Discrimination in machine learning algorithms

Discriminazione negli algoritmi di apprendimento automatico

Roberta Pappadà and Francesco Pauli

Abstract Machine learning algorithms are routinely used for business decisions which may directly affect individuals: for example, because a credit scoring algorithm refuses them a loan. It is then relevant from an ethical (and legal) point of view to ensure that these algorithms do not discriminate based on sensitive attributes (sex, race), which may occur unwittingly and unknowingly by the operator and the management. Statistical tools and methods are then required to detect and eliminate such potential biases.
Abstract *Sempre più decisioni, nei campi più vari, sono prese impiegando algoritmi a supporto o in sostituzione dell'intervento umano. Queste decisioni possono avere un effetto sulle persone che le subiscono, ad esempio quando un algoritmo di valutazione di solvibilit decide di rifiutare un prestito. Diventa quindi rilevante eticamente (e legalmente) assicurarsi che questi algoritmi non basino la loro decisione anche su caratteristiche sensibili, cosa che può avvenire senza intento e consapevolezza da parte del responsabile. Si apre dunque la necessità di studiare strumenti e metodi statistici per individuare e eventualmente eliminare queste potenziali distorsioni.*

Key words: machine learning; CEM; protected categories; sensible attribute

1 Introduction

The kind of discrimination we refer to consists in treating a person or a group depending on some sensitive attribute (s.a., S) such as race (skin color), sex, religious orientation, etc.

A human may discriminate either because of irrational prejudice induced by ignorance and stereotypes or based on statistical generalization: lacking specific in-

Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy, e-mail: rpappada@units.it, francesco.pauli@deams.units.it

formation on an individual, he is assigned the characteristics which are prevalent in the sensitive attribute category he belongs to (for example in the US lacking information on education, a black person may be assumed to have relatively low level since this is the case in general for black people in the country) [7]. When a statistical/machine learning algorithm is used in the decision process, its behavior with respect to discrimination depends on the information it is given. In particular, if the sensitive attribute is available to the algorithm (that is, it is included in the learning data and can be used for predictions), it may discriminate either because the data it is taught by contain irrational prejudice (Fig. 1*(a)*) and because the sensitive attribute is associated to an unobserved attribute which is relevant for prediction of Y, the outcome of interest (Fig. 1*(b)*). An example of the former may occur if an algorithm is sought for screening job applicants and the learning data consist of selections made in the past by humans who, in some instances, based their decisions on irrational prejudice, consciously or not; as an example of the latter consider a credit scoring algorithm where the education of applicant is unknown but the race is, then due to the fact that education is relevant to assess reliability and is associated to the race the latter may be used in the decision. For an actual example consider car insurance (RCA), where the sex of the insured used to be a relevant factor in pricing likely because it is related to the number of km driven per year which is an unobserved variable strongly related to the outcome (liabilities). We note in pass that, although this is a rare circumstance, it may also be the case that the sensitive attribute is directly related to the outcome not because of human irrational prejudice, as it happens in life insurance for the gender attribute.

If the algorithm is constrained not to use the s.a. in the final rule (it is excluded from the data from which the algorithm learns) discrimination may still occur indirectly, either because a variable which is related to the outcome is also related to S (Fig. 1*(c)*) or because a variable included in the data which is unrelated to the outcome is related to S which is related to the outcome (either directly as in Fig. 1*(d)* or mediated as in Fig. 1*(e)* or in both ways as in Fig. 1*(f)*). In the first case the collective possessing the s.a. may experience the desired outcome less frequently than the rest whenever it differs from the population in the distribution of some relevant features included in the data. In the second case the collective possessing the s.a. may experience the desired outcome less frequently than the rest even if it were equal to the population as far as the other relevant features included in the data are concerned; in this second situation the discrimination may be due to a difference in the distribution of some relevant feature not included in the data but related to S and some variable in the sample.

Whether the situations described above are instances of undesired discrimination or not depends on how the concept is defined from a legal/ethical standpoint. In particular it depends on whether we require that the groups identified by S have the same treatment (rate of positive outcome) unconditionally or that they have the same treatment conditionally on some relevant characteristics different than S and which are deemed lawful to use to discriminate.

Which definition is more appropriate is a matter of ethic/law: a requirement of unconditional equal treatment means that the groups must be given equal treatment even if they are not equal (demographic parity, disparate impact), which may seem unjust from an individual point of view [2]. On the other hand, admitting the groups to be treated differently because one of them possesses desirable characteristics possibly amounts to perpetuating past unfair (S based) discrimination which may have lead the groups to be different. In fact, the first requirement may be an instance of affirmative action since it goes in the direction of eliminating the differences between groups, differences which are at least partly admissible within the second requirement. If conditional equal treatment is sought, one must decide which characteristics other than S but possibly correlated with it are ethical/lawful to use, which may be partly dictated by law, partly uncertain (for example in a civil law system such as in the US whether a characteristic is lawful to discriminate on may be for a jury to decide).

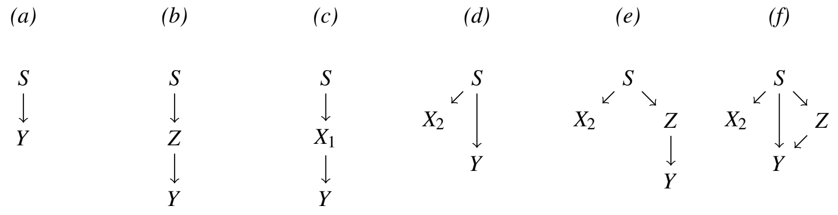

Fig. 1 Y is the outcome, S is the sensitive attribute, X_i denotes observed variables, Z denotes an unobserved variable. Example: S: race, Y: restitution of a loan, X_1: socioeconomic status, X_2: zip code residence, Z availability of family financial support.

In order to give precise definitions, from now on assume both S and Y dichotomous, $Y = 1$ be the desired outcome and $S = 1$ the belonging to a protected category. If unconditional equal treatment (demographic parity) is desired, data are apt if it is not possible to predict S from Y [1], that is, data are compatible with

$$P(Y = 1|S = 1) = P(Y = 1|S = 0).$$

If equal treatment should be conditional on X_1 then the requirement becomes

$$P(Y = 1|S = 1, X_1 = x_1) = P(Y = 1|S = 0, X_1 = x_1).$$

An extreme version of conditional equal treatment is advocated in [2], a rule is non discriminatory if the prediction errors are the same regardless of S

$$P(\hat{Y} = 1|S = 0, Y = y) = P(\hat{Y} = 1|S = 1, Y = y), \ \forall y;$$

or, which is the same, prediction and sensitive attribute are independent conditional on the outcome.

2 Measuring and avoiding discrimination by causal inference

The different strategies which can be used to avoid discrimination from an algorithm can be distinguished depending on the level at which action is taken: 1) learning data can be modified to ensure they do not imply discrimination; 2) the learning algorithm can be integrated with a non discrimination objective; 3) the final algorithm can be tampered with after it has been fitted; 4) the final predictions can be changed. The precise action to be taken depends on the definition of discrimination adopted; as outlined in the previous section, we may seek unconditional equality or conditional equality.

We focus on data preprocessing techniques. In a nutshell, this entails first establishing whether and to what extent the available data are discriminatory. If discrimination is detected data are modified in order to make them discrimination free before and then used with a standard algorithm [5]. Various ways of modifying data have been proposed including: 1) removing attributes (correlated with S); 2) changing the labels of some units; 3) weighing units; 4) resampling units We note that the precise implementation for both steps depends on the definition of discrimination which is adopted, presumably on legal/ethical grounds (see section 1).

In order to measure discrimination within a dataset it has been proposed to use causal inference techniques. Causal inference techniques aim at estimating the causal effect of a treatment–in this context, the belonging to a protected category–on an outcome. Generally speaking, causal inference methods aim at assessing the effect of a treatment based on observational data by matching treated units to untreated units which are similar with respect to their observed characteristics: a balanced dataset, suitable to draw causal inference, is built by restricting the original dataset to treated and untreated units which have been matched. Comparing a protected units outcome with the outcome of unprotected units which are similar with respect to the other observables is also a reasonable way to detect whether the unit has been discriminated (conditionally).

Following this idea, Luong *et al.* [6] propose to measure discrimination for unit (individual) i of the protected category ($S_i = 1$) as

$$\Delta_i = \frac{1}{k}\#\{j|j \neq i, x_j \in U_i^{k,1}, y_j = y_i\} - \frac{1}{k}\#\{j|j \neq i, x_j \in U_i^{k,0}, y_j = y_i\} \qquad (1)$$

where $U_i^{k,s}$ is the set of the k nearest neighbours of x_i within those units for which $S = s$, according to a Gower type distance (possibly, other measure of the difference between the two frequencies are used such as the ratio or the odds). Note that a positive Δ indicates discrimination against the protected category if y_i is the undesired outcome or discrimination in favour of the protected category if y_i is the desired outcome. The authors suggest fixing a threshold $\tau \in [0,1]$ to declare the individual i discriminated if $\Delta_i \geq \tau$, where τ. To prevent discrimination the dataset is changed by altering the value of y_i for those units fo which $\Delta_i > \tau$. τ is then a tuning parameter which regulates the trade off between residual discrimination and accuracy.

In what follows we focus on discrimination against the protected group, and so we compute a different version of the measure Δ:

$$\delta_i = \frac{1}{k}\#\{j|j \neq i, x_j \in U_i^{k,1}, y_j = 0\} - \frac{1}{k}\#\{j|j \neq i, x_j \in U_i^{k,0}, y_j = 0\} \qquad (2)$$

3 CEM based discrimination measure

Coarsened Exact Matching (CEM, [4, 3]) is based on coarsening continuous variables and match a treated unit to those untreated units which are equal with respect to the coarsened continuous variables and the categorical ones. In order to obtain a reliable estimate of the causal effect, CEM algorithm discards those units which can not be matched. Here, the objective is different in that we are not interested in a global estimate of the effect of the S, but rather whether unit i has experienced a different outcome because of it possesses the S: a comparison of unit i outcome with the outcome of units matched by CEM is a suitable measure of discrimination. In particular, let $\bar{y}_i^{(S=0)}$ be the relative frequency of positive outcomes among units matched with unit i (which possesses the s.a.) and not possessing the s.a., then $D_i = y_i - \bar{y}_i^{(S=0)}$ is a measure of discrimination which takes negative values when a unit is discriminated against and positive values when the unit is favoured (positive discrimination), so that the value of D_i is bounded between -1 and 1. However, in a standard implementation of CEM algorithm not all units are matched, which makes it unsuitable for our purpose. A possible strategy to exploit CEM technique is to apply it sequentially as follows. Suppose that all units are matched using k variables, it is possible that once an additional variable is considered in matching some units remain unmatched, we then measure the discrimination for those units based on the matching with k variables. Starting with an initial matching on 0 variables, that is, $D_i^{(0)} = y_i - \bar{y}^{(S=0)}$ where $\bar{y}^{(S=0)}$ is the relative frequency of the positive outcome for all units in the dataset not possessing the s.a., the sequential CEM allows to obtain a discrimination measure for all units. Clearly, the discrimination measures D_i depend on the order of addition of the variables, so the procedure is repeated for different (random) orders of addition and the final result is the average (See Fig. 2).

An alternative measure of discrimination based on CEM stratification which is more similar to (2), $\bar{D}_i$ in what follows is obtained by computing $\bar{D}_i = \bar{y}_i^{(S=1)} - \bar{y}_i^{(S=0)}$ instead of $\bar{D}_i$ (with obvious adaptations in the algorithm of Fig. 2). Note that, unlike D_i, $\bar{D}_i$ takes positive values when the unit is discriminated against (similar to δ).

4 Simulation experiment

We tested the procedure using the *adult* dataset as in [6]. The outcome is having an income greater than 50 000 USD, the sensitive attribute is being non white, other

```
Let
    • $x_1, \dots, x_K$ be the available variables;
    • let $\mathscr{M}^{(j_1,\dots,j_h)}$ be the set of units matched by CEM performed using variables $x_{j_1}, \dots, x_{j_h}$
      and let $C_i^{(j_1,\dots,j_h)}$ be the set of units belonging to the same CEM cell of unit $i$.
Repeat $M$ times

$-1$:  select a random permutation $i_1, \dots, i_K$ of $1, \dots, K$;
  0:   for all units $i$ let $D_i^{(0)} = y_i - \bar{y}$ where $\bar{y} = \hat{P}\{Y = 1|S = 0\} = \#\{i|y_i = 1, s_i = 0\}/\#\{i|s_i = 0\}$;
  $k$:     • for all $i \notin \mathscr{M}^{(i_1,\dots,i_k)}$ let $D_i^{(k)} = D_i^{(k-1)}$;
           • for all $i \in \mathscr{M}^{(i_1,\dots,i_k)}$ let $D_i^{(k)} = y_i - \bar{y}_i$ where $\bar{y}_i = \hat{P}\{Y = 1|S = 0, x_{i_1}, \dots, x_{i_k}\} =$
             $\#\{i|y_i = 1, s_i = 0, i \in C_i^{(i_1,\dots,i_k)}\}/\#\{i|s_i = 0, i \in C_i^{(i_1,\dots,i_k)}\}$;
$K+1$:  $D_i^{(i_1,\dots,i_k)} = D_i^{(K)}$.

Set the final discrimination scores as $D_i = \frac{1}{M} \sum_{i_1,\dots,i_K} D_i^{(i_1,\dots,i_k)}$.
```

Fig. 2 Pseudo-code for repeated sequential implementation of CEM.

variables in the dataset include age, years of education, working hours, type of occupation (sector, type of employer), family status, sex, capital gain and loss, native country (redefined as areas of the world). The *adult* dataset is comprised of 45222 observations, the s.a. is possessed by 6319 units (13.97%), the outcome is favorable in 13008 (28.76%), 31.38% among those not possessing the s.a., 12.68% among those possessing it. For the analysis, the dataset is split in a learning (30162 units) and test (15060 units) subsamples (same as in [6]).

To check the stability of the procedure, we performed it twice for 100 iterations and compared the results, which showed a very good agreement ($\rho > 0.99$).

In order to explore how well the proposed measures detect discrimination against the protected group (i.e. $D < 0$, $\bar{D} > 0$ are relevant) under different circumstances we considered three different scenarios: *(a)* presence of a variable which is related to S but unrelated to Y (Fig. 1*(e)*); *(b)* discrimination free data (simulated); *(c)* discriminating data (simulated).

It is expected that if a variable we condition on is related to the s.a., then the level of conditional discrimination be lower. We added to the data a variable having a mild correlation with S (independent of the outcome). In Fig. 3*(a)* we compare the discrimination scores D and $\bar{D}$ (first and second row respectively) estimated by the procedure with and without the correlated variable: as expected, the discrimination scores are higher if the variable is omitted ($D_i^{(-p)}$, y-axis).

Further, we modified the dataset by adding and removing discrimination against the protected class to assess the sensitivity of the discrimination scores. Starting from a classification tree estimate of the probability of a positive outcome based on all variables but the S, we built a discrimination free dataset by simulating the outcome according to the estimated probabilities and a strong discrimination dataset by changing the outcome of 2200 units for which the probability of positive outcome is between 0.3 and 0.7 (in particular, we changed the outcome to negative for 200 units having the S and a positive outcome, and we changed the outcome to positive

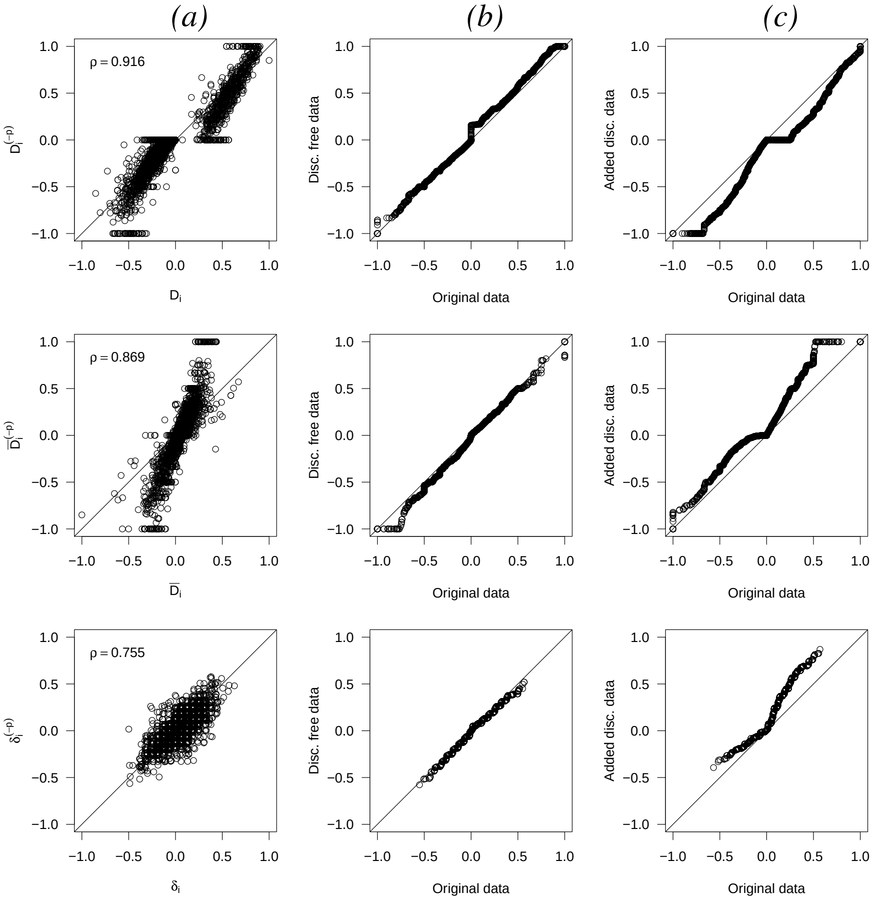

Fig. 3 Discrimination measures D_i, $\bar{D}_i$, δ_i (top, middle, bottom row respectively) performances: *(a)* scatter plots comparing discrimination measures obtained by including or omitting a conditioning variable simulated independently of the outcome to be correlated to S; *qq*-plots comparing the distributions of discrimination measures when discrimination is eliminated *(b)* and added *(c)*.

for 2000 units not having the S and a negative outcome). In Fig. 3*(b)* and *(c)* we compare the distributions of the discrimination measures using *qq*-plots.

Table 1 Classification accuracy after discrimination reductions.

% of S obs. changed	accuracy on testing data	accuracy on non S	accuracy on S
0	85.76	85.06	90.10
5	85.90	85.17	90.43
10	85.55	84.92	89.47
15	85.06	84.55	88.23
25	84.81	84.51	86.65

For comparison purposes we computed the measure δ of [6] (eq. (2) on the same datasets using 8, 16 and 32 as k values. Results are reported in Fig. 3 (bottom row) for $k = 16$, the other cases give qualitatively similar outputs. With respect to D, and to a minor extent to $\bar{D}$, the measure δ shows less variability since it is always based on groups of fixed size k, while very small (CEM) strata may contribute to the values of D ($\bar{D}$). Hence δ may be less sensitive to detect discrimination. This can be seen in scenario *(b)*, where only the D measure distribution differentiates between original and discrimination free data. Moreover, note that, contrary to D and $\bar{D}$, δ is weakly affected by the presence of a variable correlated to S.

The above discrimination measures can then be used to build a discrimination free dataset by changing the outcome of units with discrimination above a certain threshold. In Table 1 we report the performance of the classification (by a classification tree) after changing different numbers of units.

5 Conclusions

In order to detect inequality of treatment against protected classes in historical data it has been proposed to use the methods of causal inference to compare the treatment of statistical units belonging to the protected class to units not in the protected class which are similar with respect to the other observed characteristics. Similarity may be defined based on propensity scores or a distance metric. We argue that CEM stratification could be more apt to this task since it allows matching of units only if their characteristics are equal (possibly after coarsening of numerical variables), rather than relying on an overall distance. Preliminary results appear to confirm these expectations.

References

1. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268. ACM (2015)
2. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp. 3315–3323 (2016)
3. Iacus, S., King, G., Porro, G., et al.: CEM: software for coarsened exact matching. Journal of Statistical Software **30**(13), 1–27 (2009)
4. Iacus, S.M., King, G., Porro, G.: Multivariate matching methods that are monotonic imbalance bounding. Journal of the American Statistical Association **106**(493), 345–361 (2011)
5. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012)
6. Luong, B.T., Ruggieri, S., Turini, F.: k-nn as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 502–510. ACM (2011)
7. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review **29**(5), 582–638 (2014)