Fairness of decision algorithm in machine learning

Alessandro La Farciola, Leevi Rönty, Camilla Beneventano

December 1, 2024

1 Introduction

Fairness in supervised learning is an interesting in-development field of research in Machine Learning. Indeed, since Machine Learning increasingly affects decisions in domains protected by anti-discrimination laws, it is starting to be important to ensure fairness in decision algorithms.

The fundamental questions that we are going to investigate through this report are: *What is Fairness? How we can achieve it?* And more, *How much we lose if we introduce a fair definition in our predictors?*

The entire work is related to a binary classification problem with a binary protected attribute, i.e. with two groups to preserve, for example male and female or two different races, but it is easily possible to extend it to a multiclass problem. After presenting different notions of fairness, we focus on the technical procedure with whom to derive a *fair* predictor from an *unfair* one, namely a predictor that doesn't take into account any definition of fairness. Finally, we present our experiment on a specific dataset where we show the performance of several predictors and discuss the pros and cons of fairness definitions.

2 Equality of Opportunity in Supervised Learning

2.1 Definitions of fair predictors

In order to obtain a *fair* predictor, one first naïve approach might be to ignore all protected attributes from learning procedure. However, it is easy to be convinced that this methodology is not enough due to the existence of *redundant encodings*. This means that the information related to the protected attributes can be hidden into other features that are strictly correlated with them. For this reason, we need precise conceptions of non-discrimination that we are going to define and discuss.

Firstly, we present the conception of *demographic parity*, which requires that a decision be independent of the protected attribute. In particular, we introduce the following definition.

Definition 1 (Demographic parity)**.** We say that a binary predictor $\hat{Y} \in \{0,1\}$ satisfies **demographic parity** with respect to a binary protected attribute $A \in \{0,1\}$ and Y if

$$Pr\{\hat{Y} = 1|A = 0\} = Pr\{\hat{Y} = 1|A = 1\}.$$

In other words, the rate of positive predicted is the same for both protected classes. Of course, this implies that also the rate of negative predicted is the same and this means that membership in a protected class should have no correlation with the decision. Unfortunately, this definition has some problems. First of all, as we will see in our experiment, in order to ensure *demographic parity* we need to strongly modify the *unfair* predictor so long as the percentages of acceptance match, losing in accuracy. Moreover, this notion of fairness would not allow the ideal predictor $\hat{Y} = Y$, which can be considered discriminatory. These are the motivations that brought the authors of [**?**] to modify the latter definition and introduce new conceptions of fairness.

The first is *equalized odds*.

Definition 2 (Equalized odds)**.** We say that a predictor $\hat{Y}$ satisfies **equalized odds** with respect to protected attribute A and outcome Y , if $\hat{Y}$ and A are independent conditional on Y, i.e. if

$$Pr\{\hat{Y} = 1|A = 0, Y = y\} = Pr\{\hat{Y} = 1|A = 1, Y = y\}, \qquad y \in \{0,1\}.$$

In other words, we are asking that the true positive rate and the false positive rate are the same for both protected attributes. Furthermore, it is possible to relax the previous definition, asking that only the true positive rate is the same for both groups. This is *equal opportunity*.

Definition 3 (Equal opportunity)**.** We say that a binary predictor $\hat{Y}$ satisfies **equal opportunity** with respect to A and Y if

$$Pr\{\hat{Y} = 1|A = 0, Y = 1\} = Pr\{\hat{Y} = 1|A = 1, Y = 1\}.$$

2.2 Deriving from a score function

In this section, we are going to answer the main question: *How we can derive a predictor that satisfies equalized odds or equal opportunity from an unfair one?*.
To do that, we consider $R \in [0, 1]$ a **score function** and the corresponding (possibly discriminatory) predictor $\hat{Y} = \mathbb{1}\{R > t\}$. We want to find a *fair* predictor $\tilde{Y} = \mathbb{1}\{R > t_A\}$ using different thresholds for different values of A. In particular this type of predictor $\tilde{Y}$ is derived from $\hat{Y}$ (or better R) and A according to the following definition.

Definition 4 (Derived predictor)**.** A predictor $\tilde{Y}$ is **derived** from a random variable R and the protected attribute A if it is a possibly randomized function of the random variables (R, A) alone. In particular, $\tilde{Y}$ is independent of X conditional on (R, A).

It is always possible to construct a trivial predictor satisfying equalized odds or equal opportunity, but our goal is to derive predictors $\tilde{Y}$ that minimizes the expected loss $\mathbb{E}[\ell(\tilde{Y}, Y)]$, where ℓ: $\{0, 1\}^2 \to \mathbb{R}$ is a loss function.

To do this, we consider the two A-conditional ROC curves, which capture the false positive and true positive rates

$$C_a(t) := (Pr\{\hat{R} > t|A = a, Y = 0\}, Pr\{\hat{R} > t|A = a, Y = 1\});$$

and the convex hulls of the image of the conditional ROC curve

$$D_a = convhull\{C_a(t)|t \in [0, 1]\},$$

because each point in this area represents a feasible predictor, possibly obtained with a randomization technique. Now, we can proceed in the following two ways:

- **Deriving a predictor satisfying equalized odds:** we have seen that a predictor satisfies equalized odds if $Pr\{\hat{Y} = 1|A = 0, Y = y\} = Pr\{\hat{Y} = 1|A = 1, Y = y\}, y \in \{0, 1\}$, i.e. if it conditioned to A is represented by the same point in the side graph. So this point must be in the intersection of the two convex hulls and it must minimize the expected loss, therefore it will be in at least one of the two ROC curves.
- **Deriving a predictor satisfying equal opportunity:** a predictor satisfies equalized odds if $Pr\{\hat{Y} = 1|A = 0, Y = 1\} = Pr\{\hat{Y} = 1|A = 1, Y = 1\}$, i.e. if $C_0(t_0)$ and $C_1(t_1)$ agree in the second component for two suitable thresholds t_0 and t_1. So we want to find two points in the two ROC curves such that they have the same true positive rates.

Figure Roc.curves shows what is described above.

3 Experiments

3.1 Dataset and score function

The dataset consists of weighted census data collected in the United States. We are given a train-test split of 32561 and 16281 rows of data respectively. Our task is to predict whether the demographic group (one row of data) earns more or less than 50 kUSD per year. However, our goal is to make predictors fair with respect to sex and discuss the differences from the original *unfair* one. The features available for us from the data include for example age, education, race, and occupation.

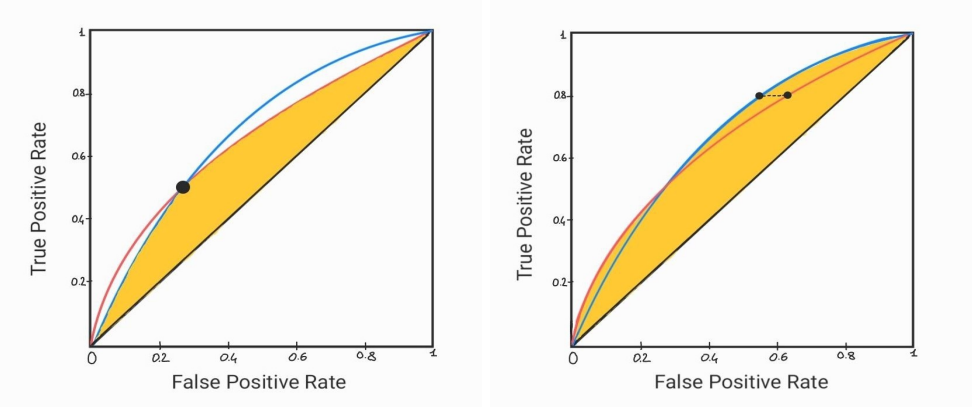

Figure 1: How to derive equalized odds (left) and equal opportunity (right) from ROC curves.

We will derive the fair predictors from a score function, that in this case is the estimated probability function from a logistic regression model. To fit the logistic regression model, we must transform categorical features to one-hot-encoded features, as the regression model requires numerical variables only.

When fitting the score function, we can see from the iteration logs, that in the end the loss does not improve anymore, indicating that the fitting process has converged. With the fitted model we can construct predictions /scores for the testing and training data. From those scores, we construct Figure fig:no-overfit. The figure demonstrates, that the model has not been overfitted, as the test loss is always close to the training loss, no matter which threshold we choose. Moreover, we can notice from the ROC curves that the predictor is more accurate for females. This is probably due to the rate of negative observations in the female class being larger, i.e. most of the females earn less than 50 kUSD per year, so being in the negative class is quite likely.

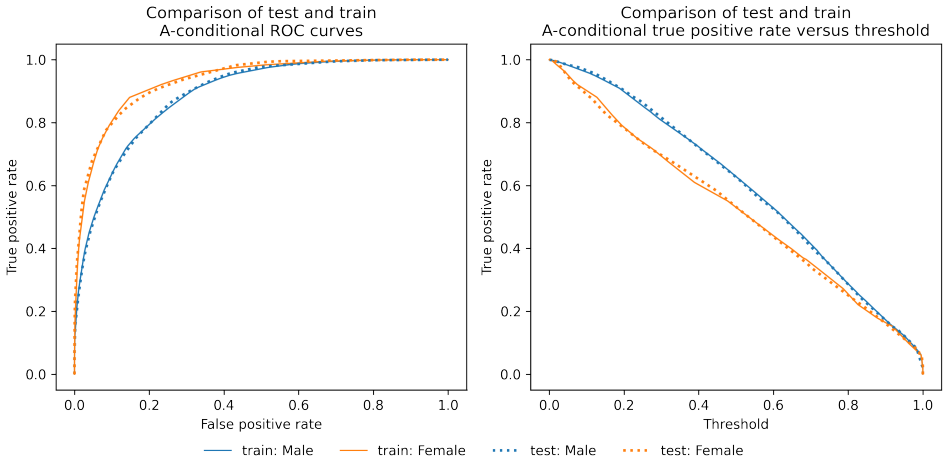

Figure 2: Demonstrating the generalization ability of the score function. On the right, we have the true positive rate vs the selected threshold. As the curves are similar, we can deduce that for a given threshold the true positive rates must be similar in train and test sets. From the figure on the right we see, that similar true positive rates imply similar false positive rates. As the prevalence of positive examples is about 24% in both train and test sets, we can conclude, that for any chosen threshold the losses must be close to each other. Thus, the score function has not been overfitted.

3.2 Models

We will consider five different definitions of fairness: **Sex Blind**: same threshold for all classes; **Max profit**: no fairness constraints, best threshold for each class; **Demographic parity**: max profit subject to positive prediction rate per class being equal; **Equal opportunity**: max profit subject to true positive rate per class being equal; and **Equalized odds**: same fraction of true positives and false positives for each group, different thresholds (possibly randomized).

The fair predictors were constructed according to the different definitions of fairness as described earlier (see Appendix Appendix for details). The resulting thresholds and predictors are visualized in Figure fig:thresholds.

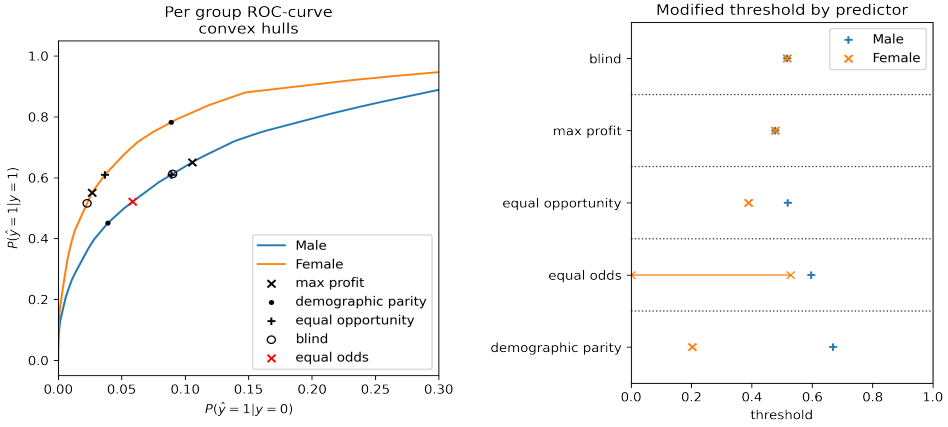

Figure 3: Resulting thresholds (right) and resulting predictors (left) on the test data A-conditional ROC curves.

3.3 Obtained losses

Figure fig:thresholds visualizes the obtained predictors, displaying the training ROC curves from which the predictors were derived, and the thresholds obtained for the groups. For the equal odds predictor, the female threshold is shown as a line to indicate that the predictor uses randomization between the thresholds at the endpoints of the line.

In Table tab:losses we list the training and testing losses of the obtained predictors. As expected, the train and test losses are close to each other, indicating that the models have not been overfitted. Max profit achieves the smallest loss, but only for the train set. The difference to the blind model is not that big, as due to redundant encodings the information about the protected attribute may have already been included. As expected, the equalized odds predictor's loss is greater than the equal opportunity predictor's loss. This is due to the equal odds condition being more restrictive. Similarly, the demographic parity predictor is performing quite poorly, as the protected classes are not equally prevalent relative to the class size in the positive examples.

predictor	train loss	test loss
blind	0.1486	0.1480
max profit	0.1446	0.1485
demographic parity	0.1647	0.1684
equal opportunity	0.1463	0.1497
equal odds	0.1598	0.1740

Table 1: Train and test losses by fair predictor type. Loss is the miss classification error.

4 Conclusion

In this report, we explored different definitions of fairness, their implementations, and the trade-off of fairness versus model accuracy. As we found out, there are multiple definitions for fairness, of which the equal opportunity and equal odds definitions make the most sense. The fair predictors do not come for free as they introduce some inaccuracy to the predictor. For this reason, according to the aim of the analysis, one could prefer a specific predictor, balancing the level of accuracy and the level of *fairness*.

5 Appendix

In this appendix, we want to explain in more detail how we derived the fair predictor. In particular, we focus on how to choose the best predictor in terms of errors still satisfying the fairness condition.

Risk minimization

Each of the fairness definitions imposed different constraints on empirical risk minimization. Nevertheless, each of the optimization problems was boiled down to minimizing 0-1 loss over a single scalar parameter. This loss minimization with respect to the scalar is described here.

For the sex blind predictor, we minimize loss w.r.t a single threshold. Knowing the true positive rate, false positive rate, and the estimated positive example rate $\hat{\pi} = |P|/(|P| + |N|)$, we can calculate the empirical loss. From the distribution of scores obtained from the score function and the training dataset, we can calculate a mapping from thresholds to true positive rate and false positive rate, and those can be again mapped to a loss value. Thus, we can construct a mapping from a threshold to a loss value.

As suggested by [?], we also chose to use ternary search to search for the minimizing scalar value. However, ternary search finds the global optimum only in convex optimization problems. Thus, we chose to work on the ROC curve convex sets instead of the ROC curves directly.

The process for other fairness definitions is very similar. For max profit, we optimize the two thresholds separately. For demographic parity, we construct a similar mapping from $\mathbb{P}(\hat{y} = 1|A = a)$, which is equal for all protected classes. For equal opportunity and equalized odds, the scalar over which to optimize was the true positive rate. In Figure fig:loss-vs-demographic we demonstrate the risk minimization problem: the search space is the reals between zero and one, and those values have a convex map to 0-1 loss.

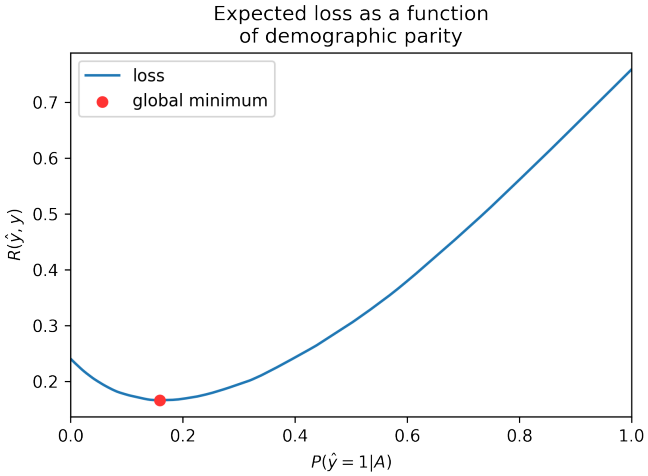

Figure 4: Demonstration of the convex optimization problem.

The specific implementations can be found here: `https://github.com/leevironty/stat-ml-project`.

References

[1] Hardt, M., Price, E., Srebro N., (2016). *Equality of opportunity in supervised learning.* In Advances in Neural Information Processing Systems (pp. 3315-3323).

[2] Adult. (1996). UCI Machine Learning Repository.