Università degli studi di
Milano-Bicocca

Advanced Machine Learning

Final Project

Fair Classification with Adversarial Debiasing

Authors:
Raffaele Anselmo - 846842- r.anselmo@campus.unimib.it
Lorenzo Pastore - 847212- l.pastore6@campus.unimib.it

May 3, 2020

Abstract

In this report a binary classification problem on income prediction is developed and analyzed in terms of classification and fairness metrics. A fair classifier based on Adversarial Debiasing is proposed, along with a Hyperparameters Optimization (HPO).

1 Introduction

The widespread use of algorithmic decision processes in sensible domains like credit ratings, justice or housing allocations have raised many questions about their transparency, accountability and fairness. Although these complex learning methods are often treated like black boxes, they should be designed in order to ensure minimum discrepancy between the outcomes related to people that share particular sensible attributes (e.g. age, sex or race) and the others.

Existing notions of fairness in the machine learning literature are largely inspired by the concept of discrimination in social sciences and law. These notions call for parity (i.e. equality) in treatment, in impact, or both [1]. A decision making process suffers from disparate treatment if its decisions are (partly) based on the subject's sensitive attribute information, and it has disparate impact if its outcomes disproportionately hurt or benefit people with certain sensitive attribute values [2].

Dealing simultaneously with both forms of unfairness is not-trivial: the exclusion of sensible attributes from the learning phase could avoid disparate treatment. However if these attributes are strongly correlated with other features, the outcomes would still depend on the sensitive attributes, and this may lead to a disparate impact. In addition, since automated decision-making systems are trained on historical data, if a group with a certain sensitive attribute was unfairly treated in the past, this unfairness may persist in future predictions through *indirect discrimination* [Pedeschi et al., 2008]. On the other hand, using sensitive attributes to avoid disparate impact would constitute disparate treatment and may also lead to *reverse discrimination* [Ricci vs DeStefano, 2009].

These definitions of fairness can relate to group of people (*group fairness*) as well as to single persons (*individual fairness*). Group fairness does not consider the individual merits and may result in choosing the less qualified members of a group, whereas individual fairness assumes a similarity metric

of the individuals for the classification task at hand that is generally hard to find [3].

In this work, the binary classification problem of income prediction has been addressed with respect to group fairness through two sensible attributes: *race* and *sex*. First, a baseline neural network classifier was developed and evaluated in terms of classification and fairness metrics. The quantification of disparate impact is an open debate. Actually there exist many numerical formula that measure the level of discrimination between sensible and protected group (or persons). As supported by the U.S. Equal Employment Opportunity Commission, the *"80%-rule"* (or more generally, the *p%-rule*) [Biddle, 2005] could be used as a good proxy for disparate impact. In particular, "the *p%-rule* states that the ratio between the percentage of subjects having a certain sensitive attribute value assigned the positive decision outcome and the percentage of subjects not having that value also assigned the positive outcome should be no less than p:100" [2]. In formal terms:

$$min\Big(\frac{P(\hat{y}=1|z=1}{P(\hat{y}=1|z=0}, \frac{P(\hat{y}=1|z=0)}{P(\hat{y}=1|z=1}\Big) \geq \frac{p}{100} \tag{1}$$

Once tested and quantified the disparate treatment between protected and unprotected groups, a "fair" classifier based on Adversarial Debiasing was developed with the aim to maximize the p-rule at the minimum possible accuracy waste. This model was optimized with a Sequential Model Based Optimization (SMBO) approach. Lastly, the performances on classification and fairness metrics were compared between the baseline classifier and the fair classifier.

2 Datasets

The experiments were conducted on the Adult Census Income Dataset [4] that contains 48842 records with 14 attributes regarding social, demographic and financial aspects of U.S. citizens. The class variable assumes value 1 when the person makes over 50K a year, 0 otherwise.

The records with *race* different than *black* or *white* were discarded from the analysis.

Of the remaining 30940 records, only 2190 (0.071%) contains missing values. Given the rarity of missing data case, they were safely removed.

The final dataframe consists 28750 samples with no missing data, divided in 3 datasets:

- X: 12 features (*age, workclass, fnlwgt, education, education_num, marital_status, occupation, relationship, capital_gain, capital_loss, hours_per_week, country*)
- Z: 2 sensitive features (*race* and *sex*)
- y: 1 class variable (*Income*)

3 The Methodological Approach

3.1 EDA

The main source of algorithmic bias is represented by the training data. This may not be related to a bad data collection, but simply because personal data act as a social mirror, where protected attributes are redundantly encoded in observables.

Given the role of the data in the training process, an explanatory data analysis was conducted first. The class variable *income* is imbalanced with favour to the 0 value (74.94%), the variable *sex* follows the same distribution with a prevalence of males (67.67%), while the variable *race* has a stronger unbalance with only 9.80% of black people.

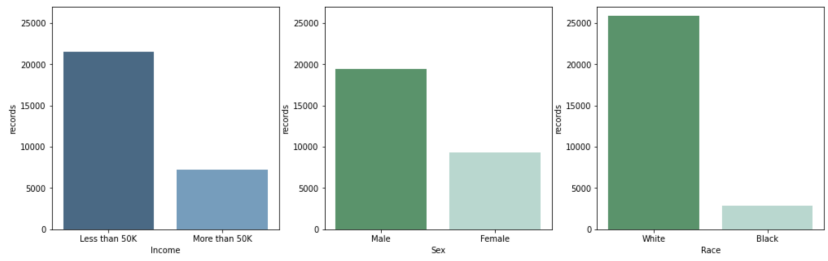

Figure 1: *Income, Sex* and *Race* values distribution

With a deeper exploration on sensitive attributes Z_i values distributions into the class variable *income* (figure 2) a strong discrepancy emerges between the positive records in the protected classes (*Female, Black*) and the positive ones in the opposite classes (*Male, White*)

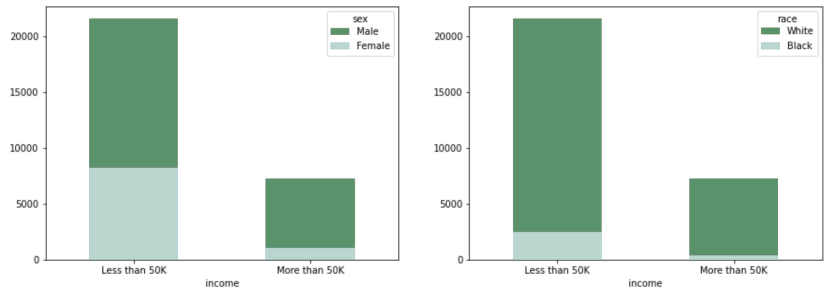

Figure 2: Protected and Unprotected classes distribution on *Income*

The rate of black people in the positive class is less than half the rate in the negative class (5.08 % vs 11.38%), while the rate of females in the positive class is almost one third of the rate in the negative class (14.66% vs 38.24%).

This discrepancy into the class values distribution could affect the algorithm into producing more likely positive predictions for the protected classes *Male* and *White* and negative predictions for the unprotected classes *Female* and *Black*.

3.2 Preprocessing

Some preprocessing operations were needed to make the data more suitable for the learning algorithm.

The attributes *education_num* and *education* carry out the same information, only the first one, that was already discretized, has been maintained.

The 5 categorical variables were transformed into dummy variables. For the *country* attribute (39 different values), that indicates the state of birth, only the 14 most represented countries were maintained, the other were grouped into *"others"* in order to restrict the number of new dummy variables without loosing a significant amount of information.

The training data X were split into training set (70%) and test set (30%). The training set, in turn, was split in training set (80%) and validation set (20%). Both processes followed a stratified sampling, due to the imbalanced target class. Finally, the features were standardized.

Before proceedings with the learning phase, the correlation matrix between the sensitive attributes and training features was analyzed (Figure 3). The *sex* attribute shows light correlation coefficients, while the correlation is

stronger between X_i and *race*, suggesting a possible indirect discrimination effect.

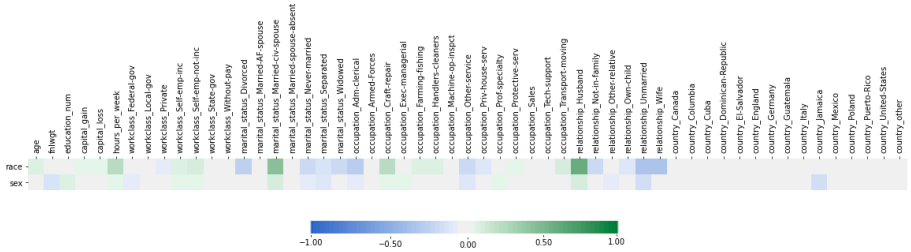

Figure 3: Correlation Matrix between sensitive attributes and X

3.3 Baseline Classifier

The first classifier used as benchmark is a feed-forward neural network. It has a simple architecture consisting of 3 hidden layers with 32 neurons each and Rectified Linear Unit activation functions (*'relu'*) alternated with 3 dropout layers (dropout range = 0.2) aimed at reduce the risk of overfitting. The output layer consists of 1 neuron with *sigmoid* activation function, in order to obtain as output for each record the probabilities of belonging to the positive class variable.

Due to the imbalanced target class, a balanced weight system was used.

The loss function utilized was the *binary crossentropy*, while *Adam* (Adaptive Moment Estimation) was the optimizer.

The model was trained over 50 epochs with an *Early Stopping* looking at the validation loss.

3.4 Fair Classifier

The procedure for developing and training a fair classifier takes inspiration from GANs: it leverages adversarial network to enforce the pivotal property on the predictive model [5]. A pivotal quantity is a function of observations and unobservable parameters so that the function's probability distribution does not depend on the nuisance parameters [6]. In this case the observable parameters are represented by the training data X_i, while the nuisance (unobservable) parameters are *race* and *sex*.

The fair classifier can be actually seen as a GAN with some key differences. The generator is replaced by the classifier itself to which an adversarial component is connected. The task of the adversarial is no longer to distinguish real from generated examples, but to predict the sensitive attributes Z_i given the predicted $\hat{y}$ from the classifier. The architecture of the fair classifier can be seen in Figure 4:

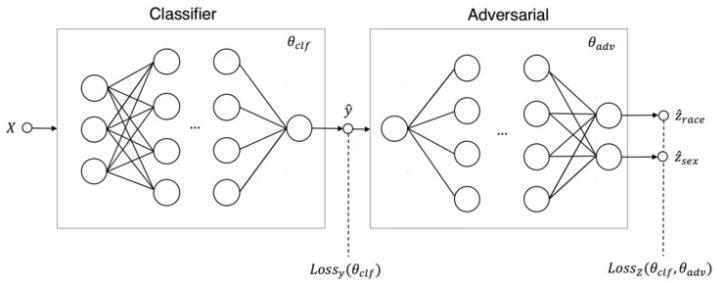

Figure 4: Fair classifier architecture
Towards fairness in ML with adversarial networks [7]

The classifier loss is denoted by $Loss_y(\theta_{clf})$, while the adversarial loss by $Loss_Z(\theta_{clf}, \theta_{adv})$. The classifier aimed to minimize its own loss, without caring about the adversarial:

$$min_{\theta clf}[Loss_y(\theta_{clf})] \tag{2}$$

The adversarial was aimed to minimize its own loss without caring about the classifier:

$$min_{\theta adv}[Loss_Z(\theta_{clf}, \theta_{adv})] \tag{3}$$

The objective was to minimize the classifier loss and maximize the adversarial loss, therefore making the best possible predictions whilst ensuring that race or sex cannot be derived from them. In formal terms:

$$min_{\theta_{clf}, \theta_{adv}}[Loss_y(\theta_{clf} - \lambda Loss_Z(\theta_{clf}, \theta_{adv})] \tag{4}$$

The classifier has the same architecture as the fair one. The architecture of the adversarial comprises 3 dense hidden layers of 32 nodes with ReLU activations, followed by an output layer of 2 nodes corresponding to the sensitive attributes.

Both models used *Adam* as optimizer and *binary crossentropy* as loss function.

After 5 epochs of "pretraining", where the classifier was trained on the full dataset and the adversarial was trained on the predictions of the pre-trained classifier, the classifier and adversarial networks were simultaneously trained: for each epoch, first the adversarial was trained while the classifier was fixed, then the classifier was trained on a minibatch with size=128 while keeping the adversarial fixed. This concurrent fitting was repeated at most for 300 epochs, unless the p-rule objective ($p = 80\%$) was reached for both *sex* and *race*.

The goal of this combined classifier-adversarial model was to learn how to make accurate and fair predictions.

3.5 Hyperoptimization

Among the many hyperparameters of the networks, the focus was on λ (4), which steers the classifier towards fairer predictions while sacrificing predictions accuracy. Given the lack of literature on this hyperparameter, λ was chosen with a Sequential Model Based Optimization into a wide range of values (1.0 to 200.0), with the help of SMAC[8] package.

First, configuration space and scenario were defined. λ_0 and λ_1, referring respectively to *race* and *sex*, were treated as *Uniform Float Hyperparameters* without any conditional hypotheses. Five initial random configurations were set; the remaining 45 runs were left to the Expected Input acquisition function. The hyperoptimization flow worked in a deterministic scenario with quality as run object.

The Bayesian Optimization used a Random Forest surrogate function.

The function to be optimized was one of the most challenging aspect of the work. Recalling that the aim was to obtain the fairest results according to the minimum possible accuracy waste, the objective function was defined as the increment on sensitive attribute's p-rule scaled by the accuracy variation(4).

$$\Big(\frac{\Delta p - rule_0/|\Delta acc|}{100} + \frac{\Delta p - rule_1/|\Delta acc|}{100}\Big) \tag{5}$$

SMAC needs an objective function to be minimized, so instead of maximizing (5), its negated was minimized.

4 Results and Evaluation

The results of the baseline and fair classifiers obtained in the test set are summarized in Table 1.

Table 1: Test set results comparison

		Baseline Model	Fair Model
classification metrics	Accuracy	0.793	0.77
	AUC	0.91	0.87
	F-1	0.67	0.64
fairness metrics	P-rule race	53%	83%
	P-rule sex	39%	81%

The baseline classifier has good performances in terms of classification metrics but it is strongly biased towards the protected classes *race* and *sex*. Actually both p-rule related to the sensitive attributes were well below the fairness treshold, placed at 80%. Also, the predictions distribution for the sensitive attributes and their opposites on test set showed a defined discrepancy as shown in Figure 5:

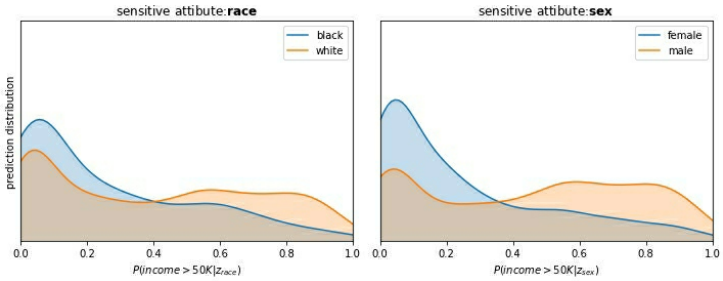

Figure 5: Baseline classifier prediction distributions

This discrepancy has been mitigated (and almost removed) by the fair classifier (Figure 6), and so the p-rules reached fairness in less than 120 epochs of training.

The hyperoptimization led to higher values for λ_0 than for λ_1 (Table 2), suggesting that the search for fairness is an harder job for *race* than for *sex*.

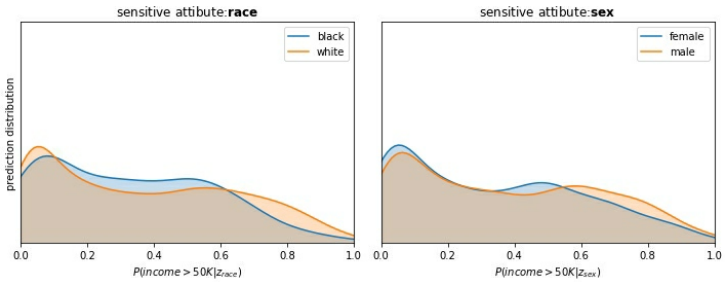

Figure 6: Fair classifier prediction distributions

Table 2: HPO best configurations

	iterations				
	2	3	8	9	18
λ_0	52.56	181.66	181.16	185.54	185.53
λ_1	28.13	125.66	60.50	60.21	60.23

The evolution of the single components of the surrogate function, $\Delta p - rule_{race}, \Delta p - rule_{sex}, \Delta accuracy$, pointed out that a single loss function where different target components are constricted is not efficient to jointly optimize them. Actually, every incumbent is dominated by a single component variation (Figure 7).

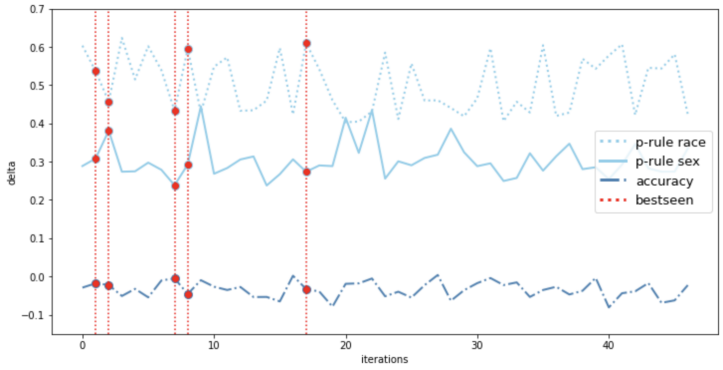

Figure 7: sensitive metrics and accuracy evolution during HPO

5 Discussion

The outcomes of the baseline classifier reflected the bias present in the data and, although the sensitive attributes were removed from the training phase, their effect was still present as hypothesized looking at their correlation coefficients with the training data (Figure 3). The use of this kind of classifier, which doesn't take into account any fairness constraint, would produce discriminatory outcomes against non protected groups. The discrimination would even increase with the collection of new data for the effect of perpetuating discrimination (feedback loop phenomenon).

The fight against discrimination with the pivotal property approach has shown that with a little accuracy waste ($\sim$ 2%), a fair classifier can be obtained but with two drawbacks:

- The λ parameters are task-specific and the results are really sensitive to these parameters. Their optimization is needed case-by-case.
- To mitigate disparate impact, a disparate treatment is needed. This approach is not legally accepted in some fields, even if accurate, well-grounded, and transparent statistical models are provided.[9]

The HPO operations were limited to the λ_i parameters due to the limited hardware resources owned and to reduce the search space. Other parameters regarding the architecture of the network should be optimized.

The objective function of the Bayesian Optimization does not distinguish the two components related to the sensitive attributes, with the risk of optimizing only one "dominant" component. The use of a stable package that allows multi-objective optimization is strongly suggested, as well as the use of some constraints concerning a minimum level of fairness.

Despite these limitations on the SMAC library, satisfying results have been obtained both on the classification side and fairness side.

6 Conclusions

This work showed that a machine learning classification model is simply a system that learns on the basis of the data used for the training. Sexist data in the training set produce a sexist classification model. Racist data in the training set produce a racist classification model.

Clean data are always desirable, but in real cases they are unlikely to be free of bias.

The disparate impact derived by biased data can be avoided with the adversarial approach, at the cost of some disparate treatment and accuracy waste.

Most of the difficulties encountered in developing a good fair classifier derived from the optimization of the different objectives: fairness and accuracy. Further studies should focus on the use of multi-objective optimization libraries.

References

[1] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification, 2017.

[2] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2015.

[3] Fairness measures. Available at "http://www.fairness-measures.org/Pages/Definitions".

[4] Ronny Kohavi and Barry Becker. UCI machine learning repository, adult census income, 2017.

[5] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks, 2016.

[6] Pivotal quantity. Available at "https://en.wikipedia.org/wiki/Pivotal$_quantity$".

[7] Towards fairness in ml with adversarial networks. Available at "https://godatadriven.com/blog/towards-fairness-in-ml-with-adversarial-networks/".

[8] Smac library. Available at "https://github.com/automl/SMAC3, 2017".

[9] Association belge des consommateurs test-achats asbl and others v conseil des ministres. Available at "http://curia.europa.eu/juris/liste.jsf?language=ennum=C-236/09".