Fairness Constraints: A Mechanism for Fair Classification

Muhammad Bilal Zafar MZAFAR@MPI-SWS.ORG
Isabel Valera IVALERA@MPI-SWS.ORG
Manuel Gomez Rodriguez MANUELGR@MPI-SWS.ORG
Krishna P. Gummadi GUMMADI@MPI-SWS.ORG

Abstract

Automated data-driven decision systems are ubiquitous across a wide variety of online services, from online social networking and e-commerce to e-government. These systems rely on complex learning methods and vast amounts of data to optimize the service functionality, satisfaction of the end user and profitability. However, there is a growing concern that these automated decisions can lead to user discrimination, even in the absence of intent.

In this paper, we introduce fairness constraints, a mechanism to ensure fairness in a wide variety of classifiers in a *principled* manner. Fairness prevents a classifier from outputting predictions correlated with certain sensitive attributes in the data. We then instantiate fairness constraints on three well-known classifiers – logistic regression, hinge loss and support vector machines (SVM) – and evaluate their performance in a real-world dataset with meaningful sensitive human attributes. Experiments show that fairness constraints allow for an optimal trade-off between accuracy and fairness.

1. Introduction

Decision making processes in online services have become increasingly automated and data-driven. By automatically analyzing a vast amount of the users' historical and current online interactions, online services are able to improve their functionality, increase their users' satisfaction, and ultimately be more profitable. For example, social networking sites rely on large-scale classifiers to detect spammers (e.g. bots) and e-commerce sites leverage recommender systems to personalize products, services, information and advertisements to match their users' interests and tastes. Remarkably, automated and data-driven decision making is also increasingly used by organizations and governments to detect and eliminate systemic biases and inefficiencies in past human-driven decision making, be it when setting goals, recruiting people or selecting strategies.

However, as automated data analysis replaces human supervision and intuition in decision making and the scale of the data analyzed becomes "big", there is a growing concern from civil organizations (EFF, 2005), governments (Podesta et al., 2014), and researchers (Hardt, 2014) about a potential loss of transparency, accountability, and fairness. For example, classifiers used in online services have become large *black boxes* that leverage hundreds or thousands of features to achieve high accuracy in classifying users. As a consequence, it is difficult, if not impossible, to understand which features the classifiers use, to quantify the weight of individual features on the classifiers decisions, and to ensure that the classifiers do not discriminate particular groups of people. Moreover, since classification algorithms are trained on the historical data, if the historical data shows correlation with sensitive attributes (e.g. certain ethnic groups are favored over the others), this bias (or discrimination) will persist in future label predictions.

In this work, we focus on the design of classifiers with fairness *guarantees*. To this end, we introduce the idea of fairness constraints, which prevent a classifier from making predictions that are correlated with certain *sensitive attributes* in the data (*e.g.*, gender or race). This framework i) is readily generalizable to a variety of classifiers; ii) does not utilize sensitive attributes during test – only during training; iii) supports sensitive attributes of any nature; and iv) provides clear mechanisms to trade-off fairness and accuracy. Finally, our proposed approach achieves *optimal* classification accuracy for a variety of classifiers under fairness constraints.

Related Work. Pedreschi et al. (Pedreschi et al., 2008) were the first to consider discrimination (or lack of fairness) in data-driven decision making. In their work, they introduced a measure of discrimination for rule-based classifiers and demonstrated its effectiveness in a credit dataset, consisting of 1000 transactions representing the good/bad credit class of bank holders. Since then there have been a few pieces of work on limiting or controlling discrimination in data-driven decision making in the context of supervised learning. These studies have typically adopted one of

the two following strategies.

The first strategy consists of pre-processing (or *massaging*) the training data to limit discrimination (Hajian & Domingo-Ferrer, 2012; Hajian et al., 2011; Kamiran & Calders, 2009; 2010). In a classification task, this means either i) changing the value of the sensitive attributes, or, ii) switching the class labels of individual items in the data. However, this strategy has two main drawbacks. First, the learning algorithm is typically considered a *black box*. As a consequence, the pre-processing can lead to unexpected, unpredictable losses in accuracy. Second, preprocessing the data is an intrusive procedure, hard to justify semantically. Moreover, previous works implementing this strategy have been often restricted to categorical sensitive attributes (Kamiran & Calders, 2010; 2009) or rule-based classifiers (Hajian & Domingo-Ferrer, 2012; Hajian et al., 2011), not easily generalizable to a wider set of learning tasks.

The second strategy consists of modifying existing learning algorithms to limit discrimination (Kamiran et al., 2012; Calders et al., 2013; Cadlers & Verwer, 2010; Kamishima et al., 2011; 2013; Pedreschi et al., 2009). However, all previous work implementing this strategy share two important limitations: (i) they need the values of the sensitive attributes during training and testing, which might not be readily available; and, (ii) they consider only categorical sensitive attributes.

In this work, we adopt the second strategy and overcome the limitations described above. In particular, we will develop a framework to incorporate fairness constraints in a *principled* manner in a variety of classifiers through fairness constraints. This framework (i) does not utilize sensitive attributes during testing, only during training; (ii) supports both categorical and continuous sensitive attributes; and, (iii) provides clear mechanisms to trade-off fairness and accuracy.

2. Our approach

In this work, we assume that in a fair decision making system, the outcomes of the decision making process should not have a disproportionately large adverse impact on a protected class of people. We translate this idea to 'fairness constraints' in our learning algorithms. Our notion of fairness stems from the Doctrine of Disparate Impact (Disparate-Impact) – a U.S. law concerning discrimination in the areas of employment, housing, etc. We are specifically inspired by an application of this law – the *"80% rule"* (Biddle, 2005), which directly relates fairness with the ratio between the subjects in both the protected and non-protected groups that have been selected as positive during decision making. For example, in a decision making process involving females and males (with females being the protected group), the ratio between percentage of all females selected in positive class (x%) and percentage of all males selected in the positive class (y%) should be close to 1:1. Specifically, according to the 80% rule, the ratio x:y should be no less than 80:100.[1] Finally, we note that the protected, or sensitive, attribute is not always categorical (like gender), but can also have continuous values (e.g: salary, body mass index). In these cases, a fair classification outcome would be the one where the output class labels do not correlate with the value of sensitive attribute. Next, we will investigate how to incorporate fairness (or non-discrimination) constraints in a variety of classifiers.

In a (binary) classification task, one needs to find a mapping function $f(\mathbf{x}_i)$ between user feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ and labels $y_i \in \{0, 1\}$. To this end, one is given a training set, $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, and needs to utilize it to construct a mapping that works *well* on *unseen* data. There are many methods to construct such a mapping, so called classifier, *e.g.*, logistic regression, support vector machines. In our work, we will enforce these mappings to be fair (or non-discriminative) with respect to a collection of sensitive attributes or variables $\mathbf{z}_i$, such as gender or race, by incorporating fairness constraints. A fair mapping is such that $f(\mathbf{x}_i)$ is not correlated with the sensitive variables $\mathbf{z}_i$.

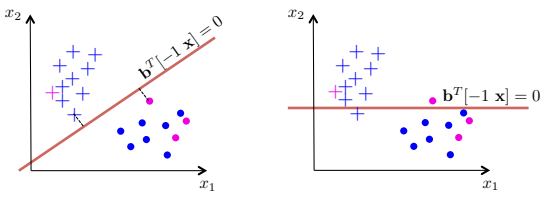
Finding this mapping often reduces to finding a hyperplane in the user feature space that separates users according to their class during training. In such scenario, one typically aims to select the hyperplane that leads to the greatest classification accuracy. However, this hyperplane may result in predicted labels that are correlated with the sensitive attributes, leading to a lack of fairness. Here, we propose to include fairness constraints during training and, as a consequence, find a hyperplane that provide an *optimal* tradeoff between fairness for accuracy, as illustrated in Figure 1. Next, we particularize our approach for three well-known classifiers: the logistic regression, the Hinge loss classifier and the SVM.

Logistic Regression. In logistic regression, one maps the feature vectors $\mathbf{x}_i$ to the labels y_i by means of a probability distribution

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}}, \qquad (1)$$

where the weights $\mathbf{b} = (b_0, \ldots, b_d)$ are obtained by applying maximum likelihood over the training set, *i.e.*, $\mathbf{b}^* = \text{argmax}_{\mathbf{b}} \sum_{i=1}^N \log p(y_i|\mathbf{x}_i))$. Then, given a feature vector $\mathbf{x}_i$, $f(\mathbf{x}_i) = 1$ if $\mathbf{b}^T[-1\ \mathbf{x}_i] \geq 0$ and $f(\mathbf{x}_i) = 0$

[1] Depending on certain factors, in some scenarios, the system administrator might stipulate that a 'fair' classification is one where the ratio x:y is no less than 50:100 (instead of 80:100). We leave the task of specifying the exact value of 'fair' ratio to the system administrator and focus on translating the given fair ratio into fairness constraints.



(a) Standard classifier. (b) Fair classifier.

Figure 1. Example of an (optimal) classifier (a) without and with (b) fairness constraints. Crosses and dots represent each class, and blue and magenta represent the value of the (binary) sensitive attribute.

otherwise. Here, we ensure fairness by solving the following constrained maximum likelihood problem:

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i) \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \geq -\mathbf{c}, \end{array} \tag{2}$$

where the fairness constraints limit the cross-covariance between the sensitive variables $\mathbf{z}_i$ and the distance to the decision boundary, $\mathbf{b}^T[-1\ \mathbf{x}_i] = 0$, which the mapping $f(\mathbf{x}_i)$ explicitly depends on. In this formulation, the constant $\mathbf{c}$ trade-offs fairness and accuracy.

Hinge Loss Classifier. The Hinge loss classifier distinguishes among classes using the linear hyperplane $\mathbf{b}^T[-1\ \mathbf{x}] = 0$, where the weights $\mathbf{b}^T$ are obtained by minimizing the Hinge loss over the training set, *i.e.*, $\mathbf{b}^* = \text{argmin}_{\mathbf{b}} \sum_{i=1}^{N} max(0, y_i(\mathbf{b}^T[-1\ \mathbf{x}_i]))$. In this case, we ensure fairness by solving the following constrained optimization problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{N} max(0, y_i(\mathbf{b}^T[-1\ \mathbf{x}_i])) \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \geq -\mathbf{c}, \end{array} \tag{3}$$

where, similarly as in logistic regression, the fairness constraints limit the cross-covariance between the sensitive variables $\mathbf{z}_i$ and the distance to the hyperplane, and $\mathbf{c}$ trade-offs fairness and accuracy.

SVM. A linear SVM classifier distinguishes among classes using also a linear hyperplane $\mathbf{b}^T[-1\ \mathbf{x}] = 0$. However, in this case, it finds the weights $\mathbf{b}$ by solving the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{b}\|^2 + C\sum_{i=1}^{N} \xi_i \\ \text{subject to} & y_i(\mathbf{b}^T[-1\ \mathbf{x}_i]) \geq 1 - \xi_i, \forall i \in \{1, \ldots, N\} \\ & \xi_i \geq 0, \forall i \in \{1, \ldots, N\}, \end{array} \tag{4}$$

where $\|\mathbf{b}\|^2$ corresponds to the margin between *support vectors* assigned to different classes, $C\sum_{i=1}^{n} \xi_i$ penalizes the number of data points that falls inside the margin. Here,

Gender	$\leq 50K$	$> 50K$
Male	68.75%	31.25%
Female	88.64%	11.36%

Table 1. Percentage of males and females belonging to each class.

Race	$\leq 50K$	$> 50K$
Amer-Indian-Eskimo	87.82%	12.18%
Asian-Pac-Islander	71.68%	28.32%
White	73.76%	26.24%
Black	87.37%	12.63%
Other	87.25%	12.75%

Table 2. Percentage people in each ethnic group belonging to each class.

we ensure fairness by solving including the fairness constraints as follows:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{b}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} & y_i(\mathbf{b}^T[-1\ \mathbf{x}_i]) \geq 1 - \xi_i, \forall i \in \{1, \ldots, n\} \\ & \xi_i \geq 0, \forall i \in \{1, \ldots, n\}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \geq -\mathbf{c}. \end{array} \tag{5}$$

where, similarly as in logistic regression and the hinge loss, the fairness constraints limit the cross-covariance between the sensitive variables $\mathbf{z}_i$ and the distance to the hyperplane, and $\mathbf{c}$ trade-offs fairness and accuracy. Note that, although for clarity we focus here on the primal form of a linear SVM, the fairness constraints can be readily added in the dual of a linear (or non-linear) SVM.

3. Evaluation

In this section, we validate the proposed classifiers in a real dataset[2], which contains a total of 45,222 subjects, each with 14 features (e.g., gender, race, age and educational level) and a binary label, which indicates whether their incomes are above (positive class) or below (negative class) 50K USD. There are 24.78% subjects in the positive class.

In our experiments, we assume the gender and race are the sensitive variables and investigate to which extent fairness constraints can tradeoff between accuracy and fairness. Tables 1 and 2 show the percentage of people of each gender and race belonging to each class in the original data. Here, we observe that the percentage of women in the positive class (earning more than 50K USD) is much lower than the percentage of males and the percentage of Asian and Pacific islander and white people in the positive class is much higher than others.

Experimental Setup. We start by training the three classifiers described in Section 2 without fairness constraints and computing the cross-covariance between

[2]http://mlr.cs.umass.edu/ml/datasets/Adult

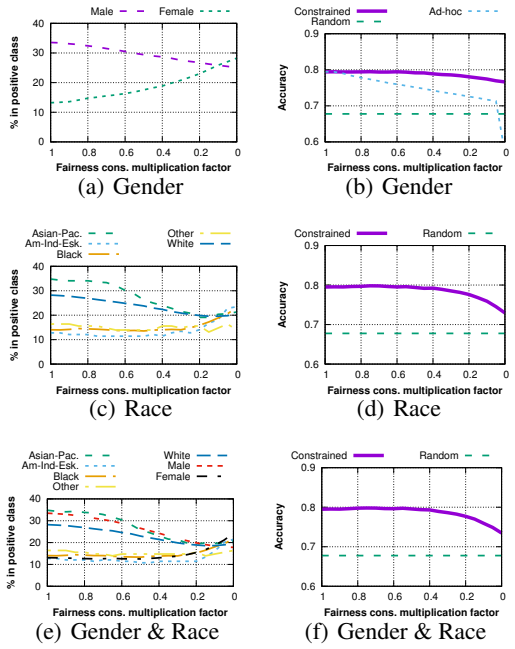


Figure 2. Logistic Regression. Trade-off between fairness and accuracy of predictions.

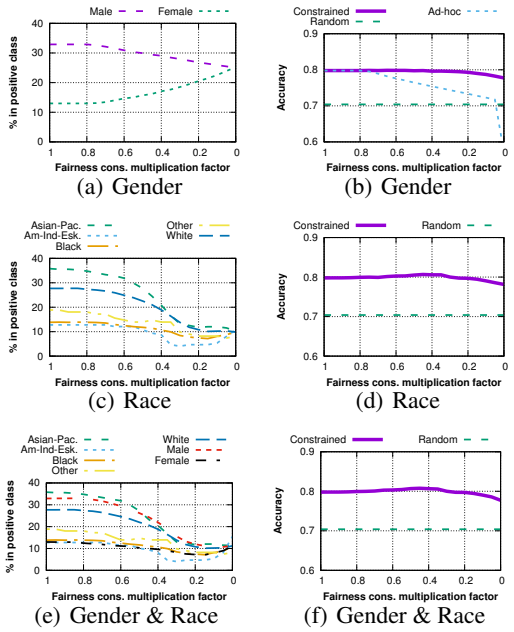


Figure 3. Hinge Loss. Trade-off between fairness and accuracy of predictions.

the predicted labels and the sensitive variables, $\mathbf{c}^* = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\,\mathbf{b}^{*T}[-1\ \mathbf{x}_i]$, in the training set. Then, we train the three classifiers with fairness constraints for decreasing values of $\mathbf{c}$, starting from the value of the cross-covariance computed on the classifiers without fairness constraints ($\mathbf{c}^*$). Specifically, we multiply $\mathbf{c}^*$ with a range of numbers from 1.0 to 0.0 to get a decreasing value of $\mathbf{c}$. We performed experiments considering the gender, the race, and both gender and race as sensitive variables.[3] Since our dataset is imbalanced, the fairness constraints always resulted in an increase in the number of subjects assigned to the negative class, because this was the optimal solution in terms of accuracy. However, in practice, this may be undesirable and we avoided this by penalizing mis-classifications in the positive class 2x more than misclassifications in the negative class.

We compare our approach with two baselines: i) a 'random' classifier, which takes the predicted labels given by the corresponding classifier, without any fairness constraints, and shuffles them randomly until the cross-covariance is zero; and ii) a 'ad-hoc' classifier, which takes the predicted labels given by the original classifier, without fairness constraints, and change the value of the predicted labels of the subjects in the discriminated group (i.e., female) to satisfy the given fairness threshold $\mathbf{c}$. We compared to the 'ad-hoc' classifier only for gender, since it was not clear how to easily implement such baseline for sensitive variables with more than two categories.

[3]For race, we introduced five different fairness constraints, one per race group, where z is a binary variable that indicates whether a subject belongs to a race group.

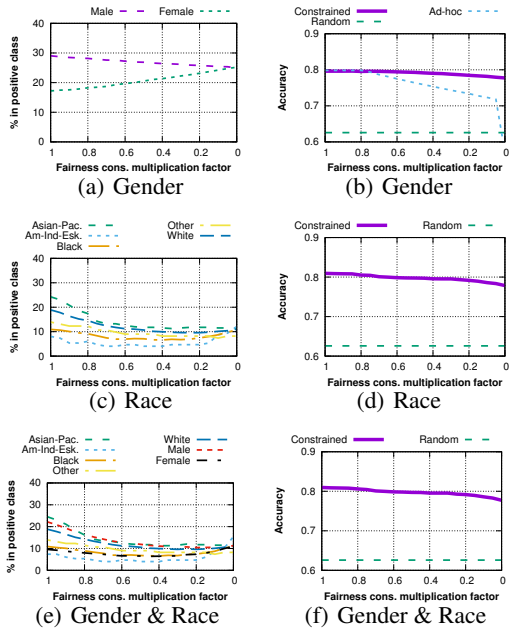
Results.

Figures 2, 3 and 4 summarize the results for the three considered classifiers, where we compare our approach with the random and ad-hoc classifiers describe above. In the figures, we show both accuracy (right) and the percentage of subjects with each sensitive variable value in the positive class (left) for different fairness constraints constants values, $\mathbf{c}$. As one may expect, as we increase fairness (i.e., decrease $\mathbf{c}$), the percentage of people in the positive class under each sensitive attribute value becomes similar and the accuracy decreases. However, for any given level of fairness, our approach always beats the baselines, achieving the highest accuracy.

4. Conclusions

In this paper, we introduced fairness constraints, a mechanism to ensure fairness in a wide variety of classifiers in a principled way, achieving an optimal trade-off between accuracy and fairness. The proposed approach can be readily applied to a wide variety of classifiers, supports binary, categorical and continuous sensitive attributes and does not require the value of the sensitive attributes in the decision making.

There are many interesting venues for future work. For example, it would be interesting to include fairness constraints in other supervised learning problems, such as regression or recommendation, and unsupervised learning problems, such as set selection or ranking problems, as well as validating our framework in other datasets.



(a) Gender (b) Gender (c) Race (d) Race (e) Gender & Race (f) Gender & Race

Figure 4. SVM. Trade-off between fairness and accuracy of predictions.

References

Biddle, D. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Gower, 2005. ISBN 9780566086540. URL https://books.google.de/books?id=q7zZ8h5X3nQC.

Cadlers, Toon and Verwer, Sico. Three Naive Bayes Approaches for Discrimination-Free Classification. 2010.

Calders, Toon, Karim, Asim, Kamiran, Faisal, Ali, Wasif, and Zhang, Xiangliang. Controlling Attribute Effect in Linear Regression. In *ICDM*, 2013.

Disparate-Impact. http://en.wikipedia.org/wiki/Disparate_impact.

EFF. https://www.eff.org/deeplinks/2005/06/websites-invade-your-privacy-charge-you-more, 2005.

Hajian, Sarah and Domingo-Ferrer, Josep. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE TKDE*, 2012.

Hajian, Sarah, Domingo-Ferrer, Josep, and Martinez-Balleste, Antoni. Rule Protection for Indirect Discrimination Prevention in Data Mining. In *MDAI*, 2011.

Hardt, Moritz. How big data is unfair: Understanding sources of unfairness in data driven decision making. *Medium*, 2014.

Kamiran, Faisal and Calders, Toon. Classifying without Discriminating. In *IC4*, 2009.

Kamiran, Faisal and Calders, Toon. Classification with No Discrimination by Preferential Sampling. In *BENELEARN*, 2010.

Kamiran, Faisal, Karim, Asim, and Zhang, Xiangliang. Decision Theory for Discrimination-aware Classification. In *ICDM*, 2012.

Kamishima, Toshihiro, Akaho, Shotaro, Asoh, Hideki, and Sakuma, Jun. Fairness-aware Classifier with Prejudice Remover Regularizer. In *PADM*, 2011.

Kamishima, Toshihiro, Akaho, Shotaro, and Sakuma, Jun. Efficiency Improvement of Neutrality-Enhanced Recommendation. In *RecSys*, 2013.

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Pedreschi, Dino, Ruggieri, Salvatore, and Turini, Franco. Discrimination-aware Data Mining. In *KDD*, 2008.

Pedreschi, Dino, Ruggieri, Salvatore, and Turini, Franco. Measuring Discrimination in Socially-Sensitive Decision Records. In *SIAM*, 2009.

Podesta, John, Pritzker, Penny, Moniz, Ernest, Holdren, John, and Zients, Jeffrey. Big data: Seizing opportunities, preserving values. *Executive Office of the President. The White House.*, 2014.