INFO-H420 Management of Data Science and Business Workflows

Assignment 4: Fairness with IBM AIF 360 (15 points)

The goal of this assignment is to study algorithmic fairness concepts using the AIF 360 tool. In this assignment, we will be using the COMPAS dataset.

Exercise 1 *(10 points)*

In group-based definitions of algorithmic fairness, we define protected groups based on values on a protected attribute, like race, sex, and then measure the discrepancy of some metric among the protected groups in some observed outcomes. For example, we might compute the difference of the positive rate between males and females.

In intersectional fairness, we are interested at what happens among groups that are defined based on intersections of attributes. For example, we might study what is the positive rate difference between males and females for those aged less than 25. Or, what is the positive rate difference between the four groups defined by race (African-American and Caucasian) and sex (males and females).

In the first exercise:

- Consider *race* to be the protected attribute, fix the bias using the reweighing preprocessing technique, and measure the bias assuming *sex* is the protected attribute.
- Consider *sex* to be the protected attribute, fix the bias using the reweighing preprocessing technique, and measure the bias assuming *race* is the protected attribute.
- Repeat these measurements considering age groups, to investigate questions like: is there unfairness with respect to either sex or race between those aged less than 25?

In all cases, you should train a simple logistic regression classifier, and measure bias on a test set. Document and present your findings in a report.

Exercise 2 *(5 points)*

Consider the Multi-Dimensional Subset Scan (MDSS) method from [1] that is implemented in AIF 360 and showcased in the "demo_mdss_classifier_metric.ipynb" example notebook. The MDSS method is able to detect unfairness instances in subpopulations.

In the second exercise:

- Examine the privileged and unprivileged groups that MDSS identifies. For each of them, measure its bias and compare it to a group that has the opposite race or sex. For example, it a group is defined as "age less than 25" and "race is Caucasian", you should compare it to the group "age less than 25" and "race is African American".

Document and present your findings in a report, where you also summarize how the MDSS method works.

[1] Zhe Zhang, Daniel B. Neill. “Identifying Significant Predictive Bias in Classifiers”. In FAT/ML 2017. https://arxiv.org/abs/1611.08292

Instructions

The assignment contributes 15% to the overall grade.

This assignment is to be made in **groups** of two persons. You are asked to form the groups via “Groups for Assignment 4” on the Université Virtuelle (UV). If you cannot find a partner, please post a request in the “Discussion Forum” on UV.

You are asked to submit a short **report** presenting your solution to the exercises, including justifications for the choices and assumptions made.

The report and any supporting files has to be uploaded to “Assignment 4” on UV before January 12, 2024.

Assignment 4 - Instructions

For Assignment 4 you need to have installed the AI Fairness 360 (AIF360) library from IBM, and have downloaded the COMPAS dataset.

Install AIF360

To install AIF360, you may follow the instructions at its github page:
https://github.com/Trusted-AI/AIF360

Alternatively, you can choose to use a conda environment, and execute these commands in your terminal:

```
conda create --name aif360 python=3.11
conda activate aif360
pip install aif360
```

Download COMPAS dataset

To download the COMPAS dataset you should follow the instructions given by AIF360 when you try to use the dataset for the first time. For example execute this Python code and follow the instructions in the error message:

```
from aif360.algorithms.preprocessing.optim_preproc_helpers.data_preproc_functions import
load_preproc_data_compas
dataset = load_preproc_data_compas()
```