Shreya Kochar
Independent Study Write Up
Algorithmic Fairness and the Adult Dataset

Introduction

In a world where machine learning and artificial intelligence have taken off in all aspects of life (elections, wars, social media), research on algorithmic fairness is necessary now more than ever. Two problems that have made this difficult with the popular existing datasets that are used for this research are defined as follows:

> On one hand, *opacity* is the result of poor documentation affecting single datasets, contributing to misunderstandings and misuse of specific resources. On the other hand, when relevant information exists but does not reach interested parties, there is a problem of documentation *sparsity*.[1]

The most used datasets for fairness currently are Adult, COMPAS, and German Credit[1], from ascending to descending popularity. At this moment in time, there are questions coming into play regarding the Adult dataset – specifically, if it is still a viable fairness gauge. This write-up will thus act as a survey of the current opinions of the Adult dataset, and where the world of machine learning and algorithms will go from here.

Background: the Adult Dataset

This dataset is a sample from the US Census Database that contains the census result of year 1994. It contains 48842 (not NaN) entries, and each entry contains the following features about a representative individual in the census record:

- **age**: (continuous, positive integer) The age of the individual.
- **workclass**: (categorical, 9 distinct values) Simplified employment status of an individual
- **fnlwgt**: (continuous, positive integer) Final weight of the record. Basically interpret[ed] as the number of people represented by this row.
- **education-num**: (categorical, 13 distinct values) The education level, in ascending positive integer value.
- **education**: (categorical, 13 distinct values) The education level. [...]
- **marital-status**: (categorical, 7 distinct values) Marital status of a person.
- **occupation**: (categorical, 15 distinct values) Rough category of the occupation.
- **relationship**: (categorical, 6 distinct values) Relationship in terms of the family. Note that we ignore this column since the semantic is somewhat covered by marital-status and gender.
- **race**: (categorical, 5 distinct values) Race of the person.
- **gender**: (boolean) gender at-birth.
- **capital-gain**: (continuous) Dollar gain of capital.
- **capital-loss**: (continuous) Dollar loss of capital.
- **hours-per-week**: (continuous positive integer) Working hours per week.
- **native-country**: (categorical, 41 distinct values) Country at birth.
- **income-bracket**: (boolean) True if ≥ 50K, otherwise False (< 50K per year).[2]

Because of the nature of this dataset, it was designed to represent the entire country's population fairly. This can be seen especially by the fnlwgt category, which allows the actual weightage of each row to be applied in terms of how many people in the country it represents proportionally to the population. In other words, if a row represents a group of people, they will be counted in that row's fnlwgt value. Yet, it should be noted that, in fact, this data is from 1994, which means that it is close to three decades old. Thus, this data is not representative of today's US population, which poses a problem in terms of the accuracy of the analysis it produces.

History of Using the Adult Dataset

As per *Algorithmic Fairness Datasets: the Story so Far* [Fabris, A., Messina, S., Silvello, G., & Susto, G. A], the Adult dataset has been used for the following fairness tasks:

1. Fair classification - equalizing some measure of interest across subpopulations, such as the recall, precision, or accuracy for different racial groups.
2. Fair clustering - an unsupervised task concerned with the division of a sample into homogenous groups.
3. Fair preference-based classification - given the choice between various sets of decision treatments or outcomes, any group of users would collectively prefer its treatment or outcomes, regardless of the (dis)parity as compared to the other groups[3]

Similarly, as per the same paper, the Adult dataset has been used for testing different settings, or challenges, for tasks, such as:

1. Rich-subgroup fairness - a setting where fairness properties are required to hold not only for a limited number of protected groups, but across an exponentially large number of subpopulations.
2. Noisy fairness - a general expression adopted to indicate problems where sensitive attributes are missing (Chen et al., 2019a), encrypted (Kilbertus et al., 2018) or corrupted by noise (Lamy et al., 2019).
3. Limited-label fairness - comprises settings with limited information on the target variable, including situations where labeled instances are few (Ji et al., 2020), noisy (Wang et al., 2021), or only available in aggregate form (Sabato and Yom-Tov, 2020).
4. Robust fairness problems - connected with work in robust machine learning, extending the stability requirements beyond accuracy-related metrics to fairness-related ones.
5. Preference-based fairness - (Zafar et al., 2017b) denotes work informed by the preferences of stakeholders.
6. Multi-stage fairness - (Madras et al., 2018b) refers to settings where several decision makers coexist in a compound decision-making process.
7. Fair few-shot learning - (Zhao et al., 2020b) aims at developing fair ML solutions in the presence of a small amount of data samples.

Analyzing the Adult Dataset Quality

Because of the age of the dataset and several other emerging concerns with Adult, those concerned with fairness and algorithms are considering retiring the Adult dataset altogether and moving onwards. For instance, another concern with the Adult dataset that has come out is as follows:

While some issues with UCI Adult are readily apparent, such as its age, limited documentation, and outdated feature encodings, a significant problem may be less obvious at first glance. Specifically, UCI Adult has a binary target label indicating whether the income of a person is greater or less than fifty thousand US dollars. This income threshold of $50k US dollars corresponds to the 76th quantile of individual income in the United States in 1994, the 88th quantile in the Black population, and the 89th quantile among women. We show how empirical findings relating to algorithmic fairness are sensitive to the choice of the income threshold, and how UCI Adult exposes a rather extreme threshold. Specifically, the magnitude of violations in different fairness criteria, trade-offs between them, and the effectiveness of algorithmic interventions all vary significantly with the income threshold. In many cases, the $50k threshold understates and misrepresents the broader picture.[3]

Works Cited

1. Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022, February 3). *Algorithmic Fairness Datasets: the Story so Far*. Retrieved April 27, 2022, from https://arxiv.org/pdf/2202.01711.pdf
2. Chen, J. (n.d.). *Feature Significance Analysis of the US Adult Income Dataset*. Retrieved April 29, 2022, from https://minds.wisconsin.edu/bitstream/handle/1793/82299/TR1869%20Junda%20Chen%203.pdf?sequence=1&isAllowed=y
3. Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P., & Weller, A. (2017, November 28). *From parity to preference-based notions of fairness in classification*. arXiv.org. Retrieved May 1, 2022, from https://arxiv.org/abs/1707.00010