Minority Report: ML Fairness in Criminality Prediction

Dominick Lim (djlim@stanford.edu) and Torin Rudeen (torinmr@stanford.edu)

Stanford University: CS 229 Final Project

Motivation

Fairness in ML refers to the study of ways to examine and adjust for discrimination and disparate impact in Machine Learning algorithms. One might assume that ML is by definition fair and impartial, but this is not necessarily true: training data (like a record of police arrests) might contain racial or other forms of discrimination which an algorithm can reproduce.

Even if the training data is assumed to be unbiased, a machine learning algorithm can have disparate impact. For example, one study[1] found that when predicting who would commit a crime, the algorithm had a vastly higher false positive rate for ethnic minorities (who did have a higher rate of ground truth positives in the data).

Dataset and Features

Our dataset is the National Longitudinal Survey of Youth 1997,[2] which followed thousands of youth over 17 years and contains:

- Demographic data
- Criminal activity (per year)
- Employment and educational status (per year)
- Much more

We decided to use incarceration during the subject's 25th year as the prediction variable, with all information up through the 24th year as features. We preprocessed the data by:

- Removing subjects who had died or stopped responding to the survey before their 25th year.
- Converting continuous-valued variables into discrete variable by bucketization.
- Splitting into Training (49%), Cross-Validation (21%), and Test (30%) sets.

The final dataset had:

- 350 features
- 2814 data points
  - 30.2% Black
  - 47.4% Non-Black/Non-Hispanic

Analyzing a Simple Model

We trained a simple Naive Bayes classifier on our data. The prediction variable (incarceration during 25th year) was chosen to mimick a hypothetical application, where a police department wanted to predict which citizens were likely to commit a crime in the next year. Here are the results:

	Black	Non-Black/Non-Hispanic
False Positive	0.195	0.102
False Negative	0.263	0.400

Thus, for Black subjects false positive rates were higher, but false negative rates were lower. This was surprising to us, but can be explained by the fact that a much greater percentage of the Black population was ground truth positive (i.e. was actually incarcerated in their 25th year):

	Black	Non-Black/Non-Hispanic
Ground-truth positive	15%	4%
Predicted positive	27%	12%

Previous approaches to fairness in ML such as [1] have focused on equalizing the positive prediction rate among different classes. We instead decided to focus on equalizing the false positive rate. We then evaluated different algorithms by how much they had to increase the false negative rate for the minority class in order to make its false positive rate equal to that for the majority class.

Threshold-Based Fairness

We first implemented a fairness method described in [1], which simply adjusted the classification threshold for Black subjects until the false positive rates were equalized on the training set. This gave the following results on the test set:

	Black	Non-Black/Non-Hispanic
False Positive	0.0930	0.102
False Negative	0.526	0.400

Feature Selection-Based Fairness

We next implemented a method of our own devising, which worked by feature selection. We wanted to devise a feature selection criterion which would attempt to reach a target false positive rate, while keeping false negatives as low as possible.

We recognized that this was similar to a constrained optimization problem, so we decided to use a **penalty method** to convert the constrained optimization problem into an unconstrained optimization problem by adding a penalty term equal to the square of the deviation from the desired constraint. This gave us the following cost function:

$$\mathrm{FN}_B + \gamma(\mathrm{FP}_B - \mathrm{FP}_W)^2$$

Where γ is a hyper parameter: Larger γ means more weight is placed on reaching the desired false positive goal. This leads to a forward feature selection algorithm where at each step we add the feature f to the current feature set $\mathcal{F}$ which satisfies the following equation (on the cross-validation set):

$$\arg\min_{f \notin \mathcal{F}} \left(\mathrm{FN}_B(\mathcal{F} \cup f) + \gamma(\mathrm{FP}_B(\mathcal{F} \cup f) - \mathrm{FP}_W)^2 \right)$$

This gave the following results on the test set (using the original feature set for prediction on Non-Black/Non-Hispanic subjects):

	Black	Non-Black/Non-Hispanic
False Positive	0.0744	0.102
False Negative	0.421	0.400

We also implemented backward feature selection on the same objective, at each step removing the feature f from $\mathcal{F}$ satisfying:

$$\arg\min_{f \in \mathcal{F}} \left(\mathrm{FN}_B(\mathcal{F} \setminus f) + \gamma(\mathrm{FP}_B(\mathcal{F} \setminus f) - \mathrm{FP}_W)^2 \right)$$

This gave the following results on the test set:

	Black	Non-Black/Non-Hispanic
False Positive	0.0698	0.102
False Negative	0.368	0.400

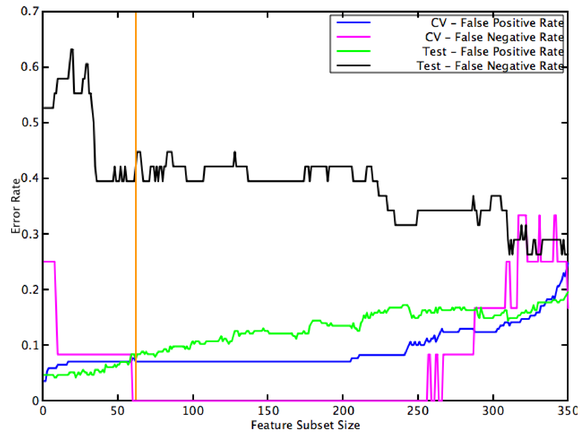

Figure 1: Forward Feature Selection Performance on Test and Cross-Validation Sets (chosen set indicated)

Discussion

We analyzed the behavior of a simple machine learning algorithm on a racially diverse dataset, and found that it had a much higher false positive rate on the minority class than the majority class.

We then tried two approaches to try to equalize false positive rates. Both approaches were successful at this, but feature selection was able to do so with a much lower false negative rate.

An interesting future investigation would be to see if the features selected by this method have the same effect when fed into a different algorithm, such as an SVM.

References

[1] F. Kamiran, A. Karim, S. Verwer, and H. Goudriaan, "Classifying Socially Sensitive Data Without Discrimination: An Analysis of a Crime Suspect Dataset" in *2012 IEEE 12th International Conference on Data Mining Workshops*, 2012.

[2] Bureau of Labor Statistics, U.S. Department of Labor, "National Longitudinal Survey of Youth 1997". Produced by the *National Opinion Research Center, the University of Chicago* and distributed by the *Center for Human Resource Research, The Ohio State University*, 2013.