Foundations of Data Science
Project: Fairness in Classification on the COMPAS dataset

Nguyen Kim Thang
(previously together with R. Couillet, E. Gaussier, O. Goga)

April 2024

General project's rules: This project should be done by groups of 3 students. All groups of fewer or more students have to be allowed by the teacher. You should constitute the groups early, and start working with your group during the lab of April 3rd. The deadline to submit your groups is **Wed April 10, 19:00**. You need to send an email to kim-thang.nguyen@univ-grenoble-alpes.fr with the subject beginning with [**DS24**] to inform us about your group.

The project is graded. We ask that you prepare a Jupyter Notebook that contains all your code (such that we can execute it) as well as your analyses, answers and comments. We encourage clean and commented codes. The project's grade will be based on the content of the notebook and the answers and explanations in the notebook. The notebook (and only the notebook, that is, the `ipynb` file) must be sent to kim-thang.nguyen@univ-grenoble-alpes.fr, again with the subject beginning with [**DS24**]. The deadline is **Monday April 29, 16:59**. Late submissions will incur a penalty that may go up to 100%. Please coordinate within the groups and do only one submission per group (but make sure that the notebook contains the names of all the group members at its top).

The project's instructions are voluntarily open and non-detailed. The objective is for you to explore different meaningful techniques on the COMPAS dataset in depth. Innovation and exploration will be rewarded, but also the ability to synthesize the results.

Timeline summary

- **Wed April 10, 19:00**: deadline to submit your groups
- **Monday April 29, 16:59**: deadline to submit your project's notebooks

1 Introduction

1.1 Dataset

You will examine the ProPublica COMPAS dataset, which consists of all criminal defendants who were subject to COMPAS screening in Broward County, Florida, during 2013 and 2014.

For each defendant, various information fields (‘features’) were also gathered by ProPublica. Broadly, these fields are related to the defendant’s demographic information (e.g., gender and race), criminal history (e.g., the number of prior offenses) and administrative information about the case (e.g., the case number, arrest date, risk of recidivism predicted by the COMPAS tool). Finally, the dataset also contains information about whether the defendant did actually recidivate or not.

The COMPAS score uses answers to 137 questions to assign a risk score to defendants—essentially a probability of recidivism. The actual output is two-fold: a risk rating of 1-10 and a “low”, “medium”, or “high” risk label.

Link to dataset: `https://github.com/propublica/compas-analysis`. The file we will analyze is: `compas-scores-two-years.csv`

The initial analysis of the data by ProPublica is summarized is the following article: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Finally, you may use the notebook on the COMPAS dataset given during the lab.

1.2 Project’s goal

The project has three parts:

1. The COMPAS scores have been shown to have biases against certain racial groups. The first goal is to analyze the dataset to highlight these biases.
2. The second goal is, based on the features in the COMPAS dataset, to train classifiers to predict who will recidivate and to (i) analyze their performance and (ii) study whether they are more or less fair than the COMPAS classifier.
3. The third goal is to build a fair classifier. Is excluding the race from the feature set enough?

1.3 References

Many research articles have discuss various fairness issues directly using the COMPAS dataset, for instance [2, 4]. Other papers have discussed other methods to build fair classifiers with different techniques using other but similar datasets [3, 5]. You can also find in the book [1] a complete survey of fairness notions in classification and discussions more directly related to the COMPAS problem. You should read them and can use them as a basis for what follows.

2 Instructions and questions

2.1 Dataset exploration

(For this part, you can use the code you received as a basis.)

Load the dataset and make a basic descriptive analysis of it (how many features, entries, missing values, distribution of labels, etc.), insisting in particular on demographic features.

Compute basic performance metrics of the COMPAS classifier for different races/genders. Do you see a difference? Are there other analyses that you could do to investigate how different races/genders are treated by the COMPAS classifier?

2.2 Standard classifiers

Train a classifier on the 2-years re-arrest label ground truth. Describe its *performance* and compare it to COMPAS for different types of classifiers. Do that for multiple different types of classifiers and comment. Which method would you recommend? Which features would you use?

Now define a meaningful notion of *fairness* (you may consider several and compare them); evaluate the fairness of your classifier and compare it to the COMPAS classifier.

2.3 Fair classifiers

Describe the different methods to obtain a fair classifier, that is a classifier that satisfies a certain notion of fairness. Discuss them in the context of the COMPAS dataset.

Implement a fair classifier. Compare its performance to the one of the classifier in Section 2.2. Discuss and conclude.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning.* fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[2] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, Special issue on Social and Technical Trade-Offs*, 2017. `https://arxiv.org/abs/1703.00056`.

[3] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, page 3323–3331, 2016. `https://arxiv.org/abs/1610.02413`.

[4] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, page 1171–1180, 2017. `https://arxiv.org/abs/1610.08452`.

[5] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 325–333, 2013. `http://proceedings.mlr.press/v28/zemel13.html`.