Ensuring Fairness under Prior Probability Shifts

Arpita Biswas
Harvard University
Cambridge, Massachusetts, USA
arpitabiswas@seas.harvard.edu

Suvam Mukherjee
Microsoft Corporation
Redmond, Washington, USA
sumukherjee@microsoft.com

ABSTRACT

Prior probability shift is a phenomenon where the training and test datasets differ structurally within population subgroups. This phenomenon can be observed in the yearly records of several real-world datasets, for example, recidivism records and medical expenditure surveys. If unaccounted for, such shifts can cause the predictions of a classifier to become unfair towards specific population subgroups. While the fairness notion called *Proportional Equality* (PE) accounts for such shifts, a procedure to ensure PE-fairness was unknown. In this work, we design an algorithm, called CAPE, that ensures fair classification under such shifts. We introduce a metric, called prevalence difference (PD), which CAPE attempts to minimize in order to achieve fairness under prior probability shifts. We theoretically establish that this metric exhibits several properties that are desirable for a fair classifier. We evaluate the efficacy of CAPE via a thorough empirical evaluation on synthetic datasets. We also compare the performance of CAPE with several state-of-the-art fair classifiers on real-world datasets like COMPAS (criminal risk assessment) and MEPS (medical expenditure panel survey). The results indicate that CAPE ensures a high degree of PE-fairness in its predictions, while performing well on other important metrics.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Social and professional topics** → *Race and ethnicity*; *Gender*.

KEYWORDS

Classification, Discrimination, Distributional shifts, Algorithmic Fairness

ACM Reference Format:
Arpita Biswas and Suvam Mukherjee. 2021. Ensuring Fairness under Prior Probability Shifts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462596

1 INTRODUCTION

Machine learning techniques are being increasingly applied in making important societal decisions, such as criminal risk assessment, school admission, hiring, sanctioning of loans, etc. Given the impact and sensitivity of such predictions, there is warranted concern regarding implicit discriminatory traits exhibited by these techniques. Such discrimination may be detrimental for certain population subgroups with a specific race, gender, ethnicity, etc, and may even be illegal under certain circumstances [2]. These concerns have spurred vast research in the area of *fair classification* [5, 8, 12–14, 17, 22, 32, 41, 52]. Most of these papers aim to establish fairness notions for a group of individuals (differentiated by their race, gender, etc.), and are known as *group fairness* notions. The popular group-fairness notions are *Disparate Impact (DI)* [10, 19, 29, 48], *Statistical Parity (SP)* [15, 31, 50], *Equalized Odds (EO)* [26, 33, 46], and *Disparate Mistreatment* [11, 47]. These group fairness notions have been used extensively to *audit* black-box classifiers for discrimination [7, 43]. An inherent requirement of such audits is to have a test dataset of (statistically) significant size.

However, *ensuring* group fairness in classifiers presents several challenges. Fairness constraints such as DI and EO turn out to be non-convex, thereby making the optimization problem—maximizing accuracy subject to fairness constraints—difficult to solve efficiently. Several papers either focus on finding near-optimal near-feasible solutions [11, 16], or provide convex surrogates of the non-convex constraints [23, 48] in order to ensure fairness. Also, there exist heuristics to solve the problem which can be categorized into pre-processing, in-processing and post-processing techniques. Pre-processing [29] and post-processing [30, 39] techniques address fairness concerns in the input (training dataset) and output (predictions) of a classifier, respectively, while leaving the classifier unmodified. In-processing techniques [51], in contrast, address fairness concerns during model generation. These solutions assume that the training and test data are identically and independently drawn from some common population distribution, and thus suffer in the presence of distributional changes (we provide empirical evidence for this in Section 5).

Existing literature has addressed the problem of fair classification based on the fairness concepts of fair allocation. Balcan et al. [3] investigated a multi-class classification problem with heterogeneous preferences over the classes, and defined an envy-free classifier to be one where the prediction labels do not cause envy among the individuals in a given dataset. They further explored the generalizability of envy-free classification, that is, whether envy-freeness on a given dataset ensures almost envy-freeness with respect to the underlying distribution with high probability. Hossain et al. [27] defined relaxations of envy-freeness for ensuring fairness among different population subgroups. On the other hand, Zafar et al. [49] and Ustun et al. [44] applied the solutions concepts of fair allocation literature to define and explore *preferential guarantees*—each population subgroup (say, women or men) prefers the set of decisions they receive over the set of decisions they would have received

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462596

had they collectively belonged to the other group. These fairness definitions differ in the way the utility of the agents or a group of agents are assumed, detailed in Section 2.1. Along similar lines, we show that an appropriate utility function, together with the definition of proportional fair share (from allocation literature), can help capture the fairness of predictions.

A possible, and less studied, cause for unfairness in predictions involve distributional changes between the training and test datasets. A recent paper by Mandal et al. [36] proposes a fair classifier that is robust to weighted perturbations of the training samples. However, disparities can be introduced when the sub-populations evolve differently over time [4]. There are important real-world scenarios where *prior probability shift*, a type of distributional change, occurs. Informally, a prior probability shift occurs when the fraction of positively labeled instances differ between the training and the test datasets (see Section 2.2 for a formal definition). A concrete example is the COMPAS dataset [40]. COMPAS stands for "Correctional Offender Management Profiling for Alternative Sanctions". The dataset contains demographic information and criminal history of defendants, and records whether they recommitted a crime within a certain period of time. We observe that, among the valid records screened in the year 2013, the fraction of Caucasian and African-American re-offenders were 0.327 and 0.486, respectively. However, in 2014, these fractions were 0.636 and 0.706, respectively. This indicates that the extent to which the prior probability differs among Caucasian and African-American defendants, between the records of 2013 and 2014, is not the same.

If such distributional changes are unaccounted for, a classifier may end up being unfair towards the population subgroups which exhibit prior probability shifts; e.g., if the rate of recidivism among a specific sensitive group reduces drastically, then a classifier trained with a higher rate of recidivism can create extreme unfairness towards individuals of that sub-population. A fairness notion called *Proportional Equality* (PE) [9, 28] appears to be the most appropriate for addressing prior probability shifts among population subgroups (see Section 3 for definition). However, their results stop short of providing a procedure to ensure PE-fair predictions.

Our Contributions. To the best of our knowledge, this paper is the first to propose an end-to-end solution for ensuring fair predictions in the presence of prior probability shifts.

(1) We design a system, called CAPE (Combinatorial Algorithm for Proportional Equality), targeted towards making PE-fair predictions (Section 4). The system uses a novel combination of quantification techniques, sampling, and an ensemble of classifiers.
(2) We introduce a metric called *Prevalence Difference* (PD), which CAPE attempts to minimize in order to ensure fairness. We theoretically establish that the PD metric exhibits several desirable properties (Theorems 4.1, 4.2)—in particular, we show that maximizing the accuracy of any subgroup is not at odds with minimizing PD. This metric also provides insights into why the predictions of CAPE show a high degree of PE-fairness (Theorem 4.4). We discuss these in Section 4.1 and 4.2.
(3) We perform a thorough evaluation of CAPE on synthetic and real-world datasets, and compare with several state-of-the-art group-fair classifiers. In Section 5, we provide empirical evidence that CAPE provides highly PE-fair predictions, while performing well on other metrics.

2 PRELIMINARIES AND NOTATIONS

Let $\hat{h} : \mathcal{X} \mapsto \mathcal{Y}$ be a prediction function, defined in some hypothesis space $\mathcal{H}$, where $\mathcal{X} \subset \mathbb{R}^m$ is the m-dimensional feature space and $\mathcal{Y} = \{0, 1\}$ is the label space. We refer to instances with label 1 as being *positively* labeled (with the remaining instances being referred to as *negatively* labeled). The goal of a classification problem is to learn the function $\hat{h}$ which minimizes a target loss function, say, misclassification error $\mathcal{P}[\hat{h}(X) \neq Y]$ (variables X and Y denote feature vectors and labels, respectively). However, if these predictions $\hat{h}(\cdot)$ are used for societal decision making, it becomes crucial to ensure lower misclassification error not only on an average, but also within each group defined by their sensitive attribute values such as race, gender, ethnicity, etc. Dropping these sensitive attributes blindly from the dataset may not be enough to alleviate discrimination, since some non-sensitive features can be closely correlated to the sensitive attributes [14, 26, 53]. Hence, most existing solutions assume access to sensitive attributes. For simplicity, assume a single sensitive attribute that partitions the instance space into $|G|$ population subgroups. Now, the goal is to learn $\hat{h} : \mathcal{X} \times [G] \mapsto \mathcal{Y}$ satisfying certain group-fairness criteria, where we reuse the symbol $\mathcal{X}$ to represent the feature space comprising non-sensitive attributes only, and $[G]$ denotes the set $\{0, 1, \ldots, G-1\}$). We use the variable $Z \in [G]$ to denote group membership. Note that one can encode multiple sensitive attributes, such race and gender, together into the set $[G]$. For example, two sensitive attributes gender and race can be encoded as four subgroups: female African-American, female Caucasian, male African-American, male Caucasian.

We assume that *training set* $D = \{(x_i, z_i, y_i)_{i=1}^N\}$ is drawn from an unknown joint distribution $\mathcal{P}$ over $\mathcal{X} \times [G] \times \mathcal{Y}$. The performance of the classifier is measured using a new set of data, referred as *test dataset* $\mathbb{D} = \{(x_j, z_j, y_j)_{j=1}^n\}$, by observing how accurate and fair the $\hat{h}(x_j)$s are with respect to the true labels y_js.

Throughout, we use variables X, Y, Z to denote feature vectors, labels, and sensitive values, respectively. In the absence of fairness concerns, the sensitive feature Z can be thought of as a dimension in the feature space $\mathcal{X}$.

2.1 Revisiting Parity-based Notions via Envy-Freeness

In this section, we show that several existing parity-based notions of fairness can be formulated in terms *envy-freeness* using appropriate utility functions. Let $u_F^z(\hat{h}, \mathbb{D})$ be the real-valued utility of a group z for a fairness notion F, computed using the predictions of a classifier $\hat{h}$ on a dataset $\mathbb{D}$. A group z envies another group z', if the following holds:

$$u_F^z(\hat{h}, \mathbb{D}) < u_F^{z'}(\hat{h}, \mathbb{D}).$$

Thus, a set of predictions $\{\hat{h}(x)\}_{x\in\mathbb{D}}$ is said to be *envy-free* if and only if, for all groups $z \in G$, the following holds:

$$u_F^z(\hat{h}, \mathbb{D}) \geq u_F^{z'}(\hat{h}, \mathbb{D}) \qquad \text{for all groups } z' \in G.$$

We now list the utility functions that appropriately defines existing parity-based notions:

(1) *Disparate Impact Free* (DI) or *Statistical Parity* (SP): the fraction of positively predicted individuals are equal in all subgroups, i.e., for all $z, z' \in G$:

$$\mathcal{P}(\hat{h}(X) = 1 \mid Z = z) = \mathcal{P}(\hat{h}(X) = 1 \mid Z = z')$$

The corresponding utility function is

$$u_{SP}^z(\hat{h}, \mathbb{D}) = \frac{|\{i \in \mathbb{D} \ : \ \hat{h}(x_i) = 1,\ z_i = z\}|}{|\{i \in \mathbb{D} \ : \ z_i = z\}|}.$$

(2) *Equal False Positive Rates* (FPR): The false positive rates are equal for all subgroups. i.e., for all $z, z' \in G$:

$$\mathcal{P}(\hat{h}(X) = 1 \mid Y = 0, Z = z) = \mathcal{P}(\hat{h}(X) = 1 \mid Y = 0, Z = z').$$

The corresponding utility function is

$$u_{FPR}^z(\hat{h}, \mathbb{D}) = \frac{|\{i \in \mathbb{D} \ : \ \hat{h}(x_i) = 1,\ y_i = 0,\ z_i = z\}|}{|\{i \in \mathbb{D} \ : \ y_i = 0,\ z_i = z\}|}.$$

(3) *Equal False Discovery Rates* (FDR): The false discovery rates are equal in all subgroups. i.e., for all $z, z' \in G$:

$$\mathcal{P}(Y = 0 \mid \hat{h}(X) = 1, Z = z) = \mathcal{P}(Y = 0 \mid \hat{h}(X) = 1, Z = z')$$

The corresponding utility function is

$$u_{FDR}^z(\hat{h}, \mathbb{D}) = \frac{|\{i \in \mathbb{D} \ : \ \hat{h}(x_i) = 1,\ y_i = 0,\ z_i = z\}|}{|\{i \in \mathbb{D} \ : \ \hat{h}(x_i) = 1,\ z_i = z\}|}.$$

2.2 Prior Probability Shift

Prior probability shift [34, 37, 42] occurs when the prior class-probability $\mathcal{P}(Y)$ (also known as *prevalence*) changes between the training and test sets, but the class conditional probability $\mathcal{P}(X|Y)$ remains unaltered. Such changes, within a sub-population, occur in many real-world scenarios, i.e., $\mathcal{P}(X|Y{=}1, Z{=}z)$ remains constant but $\mathcal{P}(Y{=}1|Z{=}z)$ changes between training and test sets.

Next, we make a brief digression to explain quantification, a technique used for estimating prior probabilities in a test dataset in the presence of prior probability shifts. We show in a later section how we leverage quantification to address the problem of fair classification under prior probability shifts.

2.3 Quantification Problem

Quantification learning (or *prevalence estimation*) is a supervised learning problem, introduced by Forman [20]. It aims to predict an aggregated quantity for a set of instances. For example, a company may be interested in estimating the percentage of users who are likely to buy their product, using tweets received in the recent few weeks [24]. The goal is to learn a function, called *quantifier* $q : \mathcal{X}^{\mathbb{N}} \mapsto [0, 1]$, that outputs an estimate of the true prevalence of a finite, non-empty and unlabeled test set $\mathbb{D} \sim \mathcal{X}^{\mathbb{N}}$. As highlighted by Forman, quantification is not a by-product of classification [24]. In fact, unlike assumptions made in classification that the training and test dataset are drawn from the same distribution, quantification techniques account for changes in prior probabilities $\mathcal{P}(Y|Z)$ within subgroups, while assuming $\mathcal{P}(X|Y, Z)$ remain the same over the training and test datasets. This allows quantifiers to perform better than naïve classify and count techniques [21]—that train a classifier using the training dataset and count the number of predicted positive labels in the test dataset.

Some commonly used algorithms to construct quantifiers are *Adjusted Classify and Count* (ACC) [21], *Scaled Probability Average* (SPA) [6], and HDy [25]. These algorithms can be used to estimate the prevalence of a particular population subgroup in the test dataset. For ease of exposition, we describe a simple quantification technique, ACC. We define some terms to aid our discussion.

- True positives: $\mathsf{TP} := \left|\{i : y_i = 1 \ \&\ \hat{h}(x_i) = 1\}\right|$
- True negatives: $\mathsf{TN} := \left|\{i : y_i = 0 \ \&\ \hat{h}(x_i) = 0\}\right|$
- False positives: $\mathsf{FP} := \left|\{i : y_i = 0 \ \&\ \hat{h}(x_i) = 1\}\right|$
- False negatives: $\mathsf{FN} := \left|\{i : y_i = 1 \ \&\ \hat{h}(x_i) = 0\}\right|$
- True positive rate: $\mathsf{TPR} := \frac{\mathsf{TP}}{(\mathsf{TP}+\mathsf{FN})}$
- False positive rate $\mathsf{FN} := \frac{\mathsf{FP}}{(\mathsf{FP}+\mathsf{TN})}$

The ACC method learns a binary classifier from the training set and estimates its true positive rate (TPR) and false positive rate (FPR) via k-fold cross-validation. Using this trained model, the algorithm counts the number of cases on which the classifier outputs positive on the test dataset, which is subsequently adjusted to obtain an estimate of the true prevalence. Let p denote the true prevalence (i.e., fraction of true positives):

$$p := \frac{\#true\_positives}{\#test\_data\_points} = \frac{\mathsf{TP} + \mathsf{FN}}{\mathsf{TP} + \mathsf{FP} + \mathsf{TN} + \mathsf{FN}}. \tag{1}$$

Let p' denote the fraction of predicted positives:

$$p' := \frac{\#predicted\_positives}{\#test\_data\_points} = \frac{\mathsf{TP} + \mathsf{FP}}{\mathsf{TP} + \mathsf{FP} + \mathsf{TN} + \mathsf{FN}}. \tag{2}$$

Equations 1 and 2 can be used to relate the fraction of predicted positives p' and the true positives p, as:

$$p' = p \cdot \mathsf{TPR} + (1 - p) \cdot \mathsf{FPR}$$

Therefore, the true fraction of positives (true prevalence) can be estimated via the equation:

$$p = \frac{p' - \mathsf{FPR}}{\mathsf{TPR} - \mathsf{FPR}}.$$

The use of TPR and FPR from the training set can be justified by the assumption that $P(X|Y)$ remains same in the training and test datasets. This simple algorithm turns out to provide good estimates of prevalences under prior probability shifts.

However, for our experiments, we use SPA [6], an algorithm similar to ACC. Instead of a classifier, SPA uses a probability estimator, and the averages are computed over the predicted posterior probabilities rather than the predicted labels and thus known as *probabilistic classify and count* (PCC). This technique turns out to be more robust to variations in the prior probability estimates when the dataset contains only a few samples [6].

3 PROPORTIONAL EQUALITY

Consider binary classification as being the problem of allocating decisions or labels among a set of individuals (data points). Now, given a test dataset $\mathbb{D}$ and the predictions $\hat{h}(\cdot)$ on $\mathbb{D}$, let the utility of group z be defined using the metric *prediction prevalence* $\hat{\rho}^z_{\mathbb{D}}$, which is the fraction of population from the subgroup z who are assigned a positive label by the classifier. Formally,

$$u^z(\hat{h}, \mathbb{D}) = \hat{\rho}^z_{\mathbb{D}} := \frac{\left|\{(x_i, z_i, y_i) \in \mathbb{D} : \hat{h}(x_i) = 1, z_i = z\}\right|}{|\{(x_i, z_i, y_i) \in \mathbb{D} \ : \ z_i = z\}|} \tag{3}$$

Assume that the label 1 is treated as "favorable" by all individuals of all groups (e.g., in a loan approval scenario, the label 1 denotes individuals whose loans are sanctioned). Hence, all individuals prefers the label 1 over the label 0. Now, a group z *envies* another group z' if $\hat{\rho}^z_{\mathbb{D}} < \hat{\rho}^{z'}_{\mathbb{D}}$.

In order to ensure an envy-free allocation, the classifier's predictions need to satisfy $\hat{\rho}^z_{\mathbb{D}} = \hat{\rho}^{z'}_{\mathbb{D}}$ for every pair of groups z, z' (recall the definition of *disparate impact free*). Forcing such equality in predictions may not be appropriate when the *true prevalences* in the test dataset differ significantly between two subgroups. *True prevalence* $\rho^z_{\mathbb{D}}$ is defined as the fraction of population from subgroup z whose true labels in the dataset $\mathbb{D}$ are positive.

$$\rho^z_{\mathbb{D}} := \frac{|\{(x_i, z_i, y_i) \in \mathbb{D} \mid y_i = 1, z_i = z\}|}{|\{(x_i, z_i, y_i) \in \mathbb{D} \mid z_i = z\}|}. \tag{4}$$

If $\rho^z_{\mathbb{D}}$ is significantly greater than $\rho^{z'}_{\mathbb{D}}$, then a set of predictions that ensures $\hat{\rho}^z_{\mathbb{D}} = \hat{\rho}^{z'}_{\mathbb{D}}$ is unfair to group z. A property, which is arguably more fair, requires $\hat{\rho}^z_{\mathbb{D}}$ to be proportionally equal to (or greater than) the prediction prevalence of group z', with respect to the ratio of their true prevalences. This concern is formalized as a fairness notion called *proportional equality* (PE) [9]. A classifier is said to be PE-fair for a subgroup z, if the following holds for all other subgroups $z' \in [G]$: $\frac{\hat{\rho}^z_{\mathbb{D}}}{\hat{\rho}^{z'}_{\mathbb{D}}} \geq \frac{\rho^z_{\mathbb{D}}}{\rho^{z'}_{\mathbb{D}}}$. Instead of defining PE as a "binary" property—either a classifier is PE-fair or not—it has been extended to define the degree of discrimination against a group z, with respect to z', as: $\mathrm{PE}^{z,z'}_{\mathbb{D}} = \left|\frac{\rho^z_{\mathbb{D}}}{\rho^{z'}_{\mathbb{D}}} - \frac{\hat{\rho}^z_{\mathbb{D}}}{\hat{\rho}^{z'}_{\mathbb{D}}}\right|$. The lower the value of $\mathrm{PE}^{z,z'}_{\mathbb{D}}$, the fairer is the classifier. The fairness definition compares the predictions with the true labels of the *same* dataset. Thus, it accounts for the fact that the true prevalences in the test dataset $\mathbb{D}$ may have undergone prior probability shifts—PE fairness requires the ratio $\hat{\rho}^z_{\mathbb{D}}/\hat{\rho}^{z'}_{\mathbb{D}}$ to match with $\rho^z_{\mathbb{D}}/\rho^{z'}_{\mathbb{D}}$ rather than $\rho^z_D/\rho^{z'}_D$, where D is the training dataset.

While Biswas and Mukherjee [9] defined an appropriate fairness notion, the problem of ensuring fair predictions under prior probability shifts remained open. In the next section, we will define an algorithm that addresses this problem. Note that any such algorithm must deal with the following key challenges: (1) $\mathrm{PE}^{z,z'}_{\mathbb{D}} \leq \epsilon$ (for a small ϵ) is a non-convex constraint. Thus, it is hard to directly optimize for accuracy subject to this constraint for all $z, z' \in [G]$. (2) The definition of PE uses true prevalences of the test datasets $\rho^z_{\mathbb{D}}$, which are unavailable to the classifier during the prediction phase. Thus, an algorithm needs to *estimate* the unknown prevalence of the test dataset.

We now provide a comprehensive solution to the fairness problem via a novel combination of quantification techniques, sampling techniques, and an ensemble of classifiers.

4 CAPE

In this section, we introduce CAPE (Combinatorial Algorithm for Proportional Equality). The algorithm has two phases: *training* and *prediction*. Figures 1 and 2 show a high level overview of the workflow of CAPE during the two phases. CAPE takes as input a training dataset D and a vector $\Theta = (\theta_1, \ldots, \theta_k) \in [0, 1]^k$. CAPE trains an ensemble of classifiers, with the desired prediction prevalence of each classifier being one of the $\theta \in \Theta$ values. Moreover, CAPE is separately trained for each group $z \in [G]$, since we hypothesize that the relationship between the non-sensitive features X and the outcome variable Y may differ across groups. Thus, each group is best served by training classifiers on datasets obtained from the corresponding group[1]. Such decoupled classifiers are also considered by Dwork et al. [18], but they do not handle prior probability shifts.

At the end of the training phase, we obtain the following output for each subgroup z:

(1) a set of $|\Theta|$ classifiers, each trained using a sampling of the training dataset obtained by the module PP-SAMPLING, which takes as input a prevalence parameter $\theta \in \Theta$ and a training dataset with N_z data points. It randomly selects, with replacement, $\theta \times N_z$ instances with Y=1 and $(1 - \theta) \times N_z$ instances with Y=0. Thus, it outputs a sample of size N_z. Each classifier is thus specialized in providing accurate predictions on datasets with particular prevalences.
(2) a quantifier $\hat{q}^z(\cdot)$, generated by the Q-ALG module, which is subsequently used in the prediction phase of CAPE to estimate the true prevalence of the test dataset, $\rho^z_{\mathbb{D}}$. Separate quantifiers are created for each group since the extent of prior probability shifts may differ across groups.

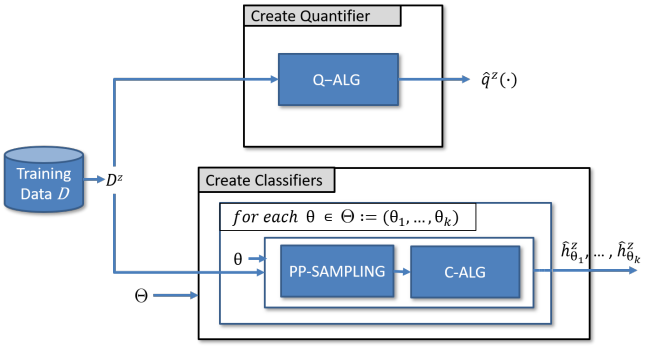


Figure 1: System diagram for the training phase of CAPE.

During the prediction phase, for each group z, an estimate of the prevalence of the test data $\mathbb{D}^z$ is obtained using $\hat{q}^z(\cdot)$. This estimate is then used to choose the classifier J_z (among the $|\Theta|$ classifiers $\{\hat{h}^z_{\theta_1}, \ldots, \hat{h}^z_{\theta_k}\}$ generated during the training phase) that minimizes

[1]Training a separate classifier for a small-sized group may be inappropriate. For the datasets we consider, this issue never arises.

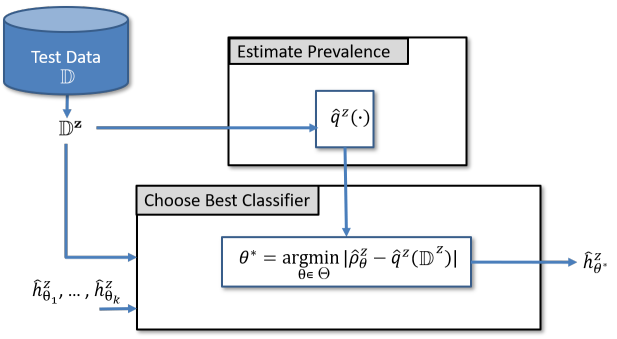


Figure 2: System diagram for the prediction phase of CAPE.

the *prevalence difference* metric (Section 4.1). Finally, CAPE outputs the predictions of the classifier J_z on the test set $\mathbb{D}^z$.

Algorithm 1 CAPE

Modules to be plugged in: C-ALG and Q-ALG.

Training Phase:

Input: A vector of prevalence parameters $\Theta := (\theta_1, \ldots, \theta_k\}$ and training dataset D.

Step 1: Partition $D = \{(x_i, z_i, y_i)_{i=1}^N\}$ based on z_i values.
$D^z \leftarrow \{(x_i, z_i, y_i) \in D \mid z_i = z\}$ for each group z.

Step 2: Create quantifiers, one for each z.
$\hat{q}^z(\cdot) \leftarrow$ Q-ALG(D^z).

Step 3: Create a set of k classifiers, for each z.
for all θ in $\{\theta_1, \ldots, \theta_k\}$ **do**
$T^z \leftarrow$ PP-SAMPLING (D^z, θ).
$\hat{h}^z_\theta(\cdot) \leftarrow$ C-ALG (T^z).
end for

Output: $\hat{q}^z$ and $(\hat{h}^z_{\theta_j})_{j=1}^k$.

Prediction Phase:

Input: Test dataset $\mathbb{D}$, and the quantifiers and classifiers obtained after the training phase.

Step 1: Partition $\mathbb{D} = \{(x_i, z_i, y_i)_{i=1}^n\}$ based on z_i values.
$\mathbb{D}^z \leftarrow \{(x_i, z_i, y_i) \in \mathbb{D} \mid z_i = z\}$ for each group z.

Step 2: Estimate prevalences $\hat{q}^z(\mathbb{D}^z)$ using the quantifiers built in the training phase.

Step 3: Compute the prediction prevalences on $\mathbb{D}^z$ obtained by each classifier $\{\hat{h}^z_{\theta_1}, \ldots, \hat{h}^z_{\theta_k}\}$.
for all θ in $\{\theta_1, \ldots, \theta_k\}$ **do**
$\hat{y}^i_\theta \leftarrow \hat{h}^z_\theta(x_i)$ for all $i \in \{1, \ldots, |\mathbb{D}^z|\}$.
$\hat{\rho}^z_\theta \leftarrow |\{i \in \mathbb{D}^z : \hat{y}^i_\theta == 1\}| / |\mathbb{D}^z|$.
end for

Step 4: Choose the best classifier in terms of estimated prevalence difference (Equation 5).
$J_z \leftarrow \underset{\theta \in \Theta}{\arg\min} |\hat{\rho}^z_\theta - \hat{q}^z(\mathbb{D}^z)|$

Output: The predictions $\hat{y}^z_{J_z}$ for group z.

Algorithm 1 provides details about CAPE. Note that CAPE provides the flexibility to plug in any classification and quantification algorithm into the modules C-ALG and Q-ALG. Key to CAPE is the *prevalence difference* (PD) metric, used in Step 3 of the prediction phase. We formalize the metric and discuss some of its properties in the next section.

4.1 Prevalence Difference

We define the *prevalence difference* (PD) metric as:

$$\Delta^z_{\mathbb{D}} := \left|\hat{\rho}^z_{\mathbb{D}} - \rho^z_{\mathbb{D}}\right| \quad \text{for each group } z. \tag{5}$$

where, $\hat{\rho}^z_{\mathbb{D}}$ and $\rho^z_{\mathbb{D}}$ denote the predicted and true prevalences of the dataset $\mathbb{D}$ (as defined in Equations 3 and 4, respectively). Hereafter, we drop the subscripts and superscripts on Δ, ρ and $\hat{\rho}$ whenever we refer to the population in aggregate.

Note that the true prevalence $\rho^z_{\mathbb{D}}$ of test set $\mathbb{D}$ cannot be used during the prediction phase. Thus, we use the value $\hat{q}^z(\mathbb{D}^z)$ to approximate $\rho^z_{\mathbb{D}}$. This allows us to use $\hat{q}^z(\mathbb{D}^z)$ in the definition of PD to pick the best classifier J^z for the group z. Also, unlike PD, other performance metrics like accuracy, FPR or FNR are not suitable for choosing the best classifier since these metrics require the *true* labels of the test datasets. We use the PD metric for: (1) choosing the best classifier in the prediction phase and (2) measuring the performance of the predictions, since a high value of Δ^z implies the inability to account for prior probability shift for the group z.

The PD metric is somewhat different from the fairness metrics aiming to capture parity between two sub-populations. Such fairness metrics may often require sacrificing the performance of a classifier towards one group to maintain parity with the other group. However PD, in itself, believes that the two groups should be treated differently since each group may have gone through a different change of prior probabilities. A high Δ^z indicates a high extent of harm caused by the predictions made towards the group z. Thus, to audit the impact of a classifier's predictions on a group z, it is important to evaluate for Δ^z, along with accuracy, FNR and FPR values within each group.

Next, we show that a *perfect classifier* (100% accurate) attains zero prevalence difference. Additionally, we show that a classifier with high accuracy on any subgroup also attains a very low Δ for that subgroup. Empirically, we observe that low Δ results in PE-fair predictions.

4.2 Theoretical Guarantees

We first show a simple result—a classifier whose predictions are exactly the ground truth also attains $\Delta = 0$, thereby satisfying our selection criterion for picking the best classifier. Note that a perfect classifier may *not* satisfy fairness notions such as disparate impact and statistical parity.

Theorem 4.1. *A perfect classifier always exhibits* $\Delta = 0$.

Proof. Let us consider a perfect classifier $\hat{h}(\cdot)$ whose predictions are equal to the ground truth i.e., $\hat{h}(x) = y(x)$ for all instances $x \in \mathcal{X}$, where $\hat{h}(x)$ is the label predicted by the classifier for the instance x. Thus, for each z, the true prevalence ρ^z is equal to the prediction prevalence $\hat{\rho}^z$, according to the definitions in Equations 3 and 4. This implies $\Delta^z = |\rho^z - \hat{\rho}^z| = 0$. □

Theorem 4.2. *If the overall accuracy of a classifier* $\hat{h}(\cdot)$ *is* $(1 - \delta)$, *where* $\delta \in (0, 1)$ *is a very small number, then the overall prevalence*

difference for the classifier $\hat{h}$ is $\Delta = \delta - 2\min\left\{\frac{\mathsf{FN}}{n}, \frac{\mathsf{FP}}{n}\right\}$, where FN *and* FP *denote number of false negatives and false positives respectively in the test dataset with n instances. This further implies that $\Delta \leq \delta$.*

Proof. Let $(\hat{h}(x_i))_{i=1}^n$ denote the predictions of a classifier $\hat{h}(x_i)$ on a test dataset $\{(x_i, y_i)_{i=1}^n\}$. Recall the following notations, which we use for the proof:

$\mathsf{TP} := \left|\{i : y_i = 1 \;\&\; \hat{h}(x_i) = 1\}\right|$ (# true positives).

$\mathsf{TN} := \left|\{i : (y_i = 0 \;\&\; \hat{h}(x_i) = 0\}\right|$ (# true negatives).

$\mathsf{FP} := \left|\{i : y_i = 0 \;\&\; \hat{h}(x_i) = 1\}\right|$ (# false positives).

$\mathsf{FN} := \left|\{i : y_i = 1 \;\&\; \hat{h}(x_i) = 0\}\right|$ (# false negatives).

Note that TP+TN+FP+FN = n. Let ρ and $\hat{\rho}$ be the true and prediction prevalences. Then, the *prevalence difference* can be written as:

$$\Delta = |\rho - \hat{\rho}| = \left|\frac{\mathsf{TP}+\mathsf{FN}}{n} - \frac{\mathsf{TP}+\mathsf{FP}}{n}\right| = \frac{|\mathsf{FN}-\mathsf{FP}|}{n} \tag{6}$$

Let the accuracy of a classifier on a test dataset be $(1-\delta)$ where $\delta \in (0,1)$. Then,

$$\frac{\mathsf{TP}+\mathsf{TN}}{n} = 1-\delta \quad \Rightarrow \frac{\mathsf{FN}+\mathsf{FP}}{n} = \delta \tag{7}$$

Without loss of generality, let us assume FN $\geq$ FP. Thus, Equation 7 can be written as:

$$\frac{\mathsf{FN}-\mathsf{FP}+2\mathsf{FP}}{n} = \delta \quad \Rightarrow \frac{\mathsf{FN}-\mathsf{FP}}{n} = \delta - \frac{2\mathsf{FP}}{n} \tag{8}$$

Similarly, assuming FP $\geq$ FN we obtain

$$\frac{\mathsf{FP}-\mathsf{FN}}{n} = \delta - \frac{2\mathsf{FN}}{n} \tag{9}$$

Combining Equation 6, 8 and 9, we get the following:

$$\Delta = \frac{|\mathsf{FP}-\mathsf{FN}|}{n} = \delta - 2\min\left\{\frac{\mathsf{FN}}{n}, \frac{\mathsf{FP}}{n}\right\} \leq \delta. \tag{10}$$

Thus, when accuracy is greater than $(1-\delta)$, the prevalence difference is at most δ. This completes the proof. □

Note that Theorem 4.2 can also be used to guarantee that highly accurate predictions for a group z implies a low value for Δ^z. This leads to Corollary 4.3.

Corollary 4.3. *If accuracy of a classifier for any sub-population z is greater than $1-\delta$, then $\Delta^z \leq \delta$.*

The next theorem provides insights into why CAPE works. Intuitively, Theorem 4.4 states that if the quantification and classification errors are bounded above by δ_1 and δ_2 respectively, then the true PD value for the predictions made by CAPE is bounded above by $\delta_1 + \delta_2 + \epsilon$. Here, ϵ is the maximum difference between two consecutive values in the vector Θ. This implies that improving Q-ALG's prevalence estimates and C-ALG's accuracies necessarily results in lower PD value for each subgroup in the predictions of CAPE. This, in turn, results in low value of $\mathsf{PE}^{z,z'}_{\mathbb{D}}$, except for a few corner cases. During our empirical evaluation on real-world datasets, such corner cases did not arise, and we observed the predictions of CAPE to exhibit low PE and PD values.

In the subsequent discussion, we drop the parameter $\mathbb{D}$ from the notations $\hat{q}$, ρ and $\hat{\rho}$ since we exclusively refer to these values in the context of the test dataset $\mathbb{D}$.

Theorem 4.4. *Let $\Theta = \{\frac{\epsilon}{2}, \frac{3\epsilon}{2}, \frac{5\epsilon}{2} \ldots, \left(k-\frac{1}{2}\right)\epsilon\}$ where $\epsilon \in (0,1)$ and $k = \left\lfloor \frac{1}{\epsilon} + \frac{1}{2} \right\rfloor$. For a group z, and test dataset $\mathbb{D}$, let the quantifier be such that $|\rho^z - \hat{q}^z| \leq \delta_1$, and the classifiers be such that $|\theta_j - \hat{\rho}^z_j| \leq \delta_2$ for all $j \in \{1,\ldots,k\}$, for small δ_1 and δ_2. Then, for the best classifier*

$$J := \underset{j\in\{1,\ldots,k\}}{\arg\min} |\hat{\rho}^z_j - \hat{q}^z|,$$

the following holds: $|\rho^z - \hat{\rho}^z_J| \leq \delta_1 + \delta_2 + \frac{\epsilon}{2}$.

Proof. For the best classifier J, the prevalence difference of a group z can be upper bounded using triangle inequality:

$$\begin{aligned} |\rho^z - \hat{\rho}^z_J| &\leq |\rho^z - \hat{q}^z| + |\hat{q}^z - \hat{\rho}^z_J| \\ &\leq \delta_1 + |\hat{q}^z - \hat{\rho}^z_J| \end{aligned} \tag{11}$$

Inequality (11) is implied by the assumption on the quantifier's performance, i.e., $|\rho^z - \hat{q}^z| \leq \delta_1$. To provide an upper bound for $|\hat{q}^z - \hat{\rho}^z_J|$, we pick J' such that

$$J' = \underset{j\in\{1,\ldots,k\}}{\arg\min} |\hat{q}^z - \theta_j|, \qquad \text{where } \theta_j = \left(j - \frac{1}{2}\right)\epsilon.$$

Since $\hat{q}^z \in [0,1]$, it is at most $\epsilon/2$ away from one of the fractional values in $\{\frac{\epsilon}{2}, \frac{3\epsilon}{2}, \frac{5\epsilon}{2} \ldots, \left(k-\frac{1}{2}\right)\epsilon\}$. Therefore,

$$|\hat{q}^z - \theta_{J'}| \leq \epsilon/2 \tag{12}$$

We use Inequality (12) to provide an upper bound to the expression $|\hat{q}^z - \hat{\rho}^z_J|$, using case-by-case analysis.

Case 1: Assume $\hat{q}^z < \hat{\rho}^z_J$. This leaves us with three possibilities for the value of $\theta_{J'}$:

(1) Assume $\theta_{J'} \geq \hat{\rho}^z_J$. Then,

$$\hat{\rho}^z_J - \hat{q}^z \leq \theta_{J'} - \hat{q}^z \leq \epsilon/2 \tag{13}$$

(2) Assume $\hat{q}^z \leq \theta_{J'} < \hat{\rho}^z_J$. Now, we bound the desired quantity using the value of $\hat{\rho}_{J'}$. Note that $|\hat{q}^z - \hat{\rho}_J| \leq |\hat{q}^z - \hat{\rho}_{J'}|$ since J is the best classifier. Thus, either $\hat{\rho}_{J'} \geq \hat{\rho}_J$ or $\hat{\rho}_{J'} \leq \hat{q}^z$.

(a) Assume $\hat{\rho}_{J'} \leq \hat{q}^z$. Then,

$$|\hat{q}^z - \hat{\rho}^z_J| \leq \hat{q}^z - \hat{\rho}^z_{J'} \leq \theta_{J'} - \hat{\rho}^z_{J'} \leq \delta_2 \tag{14}$$

(b) Assume $\hat{\rho}_{J'} \geq \hat{\rho}_J$. Then,

$$\begin{aligned} |\hat{q}^z - \hat{\rho}^z_J| &\leq (\hat{\rho}^z_{J'} - \hat{q}^z) \\ &= (\hat{\rho}^z_{J'} - \theta_{J'}) + (\theta_{J'} - \hat{q}^z) \\ &\leq \delta_2 + \epsilon/2 \end{aligned} \tag{15}$$

(3) Assume $\theta_{J'} < \hat{q}^z$. Now, we bound the desired quantity using the value of $\hat{\rho}_{J'}$, and there can be three cases.

(a) Assume $\hat{\rho}_{J'} \leq \theta_{J'}$. Then,

$$\begin{aligned} |\hat{q}^z - \hat{\rho}^z_J| &\leq \hat{q}^z - \hat{\rho}^z_{J'} \\ &\leq (\hat{q}^z - \theta_{J'}) + (\theta_{J'} - \hat{\rho}^z_{J'}) \\ &\leq \epsilon/2 + \delta_2 \end{aligned} \tag{16}$$

(b) Assume $\theta_{J'} < \hat{\rho}_{J'} \leq \hat{q}^z$. Then,

$$|\hat{q}^z - \hat{\rho}^z_J| \leq \hat{q}^z - \hat{\rho}^z_{J'} \quad \leq \quad \hat{q}^z - \theta_{J'} \leq \epsilon/2 \tag{17}$$

(c) Assume $\hat{\rho}_{J'} > \hat{\rho}_J$. Then,

$$|\hat{q}^z - \hat{\rho}^z_J| \leq \hat{\rho}^z_{J'} - \theta_{J'} \leq \delta_2 \tag{18}$$

Inequalities (13)-(18) establish the following upper bound when $\hat{q}^z < \hat{\rho}^z_J$,

$$|\hat{q}^z - \hat{\rho}^z_J| \leq \delta_2 + \epsilon/2. \tag{19}$$

Case 2: $\hat{q}^z \geq \hat{\rho}^z_J$. An analysis analogous to Case 1 gives the same inequality as (19). Combining Inequalities (11) and (19), we obtain the desired upper bound of $\delta_1+\delta_2+\epsilon/2$ on the quantity $|\rho^z - \hat{\rho}^z|$. □

5 EXPERIMENTAL EVALUATION

We first evaluate CAPE on synthetically generated datasets. We then compare it with state-of-the-art fair classifiers on the real-world COMPAS [40] and MEPS [1] datasets, where we observe possible prior-probability shifts. The performance of CAPE on a wide range of fairness-metrics, across all these datasets, enforces our proposal that CAPE should be used for predictions under prior-probability shifts.

5.1 Datasets

Synthetic: We assume a generative model with 3 features—a sensitive attribute $Z \in \{0, 1\}$, and two additional attributes U and V—along with the label $Y \in \{0, 1\}$. We assume that the overall population distribution is generated as $\mathcal{P}(U, V, Z, Y) = \mathcal{P}(U, V|Z, Y) \cdot \mathcal{P}(Z|Y) \cdot \mathcal{P}(Y)$. We further consider equal representation of the two population subgroups, $\mathcal{P}(Z = 1|Y) = \mathcal{P}(Z = 0|Y)$ for each $Y \in \{0, 1\}$. Also, U and V are conditionally independent:

$$\mathcal{P}(U, V|Z, Y) = \mathcal{P}(U, V|Y) = \mathcal{P}(U|Y) \cdot \mathcal{P}(V|Y).$$

The underlying distributions are considered to be Gaussian ($\mathcal{N}$) with the following mean and standard deviation:

- $\mathcal{P}(U|Y{=}1) \sim \mathcal{N}(15, 10)$
- $\mathcal{P}(U|Y{=}0) \sim \mathcal{N}(5, 5)$
- $\mathcal{P}(V|Y{=}1) \sim \mathcal{N}(20, 10)$
- $\mathcal{P}(V|Y{=}0) \sim \mathcal{N}(40, 10)$

We generated 50000 instances for the training dataset D with equal label distribution, i.e., $\rho^z_D = 0.5$. Also, we created multiple test datasets, each with a different value of true prevalence $\rho^z_{\mathbb{D}}$. We generated 81 different types of test datasets, each obtained by varying the prevalences for both subgroups $z \in \{0, 1\}$, such that $\rho^z_{\mathbb{D}} \in \{0.1, \ldots, 0.9\}$.

COMPAS dataset contains demographic information and criminal history for pre-trial defendants in Broward County, Florida. The dataset also contains a binary label `is_recid` that indicates whether a defendant re-offended within two years from the date of screening. The goal of learning is to predict whether an individual re-offends. We consider `is_recid` as Y labels (Y=1 denotes individuals who re-offended while Y=0 denotes individuals who did not re-offend) and `race` as the sensitive attribute (Z=1 denotes African-Americans, while Z=0 denotes Caucasians). We pre-processed the dataset to remove rows containing missing or invalid information. Our training dataset comprises 4278 records whose screening dates were in the year 2013 (of which 59.70% are African-Americans), while the test dataset comprises 1809 records screened in the year 2014 (of which 60.86% are African-Americans).

MEPS comprises medical expenditure surveys carried out on individuals, health care professions, and employers in the United States. A feature `UTILIZATION` measures the total number of trips involved in availing some sort of medical facility. The classification task involves predicting whether a person would have "high" utilization (defined as `UTILIZATION` ≥ 10, where 10 is roughly the average utilization among the respondents). Thus, we consider `UTILIZATION` as Y labels. The sensitive attribute, `RACE` is constructed as follows: 'Whites' (Z=0) is defined by the features `RACEV2X` = 1 (White) and `HISPANX` = 2 (non Hispanic); everyone else is tagged 'Non-Whites' (Z=1). The surveys for the year 2015 is our training set (with 33400 data points, of which 62.86% are 'Non-Whites'), and the surveys for 2016 is our test set (with 32006 data points, of which 61.72% are 'Non-Whites').

5.2 Other Algorithms for Comparison

We compare CAPE against an accuracy-maximizing classifier, **`Max_Acc`**, which is the same algorithm employed in the C-ALG module of CAPE. On the real-world datasets, we additionally compare CAPE with the following fair algorithms, implemented in the IBM AI Fairness 360 [7] toolkit: —Reweighing (**`Reweigh`**) [29], variants of **`Meta_fair`** [11], Adversarial Debiasing (**`AD`**) [51], Calibrated Equalized Odds Postprocessing (**`CEOP`**) [39], Reject Option Classification (**`ROC`**) [30]. None of these algorithms are designed to handle PE-fairness. We evaluate the extent to which these algorithms achieve PE-fairness and compare how they perform on a set of other metrics (such as FPR-difference, FNR-difference, Accuracy-difference, and PD). While CAPE can handle multiple sensitive attributes, we choose one sensitive attribute for all the datasets to stay consistent with the implementation in the IBM AIF360 toolkit. We ran experiments on a machine with 32 GB RAM and a quad-core Intel Core i7 processor.

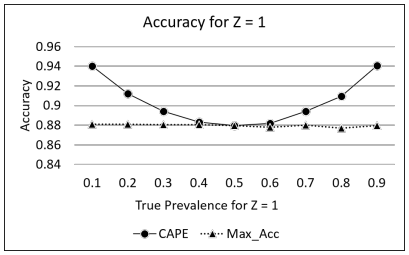
5.3 Parameters and Modules used for CAPE

- Prevalences: We set $\Theta = \{0.05, 0.15, \ldots, 0.95\}$.
- PP-SAMPLING: As described in Section 4, this module takes as input a prevalence parameter $\theta \in \Theta$ and a training dataset with N_z data points. It randomly selects, with replacement, $\theta \times N_z$ instances with Y=1 and $(1 - \theta) \times N_z$ instances with Y=0. Thus, it outputs a sample of size N_z.
- Q-ALG: *Scaled Probability Average* [6].
- C-ALG: As the synthetically generated datasets are created using simple generative models, we use generalized logistic regression (`glm`) with regularization. For COMPAS and MEPS, we use gradient boosted algorithm (`gbm`) and 10-fold cross-validation for hyper-parameter tuning.

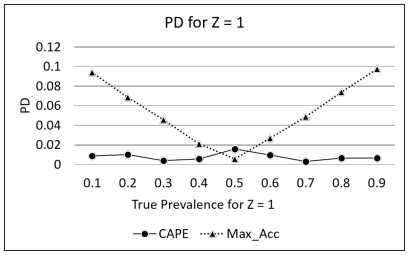
5.4 Results

Synthetic dataset: We evaluated CAPE with 81 types of test datasets, each type with a different prediction prevalence $\rho^z_{\mathbb{D}} \in \{0.1, \ldots, 0.9\}$ for each subgroup $z \in \{0, 1\}$. The general trend we observe is that CAPE outperforms `Max_Acc` whenever there is a significant shift in prior probabilities. We report three interesting sets of results here.

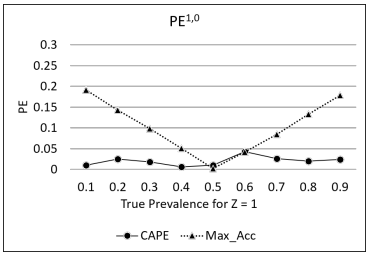
First, we consider test datasets with $\rho^0_{\mathbb{D}} = 0.5$, and $\rho^1_{\mathbb{D}}$ ranging between 0.1 and 0.9. Since CAPE accounts for prevalence changes, its accuracy on $\mathbb{D}$ for group Z=1 (Figure 3a) is consistently higher than



(a) Accuracy of $Z = 1$.



(b) Prevalence Difference for $Z = 1$.



(c) PE for group Z=1 with respect to Z=0.

Figure 3: Comparing accuracy and PD and PE on synthetic datasets with varying prevalences for group Z=1. The prevalence for group Z=0 is fixed at 0.5. The results are averaged over 20 iterations and the standard deviation is of the order 10^{-3}.

		Accuracy		**Δ**		**PE^{1,0}**		**PE^{0,1}**	
Z	$\rho^z_{\mathbb{D}}$	CAPE	Max_Acc	CAPE	Max_Acc	CAPE	Max_Acc	CAPE	Max_Acc
0	0.1	0.940	0.880	0.009	0.094	0.060	0.094	0.057	0.104
1	0.1	0.930	0.855	0.016	0.110				
0	0.2	0.894	0.855	0.017	0.084	0.199	1.432	0.012	0.142
1	0.8	0.909	0.877	0.006	0.074				
0	0.9	0.929	0.851	0.012	0.120	0.003	0.029	0.003	0.028
1	0.9	0.940	0.879	0.006	0.097				

Table 1: Accuracy, PD and PE metrics on synthetic datasets when test dataset $\mathbb{D}$ is such that $\rho^z_{\mathbb{D}} \neq 0.5$, for each $z \in \{0, 1\}$.

Max_Acc, except for the dataset with $\rho^1_{\mathbb{D}} = 0.5$ where the accuracies become nearly equal.

The prevalence difference (Figure 3b) is lower for CAPE whenever there is a prior probability shift (i.e., when $\rho^1_{\mathbb{D}} \neq 0.5$). In fact, for Max_Acc, $\hat{\rho}^1_{\mathbb{D}}$ remains around 0.5 across all the test datasets. Thus, $\Delta^1_{\mathbb{D}}$ for Max_Acc degrades linearly with increasing shift of $\rho^1_{\mathbb{D}}$ from the value 0.5.

Figure 3c shows the comparison of CAPE and Max_Acc in terms of fairness. Recall that a lower value of PE implies a greater degree of fairness. We observe that the PE value (of group $Z = 1$ with respect to group $Z = 0$) for Max_Acc increases when the true prevalence $\rho^1_{\mathbb{D}}$ deviates from 0.5. However, for CAPE, the PE value remains consistently low, across the different types of test datasets. This highlights that CAPE is better able to handle the prior-probability shifts between the training and test datasets. Moreover, note the similarity between Figure 3b and Figure 3c, which demonstrates that a high value of PD for group $Z = 1$ correlates with a high value for the PE metric for the same group. This highlights that the predictions of CAPE are more fair, compared to the purely accuracy maximizing Max_Acc.

In Table 1, we report results for scenarios where *both* $\rho^0_{\mathbb{D}}$ and $\rho^1_{\mathbb{D}}$ significantly deviate from their corresponding prevalences in the training set. The results are representative of the general trend we observed in the other test datasets—CAPE outperforms Max_Acc on the accuracy, PD, and PE metrics.

Real-world datasets: For COMPAS, columns 3 and 4 of Table 2 highlight that the true prevalences of the training (screened in the year 2013) and test (screened in the year 2014) datasets are significantly different. This is indicative of a possible prior probability shift. Column 5 shows that the Q-ALG module of CAPE makes a good estimate of the true prevalences of the test dataset. On the other hand, for MEPS, we observe a shift only for the group Z=1, between the training set (surveys in the year 2015) and test set (surveys in the year 2016). Since the differences in prevalences are rather small, this dataset is of interest—it allows us to investigate the performance of CAPE when the extent of prior probability shift is small. Though the prevalences estimated by Q-ALG seem similar to the training set, the difference in the estimates of Q-ALG and the prevalences of the test datasets are only 0.02 and 0.006, for Z=0 and Z=1 respectively, and are thus good estimates.

Table 3 summarizes the results on COMPAS and MEPS datasets for CAPE, Max_Acc, and the other fair algorithms described in Section 5.2. We consider two versions of our algorithm—CAPE-$\mathbb{D}$ and CAPE-1. The version CAPE-$\mathbb{D}$ considers the whole test dataset $\mathbb{D}$ during prediction, while CAPE-1 considers individual instances during prediction (similar to what the other algorithms do). We expect CAPE-$\mathbb{D}$ to perform better than CAPE-1 since the Q-ALG module is expected to perform better for larger test datasets.

Results on COMPAS dataset: CAPE-$\mathbb{D}$ outperforms Max_Acc on PD metric (Δ), and all the other fairness metrics (FPR-diff, FNR-diff, Accuracy-diff, and PE). The prediction prevalences of Max_Acc

	Z	Training Data True Prevalence ρ^z_D	Test Data True Prevalence $\rho^z_{\mathbb{D}}$	Quantifier's Estimate $\hat{q}(\mathbb{D}^z)$
COMPAS	**0**	0.327	0.636	0.592
	1	0.486	0.706	0.644
MEPS	**0**	0.253	0.253	0.273
	1	0.124	0.117	0.123

Table 2: The table shows prevalences and quantifier's estimates for COMPAS and MEPS datasets. Column 3 and 4 show prior probability shifts. Column 5 highlights the prevalence estimates obtained by Q-ALG module of CAPE algorithm on the test datasets.

(0.284 and 0.542) are close to the true prevalences of the *training* set (0.327 and 0.486), which highlights the inability of Max_Acc to account for the prior probability shift. One critical observation about CAPE-$\mathbb{D}$ is that FPR-diff=0.081 and FNR-diff=0.027 (both being low values), which implies that the predictions exhibit equalized odds. In comparison, these differences for Max_Acc are 0.127 and 0.289 (higher than CAPE-$\mathbb{D}$). In fact, for Max_Acc, $\text{FPR}_{Z=1}$ is almost twice than $\text{FPR}_{Z=0}$, whereas $\text{FNR}_{Z=1}$ is almost half of $\text{FNR}_{Z=0}$. This implies that Max_Acc imposes unfair higher risks of recidivism on African-American defendants, while Caucasian defendants are predicted to have lower risks than they actually do.

We observe that Δ^0 is the lowest for CAPE-$\mathbb{D}$ among all other classifiers. For $Z = 1$, Meta_fair-fdr is the only other fair classifier with a lower Δ^1 value compared to CAPE. However, the predictions of Meta_fair-fdr have high false positive rates, and low accuracies. Note that a trivial classifier, which always predicts positive labels, will have FNR-diff=0, FPR-diff=0, Accuracy-diff=0.07. However, this classifier will have high PD for both groups (Δ^0=0.364 and Δ^1=0.294), which indicates a substantial skew between the false positives and false negatives. Thus, PD is an important metric that, in addition to accuracy, captures the learning ability of the classifiers.

Moreover, the $\text{PE}^{1,0}$ and $\text{PE}^{0,1}$ values are lowest for both versions of CAPE, which reinforces our hypothesis that predictions of CAPE are highly PE-fair in the presence of a high degree of prior-probability shift.

Results on MEPS dataset: Though FPR-diff and FNR-diff are lower for Max_Acc compared to CAPE-$\mathbb{D}$, the actual FNR values are much higher than CAPE-$\mathbb{D}$ for both subgroups. The overall accuracy of CAPE-$\mathbb{D}$ is 84.9% which is slightly better than that of Max_Acc (84.3%). The accuracy for both CAPE-$\mathbb{D}$ and CEOP('fpr') are equal for $Z = 1$. However, CEOP('fpr') trivially classifies everyone in the group $Z = 1$ as being in the negative class, i.e. $\hat{h}(X) = 0$, and this issue is flagged by the corresponding Δ^1 value (which is 39 times higher than that of CAPE-$\mathbb{D}$) and $\text{PE}^{1,0}$ (which is 69.7 times higher than CAPE-$\mathbb{D}$). Both CEOP('fnr' and 'weighted') have the highest accuracy for $Z = 1$. However, they perform poorly on FNR, PD, and PE. Reweigh performs better than CAPE-$\mathbb{D}$ on FPR-diff, FNR-diff and accuracy diff, but the PD values for Reweigh are 14.3 and 62.67 times higher than CAPE-$\mathbb{D}$ for the groups $Z = 0$ and $Z = 1$ respectively.

Since the extent of the prior-probability shift is low for MEPS, the other fair algorithms are expected to perform well on the fairness metrics. We observe that both versions of CAPE perform better than other algorithms in terms of the PE and PD metrics. Thus, the predictions of CAPE are the most PE-fair, even when the degree of prior-probability shift is low. We make a final observation on our experimental results. Since both COMPAS and MEPS are real-world datasets, the distributional changes highlighted in Table 2 may not be due to prior probability shifts alone. Although CAPE is designed to handle only prior probability shifts, the good performance of both CAPE-$\mathbb{D}$ and CAPE-1 on a wide range of metrics for these real-world datasets shows the robustness of our approach.

6 CONCLUSION AND DISCUSSION

We addressed the problem of fair classification in the presence of prior probability shifts. We provided a framework, called CAPE, that combines sampling, ensemble and quantification techniques to provide fair predictions. Although several ensemble methods [38] have been proposed in the literature, they cannot be directly used for the problem we consider. We also introduced a metric called *prevalence difference* (PD), and theoretically established its compatibility with accuracy. We used the PD metric as a key component within CAPE, and established that CAPE can be used to ensure a low value of PD in the predictions.

Through extensive experimental evaluation we observed that CAPE performs well on both fairness and accuracy metrics. On synthetic datasets, we observed that the accuracy, PE, and PD metrics of Max_Acc (a classifier that maximizes accuracy on the training dataset) degrade with increasing dataset shifts. However, the performance of CAPE remained consistently better than Max_Acc under various extents of dataset shifts. On real-world datasets, we compared CAPE with state-of-the-art fair algorithms and observed that CAPE performed well with respect to PE fairness, as well as other well-studied fairness metrics.

CAPE has many advantages, as highlighted in the earlier sections of this paper. However, it is not the ultimate solution for all fairness issues. We now discuss some limitations. CAPE assumes that Y labels in the dataset are the ground-truth (eg, whether a defendant re-offends). This assumption may not be true in some datasets where the labels are human decisions (eg, jury deciding whether a defendant will re-offend). Another problem is that our techniques rely on the value of the sensitive attributes of each instance. In situations where the sensitive attribute values are unavailable or prohibited from being used, our techniques do not apply.

An obvious variant of CAPE is to use only a quantifier (but no classifier) in the training phase. In the prediction phase, the quantifier is then used to estimate the prevalence of a test dataset $\mathbb{D}$. This estimate is fed to the PP-SAMPLING module to sample the training set, which is then used to train a classifier and predict on $\mathbb{D}$. The obvious advantage is that we need to train *one* classifier instead of $|\Theta|$ classifiers, for each z. The downside of this modification is the need to persist the training data, and the prediction phase becomes computationally expensive since a classifier needs to be trained (from scratch) for every new test dataset. Moreover, such a classifier rely entirely on the estimates given by the quantifier, which varies a lot depending on the training set. We observed that

			FPR			FNR			Accuracy			Prediction Prevalences		Δ			
		Algorithms	Z = 0	Z = 1	diff	Z = 0	Z = 1	diff	Z = 0	Z = 1	diff	Z = 0	Z = 1	Z = 0	Z = 1	$PE^{1,0}$	$PE^{0,1}$
COMPAS	CAPE	CAPE-$\mathbb{D}$	0.461	0.380	0.081	0.302	0.275	0.027	0.640	0.694	0.054	0.612	0.623	0.024	0.083	0.092	0.082
		CAPE-1	0.271	0.290	0.019	0.451	0.322	0.129	0.614	0.687	0.073	0.448	0.564	0.188	0.142	0.149	0.106
		Max_Acc	0.132	0.259	0.127	0.629	0.340	0.289	0.552	0.684	0.132	0.284	0.542	0.352	0.163	0.799	0.376
	Pre	Reweigh	0.283	0.139	0.144	0.493	0.543	0.050	0.583	0.576	0.007	0.425	0.363	0.211	0.343	0.257	0.271
	In	Meta-fair-sr	0.977	0.849	0.128	0.102	0.492	0.390	0.579	0.403	0.176	0.927	0.609	0.291	0.097	0.454	0.623
		Meta-fair-fdr	0.965	0.901	0.064	0.162	0.356	0.194	0.545	0.483	0.062	0.884	0.719	0.248	0.013	0.298	0.329
		AD	0.124	0.167	0.043	0.638	0.467	0.171	0.549	0.621	0.072	0.275	0.425	0.361	0.281	0.432	0.252
	Post	CEOP-fpr	0.066	1.000	0.934	0.722	0.000	0.722	0.517	0.706	0.189	0.201	1.000	0.435	0.294	3.875	0.699
		CEOP-fnr	0.000	0.247	0.247	1.000	0.390	0.610	0.364	0.652	0.288	0.000	0.503	0.636	0.203	undef	0.900
		CEOP-weighted	0.000	0.194	0.194	1.000	0.405	0.495	0.364	0.657	0.292	0.000	0.477	0.636	0.229	undef	0.900
		ROC-aod	0.004	0.019	0.015	0.978	0.900	0.078	0.377	0.360	0.017	0.016	0.076	0.620	0.630	3.799	0.696
		ROC-eod	0.019	0.046	0.027	0.911	0.782	0.129	0.414	0.434	0.020	0.064	0.167	0.572	0.539	1.518	0.520
MEPS	CAPE	CAPE-$\mathbb{D}$	0.131	0.068	0.063	0.425	0.488	0.063	0.794	0.883	0.089	0.243	0.120	0.010	0.003	0.031	0.135
		CAPE-1	0.175	0.087	0.088	0.347	0.423	0.076	0.781	0.874	0.093	0.296	0.144	0.043	0.027	0.024	0.107
		Max_Acc	0.004	0.012	0.008	0.910	0.888	0.022	0.766	0.890	0.124	0.037	0.014	0.216	0.103	0.085	0.483
	Pre	Reweigh	0.276	0.242	0.034	0.250	0.226	0.024	0.731	0.760	0.029	0.396	0.305	0.143	0.188	0.307	0.862
	In	Meta-fair-sr	0.322	0.213	0.109	0.210	0.243	0.033	0.706	0.783	0.077	0.440	0.277	0.187	0.160	0.167	0.572
		Meta-fair-fdr	0.347	0.254	0.102	0.193	0.218	0.025	0.692	0.758	0.066	0.463	0.308	0.210	0.191	0.202	0.657
		AD	0.062	0.051	0.011	0.644	0.569	0.075	0.791	0.889	0.098	0.136	0.095	0.117	0.022	0.236	0.728
	Post	CEOP-fpr	0.078	0.000	0.078	0.573	1.000	0.427	0.797	0.883	0.086	0.166	0.000	0.087	0.117	2.160	undef
		CEOP-fnr	0.034	0.022	0.012	0.803	0.704	0.102	0.771	0.899	0.128	0.075	0.054	0.178	0.063	0.257	0.771
		CEOP-weighted	0.032	0.021	0.011	0.816	0.704	0.112	0.770	0.899	0.129	0.070	0.053	0.183	0.064	0.294	0.839
		ROC-aod	0.329	0.253	0.076	0.205	0.210	0.005	0.702	0.752	0.050	0.447	0.216	0.194	0.199	0.244	0.745
		ROC-eod	0.336	0.233	0.103	0.194	0.227	0.033	0.700	0.768	0.068	0.455	0.296	0.202	0.179	0.188	0.623

Table 3: Comparing CAPE with Max_Acc and other fair classifiers on the test-set of the real-world datasets, COMPAS and MEPS.

such a classifier performs poorly in terms of accuracy, PE, and PD. It may be worthwhile to investigate the trade-offs of CAPE with other similar modifications.

While our technique addresses prior probability shifts between the training and test datasets, it may not be applicable under *concept drifts* [35, 45], i.e. when $\mathcal{P}(X|Y,Z)$ changes. Exploring possible improvements for CAPE to handle more general distributional changes remains open.

ACKNOWLEDGMENTS

This work was done when Arpita Biswas was a PhD student at Indian Institute of Science, Bangalore, India. She gratefully acknowledges the support of a Google PhD Fellowship Award.

REFERENCES

[1] Agency for Healthcare Research & Quality. 2016. Medical Expenditure Panel Survey. https://meps.ahrq.gov/mepsweb/.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.. In *ProPublica*. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Maria-Florina F Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D Procaccia. 2019. Envy-free classification. In *Advances in Neural Information Processing Systems*. 1238–1248. http://papers.nips.cc/paper/8407-envy-free-classification

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. In *NIPS Tutorial*.

[5] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[6] Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. 2010. Quantification via probability estimators. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 737–742.

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. In *Sociological Methods & Research*. Sage Publications Sage CA: Los Angeles, CA.

[9] Arpita Biswas and Suvam Mukherjee. 2019. Fairness Through the Lens of Proportional Equality. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. 1832–1834.

[10] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *IEEE international conference on Data mining workshops, 2009. ICDMW'09*. IEEE, 13–18.

[11] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 319–328.

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

[13] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. In *arXiv preprint arXiv:1810.08810*.

[14] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. In *arXiv preprint arXiv:1808.00023*.

[15] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[16] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *International Conference on Machine Learning*. 1397–1405.

[17] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[18] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2017. Decoupled classifiers for fair and efficient machine learning. In *arXiv preprint arXiv:1707.06613*.

[19] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[20] George Forman. 2005. Counting positives accurately despite inaccurate classification. In *European Conference on Machine Learning*. Springer, 564–575.

[21] George Forman. 2006. Quantifying trends accurately despite classifier error and class imbalance. In *ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 157–166.

[22] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 329–338.

[23] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. 2016. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*. 2415–2423.

[24] Pablo González, Jorge Díez, Nitesh Chawla, and Juan José del Coz. 2017. Why is quantification an interesting learning problem? *Progress in Artificial Intelligence* 6, 1 (2017), 53–58.

[25] VíCtor GonzáLez-Castro, RocíO Alaiz-RodríGuez, and Enrique Alegre. 2013. Class distribution estimation based on the Hellinger distance. *Information Sciences* 218 (2013), 146–164.

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

[27] Safwan Hossain, Andjela Mladenovic, and Nisarg Shah. 2020. Designing Fairly Fair Classifiers Via Economic Fairness Notions. In *Proceedings of The Web Conference 2020*. 1559–1569. https://dl.acm.org/doi/abs/10.1145/3366423.3380228

[28] Nan D Hunter. 2000. Proportional Equality: Readings of Romer. *Ky. LJ* 89 (2000), 885.

[29] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[30] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In *ICDM*. IEEE Computer Society, 924–929.

[31] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[32] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2018).

[33] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*. ACM.

[34] Meelis Kull and Peter Flach. 2014. Patterns of dataset shift. In *First International Workshop on Learning over Multiple Contexts (LMCE) at ECML-PKDD*.

[35] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. 2019. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (2019), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

[36] Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. 2020. Ensuring Fairness Beyond the Training Data. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 18445–18456. https://proceedings.neurips.cc/paper/2020/file/d6539d3b57159babf6a72e106beb45bd-Paper.pdf

[37] Jose G Moreno-Torres, Troy Raeder, RocíO Alaiz-RodríGuez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45, 1 (2012), 521–530.

[38] David Opitz and Richard Maclin. 1999. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research (JAIR)* 11 (1999), 169–198.

[39] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5684–5693.

[40] ProPublica. 2016. COMPAS Recidivism Risk Score Data & Analysis. github.com/propublica/compas-analysis.

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

[42] Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation* 14, 1 (2002), 21–41.

[43] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. In *arXiv preprint arXiv:1811.05577*.

[44] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness Without Harm: Decoupled Classifiers with Preference Guarantees. In *International Conference on Machine Learning*. 6373–6382. http://proceedings.mlr.press/v97/ustun19a/ustun19a.pdf

[45] Gerhard Widmer and Miroslav Kubat. 1996. Learning in the presence of concept drift and hidden contexts. *Machine learning* 23, 1 (1996), 69–101.

[46] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*. 1920–1953.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*. 1171–1180.

[48] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.

[49] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *Advances in Neural Information Processing Systems*. 229–239.

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

[51] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES*. ACM, 335–340.

[52] Lu Zhang, Yongkai Wu, and Xintao Wu. 2018. Achieving Non-Discrimination in Prediction. In *International Joint Conference on Artificial Intelligence, IJCAI*. 3097–3103.

[53] Indre Zliobaite and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artif. Intell. Law* 24, 2 (2016), 183–201.