Balancing Classifier Fairness with Public Safety in Traffic Stops

Vikul Gupta (vikulg), Kuhan Jeyapragasan (kuhanj), Jaydeep Singh (jaydeeps)

Stanford University: CS 229 Final Project

Motivation

- The issue of algorithmic fairness has recently come to the forefront of machine learning, as classifiers increasingly propose decision rules in applications ranging from loan approval to criminal risk estimation. Expectations for group and individual fairness manifest as constraints on learned decision rules, creating an accuracy/fairness trade-off in classification.
- We expand on previous analysis by exploring police traffic stop outcomes, to analyze the trade-off between public safety/utility and fairness, utilizing data from the Stanford Open Policing Project.

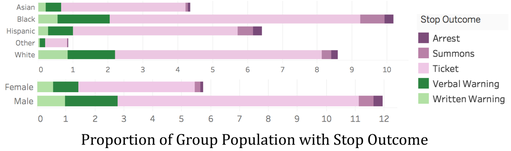
Measures of Fairness

Given a classifier D: X → [0,1] and sensitive attributes {S}, we judge fairness by the adherence to the following metrics

- **Statistical Parity:** For each s in {S}, the proportion of positively classified individuals is equal. Measured by positive classification rate (PCR)
- **Predictive Equality:** For each s in {S}, the accuracy of positive classification is equal. Measured by false positive rate (FPR)

Dataset + Features

- Data from Stanford Open Policing Project about traffic stops in Connecticut
- Each observation is a traffic stop
- Features include Stop Date/Time/Location, Driver Race/Gender/Age, and Stop Outcome
- Race binned into White, Hispanic, Black, Asian, and Other
- Gender binned into Male and Female
- Age binned into 10-year groups
- Feature created for each bin as indicator variable
- Stop Outcome (label to predict) binned into Warning and No-Warning



Proportion of Group Population with Stop Outcome

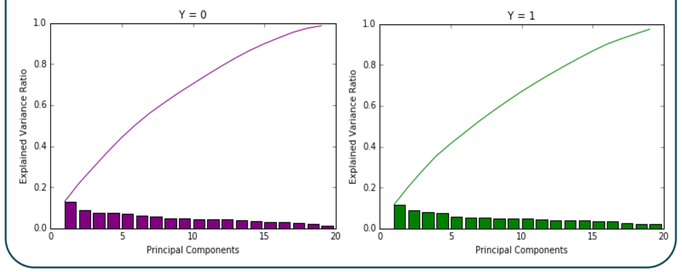
Preliminary Feature Analysis

To gain further insight into the features, we ran PCA using a linear kernel. PCA was run in four different ways:

1. All observations, with StopOutcome
2. All observations, without StopOutcome
3. Observations where StopOutcome = 1, without StopOutcome
4. Observations where StopOutcome = 0, without StopOutcome

In all 4 cases, the first principal component explained no more than 13% of variance. The latter 2 tests took 20 principal components for the cumulative explained variance to be greater than 99%. This observation suggests that each feature x_i given y is relatively independent of the other features x_j given y, justifying the Naïve Bayes assumption.

Explained Variance over 19 Principal Components



Unconstrained Models

Motivated by the results of PCA, which imply little correlation between features conditioned on outcome, we applied Naïve Bayes as a robust baseline alongside a logistic regression classifier. The higher false positive and positive classification rates for the race feature in both models suggest violations of statistical parity and predictive equality.

Fairness Metrics	Logistic Regression	Naïve Bayes
Test Classification Accuracy	0.690	0.737
Predictive Equality: (False Positive Rates)	Overall: 0.3541 Black: 0.5714 Hispanic: 0.6296 White: 0.3161 Female: 0.3161 Male: 0.3759	Overall: 0.6355 Black: 0.6295 Hispanic: 0.82222 White: 0.6261 Female: 0.64155 Male: 0.6320
Statistical Parity (Positive Classification Rates)	Overall: 0.5573 Black: 0.7958 Hispanic: 0.8330 White: 0.5012 Female: 0.5093 Male: 0.5825	Overall: 0.8009 Black: 0.8346 Hispanic: 0.9476 White: 0.7778 Female: 0.7975 Male: 0.8027

Regularization-Based Fairness

To address discrepancies in classification of a sensitive group S, we adopt a regularization approach during training. The following loss functions penalize the model for predicting divergent outcomes for individuals of sensitive groups. While the individual loss promotes equity for each training sample, the group loss minimizes discrimination averaged over a protected group. Below, **let θ refer to learned parameters, and S_i the class labels** for a fixed, sensitive group.

- **Individual Loss:**

$$L_{\text{individual}} = \frac{1}{|S_1|\ldots|S_n|} \sum_{\substack{x_i \in S_i \\ x_j \in S_j \\ y(x_i)=y(x_j) \\ i<j}} (\theta \cdot x_i - \theta \cdot x_j)^2$$

- **Group Loss:**

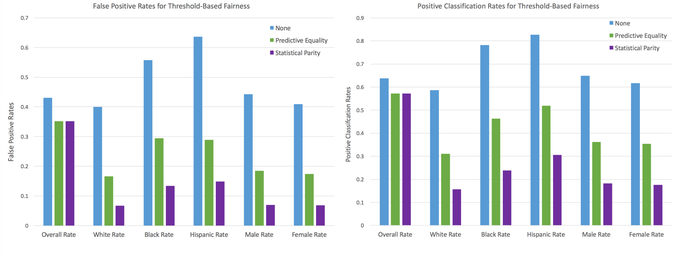
$$L_{\text{group}} = \left(\frac{1}{|S_1|\ldots|S_n|} \sum_{\substack{x_i \in S_i \\ x_j \in S_j \\ y(x_i)=y(x_j) \\ i<j}} \theta \cdot x_i - \theta \cdot x_j \right)^2$$

Yielding the following overall regularized loss:

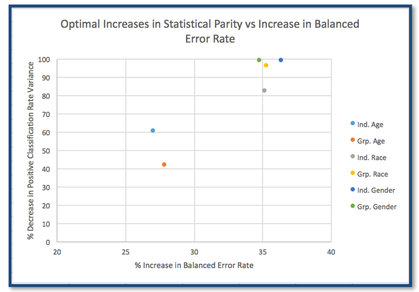
$$L_{\text{overall}} = L_{\text{logistic}} + \mu L_{\text{individual}} + \nu L_{\text{group}} + \lambda \|\theta\|^2$$

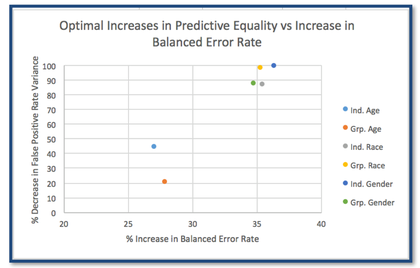
Threshold-Based Fairness

To address violations of fairness measures such as statistical parity and predictive equality, we implemented a post-training threshold approach. Given the optimal theta using unconstrained models, we define new thresholds such that one of the measures of fairness will be met in the training set.



Regularization Results





Discussion

- Interestingly, threshold values optimizing our fairness metrics were the same for five of our six model types– suggesting a link between the two measures of fairness.
- Our regularizer, despite not penalizing false positive or positive classification rate discrepancies, was found to be fairly effective at improving both metrics with proportionally smaller error rate increases.
- Further research could include analysis with other classification models, with a dataset with richer features, and on the relationships among various measures of fairness and predictive accuracy.

References

Emma Pierson, Camelia Simoui, Jan Overgoor, Sam Corbett-Davies, Vignesh Ramachandran, Cheryl Phillips, and Sharad Goel. A large-scale analysis of racial disparities in police stops across the united states. *CoRR*, 2017.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *CoRR*, abs/1706.02409, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017.