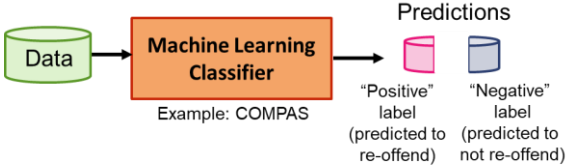
Ensuring Fairness under Prior Probability Shifts

Arpita Biswas
Harvard University
arpitabiswas@seas.harvard.edu

Suvam Mukherjee
Microsoft Corporation
sumukherjee@microsoft.com

AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

Problem Statement



Problem: COMPAS discriminates! [Angwin et al. 2016]
More likely to wrongly deny bail to an African-American.

Prior Probability Shifts

The phenomenon where the prior probability $P(Y)$ changes between the training and test datasets, but the class conditional probability $P(X|Y)$ remains unaltered. [Moreno-Torres et al., 2012]

		(2013) Training Data True Prevalence	(2014) Test Data True Prevalence
	Z	ρ^z_D	ρ^z_d
Caucasian	0	0.327	0.636
African-American	1	0.486	0.706

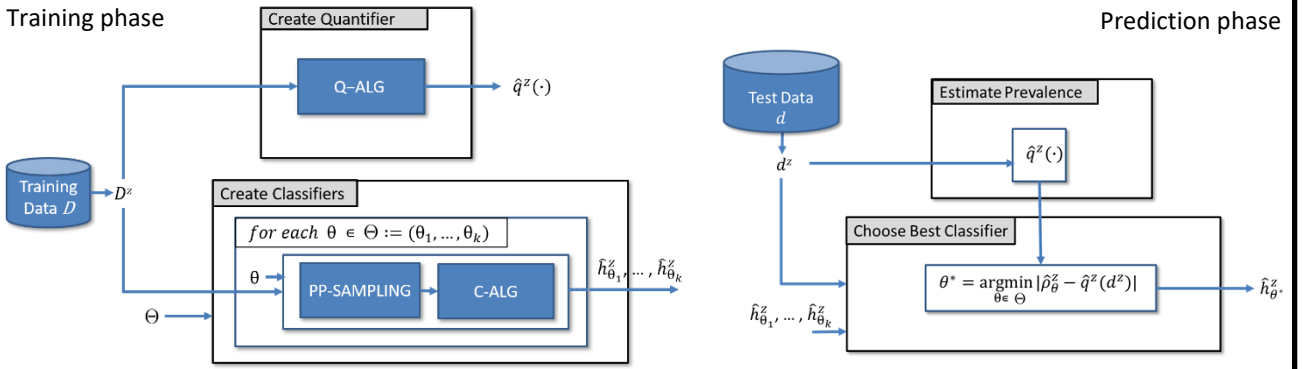
Proportional Equality [Biswas and Mukherjee 2019]

The predictions on the test dataset d is defined to be fair if, for all z, z'

$$\left| \frac{\hat{\rho}^z_d}{\hat{\rho}^{z'}_d} - \frac{\rho^z_d}{\rho^{z'}_d} \right| \leq \epsilon$$

Goal:
Maximize Accurate + fair + robust to prior probability shifts

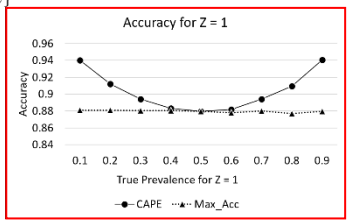
Proposed Framework: CAPE

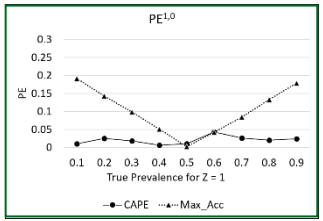


Results

Theorem 1. *Let $\Theta = \{\epsilon, 2\epsilon \ldots, k\epsilon\}$, $\epsilon \in (0,1)$, $k > \frac{1}{\epsilon} - 1$. For a group z, and test dataset d, let the quantifier be such that $|\rho^z_d - \hat{q}^z(d)| \leq \delta_1$, and the classifiers be such that $|\theta_j - \hat{\rho}^z_j| \leq \delta_2$, for small δ_1 and δ_2. Then, for the best classifier $J = \underset{j \in \{1,\ldots,k\}}{\arg\min} |\hat{\rho}^z_j - \hat{q}^z(d)|$, CAPE satisfies $|\rho^z_d - \hat{\rho}^z_d| \leq \epsilon + \delta_1 + \delta_2$.*

Synthetic Datasets





CAPE outperforms Max_Acc when there is significant shift in prior probabilities.

Max_acc degrades linearly with increasing shift, but CAPE performs well throughout.

Real-world Datasets	Method	Accuracy Z=1	Accuracy Z=0	PE (unfairness)
COMPAS	CAPE	**69**	**64**	**0.08**
	Max_acc	68	55	0.37
MEPS	CAPE	**89**	**79**	**0.13**
	Max_acc	**89**	76	0.48

CAPE does well on *both* fairness and accuracy.

More results in the paper: comparison against various state-of-the-art fair algorithms, and on several fairness metrics.

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