ECOLE POLYTECHNIQUE DE BRUXELLES

2023-2024

INFO-H420 Management of Data Science and Business Workflows

Assignment 4: Fairness with IBM AIF 360

Iyoha Peace Osamuyi (000583313)

SACHARIDIS Dimitrios

Hieu Nguyen Minh (000583782)

2023

Contents

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.

Before doing this assignment, we set-up the environment and libraries on Ubuntu 22.04 as follows. Firstly, we install AIF360 library with Miniconda by running:

```
conda create --name aif360 python=3.9
conda activate aif360
conda install -c conda-forge aif360
```

We download the COMPAS dataset (the file `compas-scores-two-years.csv`) from this Github repository. Then we put this file in the data folder of Conda environmen. Our folder is `miniconda3/envs/aif3`

1 Exercise 1

1.1 Question

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.

1.2 Solution

1.2.1 Task 1

The COMPAS dataset is widely used in criminal justice to assess the likelihood of a defendant becoming a recidivist. It contains features which include:

- `sex`: 0.0 and 1.0 indicate male and female, respectively.
- `race`: 0.0 and 1.0 indicate African-American and Caucasian, respectively.
- Age group features:
  - `age_cat=25 to 45`:
  - `age_cat=Greater than 45`
  - `age_cat=Less than 25`
- Number of prior criminal records of defendants
  - `priors_count=0`
  - `priors_count=1 to 3`
  - `priors_count=More than 3`
- The charge degree of defendants
  - `c_charge_degree=F`: Felony
  - `c_charge_degree=M`: Misdemeanor

The label for this dataset is `two_year_recid`, which is a binary variable indicating whether defendant is rearrested at within two years.

The dataset was cleaned and one-hot encoded categorical variables were transformed into numerical format. It implemented the StandardDataset class from the AIF 360 library to further preprocess the dataset. The dataset was splited into training and test dataset
A snippet of the Data Preprocessing is given below:

```
# Convert one-hot encoded features back to categorical features
age_cat = np.argmax(dataset_orig_df[['age_cat=Less than 25', 'age_cat=25 to 45', 'age_cat=Greater than 45']].values, axis=1).reshape(-1, 1)
priors_count = np.argmax(dataset_orig_df[['priors_count=0', 'priors_count=1 to 3','priors_count=More than 3']].values, axis=1).reshape(-1, 1)
c_charge_degree = np.argmax(dataset_orig_df[['c_charge_degree=M', 'c_charge_degree=F']].values, axis=1).reshape(-1, 1)

features = np.concatenate((dataset_orig_df[['sex', 'race']].values, age_cat, priors_count, c_charge_degree, dataset_orig.labels), axis=1)
feature_names = ['sex', 'race', 'age_cat', 'priors_count', 'c_charge_degree']
df = pd.DataFrame(features, columns=feature_names + ['two_year_recid'])

# Convert the DataFrame into AIF360's StandardDataset
dataset = StandardDataset(df, label_name='two_year_recid', favorable_classes=[0], protected_attribute_names=['sex', 'race'], privileged_classes=[[1], [1]], instance_weights_name=None)

# Splitting the dataset
dataset_orig_train, dataset_orig_test = dataset.split([0.7], shuffle=True, seed=0)
```

Listing 1.1: Data Preprocessing

Protected Attributes Definition:

- Race was divided into two groups: 'Caucasian' (privileged) and 'Non-Caucasian' (unprivileged).
- Sex was divided into 'Male' (privileged) and 'Female' (unprivileged) for the purpose of bias measurement.

```
female_group = [{'sex': 1}] # '1' represent female
male_group = [{'sex': 0}]   # '0' represent male

# Define race-based groups for reweighing
race_privileged_group = [{'race': 1}]  #  '1' represents Caucasian
race_unprivileged_group = [{'race': 0}]  #  '0' represents Non-Caucasian
```

Listing 1.2: Protected Attributes

Bias Fixing Technique: We employed the reweighing algorithm from AIF360 to adjust the weights of the training instances, aiming to balance the representation of the racial groups.

```
RW = Reweighing(unprivileged_groups=race_unprivileged_group,
    privileged_groups=race_privileged_group)
dataset_orig_train_rw = RW.fit_transform(dataset_orig_train)
```

Listing 1.3: Bias Fixing

Initial Bias Measurement: BinaryLabelDatasetMetric was used to calculate the mean difference in outcomes between the male and female groups for both the training and testing sets before reweighing and using the logistic regression model. This metric is called statistical parity difference and is computed as follows:

$$\text{spd} = Pr(\hat{Y} = 1|D = \text{unprivileged}) - Pr(\hat{Y} = 1|D = \text{privileged}), \tag{1.1}$$

where $\hat{Y}$ is the predicted value and D is the group considered. We calculate this metric with the following code:

```
#Initial Bias

metric_train_sex = BinaryLabelDatasetMetric(dataset_orig_train,
    unprivileged_groups=male_group, privileged_groups=female_group)
print("Train set Sex: Difference in mean outcomes between unprivileged and
    privileged groups = %f" % metric_train_sex.mean_difference())

metric_test_sex = BinaryLabelDatasetMetric(dataset_orig_test,
    unprivileged_groups=male_group, privileged_groups=female_group)
print("Test set Sex: Difference in mean outcomes between unprivileged and
    privileged groups = %f" % metric_test_sex.mean_difference())


metric_train_race = BinaryLabelDatasetMetric(dataset_orig_train,
    unprivileged_groups=race_unprivileged_group, privileged_groups=
    race_privileged_group)
print("Train set Race: Difference in mean outcomes between unprivileged and
    privileged groups = %f" % metric_train_race.mean_difference())

metric_test_race= BinaryLabelDatasetMetric(dataset_orig_test,
    unprivileged_groups=race_unprivileged_group, privileged_groups=
    race_privileged_group)
```

```
14 print("Test set Race: Difference in mean outcomes between unprivileged and
       privileged groups = %f" % metric_test_race.mean_difference())
```

Listing 1.4: Initial Bias

Model Training: A logistic regression model was trained on reweighed dataset with race protected attribute to predict the likelihood of recidivism. Key parameters included a regularization strength of 1.0 and a random state for reproducibility.

```
1 clf_reweighed = LogisticRegression(solver='lbfgs', C=1.0, penalty='l2',
       random_state=0)
2 clf_reweighed.fit(dataset_orig_train_rw.features, dataset_orig_train_rw.
       labels.flatten())
```

Listing 1.5: Model Training

Bias Measurement: We first predicted the test dataset using logistic regression and then measured bias with respect to sex using BinaryLabelDatasetMetric in AIF360, focusing on the mean difference in predicted outcomes.

```
1 dataset_bias_test_rw = dataset_orig_test.copy()
2 dataset_bias_test_rw.labels = clf_reweighed.predict(dataset_orig_test.
       features)
3 metric_test_sex_rw = BinaryLabelDatasetMetric(dataset_bias_test_rw,
       unprivileged_groups=female_group, privileged_groups=male_group)
4 print("Bias metric for Sex (Male: Privilege) on Test set (after reweighing
       based on race):", metric_test_sex_rw.mean_difference())
5
6 metric_test_sex_rw1 = BinaryLabelDatasetMetric(dataset_bias_test_rw,
       unprivileged_groups=male_group, privileged_groups=female_group)
7 print("Bias metric for Sex (Male: UnPrivilege) on Test set (after reweighing
        based on race):", metric_test_sex_rw1.mean_difference())
```

Listing 1.6: Bias Measurement

Results and Discussion

Bias Metrics Before Reweighing:

Sex Bias (mean difference): **-0.159410** for Females as privileged. The data indicates that females in the dataset exhibit a lower rate of recidivism compared to males.

Bias Metrics After Reweighing:

After reweighing based on the race attribute and employing Logistic Regression, there was a shift in the Sex Bias metric (mean difference). This metric now stands at +0.24310595065312046 when males are viewed as the privileged group and -0.24310595065312046 when considering males as the unprivileged group.

This change signifies the average difference in outcomes between males and females following the dataset reweighing. A positive value here indicates that, generally, the model's predictions tend to favor males over females in assessing recidivism risk. The size of the value, 0.24310595065312046 in this instance, denotes the level of bias present. The nearer this number is to zero, the lesser the bias. Therefore, a value of 0.24310595065312046 suggests **a moderate bias in favor of males**.

The negative sign indicates that when considering males as unprivileged, the model's predictions are less favorable towards them compared to females.

1.2.2 Task 2

Same approach as task 1, only few notable changes were made:

Protected Attributes Definition:

- Sex was divided into 'Male' (privileged) and 'Female' (unprivileged) for the purpose of bias measurement.
- Race was divided into two groups: 'Caucasian' (privileged) and 'Non-Caucasian' (unprivileged).

Bias Fixing Technique: We employed the reweighing algorithm from AIF360 to adjust the weights of the training instances, aiming to balance the representation of the **sex groups.**

```
RW_sex = Reweighing(unprivileged_groups=male_group, privileged_groups=
    female_group)
dataset_orig_train_rw_sex = RW_sex.fit_transform(dataset_orig_train)
```

Listing 1.7: Bias Fixing

Model Training: A logistic regression model was trained on reweighed dataset for sex protected attribute to predict the likelihood of recidivism. Key parameters included a regularization strength of 1.0 and a random state for reproducibility.

```
clf_reweighted = LogisticRegression(solver='lbfgs', C=1.0, penalty='l2',
    random_state=0)
clf_reweighted.fit(dataset_orig_train_rw_sex.features,
    dataset_orig_train_rw_sex.labels.ravel())
```

Listing 1.8: Model Training

Bias Measurement: We first predicted the test dataset using logistic regression and then measured bias with respect to sex using BinaryLabelDatasetMetric in AIF360, focusing on the mean difference in predicted outcomes.

Results and Discussion

Bias Metrics Before Reweighing:

Sex Bias (mean difference): **-0.159410** for Females as privileged. The data indicates that females in the dataset exhibit a lower rate of recidivism compared to males.

Bias Metrics After Reweighing:

Bias metric for Race (Caucasian: privileged) on Test set: **-0.30360576923076926**. This metric measures the mean difference in outcomes between the Caucasian group (considered privileged) and the non-Caucasian group (considered unprivileged) after reweighing the dataset to address bias in terms of sex. A negative value indicates that the model's predictions are, on average, less favorable towards the Caucasian group compared to the non-Caucasian group. This suggests that, despite Caucasians being deemed the privileged group, the model exhibits a bias against them following the sex-based reweighing. The magnitude of this value (-0.30360576923076926) is relatively significant, suggesting a noticeable level of bias against Caucasians in the model's predictions after sex-based reweighing.

1.2.3 Task 3

Data Filtering: The dataset was filtered to focus on individuals aged less than 25. This subgroup was extracted to specifically analyze bias in a younger demographic. And it was splitted to both train and test dataset

```
# Filter the dataset for individuals aged less than 25
age_less_than_25_index = dataset_orig_df['age_cat=Less than 25'] == 1
dataset_age_less_than_25 = dataset.subset(age_less_than_25_index)

# Split the filtered dataset into training and testing sets
dataset_train_age_less_than_25, dataset_test_age_less_than_25 =
    dataset_age_less_than_25.split([0.7], shuffle=True, seed=0)
```

Listing 1.9: Data Filtering

Protected Attributes and Groups:

- Race: Defined as 'Caucasian' (privileged) and 'Non-Caucasian' (unprivileged).
- Sex: Defined as 'Male' (privileged) and 'Female' (unprivileged).

Model Training: A logistic regression model was trained on the full dataset to predict recidivism. Parameters included a standard logistic loss function and a regularization parameter of 1.0.

```
model = LogisticRegression(solver='lbfgs', C=1.0, penalty='l2', random_state
    =0)
model.fit(dataset_train_rw.features, dataset_train_rw.labels.ravel(),
    sample_weight=dataset_train_rw.instance_weights)
return model
```

Listing 1.10: Data Filtering

Bias Mitigation Technique: We employed the Reweighing algorithm from the AIF360 toolkit, a preprocessing technique that adjusts instance weights in the training set to balance representation across racial groups.

```
RW = Reweighing(unprivileged_groups=unprivileged_groups, privileged_groups=
    privileged_groups)
dataset_train_rw = RW.fit_transform(dataset_train)
```

Listing 1.11: Data Filtering

Bias Measurement: We first predicted the test dataset using logistic regression and then measured bias with respect to sex and race using BinaryLabelDatasetMetric in AIF360, focusing on the mean difference in predicted outcomes.

```
ataset_test_bias_sex = dataset_test_age_less_than_25.copy()
dataset_test_bias_sex.labels = model_reweighed_sex.predict(
    dataset_test_age_less_than_25.features)
metric_sex_bias_race = BinaryLabelDatasetMetric(dataset_test_bias_sex,
    unprivileged_groups=race_unprivileged_group, privileged_groups=
    race_privileged_group)

metric_sex_bias_race1 = BinaryLabelDatasetMetric(dataset_test_bias_sex,
    unprivileged_groups=race_privileged_group, privileged_groups=
    race_unprivileged_group)

# Measure bias with respect to sex (after reweighing based on race)
dataset_test_bias_race = dataset_test_age_less_than_25.copy()
dataset_test_bias_race.labels = model_reweighed_race.predict(
    dataset_test_age_less_than_25.features)
metric_race_bias_sex = BinaryLabelDatasetMetric(dataset_test_bias_race,
    unprivileged_groups=male_group, privileged_groups=female_group)

metric_race_bias_sex1 = BinaryLabelDatasetMetric(dataset_test_bias_race,
    unprivileged_groups=female_group,privileged_groups=male_group)
```

Listing 1.12: Data Filtering

Results and Discussion

Race Bias Metrics Before Reweighing under age 25:

Sex Bias (mean difference): **-0.150392** for Females as privileged. The data indicates that females in the dataset exhibit a lower rate of recidivism compared to males.

Race Bias Metrics After Reweighing:

Bias metric for Race (Caucasian: privileged) on Test set under age 25: **-0.07191752577319588**. This metric measures the mean difference in outcomes between Caucasian individuals (privileged) and non-Caucasian individuals (unprivileged) in the under-25 age group, after addressing bias based on sex. A negative value (-0.07191752577319588) indicates that, on average, the model's

predictions are less favorable for Caucasians compared to non-Caucasians in this age group. The bias is against the group presumed to be privileged (Caucasians). The value -0.07191752577319588 is modest, suggesting a relatively small level of bias against Caucasians

Sex Bias Metrics Before Reweighing under age 25:
Sex Bias (mean difference): **-0.277972** for Females as privileged. The data indicates that females in the dataset exhibit a lower rate of recidivism compared to males.

Sex Bias Metrics After Reweighing:

Bias metric for Sex (Caucasian: privileged) on Test set under age 25: **-0.21497850214978503**. This metric assesses the mean difference in results between male (viewed as privileged) and female (viewed as unprivileged) individuals under the age of 25, after making adjustments for bias related to race.

The negative value (-0.21497850214978503) indicates that the model's forecasts are generally less favorable for males compared to females in this particular age group, signifying a bias against males, who are conventionally seen as the privileged group.

Moreover, the value's magnitude, at -0.21497850214978503, in comparison to the race bias metric, denotes a more significant bias against males in this particular context.

2 Exercise 2

2.1 Question

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

2.2 Solution

The MDSS approach is tailored to uncover biases within particular subgroups, offering an in-depth perspective on algorithmic fairness by comparing model predictions to established fairness benchmarks for different combinations of attributes. This method differs from conventional approaches that evaluate bias over the entire dataset, as MDSS zeroes in on specific subpopulations for a more detailed fairness analysis.

How MDSS Works: MDSS methodically explores various attribute combinations, such as age, race, and sex, to pinpoint specific subgroups that exhibit notable prediction discrepancies. It assesses

the degree to which the model's predictions for these identified subgroups vary from the expected outcomes under fair prediction conditions.

2.2.1 Methodology

Data Preparation

The dataset was prepared to align with AI Fairness 360 standards, focusing on race, sex, and age attributes.

```
from aif360.datasets import StandardDataset
dataset = StandardDataset(df, label_name='two_year_recid', favorable_classes
    =[0], protected_attribute_names=['sex', 'race','age_cat'],
    privileged_classes=[[1], [1], [0]], instance_weights_name=None)
```

Listing 2.1: Data Preparation

MDSS Analysis

Using the MDSSClassificationMetric, we conducted an in-depth analysis to identify subgroups with significant disparities in model predictions.

```
mdss_classified = MDSSClassificationMetric(dataset_orig_test,
    dataset_bias_test, unprivileged_groups=female_group, privileged_groups=
    male_group)

male_privileged_score = mdss_classified.score_groups(privileged=True)
male_privileged_score

female_unprivileged_score = mdss_classified.score_groups(privileged=False)
female_unprivileged_score
```

Listing 2.2: MDSS Analysis

Groups Identified

- Sex group

- Race group
- Age less than 25 group with different sex
- Age less than 25 group with different race
- Race with different sex

2.2.2 Results and Discussion

Sex-Based Analysis

Male Privileged, Female Unprivileged: The analysis indicated a moderate bias in favor of males (privileged score: 0.63) and a substantial bias against females (unprivileged score: 1.1769).

Female Privileged, Male Unprivileged: In this configuration, no significant bias was detected in favor of females or against males (scores of -0.0).

Race-Based Analysis

Caucasian vs. Non-Caucasian Groups: The initial race-based analysis showed no significant bias, with privilege scores of -0.0 for both groups.

Non-Caucasian Privileged, Caucasian Unprivileged: In this reversed scenario, the MDSS analysis revealed a privilege score of 0.1636 for non-Caucasians (privileged) and 0.0138 for Caucasians (unprivileged). This indicates a modest bias in favor of non-Caucasians when they are considered the privileged group.

Age and Sex Intersectional Analysis

Sex Bias in the Under-25 Group: In this demographic, young females showed pronounced biases against them (unprivileged score: 3.1218), while young males benefited (privileged score: 3.0914). The inverse gender roles showed no significant bias.

Age and Race Intersectional Analysis

Under 25, Race-Based: Significant biases were detected, particularly against non-Caucasians under 25 (privilege score of 0.5618), and in favor of Caucasians under 25 (privilege score of 2.1324).

Race and Sex Intersection Analysis

Non-Caucasian Males vs. Caucasian Males **Non-Caucasian Males Privilege Score:** The MDSS analysis for non-Caucasian males (unprivileged) versus Caucasian males (privileged) yielded a score of -0.0, indicating no significant bias in either direction. **Caucasian Males Privilege Score:** Similarly, the privilege score for Caucasian males was -0.0, further suggesting no detectable bias within this subgroup configuration.

Non-Caucasian Female vs. Caucasian Females **Non-Caucasian Females Privilege Score:** The MDSS analysis for non-Caucasian females (unprivileged) versus Caucasian females (privileged) yielded a score of -0.0, indicating no significant bias in either direction. **Caucasian Females Privilege Score:** Similarly, the privilege score for Caucasian females was -0.0, further suggesting no detectable bias within this subgroup configuration.