm that aims to remove disparate impact. It equalises the marginal distributions of the subsets of each attribute with each sensitive value (Feldman et al. 2015). The resulting "repaired" data is then used to train scikit-Learn classifiers - Decision Trees (`DT`) and Gaussian naive Bayes (`NB`).
- `Kamishima`: An in-processing method that introduces a regularisation term to logistic regression to enforce independence of labels from the sensitive feature (Kamishima et al. 2012).
- `ZafarAccuracy`, `ZafarFairness`: An in-processing algorithm that applies fairness constraints to convex margin-based classifiers (Zafar et al. 2017) . Specifically, we test two variations of a modified logistic regression classifier: The first maximises accuracy subject to fairness (disparate impact) constraints, while the latter prioritises removing disparate impact.
- `2NB`: Calders and Verwer's original algorithm, using the same GaussianNB sub-estimator as NNB.
- `NNB-Parity`, `NNB-DF`: N-naive-Bayes tuned to satisfy statistical parity using Algorithm 1, and DF-parity using Algorithm 2.

For the comparison we use the benchmark provided by (Friedler et al. 2019). The fairness-aware algorithms are tuned via grid-search to optimise accuracy. The performance of the algorithms is then measured over ten random train-test splits of the data.

Second Experiment This experiment demonstrates how NNB performs in finer detail. We consider `GaussianNB`, `NNB-Parity`, and `NNB-DF` as before, and we further include `2NB`, the original two-naive-Bayes algorithm implemented identically to NNB. Finally, we include `Perfect` as a secondary baseline, to illustrate the scores that would be achieved by a perfect classifier.

To evaluate the performance of the above algorithms, we note the mean and variance of the following measures over 10 random train-test splits: accuracy, AUC, disparate impact score (mean of the DI between all privileged and non-privileged groups), statistical parity score (as defined in 2),

---

[6] https://www.census.gov/programs-surveys/acs/microdata/documentation.html

[7] https://github.com/zykls/folktables

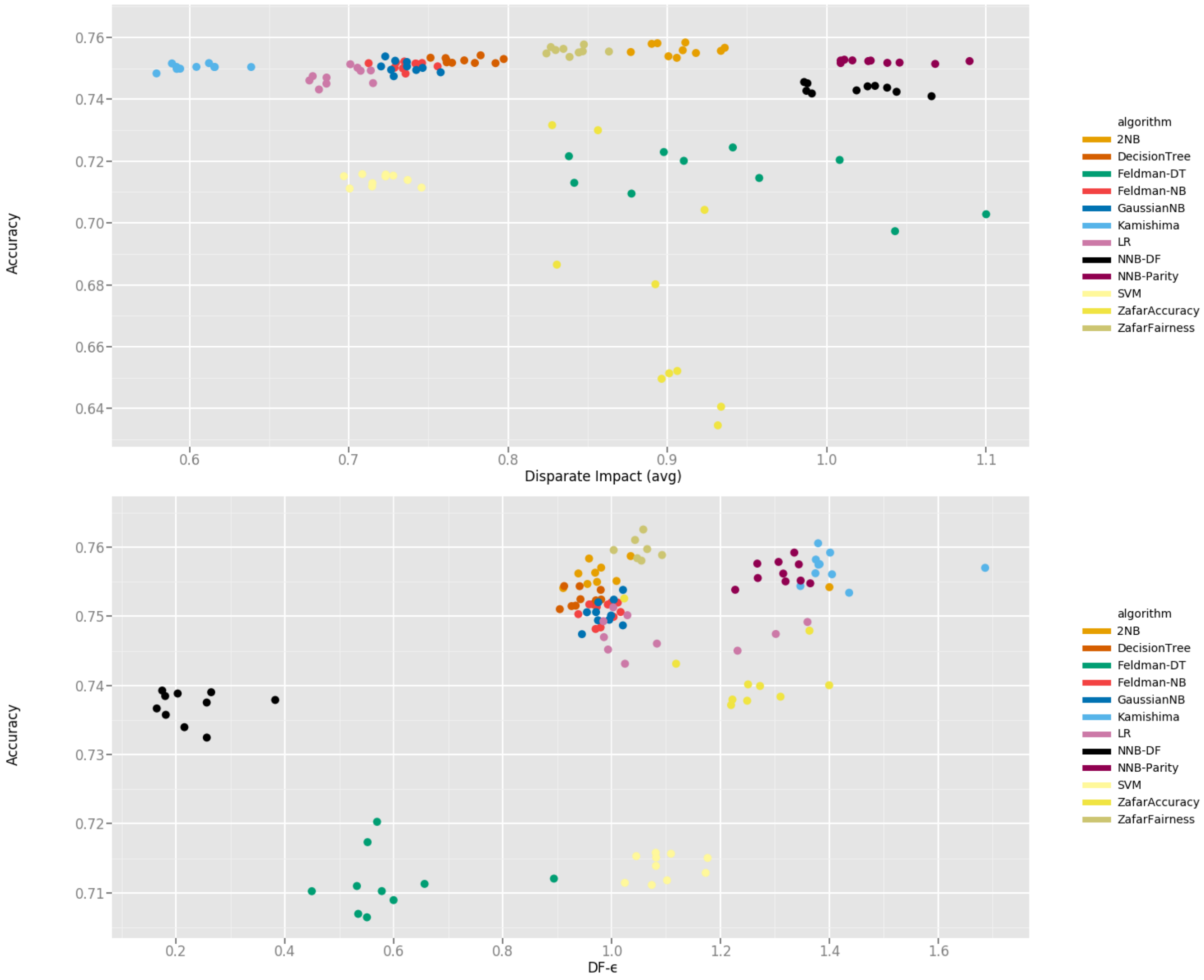

Figure 1: Scatter plots of accuracy vs. disparate impact for Income-Race and vs. ϵ-score for Income-Race-Sex

DF-ϵ (as defined in 4), DF-bias amplification score (as defined in 5). We also compare the resultant distribution of labels over groups of S on a single random train-test split.

Results

First Experiment Figure 1 gives the accuracy vs. the disparate impact and DF-ϵ scores on the Income-Race and Income-Race-Sex tasks. Figure 2 shows the same for Employment-Race and Employment-Race-Sex. It can be seen that on Income-Race, NNB results in a higher DI score than 2NB and has often over-favoured non-privileged groups causing a score > 1. Its accuracy is on-par with 2NB and the baseline naive Bayes, DT, and LR models. Feldman's algorithm with Decision Trees results similar disparate impact score in some splits, but lower accuracy. The same is true for the DF-ϵ score on this task. On Income-Race-Sex, `NNB-DF` beats out all other algorithms in achieving DI ~ 1, however `NNB-Parity` has higher accuracy than both `NNB-DF` and naive Bayes. `NNB-DF` is also the most successful at minimising the ϵ-score for this task, though again this comes at the cost of lower accuracy than the baseline model.

On Employment-Race all naive Bayes models achieve similar accuracy, while DT and LR-based models rank higher, and SVM the highest. The same can be observed for Employment-Race-Sex, and for both tasks `NNB-DF` again gives the ϵ-scores closest to zero.

Second Experiment Table 2 gives the scores achieved on the `Income-Race` task, and Table 3 gives the same `Employment-Race-Sex`. On `Income-Race`, both NNB models gave an improved parity score compared to the perfect classifier and `GaussianNB`. NNB and 2NB also gave improved disparate impact scores over the baseline models, but 2NB under-corrected while the NNB models gave a score > 1 indicating they favoured the non-privileged groups over the privileged group.

`NNB-Parity` and `NNB-DF` gave similar disparate impact scores, but the former gave higher accuracy while the latter produced a narrower range of positive label proportions, and thus better parity, ϵ, and DF-bias amplification scores. The evident accuracy trade-off is more pronounced in the latter task, with `NNB-Parity` achieving an accuracy of 0.7445 ± 0.00, and `NNB-DF` achieving 0.7199 ± 0.00.

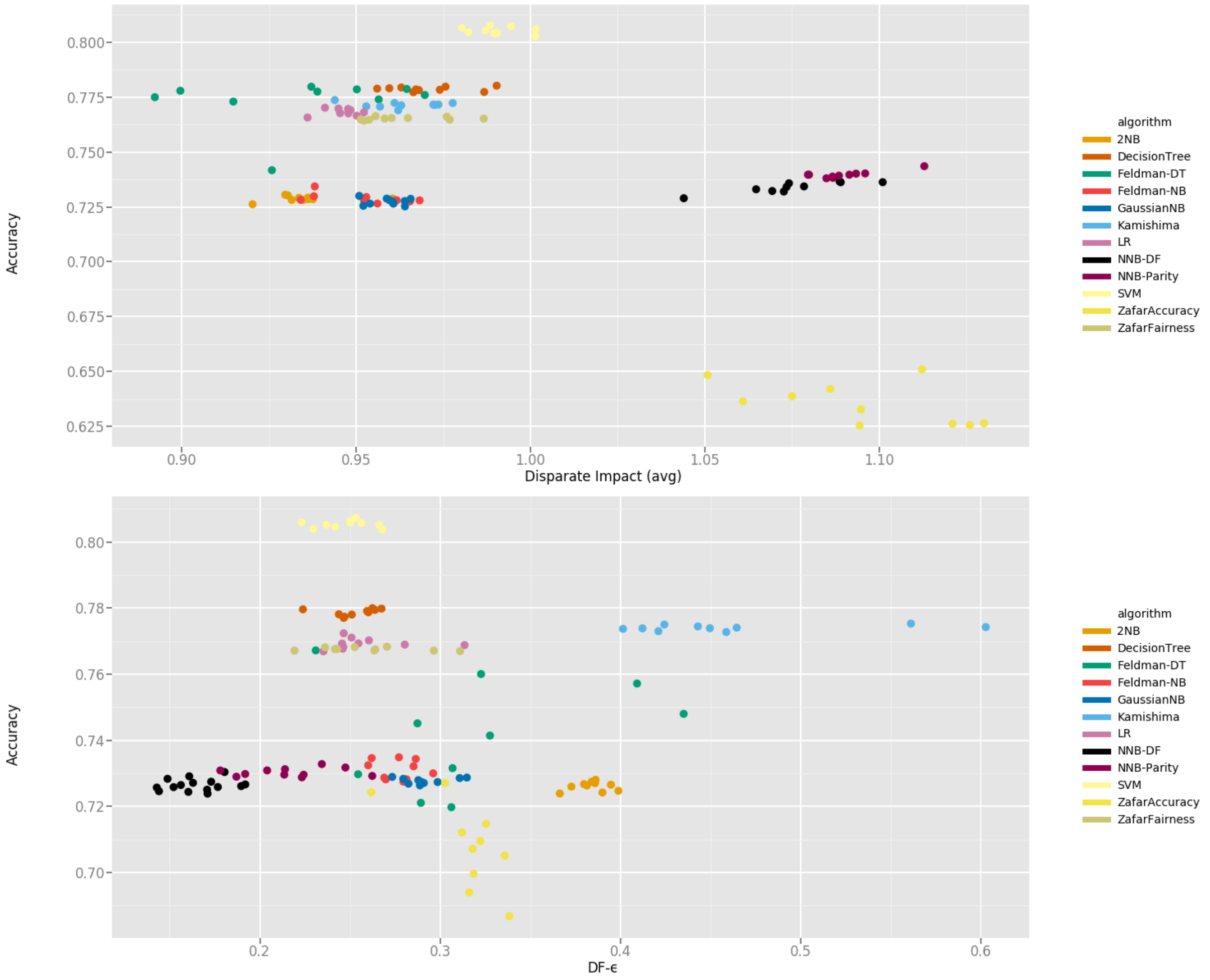

Figure 2: Scatter plots of accuracy vs. disparate impact for Employment-Race and vs. ϵ-score for Employment-Race-Sex

On `Employment-Race-Sex`, `NNB-DF` outperformed `NNB-Parity` on all scores. This was also the case for `Employment-Race`, where both models had similar accuracy but `NNB-DF` displayed less over-correction in its disparate impact score (1.0336 ± 0.0001 versus 1.2760 ± 0.0002), in addition to the expected improvement in ϵ-score (0.1068 ± 0.001 versus 0.3434 ± 0.0001). This suggests the DF balancing routine is better suited for the `Employment` task than the parity-based routine.

4 Discussion

In this work we presented an extension of the two-naive-Bayes algorithm, adapting it to suit datasets with multiple, polyvalent sensitive features. We applied the proposed N-naive-Bayes structure to intersectionality and differential fairness by giving an alternative probability-balancing routine. Our experiments on real-world datasets yielded favourable results and demonstrated the effectiveness and the differences between the parity and DF-based approaches.

We conclude by laying out key considerations users should take into account before using N-naive-Bayes:

Statistical Parity as a Fairness Criterion Statistical parity stands opposed to the (aggregate) accuracy of a classifier, except in degenerate cases where the data is already fair, so it is recommended that a balance between the two is pursued (Hertweck, Heitz, and Loi 2021). This also applies to the extended, but still parity-based, DF measure that was explored in Section 2. In their worldview-based analysis, Yeom and Tschantz caution us that even under WAE, blind enforcement of statistical parity can introduce new discrimination into the system (Yeom and Tschantz 2021). Thus, users must be aware of the ethical implications of using parity as a core fairness constraint, the possible impact it may have on individuals, and the moral objections these individuals may justifiably raise.

We recommend further reading on the advantages and disadvantages of group fairness in general (Räz 2021; Dwork et al. 2012; Heidari et al. 2019), as well as parity specifically (Hertweck, Heitz, and Loi 2021; Yeom and Tschantz 2021), so users can make informed decisions on how to apply statistical parity and N-naive-Bayes to their application.

	AUC	Accuracy	DI	Parity	DF-ϵ	DF-amp
`GaussianNB`	0.8270 ± 0.00	0.7503 ± 0.00	0.6304 ± 0.0001	0.4222 ± 0.0000	1.4100 ± 0.0012	0.4680 ± 0.0045
`2NB`	0.8223 ± 0.00	0.7577 ± 0.00	0.8930 ± 0.0013	0.3606 ± 0.0000	0.9774 ± 0.0016	0.0353 ± 0.0059
`NNB-Parity`	0.8114 ± 0.00	0.7480 ± 0.00	1.0810 ± 0.0006	0.1984 ± 0.0005	0.4580 ± 0.0045	-0.4840 ± 0.0041
`NNB-DF`	0.8138 ± 0.00	0.7380 ± 0.00	1.0636 ± 0.0007	0.1530 ± 0.0008	0.3112 ± 0.0048	-0.6308 ± 0.0035
`Perfect`	1.0000 ± 0.00	1.0000 ± 0.00	0.6975 ± 0.0005	0.2950 ± 0.0001	0.9420 ± 0.0048	0.0000 ± 0.0000

Table 2: Scores Achieved on Income with Race as the Sensitive Feature

	AUC	Accuracy	DI	Parity	DF-ϵ	DF-amp
`GaussianNB`	0.8159 ± 0.00	0.7273 ± 0.00	1.0228 ± 0.0001	0.3000 ± 0.0001	0.4994 ± 0.0003	0.0922 ± 0.0016
`2NB`	0.8112 ± 0.00	0.7202 ± 0.00	0.9352 ± 0.0001	0.2951 ± 0.0001	0.4818 ± 0.0002	0.0746 ± 0.0015
`NNB-Parity`	0.7820 ± 0.00	0.7241 ± 0.00	1.2990 ± 0.0004	0.2478 ± 0.0007	0.3971 ± 0.0013	-0.0101 ± 0.0005
`NNB-DF`	0.7909 ± 0.00	0.7251 ± 0.00	1.0601 ± 0.0002	0.1272 ± 0.0009	0.1840 ± 0.0018	-0.2232 ± 0.0011
`Perfect`	1.0000 ± 0.00	1.0000 ± 0.00	0.8643 ± 0.0001	0.1782 ± 0.0002	0.4072 ± 0.0014	0.0000 ± 0.0000

Table 3: Scores Achieved on Employment with Race and Sex as the Sensitive Features

Limitations of NNB N-naive-Bayes (as with two-naive-Bayes) has inherent limitations. The algorithm does not automatically make a classification task fair when it is applied. This is only considered to be possible by doing extensive domain-specific investigation (Hardt, Price, and Srebro 2016). Rather, the algorithm introduces a form of affirmative action to the task, increasing and decreasing the likelihood of different groups receiving a positive label in an attempt to satisfy the given parity constraint. This intentional manipulation of the original distribution over the data can be done to correct for structural biases in the data, for the purposes of compliance with regulations, or even as part of an effort to counteract historical inequalities.

Users should always consider the implications of estimating probability distributions for each group separately (as is done at the beginning of the training stage), as well as the mechanism behind any post-facto probability tuning they decide on. Further, users should understand the implications of affirmative action, its downstream effects, and ensure it is appropriate to their application. As a starting point for further reading, see (Dwork et al. 2012; Kannan, Roth, and Ziani 2019). Sociological and legal works such as (Kalev, Dobbin, and Kelly 2006; Anderson 2003) are also recommended.

Finally, the explicit choice of sensitive features to consider when enforcing statistical parity is a simplification of the real world and should be done carefully. One should consider the ontology behind observed values in the dataset: race, for example, has varying definitions, each of which comes with its own assumptions. Further, identifying groups in the data using a set of observable qualities, whatever those may be, also carries implicit assumptions about how all the factors involved interact with each other and the validity of decomposing them into discrete features (Barocas, Hardt, and Narayanan 2019, Ch. 5).

References

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 60–69. Stockholm, Sweden: PMLR.

Anderson, E. 2003. Integration, Affirmative Action, and Strict Scrutiny. *New York University Law Review*, 77(5): 1195–1271.

Awasthi, P.; Cortes, C.; Mansour, Y.; and Mohri, M. 2020. Beyond Individual and Group Fairness. *CoRR*, abs/2008.09490.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Blum, A.; Gunasekar, S.; Lykouris, T.; and Srebro, N. 2018. On Preserving Non-Discrimination When Combining Expert Advice. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 8386–8397. Red Hook, NY, USA: Curran Associates Inc.

Calders, T.; and Verwer, S. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2): 277–292.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *CoRR*, abs/2108.04884.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical Risk Minimization under Fairness Constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 2796–2806. Red Hook, NY, USA: Curran Associates Inc.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through Awareness. In *Proceedings of* *the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 214–226. New York, NY, USA: Association for Computing Machinery. ISBN 9781450311151.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 259–268. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336642.

Feng, X.; Li, S.; Yuan, C.; Zeng, P.; and Sun, Y. 2018. Prediction of Slope Stability using Naive Bayes Classifier. *KSCE Journal of Civil Engineering*, 22(3): 941–950.

Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020. An Intersectional Definition of Fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1918–1921. Dallas, Texas, USA: IEEE.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (im)possibility of fairness. *CoRR*, abs/1609.07236: 16.

Friedler, S. A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E. P.; and Roth, D. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 329–338. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3323–3331. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Heidari, H.; Loi, M.; Gummadi, K. P.; and Krause, A. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 181–190. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Hertweck, C.; Heitz, C.; and Loi, M. 2021. On the Moral Justification of Statistical Parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 747–757. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Jiang, L. 2011. Random one-dependence estimators. *Pattern Recognition Letters*, 32(3): 532–539.

Kalev, A.; Dobbin, F.; and Kelly, E. 2006. Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies. *American Sociological Review*, 71(4): 589–617.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Flach, P. A.; De Bie, T.; and Cristianini, N., eds., *Machine Learning and Knowledge Discovery in Databases*, 35–50. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-33486-3.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2013. The Independence of Fairness-Aware Classifiers. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops*, ICDMW '13, 849–858. USA: IEEE Computer Society. ISBN 9781479931422.

Kannan, S.; Roth, A.; and Ziani, J. 2019. Downstream Effects of Affirmative Action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 240–248. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2564–2572. Stockholmsmassan, Stockholm, Sweden: PMLR.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H., ed., *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 43:1–43:23. Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-029-3.

Mhasawade, V.; and Chunara, R. 2021. Causal Multi-Level Fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 784–794. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Morina, G.; Oliinyk, V.; Waton, J.; Marusic, I.; and Georgatzis, K. 2019. Auditing and Achieving Intersectional Fairness in Classification Problems. arXiv:1911.01468.

Niazi, K. A. K.; Akhtar, W.; Khan, H. A.; Yang, Y.; and Athar, S. 2019. Hotspot diagnosis for solar photovoltaic modules using a Naive Bayes classifier. *Solar Energy*, 190: 34–43.

Oneto, L.; Donini, M.; and Pontil, M. 2020. General Fair Empirical Risk Minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. Glasgow, United Kingdom: IEEE.

Räz, T. 2021. Group Fairness: Independence Revisited. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 129–137. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Ritov, Y.; Sun, Y.; and Zhao, R. 2017. On conditional parity as a notion of non-discrimination in machine learning. arXiv:1706.08519.

Singh, H.; Singh, R.; Mhasawade, V.; and Chunara, R. 2021. Fairness Violations and Mitigation under Covariate Shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 3–13. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Valdiviezo-Diaz, P.; Ortega, F.; Cobos, E.; and Lara-Cabrera, R. 2019. A Collaborative Filtering Approach Based on Naïve Bayes Classifier. *IEEE Access*, 7: 108581–108592.

Yeom, S.; and Tschantz, M. C. 2021. Avoiding Disparity Amplification under Different Worldviews. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 273–283. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness Constraints: Mechanisms for Fair Classification. In Singh, A.; and Zhu, J., eds., *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 962–970. Ft. Lauderdale, FL, USA: PMLR.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *CoRR*, abs/1707.09457.