On the relation between accuracy and fairness in binary classification

Indrė Žliobaitė INDRE.ZLIOBAITE@AALTO.FI
Aalto University and Helsinki Institute for Information Technology HIIT, Finland

Abstract

Our study revisits the problem of accuracy-fairness tradeoff in binary classification. We argue that comparison of non-discriminatory classifiers needs to account for different rates of positive predictions, otherwise conclusions about performance may be misleading, because accuracy and discrimination of naive baselines on the same dataset vary with different rates of positive predictions. We provide methodological recommendations for sound comparison of non-discriminatory classifiers, and present a brief theoretical and empirical analysis of tradeoffs between accuracy and non-discrimination.

1. Introduction

Discrimination-aware machine learning is an emerging research area, which studies how to make predictive models free from discrimination, when historical data, on which they are built, may be biased, incomplete, or even contain past discriminatory decisions. Research assumes that the protected grounds, against which discrimination is forbidden, are given by legislation. The goal for machine learning is to develop algorithmic techniques for incorporating those non-discriminatory constraints into predictive models.

A number of studies in discrimination-aware machine learning and data mining (Pedreschi et al., 2009; Kamiran et al., 2010; Calders & Verwer, 2010) focus on achieving equal acceptance rates (proportions of positive decisions) for favored and protected groups of individuals in binary classification. Forcing acceptance rates to be equal without taking into account other characteristics of individuals can be seen as an affirmative action, which introduces positive discrimination promoting the protected community. This may be desired for legal and political reasons.

We revisit this popular scenario of discrimination aware machine learning, and identify some pitfalls to avoid when comparing the performance of such classifiers, that is, a comparison may be misleading if the proportions of positive predictions of the classifiers are different. We provide methodological recommendations for sound comparison, and present a brief theoretical and empirical analysis of tradeoffs between accuracy and non-discrimination.

2. Problem setting and assumptions

Given a dataset that contains discrimination the goal is to build a classifier that would be as accurate as possible, and obey non-discrimination constraints. For example, a model could decide upon granting a loan given demographic information and financial standing, and considering ethnicity of an applicant (native, foreign) as the protected ground. We assume that the values of the target variable (labels) in the historical dataset are objectively correct, e.g. whether the loan has been repaid or not. For discrimination to happen the target variable needs to be polar, that is, one outcome (accept) should be preferred over the other (reject).

Let X denote a set of input variables (e.g. salary, assets), s denote the protected characteristic (e.g. ethnicity: native (w) or foreign (b)), and y denote the target variable (e.g. loan decision: accept ($+$) or reject ($-$)). A classifier maps X to y, that is, $\hat{y} = f(X)$. Even though s is not among the input variables, some variables in X may be correlated with s (e.g. social security payment history may be shorter for foreigners, because they have arrived recently), and, as a result, classifier f may capture the protected characteristics, and induce indirect discrimination in decision making.

Let discrimination be measured as the difference in rates of acceptance: $d = p(+|w) - p(+|b)$. Suppose that discrimination in the historical dataset is $d_0 = \delta$, the desired discrimination in the classifier output is $d^\star$, the proportion of favored individuals in the data is $p(w) = \alpha$, the prior probability of acceptance in the data is $p(+) = \pi_0$, and the rate of acceptance in the classifier output is $p_f(+) = \pi$.

Many classifiers produce probability scores (such as Naive Bayes or logistic regression). Typically, a probability score can be computed for non-probabilistic classifiers as well (such as kNN, SVM, decision trees). Individuals scoring above a threshold, which by default is typically 0.5, will

The 2^{nd} Workshop on Fairness, Accountability, and Transparency in Machine Learning, Lille, France. Copyright 2015 by the author.

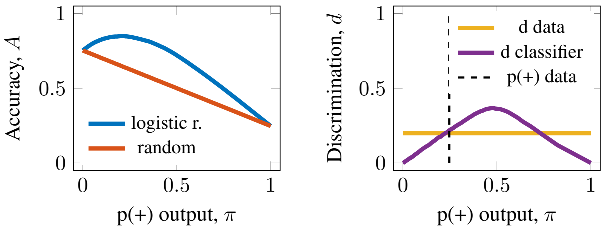

Figure 1. Accuracy and discrimination measured directly.

get a positive decision. Considering available resources a decision maker can choose a different threshold. Suppose that the objective is to keep discrimination at the desired level $d^\star$ (typically zero), and at the same time maximize the prediction accuracy. Effectively, by choosing the threshold a decision maker chooses the acceptance rate π.

3. Accuracy and fairness

The performance of discrimination-aware classifiers is typically compared by plotting discrimination vs. accuracy. An attempt to remove discrimination can easily produce classifiers with different acceptance rates π from those in the original dataset, especially when using off-the-shelve classifier implementations (e.g. WEKA[1]), which simply round the numerical probability scores without any constraints on the positive output rates.

Our main message is that evaluation of non-discriminatory classifiers must take into account rates of acceptance, otherwise classifier performance is not comparable, because changing the acceptance rate changes baseline accuracy and baseline discrimination.

A small experiment with a benchmark dataset (Adult from UCI[2] repository) illustrates the situation. The target variable describes whether a person has high income or low. The protected characteristic (gender) is not among the inputs. We randomly split the dataset into two halves: training and testing. We train a logistic regression (similar results have been obtained with Naive Bayes and decision tree J48) on a train set, output class probability scores for the test set, and vary the classification threshold from 0 to 1, which changes the acceptance rate π. We also plot the accuracy of a random classifier that does not use any inputs, but randomly decides upon the outcome given the probability of acceptance π. Figure 1 presents the results.

From the left plot we see that the more extreme the acceptance rate is (either all reject, or all accept), the closer the performance of an intelligent classifier (logistic regression) is to that of a random classifier, which assigns labels at random. Therefore, better observed accuracy does not necessarily mean better classification ability, if the acceptance rates of the two classifiers are different. In order to be able to compare such classifiers we could normalize the accuracy with respect to π. Therefore, we suggest using for comparison a normalized accuracy, such as Cohen's Kappa (Cohen, 1960), which indicates by how much a classifier in question is better than a random classifier:

$$\kappa = \frac{A-R}{1-R}, \tag{1}$$

where A is the accuracy of the classifier in question, and R is the accuracy of a random classifier, in our case
$R = \pi_0\pi + (1-\pi_0)(1-\pi)$. Note, that $\kappa \in [0,1]$, where 1 means the ideal accuracy, and 0 indicates a random result[3].

In the right plot we see how discrimination varies with different acceptance rates. There is no discrimination if everybody is accepted, or nobody is accepted, and the closer the acceptance rate π gets to these extremes, the smaller is d. This is not due to a better fairness of the classifier, because the classifier is exactly the same, and its output is the same, just the classification threshold varies. We would like to assess the fairness of the classifier, therefore, similarly to the accuracy, we need to normalize the result with respect to π.

We propose to normalize d by the maximum possible d_{max} at each π. Discrimination would be at its maximum if a classifier ranks candidates in such a way that first everyone from the favored community is accepted, and only then candidates from the protected community start to be accepted[4]. In such a case the maximum discrimination is

$$d_{max} = \min\left(\frac{\pi}{\alpha}, \frac{1-\pi}{1-\alpha}\right), \tag{2}$$

where α is the proportion of the favored community individuals in the data, and π is the acceptance rate.

We propose to normalize the discrimination measure by the maximum possible discrimination.

$$\delta = \frac{p(+|w) - p(+|b)}{d_{max}}, \tag{3}$$

where d_{max} given in Eq. (2) is the maximum possible discrimination at a given acceptance rate. The maximum

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://archive.ics.uci.edu/ml/

[3] One could consider other accuracy measures for imbalanced data, such as F-score. We prefer Cohen's Kappa, since F-score does not behave consistently at the extreme acceptance rates, and, therefore, is more difficult to interpret. F-score of a classifier that accepts everybody would be equal to π_0, which varies depending on the dataset, while Kappa always gives 1 in this case.

[4] It can be compared to a (supposedly fictional) evacuation procedure from the Titanic. Passengers are put in a queue, where all the first class passengers have a priority over third class passengers. Then as many passengers are evacuated, as there are boats.

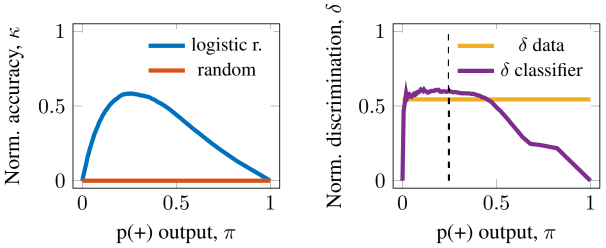

Figure 2. Normalized accuracy and discrimination.

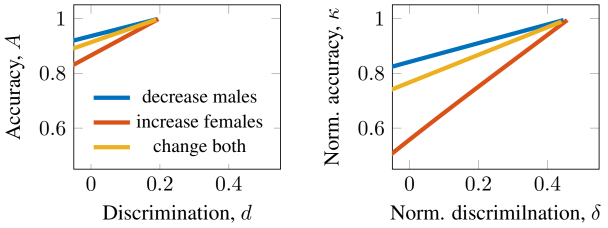

Figure 3. Oracle.

value of δ is 1, which means the worst possible discrimination, where the favored community has a complete priority, $\delta = 0$ means no discrimination where people from the favored and protected communities fully mix in the queue. δ can be negative, indicating a reverse discrimination.

Figure 2 plots normalized accuracy κ and normalized discrimination δ of the logistic regression in our experiment. Large part of discrimination appears to be flat and closely in line with the discrimination in the data. The results now make sense, since the classifier in the experiment does not have any mechanisms for discrimination removal. At the extreme ends, where everybody is accepted, or everybody is rejected, intuitively, there is no discrimination, and the normalized measure correctly shows no discrimination.

4. Baselines and tradeoffs

It has been observed (Kamiran et al., 2010) that, assuming the labels in data are correct, discrimination removal comes at a cost – it reduces prediction accuracy. The authors have found given no constraints on the acceptance rates, that the maximum possible accuracy decreases linearly with reducing difference in rates of acceptance. We revisit the problem of accuracy-fairness tradeoff to see if the normalized measures would show similar relations.

An oracle is a fictional baseline classifier that has the maximum possible intelligence (as if it knows the true labels), and strives to satisfy non-discrimination constraints. A random classifier is the opposite, it does not use any intelligence. For each individual a random classifier makes a random prediction with the probability of acceptance π.

The accuracy of the oracle will be $A_0 = 1$, kappa will be $\kappa_0 = 1$, the discrimination would be as in the data d_0 and δ_0. The random classifier defines the other baseline of performance with $A = \pi_0\pi + (1 - \pi_0)(1 - \pi)$, $\kappa = 0$, and $d = \delta = 0$. With $\pi = 0$ (or $\pi = 1$) the random classifier turns into the majority class classifier.

Suppose, a decision maker aims at removing all discrimination such that $d^\star = 0$ and $\delta^\star = 0$. As suggested in (Kamiran et al., 2010), the oracle would either reduce the acceptance rate for the favored community (if $\alpha \leq 0.5$), or increase the acceptance rate for the protected community (if $\alpha > 0.5$). The resulting decrease in classification accuracy would be linearly proportional to the discrimination in the data $(A_0 - A) = \min(\alpha, (1 - \alpha))\,(d_0 - d)$.

We find that if the rate of acceptance is to be fixed, that is $\pi = \pi_0$, then the normalized accuracy of the oracle decreases linearly with decrease in normalized discrimination

$$(\kappa_0 - \kappa) = \min\left(\frac{\alpha}{\pi_0}, \frac{1 - \alpha}{1 - \pi_0}\right)(\delta_0 - \delta). \qquad (4)$$

If the rate of acceptance does not need to be fixed, the optimal strategy is still the same – either to reduce acceptance for the favored community (”decrease males”), or to increase acceptance for the protected community (”increase females”), but the choice now depends not only on α, but also on π_0 and $\delta^\star$. We do not have a closed form solution at the moment, but Figure 3 presents simulated results of the oracle classifier on the benchmark dataset (Adult). ”Change both” is the solution where the acceptance rate is kept the same as in the original data. These experiments show the maximum possible accuracy, given the discrimination constraints. We can see that when using the normalized measures for accuracy and discrimination the upper bounds remain linear.

5. Interesting cases

We wrap up our study with an experiment to illustrate the difference between the raw and normalized measures when comparing non-discriminatory classifiers.

The experiment compares the performance of three classifiers (logistic regression, Naive Bayes and decision tree J48 from WEKA) trained using three different strategies: including the protected characteristic among classifier inputs, excluding the protected characteristic from classifier inputs, and excluding the protected characteristic from classifier inputs plus massaging the labels of the training data. Massaging is perhaps the simplest discrimination removal strategy, it has been introduced in (Kamiran & Calders, 2009). Training labels are converted from binary to numeric using a ranker function, we use a logistic regression

Table 1. Performance of classifiers, everything $\times 10^{-2}$

	p(+) π	Acc. A	Disc. d	N. acc. κ	N. disc. δ
Data/oracle	24.7	100	19.9	100	54.4
Logistic with s	20.2	84.9	18.3	56.7	**61.4**
Logistic no s	20.1	84.9	17.6	56.6	**59.6**
Logistic massage	22.1	83.5	6.9	53.9	21.3
NB with s	15.4	81.9	13.5	44.2	**59.7**
NB no s	14.4	81.4	10.9	41.7	51.3
NB massaged	15.4	81.5	6.8	43.3	29.7
Tree J48 with s	19.6	85.1	17.9	56.9	**61.9**
Tree J48 no s	19.6	85.0	17.9	56.7	**61.8**
Tree massage	22.9	83.5	6.1	54.6	18.1

fit on the same training data. A number of lowest ranked males who have a positive label are changed to negative, and the same number of highest ranked females, who have a negative label, are changed to positive such that the positive rate remains the same as in the original data, but the discrimination is zero. Then a classifier is learned on this modified training data. Testing data is not modified. Table 1 presents the results measured on the testing data.

We can make several interesting observations. First, all classifiers tend to output lower acceptance rates than that in the original data. At the same time, if the protected characteristic is used, the discrimination measure d may show a decrease in the nominal discrimination as compared to the original data, but the normalized discrimination δ by all three classifiers is even higher than in the data. Apparently, a classifier learned on discriminatory data without any protective measures amplifies discrimination.

Removing the protected characteristic (no s) indicates little improvement in discrimination. This is due to, so called, redlining effect. A number of features in the data are correlated with the protected characteristic, therefore, discrimination is still captured, and, in cases of logistic regression and decision tree, is still higher than in the original dataset.

Interestingly, massaging strategy outputs higher acceptance rates than removing the protected characteristic. The acceptance rates of massaging are closer to the positive rates in the original data, and discrimination is lower, as expected. This suggests, that when discrimination is present in the training data, but usage of the protected characteristic is not allowed, classifiers tend to decrease the acceptance rate, which may show better nominal discrimination figures, but the real underlying discrimination (measured by normalized δ) remains.

Finally, Figure 4 presents normalized accuracies and discriminations at different acceptance rates. Overall we can see that massaging does remove some of discrimination, but at many acceptance rates the removal is not very precise, and sometimes even overshoots introducing a reverse discrimination. This calls for a revision of the massaging, and possibly other discrimination removal techniques, taking into consideration possibility of different acceptance rates and normalized measures of discrimination.

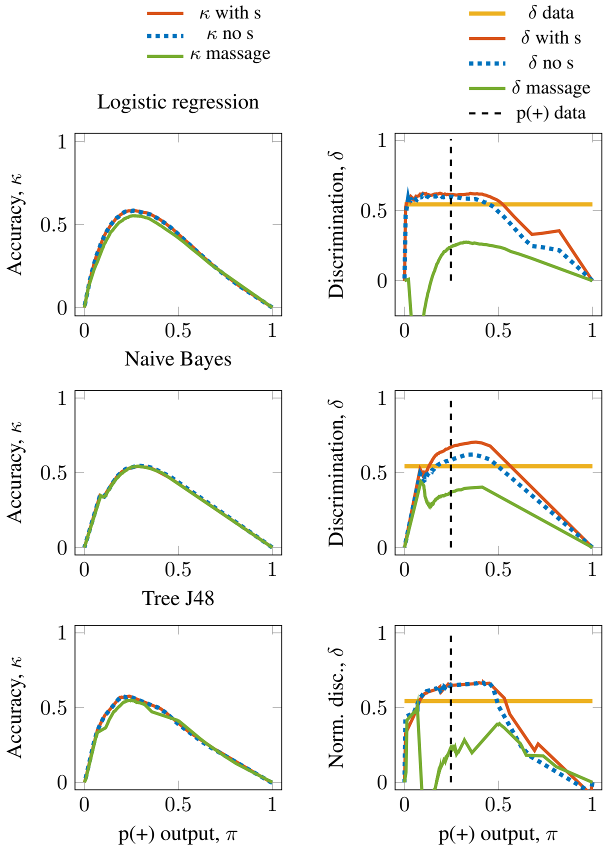

Figure 4. Performance of baseline classifiers.

6. Conclusion

Evaluation of non-discriminatory classifiers needs to take into account positive output rates, otherwise the comparison may be misleading and conclusions about comparative performance may be invalid.

We have introduced a normalization factor for discrimination measure, considering the maximum possible discrimination at a given acceptance rate. The maximum discrimination is present when the protected individuals start to be accepted only after everybody from the favored community is accepted.

Acceptance rates may be constrained by resources, and not freely available to choose for decision makes. If the acceptance rate in the data and in the classifier outputs is fixed, then classifiers are comparable in terms of A and d, otherwise they need to be compared in terms of κ and δ.

References

Calders, Toon and Verwer, Sicco. Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21(2):277–292, 2010.

Cohen, Jacob. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1):37–46, 1960.

Kamiran, Faisal and Calders, Toon. Classification without discrimination. In *Proc. of the 2nd IC4 conf. on Computer, Control and Communication*, pp. 1–6, 2009.

Kamiran, Faisal, Calders, Toon, and Pechenizkiy, Mykola. Discrimination aware decision tree learning. In *Proc. of the 2010 IEEE International Conference on Data Mining*, ICDM, pp. 869–874, 2010.

Pedreschi, Dino, Ruggieri, Salvatore, and Turini, Franco. Measuring discrimination in socially-sensitive decision records. In *Proc. of the SIAM Int. Conf. on Data Mining*, SDM, pp. 581–592, 2009.