Fairness in Machine Learning Algorithms

Jayron David Martinez: JDM21582@gmail.com
Synergy Quantum Academy, Class of 2023
Prof. Vatsal Sharan, Theoretical Computer Science Group
USC Viterbi Department of Computer Science Engineering, SHINE 2023

USC Viterbi School of Engineering

SHINE Summer High School Intensive in Next-Generation Engineering

USC-Meta Center for Research and Education in AI and Learning

Introduction

Machine learning (ML) is a set of methods in computer science for learning patterns from data. Our goal is to analyze the UC Irvine Adult dataset [1] which is a subset of 1994 census data. Each person (datapoint) has 12 different features, consisting of sex, age, race, martial status, etc. We attempt to use machine learning methods to predict whether each person in the dataset makes above or below $50k annually.

Recently, the "fairness" of ML algorithms has been under examination. The fairness of ML algorithms today is key to the advancement of technology due to incorporation of algorithms that help decide results. If an algorithm turns out biased, it will potentially affect someone negatively do to system's error. We will analyze our work to see if the algorithm we train is biased in any way. All our experiments are run with python and the scikit learn package [2]. Data preprocessing was done identically to [3].

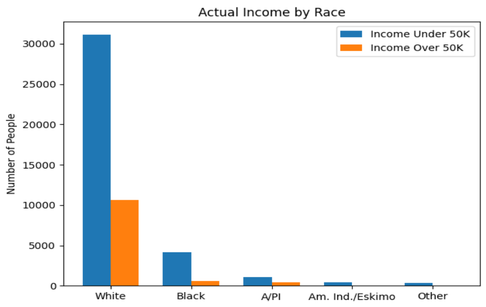


Figure 1: Actual income by racial group in UCI Adult Dataset.

Objective & Impact of Professor's Research

Work was performed under the guidance of Professor Vatsal Sharan and PhD student Siddartha Devic. Professor Sharan works on machine learning, statistics, and theoretical computer science. His research aims to understand how to solve learning and inference tasks in the face of various computational and statistical constraints, such as limited memory or too little data. He is also broadly interested in additional desiderata like robustness or fairness of ML algorithms.

Methods and Results

Our data consists of 48k people, and a label for each corresponding to whether they make over or under $50k. We randomly split our data into a training and test set. After the model was trained with these, the test data points and labels were tested on the trained model. This would be the most crucial point to spot biasness against certain races, sex, or features of a person (see Fig. 1 for distribution of labels by race). We tested many different ML algorithms. The following figure shows the resulting accuracy of each.

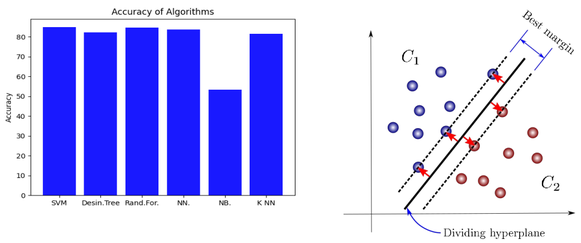
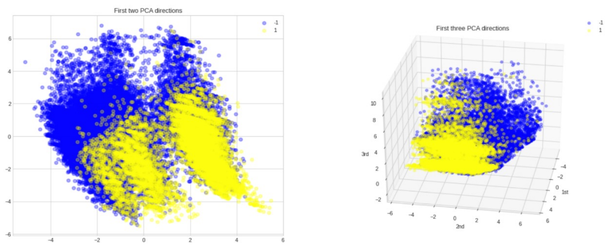


Figure 2: (Left) Accuracy by algorithm. (Right) SVM illustration.

Due to the relatively low accuracy of all methods, we inspect the 2- and 3-dimensional principal component analysis (PCA) plot to determine if clean separation is possible.



Notice the significant overlap in positive and negatively labeled examples (yellow and blue respectively). Therefore, we expect that we are doing quite well above 80% accuracy.

We choose the Support Vector Machine (SVM) with linear kernel algorithm for bias analysis due to its relatively good performance as well as method simplicity (see Fig. 2, Right). Nonetheless, the results generally hold for all methods tested. The SVM algorithm learns to separate different data points into two classes using a hyperplane. SVM achieves an overall **accuracy of 0.8489**.

Attempting to De-bias the Classifier

The following figure shows the error rate by race. Interestingly, it appears that our learned classifier has the *highest* error rate amongst people who self-report as "White".

We attempt to fix this imbalance in the following manner. We train an SVM on only self-reporting White people, and then train *another* SVM on everyone else. We report the resulting accuracies in the below table.

Dataset	Accuracy of SVM trained only on White	Accuracy of SVM trained only on non-White
White only	0.8425	0.8436
Non-white only	0.8782	0.8960

We can calculate the resulting accuracy of the *aggregate classifier* in the following way:
(% White) * (Acc of SVM trained on White)
+ (% non-White) * (Acc of SVM trained non-white)
= (41762 / 48842) * 0.843 + (7080 / 48842) * 0.896
= **0.8503**

The accuracy of the new aggregate model is not much better than before. This indicates that most of the complexity is in the self-reported "White" subgroup of the dataset. However, we achieve quite good performance on Non-white individuals even if only training on the White ones. This means that somehow the race feature seems less useful (in isolation) for this prediction task. Future work could attempt to use more recent methods from the ML fairness literature to achieve better performance on the White subgroup, however it seems that this is a difficult task, and it is not clear what much more can be done.

Relation to Coursework, Advice for Future SHINE Students

With exploring the computer science department, SHINE was able to show me a new concept, machine learning. I was able to understand and work with machine learning even though it took extensive critical thinking to understand the concept. STEM is used in different areas across different majors/careers. Having said that, it is important to know the basics of machine learning as it is affecting you one way or another without you knowing.

If you are open to learn and are able to assimilate to the work habitat, than it will become easier. But it is not only about learning in SHINE, it is about the people that you meet, especially since everyone comes from different background, you gain experience from just meeting people here. These type of connections are essential if you want to pursue something big. Questions will be important to understand, and it is key to being here in this program.

Citations

[1]: Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.
https://archive.ics.uci.edu/ml/datasets/adult
[2]: Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[3]: Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. 2020. Classification under misspecification: halfspaces, generalized linear models, and evolvability. Neurips 2020.

Acknowledgements

I would like to thank Professor Vatsal Sharan who gave me the opportunity, with the help of my mentor, Sid. In addition, Dr. Mills who kept SHINE running and gave us opportunities and Monica for helping with arrangements that were done with SHINE. Thank you to the Meta USC center, Professors. Murali and Meisam. Lastly, thanks to everyone in SHINE.