An Efficient Method to Impose Fairness in Linear Models

Michele Donini, Shai Ben-David, Massimiliano Pontil, John Shawe-Taylor

October 27, 2017

Abstract

We present a method for learning linear models which meet a fairness requirement with respect to a group of sensitive attributes. The method involves adding a set of linear constraints into the optimization problem such as SVMs or logistic regression. Preliminary experiments indicate that the method performs favorably over a state-of-the-art approach.

1 Introduction and overview

We address the question of fairness in machine learning. This question is concerned with whether the value of a sensible (binary) variable or variables are 'unfairly' influencing the outcome of a classifier. Perhaps giving an example of such a situation will make the idea intuitively clear. Consider that we wish to decide which of a set of applicants should be offered cheap health insurance. We are given a training set with examples of people insured in the past annotated by the costs arising from that insurance. Suppose the sensible variable under consideration is whether the applicant is a smoker. It would probably be necessary to take this variable into account when deciding whether to offer cheaper insurance since there are many studies that suggest correlations between smoking and ill-health. For example, if we require that an equal proportion of smokers and non-smokers were to receive cheaper insurance, it would likely be unfair on the non-smokers as some of them would have to be excluded from cheaper insurance in order to make space for smokers who were actually at a greater risk.

We consider the following model of fairness, first introduced by [2]. It defines fairness as independence between the sensible variable and the output of the classifier given the correct classification of the example. Intuitively, this says that if you know people are not going to cause additional health costs (e.g. in a validation set) then the fact that they smoke should not make them more likely to have to pay more. It does, however, allow for the possibility that smoking is a predictor of additional health costs and its being used in the classification, just not overused.

The paper [2] translates their definition of fairness into a postprocessing adaptation of the classification to mandate fairness. This has the advantage that it can be applied to any underlying classifier, but has the drawback that it can negatively impact its accuracy at the same time as invalidating any generalisation analysis that might have been obtained for the original classifier. It leaves open the question of whether the original classifier could have been selected more optimally if the fairness constraint were enforced during training rather than as a postprocessing step.

One of the main contributions of this paper is to consider applying this approach to any linear classification framework, e.g. Linear Support Vector Machines translating the requirement into an preprocessing step that enforces the fairness constraint in any linear classification algorithm in order to achieve good performance while respecting the fairness constraint.

2 Correlation with respect to the sensible features

Our enforced constraint encodes the requirement that given that the true label is positive, the underlying linear function is not correlated with the sensible feature on the validation set. We are using this as a proxy for the requirement that the value of the sensible feature does not give us information about the label predicted by the classifier. In order to formalize this requirement, we firstly introduce the following definitions.

Let V be a set of pairs $(\mathbf{x}, y)$ (e.g. the training set), where $\mathbf{x} \in \mathbb{R}^d$ is an example and $y \in \{-1, 1\}$ is its label. For each $\mathbf{x}$, the sensible feature has two different categorical values and it is encoded using a pair of bits as $\mathbf{x}_j^t = (x_{j_1}, x_{j_2}) \in \{(1,0) =: A^t, (0,1) =: B^t\}$ with $j = (j_1, j_2)$.

Then, we define the following subsets and their cardinalities: $V_1 = \{\mathbf{x}, y \in V : y = 1\}$ with cardinality $|V_1| = n$, $V_A = \{\mathbf{x}, y \in V_1 : \mathbf{x}_j = A\}$ ($|V_A| = n_A$) and $V_B = \{\mathbf{x}, y \in V_1 : \mathbf{x}_j = B\}$ ($|V_B| = n_B$).

Finally, we define the averages of the examples in V_A and V_B by $\widehat{\mathbf{x}}(A)$ and $\widehat{\mathbf{x}}(B)$, respectively.

As stated before, our goal is to enforce in our model the following property (i.e. the equal opportunity introduced in [2]):

$$\mathbb{P}[f(\mathbf{x})|\mathbf{x}_j = A, y = 1] = \mathbb{P}[f(\mathbf{x})|\mathbf{x}_j = B, y = 1]. \tag{1}$$

We relax this constraint encoding the not correlation between our linear model and the sensible feature, given the fact that the true label is positive.

In order to achieve the above goal, we enforce the following relaxed Equal Opportunity constraint:

$$\mathbb{E}[f(\mathbf{x}) \odot \mathbf{x}_j | y = 1] = \mathbb{E}[f(\mathbf{x})|y = 1] \odot \mathbb{E}[\mathbf{x}_j|y = 1]. \tag{2}$$

In the case of a linear model in the form $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$, we have that Eq. 2 is equivalent to

$$\mathbf{w} \odot \begin{pmatrix} \sum_{\mathbf{x} \in V_A} \mathbf{x} - \sum_{\mathbf{x} \in V_1} \mathbf{x} \frac{n_A}{n} \\ \sum_{\mathbf{x} \in V_B} \mathbf{x} - \sum_{\mathbf{x} \in V_1} \mathbf{x} \frac{n_B}{n} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{3}$$

Observing that $\sum_{\mathbf{x} \in V_A} \mathbf{x} = n_A \widehat{\mathbf{x}}(A)$ and $\sum_{\mathbf{x} \in V_B} \mathbf{x} = n_B \widehat{\mathbf{x}}(B)$, it is possible to simplify the equation as

$$\mathbf{w} \odot \frac{n_A n_B}{n} \begin{pmatrix} \widehat{\mathbf{x}}(A) - \widehat{\mathbf{x}}(B) \\ \widehat{\mathbf{x}}(B) - \widehat{\mathbf{x}}(A) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \mathbf{w} \perp \widehat{\mathbf{x}}(A) - \widehat{\mathbf{x}}(B) =: \mathbf{u}. \tag{4}$$

Finally, due to the particular form of our model, we can consider that

$$\mathbf{w} \perp \mathbf{u} \iff \sum_{i=1, i \notin j}^{d} w_i u_i = -w_{j_1} u_{j_1} - w_{j_2} u_{j_2}. \tag{5}$$

It is important to note that $u_{j_1} = 1$ and $u_{j_2} = -1$ by definition and – without loss of generality – we can fix $w_{j_1} = 1$. Now, we can solve the equation with respect to w_{j_2} as

$$\mathbf{w} \perp \mathbf{u} \iff 1 + \sum_{i=1, i \notin j}^{d} w_i u_i = w_{j_2}. \tag{6}$$

From this observation, we have a new fair representation for each example $\phi(\mathbf{x}) \in \mathbb{R}^{d-2}$ of our data that enforces the constraint in Eq. 2. The $d - 2$ entries of the fair transformation of $\mathbf{x}$ are defined as follow:

$$\phi(\mathbf{x})_r = x_r + \mathbb{I}_{\{\mathbf{x}_j^t = B\}} u_r, \quad \forall\, 1 \leq r \leq d, r \neq j_1, j_2. \tag{7}$$

Consequently, we can use this new representation in order to learn a fair model $\widetilde{\mathbf{w}} \in \mathbb{R}^{d-2}$ (plus an entry equals to 1 for the bias).

It is important to remark that, after this transformation, we are free to apply any standard machine learning algorithm in order to learn $\widetilde{\mathbf{w}}$ (for example by using a linear SVM). In other words, we are able to obtain a fair linear model without any other constraint and by using a representation that has 2 features less than the original one.

3 Experiments

We compare our method with a vanilla linear SVM and the Hardt method [2] (applied to the linear SVM model). At this stage, both our and Hardt methods exploit the training data (and not a validation set) in order to optimize the fairness of the final model. The code of our method is available at: `https://github.com/jmikko/fairnessML`.

We consider the following three datasets:

- *Diabetes:* 10 features (age, sex, body mass index, average blood pressure, and six blood serum measurements) obtained for 442 patients [1]. The target is a quantitative measure of disease progression one year later. In our binary task, the label is True if this measure is high (≥ 140), False otherwise.
- *Heart:* this dataset, from UCI repository [3], contains 13 attributes concerning clinical and demographic information of 270 patients. The goal is to detect the presence of the heart disease.
- *Default:* 23 features from 5000 Taiwanese credit card users (a random subset from the original dataset of 30000 users) concerning payments data and demographic information. The goal is to predict whether an individual will default on payments [4].

For all the three datasets the selected sensible feature is the gender of the person. We collect statistics concerning the classification balanced accuracy on the test set and the (empirical) difference between the True Positive Rate among the different sensible groups (i.e. groups with different value for the sensible feature). Formally, we estimate the difference in Equal Opportunity[2], namely DEO:

$$\widetilde{\mathbb{P}}[\mathbf{w}^* \cdot \mathbf{x} | \mathbf{x}_j = A, y = 1] - \widetilde{\mathbb{P}}[\mathbf{w}^* \cdot \mathbf{x} | \mathbf{x}_j = B, y = 1],$$

where $\mathbf{w}^*$ is the learned model and $\widetilde{\mathbb{P}}$ is the empirical estimation of the probability.

We selected randomly 80% of the dataset as training set and 20% for test. On the training set, we performed a 5-fold CV to select the best hyper-parameters[1]. We tested the final models on the test set. This procedure is repeated 50 times, and we reported the average and the standard deviation.

Table 3 shows our experimental results. In all the three datasets the behavior is similar. The best balanced accuracy is reached by using the vanilla Linear SVM without any fairness constraint. Consequently, the DEO value for the SVM is high, i.e. the generated models are unfair with respect to the sensible feature. Hardt method is able to obtain a good level of fairness on the training set but its generalization of the fairness on the test set is not very satisfactory. Moreover, the price of this fair model is a large decrease of the balanced accuracy. Our method obtains similar results concerning the DEO on the training set but better performance in generalization (both in balanced accuracy and DEO). From a first experimental analysis, we observe that our method penalizes the examples near to the classification surface (i.e. where the generated classifier is less confident) in order to achieve the fairness constraint.

[1] C (for both SVM and our method) in $\{0.1, 0.5, 1, 10, 100\}$, kernel can be Linear or RBF (i.e. for two examples $\mathbf{x}$ and $\mathbf{z}$, the RBF kernel is $e^{-\gamma||\mathbf{x}-\mathbf{z}||^2}$) with $\gamma \in \{0.001, 0.01, 0.1, 1\}$.

Diabetes dataset			
Method	Balanced Accuracy	DEO on test	DEO on train
Linear SVM	0.743 ± 0.039	0.258 ± 0.096	0.242 ± 0.078
Hardt	0.690 ± 0.170	0.151 ± 0.148	0.076 ± 0.145
Our method	0.705 ± 0.044	0.100 ± 0.082	0.053 ± 0.041
Heart dataset			
Method	Balanced Accuracy	DEO on test	DEO on train
Linear SVM	0.862 ± 0.011	0.298 ± 0.116	0.230 ± 0.072
Hardt	0.792 ± 0.074	0.141 ± 0.152	0.058 ± 0.037
Our method	0.857 ± 0.012	0.099 ± 0.108	0.062 ± 0.043
Default dataset			
Method	Balanced Accuracy	DEO test	DEO train
Linear SVM	0.701 ± 0.022	0.186 ± 0.012	0.163 ± 0.026
Hardt	0.695 ± 0.040	0.039 ± 0.039	0.044 ± 0.043
Our method	0.699 ± 0.033	0.027 ± 0.020	0.048 ± 0.034

Table 1: Results (average $\pm$ standard deviation) on all the three datasets, concerning balanced accuracy and the value of DEO on training and test set. The sensible feature is the gender (female=A, male=B).

References

[1] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[2] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[3] M. Lichman. UCI machine learning repository, 2013.

[4] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.