Project Writeup for IEOR4579

Tian Xie(tx2221), Yanchen Liu(yl4637), Zhiwen Huang(zh2387)

April 2022

1 Introduction

As machine learning grows rapidly, automated algorithms tend to replace humans to make decisions. We are interested in the definition of "fair algorithms" and whether it is practical to balance accuracy and fairness. Therefore, we choose a famous paper[1] on fair machine learning to replicate. This paper provides researchers with valuable understandings on trade-off between fairness and accuracy. It proposes a fairness metric derived from "disparate mistreatment" and visualizes the trade-off between fairness and accuracy on different data sets including synthesized data and real-world data. We replicate the experiments of the paper and find most of the results convincing although minor differences exist because of randomness of data generation. Moreover, we also use the code framework to do one more experiment using the credit approval data from Kaggle[2].

2 Fairness Metric: Disparate Mistreatment

Disparate mistreatment means with respect to a sensitive attribute, the model would have misclassification rates differ for groups of people having different values of that sensitive attribute (e.g., blacks and whites in races). There are also multiple ways to measure the so-called misclassification rate, among which the paper discusses false positive rate and false negative rate.

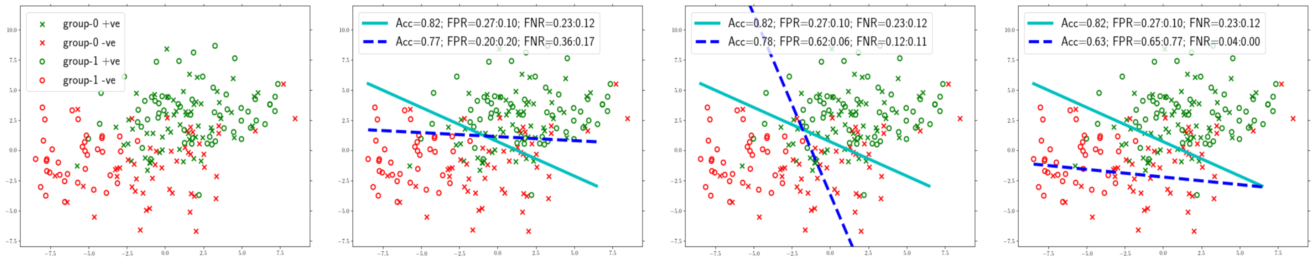
3 Replication

We replicate the same experiments on synthesized data and real-world data as the paper. The paper has its github but the codes[3] are out-dated. We rewrote the code from Python 2 to Python 3. More importantly, we integrate all experiments in one Jupyter notebook to provide better visualizations instead of the original .py files.

3.1 Synthesized Data

This part generates a dataset with a multivariate normal distribution. The data set consists of two non-sensitive features and one sensitive feature. We generate different multivariate normal distributions for non-sensitive features when sensitive features are different (i.e. different covariance matrices and means).

With this technique, we generate a dataset suffering from different disparate mistreatment effect both on FPR and FNR, which means for different values of z (the sensitive attribute), the classifier aiming to optimize accuracy would have different FPR and FNR.Then we come to the visualization of original decision boundary versus decision boundaries with fairness constraints on FPR, FNR, and FPR + FNR:



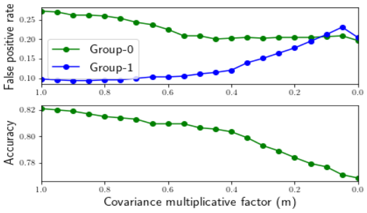
(a) Data Set Visualization (b) Constraint on FPR (c) Constraint on FNR (d) Constraint on Both

3.2 Real-world Data

We then apply the constraints on actual data: ProPublica COMPAS data set. Basically, this data set is about the crime offenders for different demographic groups. Just as the paper, we build a Logistic Regression and impose a fairness constraint on **FPR** with respect to race.
We also examine the constraint on a different data set[2]. It contains personal information and data submitted by credit card applicants. We build a Logistic Regression to predict whether the client has good credit score (paid off that month), and impose a fairness constraint on **FPR** with respect to gender. The results for the above two datasets are demonstrated in our Jupyter Notebook[4].

3.3 Accuracy-Fairness Trade-off While relaxing the Threshold

The last part of the experiment demonstrates the accuracy-fairness when we gradually apply fairness constraints more seriously on the synthetic data. The paper proposes to measure fairness using the covariance between the users sensitive attributes and the signed distance between the feature vectors of misclassified users and the classifier decision boundary (details in page 4). Thus, we can vary the threshold of this covariance. The closer this threshold to 0, the stricter the fairness constraint would be. We vary the threshold from 0 to 1 as the paper suggests and obtain similar results as the paper:



4 Conclusion

Through replicating the paper, we visualize the fairness-accuracy trade-off, and find the trend is very similar to the paper with only minor differences. All codes and detailed description can be found in our Jupyter notebook[4].

5 References

[1] Zafar, Muhammad Bilal, et al. ”Fairness beyond disparate treatment disparate impact: Learning classification without disparate mistreatment.” Proceedings of the 26th international conference on world wide web. 2017.
[2] `https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction`
[3] `https://github.com/mbilalzafar/fair-classification`
[4] `https://github.com/TianXie1999/Replicate-Paper/blob/main/fair-classification/disparate_mistreatment/Paper%20Replication.ipynb`