BIAS IN MACHINE LEARNING

ELYSSA SLIHEET, WILL PRICE, JACOB SCHRUM
SOUTHWESTERN UNIVERSITY, DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

APRIL 3rd, 2019

Abstract

This project aims to investigate current biases in machine learning. Machine learning uses statistical techniques to give computer systems the ability to "learn" from data by generating and refining models. These models are used to make classifications or predictions. We investigate the ways in which machine learning algorithms, specifically the Logistic Regression model and the Support Vector Machine, can encode and reinforce societal biases.

Motivation

Machine learning models are used in many areas but few are as significant as the judicial system. Algorithms for predicting a criminal’s likelihood of reoffending (known as recidivism) are currently in use. We aim to demonstrate potential inaccuracies that could be the difference in a number of important judicial decisions.

Methodology and Dataset

Models:

- Logistic Regression - predictive model that uses the logistic function for binary classification problems
- Support Vector Machine - supervised learning model that maps input to a higher dimensional space and trains on the transformed data

Dataset Attributes - Sex, age, **COMPAS** score, number of previous offenses, juvenile felony account, type of crime, etc. Race was not included in the training of either model.

Confusion Matrix - A confusion matrix is a method for testing the validity of a machine learning model. One axis of the matrix represents model predictions, while the other axis represents actual observations. The confusion matrix can be used to tally true positives and negatives as well as false positives and negatives.

Results

African American

Support Vector Machine

Total: 2363	P-NR	P-R
A-NR	648.4 27.4%	438.6 18.6%
A-R	330.0 14.0%	947.0 40.1%

Accuracy: 67.5%

Logistic Regression

Total 2364	P-NR	P-R
A-NR	595.4 25.2%	484.6 20.5%
A-R	326.6 13.8%	956.8 40.5%

Accuracy: 65.7%

Caucasian

Support Vector Machine

Total: 1447	P-NR	P-R
A-NR	634.6 43.9%	188.2 13.0%
A-R	306.4 21.2%	318 22.0%

Accuracy: 65.8%

Logistic Regression

Total: 1419	P-NR	P-R
A-NR	622.0 43.8%	193.6 13.6%
A-R	281.6 19.8%	321.8 22.7%

Accuracy: 66.5%

Hispanic

Support Vector Machine

Total: 326	P-NR	P-R
A-NR	165.4 50.1%	42.6 13.1%
A-R	61.0 18.7%	57.4 17.6%

Accuracy: 68.3%

Logistic Regression

Total: 350	P-NR	P-R
A-NR	176.6 50.5%	45.2 12.9%
A-R	60.8 17.4%	67.4 19.3%

Accuracy: 69.7%

The tables above contain values averaged across five runs of each model. P stands for predicted, A stands for actual. R stands for recidivism, and NR stands for no recidivism.

Discussion

The upper right corner of each matrix represents the number of times that the model, either SVM or LR, predicted that an individual of a given race would commit a crime after their release, but they actually did not. This is much higher for African Americans at 18.6% (for SVM) and 20.5% (for LR) than for Caucasians and Hispanics. The lower left corner of each matrix represents the number of times that an individual was predicted to not commit another crime after their release, but they actually did. This is greatest for Caucasians at 21.2% (for SVM) and 19.8% (for LR), then for Hispanics at 18.7% (for SVM) and 17.4% (for LR), and lowest for African Americans at 14.0% (for SVM) and 13.8% (for LR). From this we see that Caucasians are more likely than Hispanics and African Americans to not be classified as future reoffenders. On the other hand, African Americans are more likely than Hispanics and Caucasians to be classified as future reoffenders. These classifications have dire consequences. Another notable result is that these models perform at roughly the same rate when trained on datasets that include race as opposed to those that do not. Therefore, the generated models are racially biased despite being unaware of the race of the individuals used in training.

Further Development

- Equal representation of races in the training data set
- Control for prior felony counts when training the model
- Training separate models for different racial categories
- Precision and recall, F1 score, not just accuracy

Acknowledgements

Dr. Joey King
Dr. Jacob Schrum
https://www.kaggle.com/danofer/compass
https://github.com/Squillyprice01/KingCreativtiy2019