January 10, 2023

Seminar Quality Assurance for Machine Learning
FG UNIML, TU Berlin

Contact person: Yasemin Bozkurt V.[1]

Homework 3: Fairness

Two datasets are provided to you. *Dataset law* contains records of law school admissions. The classification task involves predicting whether a candidate will pass the bar exam. The race of the students will be examined in order to determine if there has been any discrimination. The *Dataset credit* contains samples of bank account holders, and is used for risk assessment prediction, i.e., determining whether or not to grant credit to a particular individual. We will investigate if there is a discrimination on the age of customer. Tables 2 and 3 provide more information about the datasets.

Dataset ID	Protected attribute	Class attribute
Law	Race: {non-white, white}	Pass the bar exam $\in \{0, 1\}$: Class 0 (1) is fail (pass).
Credit	Age: $\{\leq 25, > 25\}$	Class label: Class 0 (1) is low (high) risk hence good (bad) customer

Table 1: Protected attributes and targets

Class attributes are denoted by $y \in \{y_0, y_1\}$. Let G be a binary protected attribute with $G \in \{g_p, g_{np}\}$, in which g_p is the protected (discriminated) group, and g_{np} is the non-protected (non-discriminated) group. Protected groups are non-white and younger people in *Dataset law* and *Dataset credit*, respectively.

Fairness metrics

Fairness can be measured in a few different ways. However, there is no single fairness measure that is suitable for all situations. This homework examines three commonly used definitions: statistical parity, equalized odds, and absolute between receiver operating characteristics area (ABROCA).

1. **Statistical parity (SP)** (also called as demographic parity and acceptance rate parity) condition is satisfied if the difference in predicted outcome ($\hat{y}$) between non-protected and protected groups under study (i.e., g_{np} and g_p) is up to a predefined threshold ϵ:

$$SP : P(\hat{y} = y_1 | G = g_{np}) - P(\hat{y} = y_1 | G = g_p) \leq \epsilon \tag{1}$$

   Violation of SP can be measured as the difference in probability of being assigned to the positive predicted class:

$$SP_{viol} = P(\hat{y} = y_1 | G = g_{np}) - P(\hat{y} = y_1 | G = g_p) \tag{2}$$

2. The classifier meets **Equalized Odds (EO)** condition if the TPR and FPR of the protected and non-protected groups are equal, satisfied by the following formula:

$$EO : P(\hat{y} = y_1 | G = g_{np}, Y = y) = P(\hat{y} = y_1 | G = g_p, Y = y) \tag{3}$$

   Violation of EO can be measured as:

$$EO_{viol} = \sum_{y \in \{y_0, y_1\}} |P(\hat{y} = y_1 | G = g_{np}, Y = y) - P(\hat{y} = y_1 | G = g_p, Y = y)|. \tag{4}$$

[1] yasemin.bozkurt.varolguenes@campus.tu-berlin.de

3. **ABROCA** measures the divergence between the non-protected ($ROC_{g_{np}}$) andprotected (ROC_{g_p}) curves across all possible thresholds $t \in [0, 1]$ of TPR and FPR.

$$\int_{t=0}^{1} |ROC_{g_{np}}(t) - ROC_{g_p}(t)| \, dt \tag{5}$$

Tasks

- Train 2 different classifiers of your choice (e.g., logistic regression) and obtain confusion matrices. (You need to choose a categorical data encoding method.) Based on the confusion matrix and the definitions given below

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6a}$$

$$\text{TPR on protected group} = \frac{TP_p}{TP_p + FN_p} \tag{6b}$$

$$\text{TPR on non-protected group} = \frac{TP_{np}}{TP_{np} + FN_{np}} \tag{6c}$$

$$\text{TNR on protected group} = \frac{TN_p}{TN_p + FP_p} \tag{6d}$$

$$\text{TNR on non-protected group} = \frac{TN_{np}}{TN_{np} + FP_{np}} \tag{6e}$$

  calculate: Accuracy, TPR on protected group, TPR on non-protected group, TNR on protected group, TNR on non-protected group, Statistical parity, Equalized odds, ABROCA values [also plot the abroca slice using the utility function given].

- Using these calculated metrics, assess discrimination based on protected attribute within each dataset.
- Which modifications would you make to your classifiers based on the fairness analysis so that they become *fair* in terms of the protected attribute?
- Comment on the following:
  (i) What is the range of values that each fairness metric can take?
  (ii) In terms of the protected attribute, what is the overall balance of the datasets?
  (iii) What might be the advantages and disadvantages of each of the fairness metrics?
  (iv) What is the level of consistency between the fairness metrics?
  (v) Is there a significant difference between the datasets in terms of predictive performance and fairness measures?

Please submit a standalone, well-documented, Jupyter notebook with your solution methodology, answers and comments.

Column name	Type	Value	Description
decile1b	Numerical	[1-10]	The student's decile in the school given his grades in Year 1
decile3	Numerical	[1-10]	The student's decile in the school given his grafes in Year 3
lsat	Numerical	[11-48]	The student's LSAT score
ugpa	Numerical	[1.5-4]	The student's undergraduate GPA
zfygpa	Numerical	[-3.35-3.48]	The first year law school GPA
fulltime	Binary	{1, 2}	Whether the student will work full-time or part-time
fam_inc	Categorical	{1, 2, 3, 4, 5}	The student's family income bracket
male	Binary	{0, 1}	Whether the student is a male or female
tier	Categorical	{1, 2, 3, 4, 5, 6}	Tier
race	Categorical	{White, Non-White}	Race
pass_bar	Binary	{0,1}	Whether the student passed the bar exam on the first try

Table 2: Law dataset information

Column name	Type	Value	Description
checking-account	Categorical	4 values	The status of existing checking account
duration	Numerical	[4–72]	The duration of the credit (month)
credit-history	Categorical	5 values	The credit history
purpose	Categorical	10 values	Purpose (car, furniture, education, etc.)
credit-amount	Numerical	[250–18,424]	Credit amount
savings-account	Categorical	5 values	Savings account/bonds
employment-since	Categorical	5 values	Present employment since
installment-rate	Numerical	[1–4]	The installment rate in percentage of disposable income
other-debtors	Categorical	3 values	Other debtors/guarantors
residence-since	Numerical	[1–4]	Present residence-since
property	Categorical	4 values	Property
age	Numerical	[19–75]	The age of the individual
other-installment	Categorical	3 values	Other installment plans
housing	Categorical	3 values	Housing (rent, own, for free)
existing-credits	Numerical	[1–4]	Number of existing credits at this bank
job	Categorical	4 values	Job (unemployed, (un)skilled, management)
number-maintenance	Numerical	[1–2]	Number of people being liable to provide maintenance for
telephone	Binary	{yes, none}	Telephone number
foreign-worker	Binary	{yes, no}	Is the individual a foreign worker?
sex	Categorical	{female, male}	Sex of individual
marital-status	Categorical	2 values	Marital status of an individual
class-label	Binary	{0, 1}	Class

Table 3: Credit dataset information