Taking Advantage of Multitask Learning for Fair Classification

Abstract

A central goal of algorithmic fairness is to reduce bias in automated decision making. An unavoidable tension exists between accuracy gains obtained by using sensitive information (e.g., gender or ethnic group) as part of a statistical model, and any commitment to protect these characteristics. Often, due to biases present in the data, using the sensitive information in the functional form of a classifier improves classification accuracy. In this paper we show how it is possible to get the best of both worlds: optimize model accuracy and fairness without explicitly using the sensitive feature in the functional form of the model, thereby treating different individuals equally. Our method is based on two key ideas. On the one hand, we propose to use Multitask Learning (MTL), enhanced with fairness constraints, to jointly learn group specific classifiers that leverage information between sensitive groups. On the other hand, since learning group specific models might not be permitted, we propose to first predict the sensitive features by any learning method and then to use the predicted sensitive feature to train MTL with fairness constraints. This enables us to tackle fairness with a three-pronged approach, that is, by increasing accuracy on each group, enforcing measures of fairness during training, and protecting sensitive information during testing. Experimental results on two real datasets support our proposal, showing substantial improvements in both accuracy and fairness.

Introduction

In recent years there has been a lot of interest in the problem of enhancing learning methods with "fairness" requirements, see (Pleiss et al. 2017; Beutel et al. 2017; Hardt, Price, and Srebro 2016; Feldman et al. 2015; Agarwal et al. 2017; 2018; Woodworth et al. 2017; Zafar et al. 2017a; Menon and Williamson 2018; Zafar et al. 2017c; Bechavod and Ligett 2018; Zafar et al. 2017b; Kamishima, Akaho, and Sakuma 2011; Kearns et al. 2017; Pérez-Suay et al. 2017; Dwork et al. 2018; Berk et al. 2017; Alabi, Immorlica, and Kalai 2018; Adebayo and Kagal 2016; Calmon et al. 2017; Kamiran and Calders 2009; Zemel et al. 2013; Kamiran and Calders 2012; 2010) and references therein. The general aim is to ensure that sensitive information (e.g. knowledge about gender or ethnic group of an individual) does not "unfairly" influence the outcome of a learning algorithm. For example, if the learning problem is to predict what salary a person should earn based on her skills and previous employment records, we would like to build a model which does not unfairly use additional sensitive information such as gender or race.

A central question is how sensitive information should be used during the training and testing phases of a model. From a statistical perspective, sensitive information can improve model performance: removing this information may result in a less accurate model, without necessarily improving the fairness of the solution, (Dwork et al. 2018; Zafar et al. 2017a; Pedreshi, Ruggieri, and Turini 2008). However, it is well known, that in some jurisdictions using different classifiers, either explicitly or implicitly, for members of different groups, may not be permitted, we refer to the remark at page 3 in (Dwork et al. 2018) and references therein. These imply that we can access the sensitive information during the training phase of a model but not during the testing phase. Our principal objective is then to optimize model accuracy while still protecting sensitive information in the data.

As a first step towards not discriminating minority groups we focus on maximizing average accuracy with respect to each group as opposed to maximizing the overall accuracy (Chouldechova 2017). For the underlying generic learning method, we consider both Single Task Learning (STL) and Independent Task Learning (ITL). While the latter independently learns a different function for each group, the former aims to learn a function that is common between all groups. A well-known weakness of these methods is that they tend to generalize poorly on smaller groups: while STL may learn a model which better represents the largest group, ITL may overfit minority groups (Baxter 2000). A common approach to overcome such limitations is offered by Multitask Learning (MTL), see (Baxter 2000; Caruana 1997; Evgeniou and Pontil 2004; Bakker and Heskes 2003; Argyriou, Evgeniou, and Pontil 2008) and references therein. This methodology leverages information between the groups (tasks) to learn more accurate models. Surprisingly, to the best of our knowledge, MTL has received little attention in the algorithmic fairness domain. We are only aware of the work (Dwork et al. 2018) which proposes to learn different classifiers per group, combined with MTL to ameliorate the issue of potentially having too little data on minority groups.

We build upon a particular instance of MTL which jointly learns a shared model between the groups as well as a specific model per group. We show how fairness constraints, measured with Equalized Odds or Equal Opportunities introduced in (Hardt, Price, and Srebro 2016), can be built in MTL directly during the training phase. This is in contrast to other approaches which impose the fairness constraint as a post-processing step (Pleiss et al. 2017; Beutel et al. 2017; Hardt, Price, and Srebro 2016; Feldman et al. 2015) or by modifying the data representation before employing stan-

dard machine learning methods (Adebayo and Kagal 2016; Calmon et al. 2017; Kamiran and Calders 2009; Zemel et al. 2013; Kamiran and Calders 2012; 2010). In many recent works (Donini et al. 2018; Agarwal et al. 2017; 2018; Woodworth et al. 2017; Zafar et al. 2017a; Menon and Williamson 2018; Zafar et al. 2017c; Bechavod and Ligett 2018; Zafar et al. 2017b; Kamishima, Akaho, and Sakuma 2011; Kearns et al. 2017; Pérez-Suay et al. 2017; Dwork et al. 2018; Berk et al. 2017; Alabi, Immorlica, and Kalai 2018; Dwork et al. 2018) it has been shown how to enforce these constraints during the learning phase of a classifier. Here we opt for the approach proposed in (Donini et al. 2018) since it is convex, theoretically grounded, and performs favorably against state-of-the-art alternatives. We present experiments on two real-datasets which demonstrate that the shared classifier learned by MTL works better than STL and in turn MTL's group specific classifiers perform better than both ITL as well as the shared MTL model. These results are in line with previous studies on MTL, which suggest the benefit offered by this methodology, see (Khosla et al. 2012; Evgeniou and Pontil 2004; Donini et al. 2016) and references therein. Moreover, we observe that the fairness constraint is effective in controlling the fairness measure.

Unfortunately, as remarked before, all the models which employ the sensitive feature in the testing phase may not be adoptable. Independent models cannot be employed since we are using different classifiers for members of different groups. Even the shared model may not be a feasible option, if the sensitive feature is used as a predictor (e.g. if the model is linear, including the sensitive feature entails using a group specific threshold). Therefore, the only feasible[1] option would be to learn a shared model based on the non-sensitive features. This constraint may limit our ability to learn classifiers of high generalization ability. In order to

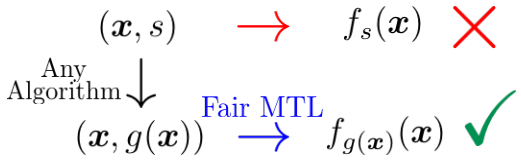

Figure 1: Our proposal in a graphical abstract: rather than using the sensitive feature s as a predictor we propose to learn, with any learning algorithm, a function g, which captures the relationship between $\boldsymbol{x}$ and s, and then use $g(\boldsymbol{x})$, instead of s, to learn group specific models via MTL.

overcome such limitations, we propose to first use the non-sensitive features to predict the value of the sensitive one and then use the predicted sensitive feature to learn group specific models via MTL. The proposal is depicted in the graphical abstract of Figure 1. We experimentally demonstrate that the proposed approach matches the classification accuracy of the best performing model which uses the sensitive information during testing, in addition to further improving upon measures of fairness.

The rest of the paper is organized as follows. We firsr present some preliminary definitions and notions concerning the fair classification framework. Then we outline the central problem that we face in the paper: exploiting the sensitive feature while still treating different groups equally. We continue with the presentation of our proposal: predicting the sensitive feature based on the non-sensitive ones and then exploiting MTL with fairness constraints in order to increase both accuracy and fairness measures (see Figure 1). Later we test the proposal on two well known fairness related datasets (Adult and COMPAS) demonstrating the potentiality of it. Finally we conclude the paper with a brief discussion.

[1]The sensitive feature may not be available in the testing phase or it might not be possible to use it as a predictor in the model due to legal requirements (Dwork et al. 2018).

Preliminaries

We let $\mathcal{D}=\{(\boldsymbol{x}_1, s_1, y_1), \ldots, (\boldsymbol{x}_n, s_n, y_n)\}$ be a training set formed by n samples drawn independently from an unknown probability distribution μ over $\mathcal{X}\times\mathcal{S}\times\mathcal{Y}$, where $\mathcal{Y}=\{-1,+1\}$ is the set of binary output labels, $\mathcal{S}=\{1,\ldots,k\}$ represents group membership, and $\mathcal{X}$ is the input space.

For every $t\in\mathcal{S}$ and operator $\diamond\in\{-,+\}$, we define the subset of training points with sensitive feature equal to t as $\mathcal{D}_t=\{(\boldsymbol{x},s,y) : (\boldsymbol{x},s,y)\in\mathcal{D}, s=t\}$ and the subset of training point negatively and positively labeled with sensitive feature equal to t as $\mathcal{D}_t^\diamond=\{(\boldsymbol{x},s,y) : (\boldsymbol{x},s,y)\in\mathcal{D}, s=t, y=\diamond 1\}$. We also let $n_t=|\mathcal{D}_t|$ and for $\diamond\in\{-,+\}$, we let $n_t^\diamond=|\mathcal{D}_t^\diamond|$.

Let us consider a function (or model) $f : \mathcal{X}\times\mathcal{S}\to\mathbb{R}$ chosen from a set $\mathcal{F}$ of possible models. The error (risk) of f is measured by a prescribed loss function $\ell : \mathbb{R}\times\mathcal{Y}\to\mathbb{R}$. The average accuracy with respect to each group of a model $L(f)$, together with its empirical counterparts $\widehat{L}(f)$, are defined respectively as

$$L(f)=\frac{1}{k}\sum_{t\in\mathcal{S}} L_t(f),\ L_t(f)=\mathbb{E}\left[\ell(f(\boldsymbol{x}),s,y)|s=t\right],\ t\in\mathcal{S},$$

and

$$\widehat{L}(f)=\frac{1}{k}\sum_{t\in\mathcal{S}}\widehat{L}_t(f),\ \widehat{L}_t(f)=1/n_t\sum_{(\boldsymbol{x},s,y)\in\mathcal{D}_t}\ell(f(\boldsymbol{x},s),y),\ t\in\mathcal{S}.$$

The fairness of the model can be measured w.r.t. many notions of fairness as mentioned in the introduction. In this work we choose to opt for the Equal Opportunity (EOp) and the Equal Odds (EOd). For $\diamond\in\{-,+\}$, the EOp$^\diamond$ constraint is defined as (Hardt, Price, and Srebro 2016)

$$\begin{aligned}&\mathbb{P}\{f(\boldsymbol{x},s)>0|s=1,y=\diamond 1\}=\\&\ldots=\mathbb{P}\{f(\boldsymbol{x},s)>0|s=k,y=\diamond 1\},\end{aligned}\tag{1}$$

where $\diamond\in\{-,+\}$. The EOd, instead is just the concurrent verification of the EOp$^+$ and EOp$^-$, then $\forall\diamond\in\{-,+\}$

$$\begin{aligned}&\mathbb{P}\{f(\boldsymbol{x},s)>0 \mid s=1,y=\diamond 1\}=\\&\ldots=\mathbb{P}\{f(\boldsymbol{x},s)>0 \mid s=k,y=\diamond 1\}.\end{aligned}\tag{2}$$

Since a model f, in general, will not be able to exactly fulfill the EOp$^+$ with $\diamond\in\{-,+\}$ nor the EOd constraints we define the Difference of EOp$^\diamond$ (DEOp$^\diamond$) with $\diamond\in\{-,+\}$ as

$$\sum_{t\in\mathcal{S}}\left|\mathbb{P}\{\hat{y}=y|s=t,y=\diamond 1\}-\tfrac{1}{|\mathcal{S}|}\textstyle\sum_{t'\in\mathcal{S}}\mathbb{P}\{\hat{y}=y|s=t',y=\diamond 1\}\right|,$$

where $\hat{y}=\operatorname{sign}(f(\boldsymbol{x},s))$. Finally, the Difference of EOd (DEOd) is defined as

$$\text{DEOd}=(\text{DEOp}^+ + \text{DEOp}^-)/2.$$

Paradigm

A central problem, when learning a model f from data under fairness requirements, is that using a different classification method, or even using different weights on attributes for members of different groups may not be allowed for certain classification tasks (Dwork et al. 2018). In other words, it may not be permitted to use the sensitive feature explicitly or implicitly in the functional form of the model[2]. This means that f should be a function of $\boldsymbol{x}$ only, that is, $f(\boldsymbol{x}, s) = f(\boldsymbol{x})$.

For instance, if $\mathcal{X}{=}\mathbb{R}^d$ and the sensitive feature is encoded with a one-hot encoding, and we use a linear classifier then

$$f(\boldsymbol{x}, s) = \boldsymbol{w} \cdot \boldsymbol{x} + b_s, \quad \boldsymbol{w}{\in}\mathbb{R}^d,\ b_s{\in}\mathbb{R},$$

which is forbidden since the model involves a different bias for each of the sensitive groups. The problem is even more apparent when we use a different model per each group, namely we set

$$f(\boldsymbol{x}, s) = \boldsymbol{w}_s \cdot \boldsymbol{x} + b_s, \quad \boldsymbol{w}_s{\in}\mathbb{R}^d,\ b_s{\in}\mathbb{R}. \tag{3}$$

Unfortunately, the above requirement can be highly constraining, resulting in a model with poor accuracy. In practice, due to bias present in the data, learning a model which involves the sensitive feature in its functional form may substantially improve model accuracy.

Our proposal to overcome the above limitation is to use the input $\boldsymbol{x}$ to predict the sensitive group s. That is, we learn a function $g : \mathcal{X}{\to}\mathcal{S}$, such that $\hat{s} = g(\boldsymbol{x})$ is the prediction of the sensitive feature of $\boldsymbol{x}$. Therefore, our method replaces the specific model $f(\boldsymbol{x}, s)$ with the composite model $h(x) \equiv f(\boldsymbol{x}, g(\boldsymbol{x}))$, thereby treating different individuals equally. Indeed if (x, t) and (x', t') are two instances, then $h(x) \approx h(x')$ provided $x \approx x'$ irrespective of the values of t and t'. Hence, we can freely use $\hat{s}$ in the functional since, during the testing phase, we do not require any knowledge of s. As we shall see, on the one hand, in the regions of the input space where the classifier g predicts well, this approach allows us to exploit MTL to learn group specific models. On the other hand, when the prediction error is high, this approach acts as a randomization procedure[3] which, as we will empirically show, improves the fairness measure of the overall model.

In this paper we investigate (i) the effect of having the sensitive feature as part of the functional form of the model, (ii) the effect of using a shared model between the groups or a different model per group, (iii) the effect of learning a shared model with STL or MTL and the effect of learning group specific models with ITL or MTL, and (iv) the effect of using the predicted sensitive feature instead of its actual value inside the functional form of the model. Then we will show that it is possible to take the best result of the different approaches with substantial benefits in terms of both model accuracy and fairness, while still treating different individuals equally.

[2]Note that, for clarity, the above limitation is imposed only when making predictions with f. During the training phase, the sensitive information can and should be used to guide the choice of model parameters.

[3]A random prediction $\hat{s}$ of s is substituted in the functional form of Eq. (3) which then randomly selects one of the group specific models, transforming the function form in a randomized shared model. Suppose we have many classifiers $f(\cdot, s)$ and a function g which chooses which classifier to use. If one assumes that $g(\boldsymbol{x})$ is purely random, then $f(\cdot, \boldsymbol{x})$ is a randomized classifier. Therefore if g has a high error rate, g is unable to predict the sensitive feature. Consequently $f(\cdot, \boldsymbol{x})$ is just a shared classifier composed of many functions chosen at random by g.

Methodology

In this section, we describe our approach to learning fair and accurate models and highlight the connection to MTL (Evgeniou and Pontil 2004). We consider the following functional form

$$f(\boldsymbol{x}, s) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, s), \quad (\boldsymbol{x}, s){\in}\mathcal{X}{\times}\mathcal{S}, \tag{4}$$

where " $\cdot$ " is the inner product between two vectors in a Hilbert space[4] $\mathbb{H}$, $\boldsymbol{w} \in \mathbb{H}$ is a vector of parameters, and $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{S} \to \mathbb{H}$ is a prescribed feature mapping[5].

We can then learn the parameter vector $\boldsymbol{w}$ by regularized empirical risk minimization, using the square Euclidean norm of the parameter vector $\|\boldsymbol{w}\|^2$ as the regularizer. The generality of this approach comes from the general form of the feature mapping $\boldsymbol{\phi} : \mathcal{X}{\times}\mathcal{S}{\to}\mathbb{H}$ which may be implicitly defined by a kernel function, see e.g. (Shawe-Taylor and Cristianini 2004; Smola and Schölkopf 2001) and references therein. In the following, first we will briefly discuss three approaches for learning the parameter vector which correspond to the three methods investigated in this paper. Then, we will explain how these methods can be enhanced with fairness constraints.

Single Task Learning. As we argued above, we may not be allowed to explicitly use the sensitive feature in the functional form of the model. A simple approach to overcome this problem, would be to train a shared model between the groups, that is, we choose $\boldsymbol{\phi}(\boldsymbol{x}, s){=}\boldsymbol{\varphi}(\boldsymbol{x})$ and $\boldsymbol{w}{=}\boldsymbol{w}_0$ in Eq. (4), where $\boldsymbol{\varphi} : \mathcal{X}{\to}\mathbb{H}$ and $\boldsymbol{w}_0{\in}\mathbb{H}$, so that $f(\boldsymbol{x}, s){=}\boldsymbol{w}_0 \cdot \boldsymbol{\varphi}(\boldsymbol{x})$ (a potentially unregularized threshold may be built in the feature map to include a bias term). We learn the model parameters by solving the Tikhonov regularization problem[6]

$$\min_{\boldsymbol{w}_0 \in \mathbb{H}} \quad \hat{L}(\boldsymbol{w}_0) + \rho\|\boldsymbol{w}_0\|^2, \tag{5}$$

where $\rho{\in}[0, \infty)$ is a regularization parameter. This method, which we will call Single Task Learning (STL), searches for the linear separator which minimizes a trade-off between the empirical average risk per group and the complexity (smoothness) of the models.

As we shall see in our experiments below, STL performs poorly, because it does not capture variance across groups. A slight variation which may improve performance is to introduce group specific thresholds. However, we remark again that this approach may not be permitted. Specifically, we choose $\boldsymbol{\phi}(\boldsymbol{x}, s){=}(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{e}_s)$ and $\boldsymbol{w}{=}(\boldsymbol{w}_0, \mathbf{b})$ where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_S$ are the canonical basis vectors in $\mathbb{R}^k$ and $\mathbf{b}{=}(b_1, \ldots, b_k){\in}\mathbb{R}^k$, so that $f(\boldsymbol{x}, s) = \boldsymbol{w}_0 \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + b_s$.

[4]For all intents and purposes, one may also assume throughout that $\mathbb{H} = \mathbb{R}^d$, the standard d-dimensional vector space, for some positive integer d.

[5]In practice, a bias term (threshold) can be added to $f(\boldsymbol{x}, s)$ (which may depend on s) but to ease our presentation we do not include it if not necessary.

[6]With a little abuse of notation we replace in the risk definitions the function with its parameter vector.

Independent Task Learning. An approach to overcome the potentially underfitting performance of STL is to learn different models for each of the groups, we refer to this approach as independent task learning (ITL). It corresponds to setting $\boldsymbol{\phi}(\boldsymbol{x},s)=(\mathbf{0}_{s-1},\boldsymbol{\varphi}(\boldsymbol{x}),\mathbf{0}_{k-s})$ and $\boldsymbol{w}=(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_k)$ in Eq. (4), where $\boldsymbol{\varphi}:\mathcal{X}\to\mathbb{H}$ and $\boldsymbol{w}_s\in\mathbb{H}\ \forall s\in\mathcal{S}$, so that $f(\boldsymbol{x},s)=\boldsymbol{w}_s\cdot\boldsymbol{\varphi}(\boldsymbol{x})$. As before, the feature map may account for a constant component to accommodate a threshold for each of the groups. To find the vectors $\boldsymbol{w}_s$ we solve k independent Tikhonov regularization problems of the form

$$\min_{\boldsymbol{w}_s\in\mathbb{H}}\quad \hat{L}_s(\boldsymbol{w}_s)+\rho\|\boldsymbol{w}_s\|^2. \tag{6}$$

Note that, similar to STL, if we substitute $\hat{s}$ to s in this last functional form then the method treats members of different groups equally, since , as we mentioned before, learning independent models may not be allowed. Furthermore, we remark that from a statistical point of view, minority groups (small sample sizes) will be prone to overfitting. Nevertheless, as we shall see, ITL works better than STL in our experiments, suggesting that there is a lot of bias in the data. Still one would expect that by leveraging similarities between the groups ITL can be further improved. We discuss this next.

Multitask Learning. Let us now discuss the multitask learning approach used in the paper, which is based on regularization around a common mean (Evgeniou and Pontil 2004). We choose $\boldsymbol{\phi}(\boldsymbol{x},s)=(\boldsymbol{\varphi}(\boldsymbol{x}),\mathbf{0}_{s-1},\boldsymbol{\varphi}(\boldsymbol{x}),\mathbf{0}_{k-s})$ and $\boldsymbol{w}=(\boldsymbol{w}_0,\boldsymbol{v}_1,\ldots,\boldsymbol{v}_k)$ in Eq. (4), where $\boldsymbol{w}_0\in\mathbb{H}$ and $\boldsymbol{v}_s\in\mathbb{H}$ $\forall s\in\mathcal{S}$, so that $f(\boldsymbol{x},s)=\boldsymbol{w}_0\cdot\boldsymbol{\varphi}(\boldsymbol{x})+\boldsymbol{v}_t\cdot\boldsymbol{\varphi}(\boldsymbol{x})$. MTL jointly learns a *shared* model $\boldsymbol{w}_0$ as well as *task specific* models $\boldsymbol{w}_s=\boldsymbol{w}_0+\boldsymbol{v}_s\in\mathbb{H}\ \forall s\in\mathcal{S}$ by encouraging the specific models and the shared model to be close to each other. To this end, we solve the following Tikhonov regularization problem

$$\begin{aligned}\min_{\boldsymbol{w}_0,\boldsymbol{w}_1,\ldots,\boldsymbol{w}_S\in\mathbb{H}}\quad & \theta\hat{L}(\boldsymbol{w}_0)+(1-\theta)\tfrac{1}{k}\textstyle\sum_{s=1}^{k}\hat{L}_s(\boldsymbol{w}_s)\\ &+\rho\left[\lambda\|\boldsymbol{w}_0\|^2+(1-\lambda)\tfrac{1}{k}\textstyle\sum_{s=1}^{k}\|\boldsymbol{w}_s-\boldsymbol{w}_0\|^2\right],\end{aligned} \tag{7}$$

where the parameter $\lambda\in[0,1]$ forces the dependency between shared and specific models and the parameter $\theta\in[0,1]$ captures the relative importance of the loss of the shared model and the group-specific models. This MTL approach is general enough to include STL and ITL, which are recovered by setting $\lambda=\theta=1$ and $\lambda=\theta=0$, respectively. Similar to STL and ITL, regularized group specific thresholds could be added in the shared model and in the group specific models.

Again, note that the group specific models trained by MTL may not be permitted. Likewise the shared model trained by MTL may not be permitted if we include the sensitive variable to the input. However if the sensitive variable is predicted from an external classifier and then MTL retrained with the predicted values, then this model treats different groups equally (see Figure 1).

Adding Fairness Constraints

Note that both STL, ITL and MTL problems are convex provided the the loss function used to measure the empirical errors $\hat{L}$ and $\hat{L}_s$ in Eqns. (5), (6), and (7) are convex. Since we are dealing with binary classification problems, we will use the hinge loss (see e.g. (Shalev-Shwartz and Ben-David 2014)), which is defined as $\ell(f(\boldsymbol{x},s),y)=\max\big(0,1-yf(\boldsymbol{x},s)\big)$.

In many recent papers (Pleiss et al. 2017; Beutel et al. 2017; Hardt, Price, and Srebro 2016; Feldman et al. 2015; Agarwal et al. 2017; 2018; Woodworth et al. 2017; Zafar et al. 2017a; Menon and Williamson 2018; Zafar et al. 2017c; Bechavod and Ligett 2018; Zafar et al. 2017b; Kamishima, Akaho, and Sakuma 2011; Kearns et al. 2017; Pérez-Suay et al. 2017; Dwork et al. 2018; Berk et al. 2017; Alabi, Immorlica, and Kalai 2018; Adebayo and Kagal 2016; Calmon et al. 2017; Kamiran and Calders 2009; Zemel et al. 2013; Kamiran and Calders 2012; 2010; Donini et al. 2018) it has been shown how to enforce EOp$^\diamond$ constraints for $\diamond\in\{-,+\}$, during the learning phase of the model $f\in\mathcal{F}$. Here we build upon the approach proposed in (Donini et al. 2018) since it is convex, theoretically grounded, and showed to perform favorably against state-of-the-art alternatives. To this end, we first observe that

$$\begin{aligned}&\mathbb{P}\{f(\boldsymbol{x},s)>0\mid s=t,y=\diamond 1\}\\ &=1-\mathbb{E}\left\{\ell_h(f(\boldsymbol{x},s),y)\mid s=t,y=\diamond 1\right\}\\ &=1-L_t(f),\quad t\in\mathcal{S},\end{aligned} \tag{8}$$

where $\ell_h(f(\boldsymbol{x},s),y)=[yf(\boldsymbol{x},s)\leq 0]$ is the hard loss function. Then, by substituting Eq. (8) in Eqs. (1) and (2), replacing the deterministic quantities with their empirical counterpart, and by approximating the hard loss function ℓ_h with the linear one $\ell_l=(1-yf(\boldsymbol{x},s))/2$ we have that the convex EOp$^\diamond$ constraints with $\diamond\in\{-,+\}$ is defined as follows

$$\tfrac{1}{n_1^\diamond}\textstyle\sum_{(\boldsymbol{x},s,y)\in\mathcal{D}_1^\diamond}f(\boldsymbol{x},s)=\ldots=\tfrac{1}{n_k^\diamond}\sum_{(\boldsymbol{x},s,y)\in\mathcal{D}_k^\diamond}f(\boldsymbol{x},s), \tag{9}$$

while for the EOd we just have to enforce both the EOp$^+$ and EOp$^-$ constraints.

In order to plug the constraint of Eq. (9) inside STL, ITL and MTL we first define the quantities

$$\boldsymbol{u}_t^\diamond=\tfrac{1}{n_t^\diamond}\textstyle\sum_{(\boldsymbol{x},s)\in\mathcal{D}_t^\diamond}\boldsymbol{\varphi}(\boldsymbol{x}),\quad t\in\mathcal{S},\ \diamond\in\{-,+\}. \tag{10}$$

It is then straightforward to show that if we wish to enforce the EOp$^\diamond$ constraint onto the shared model one has to add these $(k-1)$ constraints to the STL and MTL

$$\boldsymbol{w}_0\cdot(\boldsymbol{u}_1^\diamond-\boldsymbol{u}_2^\diamond)=0\ \wedge\ \ldots\ \wedge\ \boldsymbol{w}_0\cdot(\boldsymbol{u}_1^\diamond-\boldsymbol{u}_k^\diamond)=0. \tag{11}$$

We remark again that for the EOd constraints we just have to insert EOp$^+\wedge$ EOp$^-$ which means $2(k-1)$ constraints.

If, instead, we want to enforce the EOp$^\diamond$ constraint onto group specific models we have to add these $(k-1)$ constraints to the MTL and ITL

$$\boldsymbol{w}_1\cdot\boldsymbol{u}_1^\diamond=\boldsymbol{w}_2\cdot\boldsymbol{u}_2^\diamond\ \wedge\ \ldots\ \wedge\ \boldsymbol{w}_1\cdot\boldsymbol{u}_1^\diamond=\boldsymbol{w}_k\cdot\boldsymbol{u}_k^\diamond, \tag{12}$$

while for the EOd we just have to insert EOp$^+\wedge$ EOp$^-$.

Al last we note that by the representer theorem, as shown in (Donini et al. 2018), it is straightforward to derive the kernelized version of the fair STL, ITL, and MTL convex problems which can be solved with any solver, in our case CPLEX (IBM 2018).

Experiments

The aim of the experiments is to address the questions raised before. Namely, we wish to: (a) study the effect of using the sensitive feature as a way to bias the decision of a common model or to learn group specific models, (b) show the advantage of training either the shared or group specific models via MTL, and (c) show that MTL can be effectively used even when the sensitive feature is not available during testing by predicting the sensitive feature based on the non-sensitive ones.

Datasets and Setting

We employed the Adult dataset from the UCI repository[7] and the Correctional Offender Management Profilingfor Alternative Sanctions (COMPAS) dataset[8].

The Adult dataset contains 14 features concerning demographic characteristics of 45222 instances (32561 for training and 12661 for testing), 2 features, Gender (G) and Race (R), can be considered sensitive. The task is to predict if a person has an income per year that is more (or less) than $50,000\$$. Some statistics of the adult dataset with reference to the sensitive features are reported in Table 7 in the appendix.

The COMPAS dataset is constructed by the commercial algorithm COMPAS, which is used by judges and parole officers for scoring criminal defendants likelihood of reoffending (recidivism). It has been shown that the algorithm is biased in favor of white defendants based on a 2-years follow up study. This dataset contains variables used by the COMPAS algorithm in scoring defendants, along with their outcomes within two years of the decision, for over 10000 criminal defendants in Broward County, Florida. In the original data, 3 subsets are provided. We concentrate on the one that includes only violent recividism. Table 7, reports the statistics with reference to the sensitive features.

In all the experiments, we compare STL, ITL, and MTL in different settings. Specifically we test each method in the following cases: when the models use the sensitive feature (S=1) or not (S=0), when the fairness constraint is active (F=1) or not (F=0), when we consider the group specific models (D=1) or the shared model between groups (D=0), and when we use the true sensitive feature (P=1) or the predicted one (P=0). Note that when D=0 we can only compare STL with MTL, since only these two models produce a shared model between the groups, and furthermore, when D=1 we can only compare ITL with MTL, since these produce group specific models.

We collect statistics concerning the classification average accuracy per group in percentage (ACC) on the test set, difference of equal opportunities on both the positive and negative class (denoted as DEO^+ and DEO^-, respectively), and the difference of equalized odds (DEOd) of the selected model - see the preliminaries for a definition of these quantities.

We selected the best hyperparameters[9] by the two steps 10-fold cross validation (CV) procedure described in (Donini et al. 2018). In the first step, the value of the hyperparameters with highest accuracy is identified. In the second step, we shortlist all the hyperparameters with accuracy close to the best one (in our case, above 97% of the best accuracy). Finally, from this list, we select the hyperparameters with the lowest fairness measure. This validation procedure, ensures that fairness cannot be achieved by a mere modification of hyperparameter selection procedure.

Results

The results for all possible combinations described above, are reported in the appendix.

Table 1 presents the performance of the shared model trained with STL or MTL, with or without the sensitive feature as a predictor, and with or without the fairness constraint. From Table 1 it is possible to see that MTL reaches higher accuracies compared to STL while the fairness measure is mostly comparable, this means that there is a relation between the tasks which can be captured with MTL. This hypothesis is also supported by the results of Figure 2 in the appendix, in which we check how the accuracy and fairness, as measured with the EOd, varies by varying λ. Figure 2 shows that there are commonalities between the groups which increase by increasing the number of groups: the optimal parameter λ it is smaller than one when we consider the shared model (D=0) and it is larger than zeros when we consider group specific models (D=1). Moreover, the fairness constraint has a negative impact on the accuracy (less strong for MTL) whilst having a highly positive impact on fairness. Having the sensitive feature as a predictor increases accuracy, but decreases fairness measure, as expected.

Table 2 reports the case when the group specific models are trained with ITL or MTL, the same setting as Table 1. MTL notably improves both accuracy and fairness. The fairness constraints do not affect the accuracy too much, while giving remarkable improvements in fairness. ITL and MTL are not affected by not including or including the sensitive feature predictor, as expected from the theory given that the models already have already different biases. Table 3 reports a comparison between STL, ITL, and MTL on the Adult dataset, showing the accuracy on each group for the different models for the case that P=0, F=0, and S=0. These results clearly demonstrate that STL and ITL tend to generalize poorly on smaller groups, whereas MTL generalizes better. Results on COMPAS datasets are analogous.

Table 5 reports the comparison between the most accurate, fair and legal[10] model (the shared model trained with MTL, with fairness constraint, and no sensitive feature in the predictors) and the most accurate, fair and illegal model (the group specific models trained with MTL, with fairness constraint, the sensitive feature used as predictor). From the table one can note that the illegal model remarkably improves over the legal one in terms of accuracy and in some cases it is even better than the legal one in terms of fairness. Based on the result of Table 5 we would like to be able to use the 'illegal' model'. In order to do so make use of the trick described in the previous sections, namely we use the predicted sensitive feature based on the non-sensitive features, instead of the true one. For this purpose we used a Random Forests model (Breiman 2001) where we weighted the errors differently based on the group membership. Table 8 in the appendix reports the confusion matrices computed on the test set.

Finally, in Table 6 we report a comparison between the best illegal model and the same model, but for which uses we used the predicted sensitive feature, instead of the true one, both in training and in testing. Notably, Table 6 shows that using the predicted sensitive feature in place of the true one preserves the accuracy of the learned model, but with a notable improvement in fairness. In attempt to explain this phenomena, in Table 4 we report the average group accuracy for predicting the sensitive features gender and race, as a function of the distance from the group specific models separators trained with MTL on the Adult dataset. Table 4

[7] archive.ics.uci.edu/ml/datasets/adult

[8] github.com/propublica/compas-analysis

[9] The ranges of hyperparameters used in the validation procedure of STL, MTL, and ITL are $\rho \in \{10^{-6.0}, 10^{-5.5}, \ldots, 10^{+6.0}\}$ and $\lambda, \theta \in \{0, 2^{-15}, 2^{-14}, \ldots, 2^{-1}, 1-2^{-2}, \ldots, 1-2^{-15}, 1\}$.

[10] From now, for sake of simplicity, we use the word illegal (legal) to define a model which uses (not-uses), either implicitly or explicitly, the sensitive feature as part of its functional form.

		Adult Dataset														COMPAS Dataset											
		– 0 – –				STL ITL		MTL		STL ITL		MTL		STL ITL		MTL		STL ITL		MTL		STL ITL		MTL		STL ITL		MTL	
		– 1 – –																									
		P	D	F	S	ACC	DEOp$^+$	ACC	DEOp$^+$	ACC	DEOp$^-$	ACC	DEOp$^-$	ACC	DEOd	ACC	DEOd	ACC	DEOp$^+$	ACC	DEOp$^+$	ACC	DEOp$^-$	ACC	DEOp$^-$	ACC	DEOd	ACC	DEOd
G		0	0	0	0	80.2	0.11	83.4	0.13	80.4	0.09	84.3	0.12	80.3	0.10	83.6	0.13	76.1	0.15	78.1	0.12	76.3	0.14	78.0	0.11	76.2	0.13	77.3	0.10
		0	0	1	0	75.7	0.03	81.8	0.06	75.8	0.02	82.7	0.05	75.7	0.03	82.0	0.06	71.5	0.03	76.5	0.03	71.7	0.03	76.4	0.03	71.6	0.03	75.7	0.03
		0	0	1	1	78.6	0.06	82.4	0.04	78.8	0.05	83.3	0.04	78.7	0.05	82.6	0.04	74.4	0.05	77.4	0.05	74.6	0.05	77.3	0.04	74.5	0.05	76.6	0.04
R		0	0	0	0	80.3	0.08	84.2	0.07	80.5	0.07	85.1	0.06	80.4	0.08	84.4	0.07	80.2	0.09	84.2	0.08	80.4	0.08	85.1	0.07	80.3	0.09	84.4	0.08
		0	0	1	0	75.3	0.02	82.6	0.01	75.5	0.02	83.5	0.01	75.4	0.02	82.8	0.01	75.5	0.04	82.4	0.03	75.7	0.04	83.3	0.03	75.6	0.04	82.6	0.03
		0	0	1	1	78.4	0.03	83.4	0.03	78.6	0.03	84.3	0.02	78.5	0.03	83.6	0.03	78.5	0.05	83.5	0.02	78.7	0.04	84.4	0.02	78.6	0.05	83.7	0.02
G+R		0	0	0	0	80.2	0.16	84.6	0.14	80.4	0.14	85.3	0.14	80.3	0.15	84.9	0.14	80.2	0.16	84.8	0.14	80.4	0.14	85.5	0.14	80.3	0.15	85.1	0.14
		0	0	1	0	75.2	0.05	83.2	0.04	75.3	0.04	83.9	0.04	75.3	0.05	83.5	0.04	75.3	0.05	83.1	0.05	75.5	0.04	83.8	0.05	75.4	0.05	83.4	0.05
		0	0	1	1	78.5	0.05	83.9	0.05	78.7	0.04	84.6	0.05	78.6	0.05	84.2	0.05	78.6	0.06	84.1	0.04	78.8	0.05	84.7	0.04	78.7	0.06	84.3	0.04

Table 1: Shared model trained with STL and MTL with or without sensitive feature as predictor and/or fairness constraint.

		Adult Dataset														COMPAS Dataset											
		– 0 – –				STL ITL		MTL		STL ITL		MTL		STL ITL		MTL		STL ITL		MTL		STL ITL		MTL		STL ITL		MTL	
		– 1 – –																									
		P	D	F	S	ACC	DEOp$^+$	ACC	DEOp$^+$	ACC	DEOp$^-$	ACC	DEOp$^-$	ACC	DEOd	ACC	DEOd	ACC	DEOp$^+$	ACC	DEOp$^+$	ACC	DEOp$^-$	ACC	DEOp$^-$	ACC	DEOd	ACC	DEOd
G		0	1	0	0	74.5	0.18	90.0	0.14	74.7	0.15	91.0	0.13	74.6	0.17	90.2	0.14	70.7	0.19	84.5	0.15	70.9	0.17	84.4	0.14	70.8	0.16	83.6	0.13
		0	1	1	0	69.7	0.08	88.3	0.04	69.9	0.07	89.2	0.04	69.8	0.08	88.5	0.04	66.1	0.08	83.0	0.04	66.3	0.08	82.8	0.04	66.2	0.07	82.1	0.04
		0	1	1	1	69.7	0.08	88.1	0.03	69.9	0.07	89.1	0.03	69.8	0.08	88.3	0.03	66.1	0.09	82.9	0.07	66.3	0.08	82.8	0.06	66.2	0.08	82.1	0.06
R		0	1	0	0	67.4	0.13	91.8	0.10	67.6	0.11	92.8	0.08	67.5	0.13	92.0	0.10	67.3	0.12	91.7	0.08	67.5	0.11	92.7	0.07	67.4	0.12	92.0	0.08
		0	1	1	0	62.5	0.05	90.0	0.03	62.7	0.05	90.9	0.03	62.6	0.05	90.2	0.03	62.4	0.07	90.1	0.02	62.6	0.06	91.0	0.02	62.5	0.07	90.3	0.02
		0	1	1	1	62.6	0.06	90.4	0.03	62.7	0.05	91.3	0.03	62.6	0.06	90.6	0.03	62.4	0.07	90.0	0.03	62.5	0.07	91.0	0.03	62.4	0.07	90.2	0.03
G+R		0	1	0	0	64.0	0.23	91.5	0.15	64.2	0.20	92.2	0.15	64.1	0.22	91.8	0.15	64.2	0.24	91.4	0.16	64.3	0.21	92.2	0.16	64.3	0.22	91.7	0.16
		0	1	1	0	59.3	0.14	89.8	0.05	59.4	0.12	90.6	0.05	59.3	0.13	90.1	0.05	59.2	0.13	90.1	0.05	59.4	0.11	90.8	0.05	59.3	0.12	90.4	0.05
		0	1	1	1	59.2	0.13	90.0	0.05	59.4	0.11	90.8	0.05	59.3	0.12	90.3	0.05	59.4	0.13	89.9	0.05	59.5	0.11	90.7	0.05	59.5	0.12	90.3	0.05

Table 2: Group specific models with ITL and MTL with or without sensitive feature as predictor and/or fairness constraint.

Sens.	Group	D=0 STL	D=0 MTL	D=1 ITL	D=1 MTL
G	M	85.4	88.5	78.8	92.8
	F	81.2	85.9	74.2	91.0
R	W	86.7	89.8	89.7	93.2
	B	83.5	88.9	83.5	92.8
	API	82.3	87.9	65.2	92.1
	AIE	82.1	87.6	48.5	92.0
	O	81.2	86.9	47.5	92.1
G+R	W&M	87.8	92.8	85.8	94.7
	W&F	85.6	89.5	84.7	93.2
	B&M	84.4	89.9	66.3	93.2
	B&F	82.4	88.1	64.6	92.1
	API&M	83.6	89.2	67.3	93.0
	API&F	81.8	88.0	63.5	92.8
	AIE&M	83.0	88.8	50.2	92.7
	AIE&F	81.9	87.3	45.1	92.5
	O&M	81.7	88.3	50.7	93.1
	O&F	81.1	86.6	43.3	92.1

Table 3: Adult dataset: ACC of STL, ITL, and MTL when P=0, F=0, and S=0.

	Margin Distance 1/10	1	∞
G	75.4	83.3	87.3
R	69.9	80.7	84.7

Table 4: Adult dataset: accuracy in % of prediction based on the distance from the MTL separator which uses the predicted sensitive feature (see Table 6).

	P	D	F	S	MTL ACC	MTL DEOp$^+$	MTL ACC	MTL DEOp$^-$	MTL ACC	MTL DEOd
Adult Dataset										
G	0	0	1	0	81.8	0.06	82.7	0.05	82.0	0.06
	0	1	1	1	88.1	0.03	89.1	0.03	88.3	0.03
R	0	0	1	0	82.6	0.01	83.5	0.01	82.8	0.01
	0	1	1	1	90.4	0.03	91.3	0.03	90.6	0.03
G+R	0	0	1	0	83.2	0.04	83.9	0.04	83.5	0.04
	0	1	1	1	90.0	0.05	90.8	0.05	90.3	0.05
COMPAS Dataset										
G	0	0	1	0	76.5	0.03	76.4	0.03	75.7	0.03
	0	1	1	1	82.9	0.07	82.8	0.06	82.1	0.06
R	0	0	1	0	82.4	0.03	83.3	0.03	82.6	0.03
	0	1	1	1	90.0	0.03	91.0	0.03	90.2	0.03
G+R	0	0	1	0	83.1	0.05	83.8	0.05	83.4	0.05
	0	1	1	1	89.9	0.05	90.7	0.05	90.3	0.05

Table 5: The most accurate, fair and legal model (MTL shared model, with fairness constraint, no sensitive feature in the predictor) and the most accurate, fair and illegal model (MTL group specific models, with fairness constraint, sensitive feature exploited as predictor).

	P	D	F	S	MTL ACC	MTL DEOp$^+$	MTL ACC	MTL DEOp$^-$	MTL ACC	MTL DEOd
Adult Dataset										
G	0	1	1	1	88.1	0.03	89.1	0.03	88.3	0.03
	1	1	1	1	87.4	0.01	88.3	0.01	87.6	0.01
R	0	1	1	1	90.4	0.03	91.3	0.03	90.6	0.03
	1	1	1	1	89.2	0.01	90.2	0.01	89.4	0.01
G+R	0	1	1	1	90.0	0.05	90.8	0.05	90.3	0.05
	1	1	1	1	89.0	0.01	89.8	0.01	89.3	0.01
COMPAS Dataset										
G	0	1	1	1	82.9	0.07	82.8	0.06	82.1	0.06
	1	1	1	1	82.1	0.01	82.0	0.01	81.3	0.01
R	0	1	1	1	90.0	0.03	91.0	0.03	90.2	0.03
	1	1	1	1	89.0	0.01	89.9	0.01	89.2	0.01
G+R	0	1	1	1	89.9	0.05	90.7	0.05	90.3	0.05
	1	1	1	1	89.0	0.01	89.8	0.01	89.3	0.01

Table 6: Comparison between the group specific models trained with MTL, with fairness constraint, and the true sensitive feature exploited as a predictor, against the same model when the predicted sensitive feature exploited as predictor.

shows that the accuracy in predicting the sensitive feature decreases as we get closer to the separator. This can be understood as allowing the group specific model to randomize which specific classifier to use, reducing overall unfairness of the decision. Results on COMPAS dataset are analogous.

Discussion

We have presented two novel, but related, ideas in this work. Firstly, to resolve the tension between accuracy gains obtained by using a sensitive feature as part of the model, and the potential inapplicability of such an approach, we have suggested to first predict the sensitive feature based on the non-sensitive features, and then use the predicted value in the functional form of a model, allowing to treat people belonging to different groups, but having similar non-sensitive features, equally. Furthermore, we have demonstrated how the predicted sensitive feature can then used in a fairness constrained MTL framework. We confirmed the validity of the above approach empirically, giving us substantial improvements in both accuracy and fairness, compared to STL and ITL. We believe this to be a fruitful area of possible future research. Of course, a non-linear extension of the above framework would be interesting to study, although we did not notice any substantial improvements on the Adult and COMPAS datasets considered in this work. Moreover, it would be interesting to see if the above framework can be extended to include other fairness definitions, apart from the EOp and EOd that we have tested. Finally, it would be valuable to provide theoretical conditions on the data distribution for which our approach provably works.

References

Adebayo, J., and Kagal, L. 2016. Iterative orthogonal feature projection for diagnosing bias in black-box models. In *Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML)*.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; and Langford, J. 2017. A reductions approach to fair classification. In *FATML*.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Alabi, D.; Immorlica, N.; and Kalai, A. T. 2018. When optimizing nonlinear objectives is no harder than linear objectives. *arXiv preprint arXiv:1804.04503*.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3):243–272.

Bakker, B., and Heskes, T. 2003. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research* 4:83–99.

Baxter, J. 2000. A model of inductive bias learning. *Journal of artificial intelligence research* 12:149–198.

Bechavod, Y., and Ligett, K. 2018. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044v3*.

Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. In *FATML*.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems (NIPS)*.

Caruana, R. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2):153–163.

Donini, M.; Martinez-Rego, D.; Goodson, M.; Shawe-Taylor, J.; and Pontil, M. 2016. Distributed variance regularized multitask learning. In *International Joint Conference on Neural Networks*.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *NIPS*.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. D. M. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency (FAT)*.

Evgeniou, T., and Pontil, M. 2004. Regularized multitask learning. In *ACM SIGKDD international conference on Knowledge discovery and data mining*.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NIPS*.

IBM. 2018. User-Manual CPLEX 12.7.1. IBM Software Group.

Kamiran, F., and Calders, T. 2009. Classifying without discriminating. In *International Conference on Computer, Control and Communication*.

Kamiran, F., and Calders, T. 2010. Classification with no discrimination by preferential sampling. In *Machine Learning Conference*.

Kamiran, F., and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1):1–33.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *International Conference on Data Mining Workshops*.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

Khosla, A.; Zhou, T.; Malisiewicz, T.; Efros, A. A.; and Torralba, A. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision*.

Menon, A. K., and Williamson, R. C. 2018. The cost of fairness in binary classification. In *FAT*.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *ACM SIGKDD international conference on Knowledge discovery and data mining*.

Pérez-Suay, A.; Laparra, V.; Mateo-García, G.; Muñoz-Marí, J.; Gómez-Chova, L.; and Camps-Valls, G. 2017. Fair kernel learning. In *Machine Learning and Knowledge Discovery in Databases*.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *NIPS*.

Shalev-Shwartz, S., and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shawe-Taylor, J., and Cristianini, N. 2004. *Kernel methods for pattern analysis*. Cambridge University Press.

Smola, A. J., and Schölkopf, B. 2001. *Learning with Kernels*. MIT Press.

Woodworth, B.; Gunasekar, S.; Ohannessian, M. I.; and Srebro, N. 2017. Learning non-discriminatory predictors. In *Computational Learning Theory*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics*.

Zafar, M. B.; Valera, I.; Rodriguez, M.; Gummadi, K.; and Weller, A. 2017c. From parity to preference-based notions of fairness in classification. In *NIPS*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International Conference on Machine Learning*.

Appendix

Dataset Statistics

In Table 7 we reported the Adult and the COMPAS datasets statistics with reference to the sensitive features.

Adult dataset		
Sens.	Group	%
G	Male (M)	66.9
	Female(F)	33.2
R	White (W)	85.5
	Black (B)	9.6
	Asian-Pac-Islander (API)	3.1
	Amer-Indian-Eskimo (AIE)	1.0
	Other (O)	0.8
G+R	W&M	58.8
	W&F	26.7
	B&M	4.9
	B&F	4.7
	API&M	2.1
	API&F	1.1
	AIE&M	0.6
	AIE&F	0.4
	O&M	0.5
	O&F	0.3

COMPAS dataset		
Sens.	Group	%
G	Female (F)	19.34
	Male (M)	80.66
R	African-American (AA)	51.23
	Asian (A)	0.44
	Caucasian (C)	34.02
	Hispanic (H)	8.83
	Native American (NA)	0.25
	Other (O)	5.23
G+R	Female African-American	9.04
	Female Asian	0.03
	Female Caucasian	7.86
	Female Hispanic	1.48
	Female Native American	0.06
	Female Other	0.93
	Male African-American	42.20
	Male Asian	0.45
	Male Caucasian	26.16
	Male Hispanic	7.40
	Male Native American	0.19
	Male Other	4.30

Table 7: Adult and COMPAS datasets: statistics with reference to the sensitive features.

Predicting the sentitive feature

Table 8 reports confusion matrices in percentage (true class in columns and predicted classes in rows) obtained by predicting Gender and Race from the other non-sensitive features using Random Forests for both Adult and COMPAS datasets.

The effect of the hyperparameter λ

In Figure 2 we check how the accuracy and fairness, as measured with the EOd, varies with λ.

Complete Set of Results

In this section we report the complete set of results. The results for all the possible combinations described in the experimental section, are reported in Table 9. In Figures 3, 4, and 5, we present a visualization of Table 9 for the Adult dataset (results are analogous for the COMPAS one). Where both the error (i.e., 1-ACC), and the EOd are normalized to be between 0 and 1, column-wise. The closer a point is to the origin, the better the result.

Adult dataset		
G	M	F
M	58.2	3.8
F	8.7	29.4

Adult dataset					
R	W	B	API	AIE	O
W	78.5	1.7	0.5	0.2	0.1
B	4.6	7.8	0.1	0.0	0.0
API	0.5	0.0	0.8	0.0	0.0
AIE	1.5	0.1	0.0	2.6	0.0
O	0.4	0.0	0.0	0.0	0.7

COMPAS dataset		
G	M	F
M	16.7	8.6
F	2.6	72.1

COMPAS dataset						
R	AA	A	C	H	NA	O
AA	44.8	0.0	3.4	0.6	0.0	0.3
A	0.1	0.3	0.0	0.0	0.0	0.0
C	4.4	0.0	29.6	0.4	0.0	0.2
H	1.2	0.0	0.6	7.7	0.0	0.1
NA	0.0	0.0	0.0	0.0	0.2	0.0
O	0.7	0.0	0.4	0.1	0.0	4.6

Table 8: Adult (up) and COMPAS (down) datasets: confusion matrices in percentage (true class in columns and predicted classes in rows) obtained by predicting Gender and Race from the other non-sensitive features using Random Forests.

	P	D	F	S	Adult Dataset: STL ITL ACC	STL ITL DEOp$^+$	MTL ACC	MTL DEOp$^+$	STL ITL DEOp$^-$	MTL DEOp$^-$	STL ITL DEOd	MTL DEOd	COMPAS Dataset: STL ITL ACC	STL ITL DEOp$^+$	MTL ACC	MTL DEOp$^+$	STL ITL DEOp$^-$	MTL DEOp$^-$	STL ITL DEOd	MTL DEOd
G	0	0	0	0	80.2	0.11	83.4	0.13	80.4	0.09	84.3	0.12	80.3	0.10	83.6	0.13	76.1	0.15	78.1	0.12	76.3	0.14	78.0	0.11	76.2	0.13	77.3	0.10
	0	0	0	1	83.3	0.14	83.9	0.13	83.5	0.12	84.8	0.12	83.4	0.13	84.1	0.13	79.3	0.15	79.2	0.13	79.5	0.14	79.1	0.12	79.4	0.13	78.4	0.11
	0	0	1	0	75.7	0.03	81.8	0.06	75.8	0.02	82.7	0.05	75.7	0.03	82.0	0.06	71.5	0.03	76.5	0.03	71.7	0.03	76.4	0.03	71.6	0.03	75.7	0.03
	0	0	1	1	78.6	0.06	82.4	0.04	78.8	0.05	83.3	0.04	78.7	0.05	82.6	0.04	74.4	0.05	77.4	0.05	74.6	0.05	77.3	0.04	74.5	0.05	76.6	0.04
	0	1	0	0	74.5	0.18	90.0	0.14	74.7	0.15	91.0	0.13	74.6	0.17	90.2	0.14	70.7	0.19	84.5	0.15	70.9	0.17	84.4	0.14	70.8	0.16	83.6	0.13
	0	1	0	1	74.6	0.17	89.7	0.14	74.7	0.15	90.7	0.13	74.7	0.16	90.0	0.14	70.9	0.19	84.5	0.14	71.1	0.18	84.4	0.13	71.0	0.16	83.6	0.12
	0	1	1	0	69.7	0.08	88.3	0.04	69.9	0.07	89.2	0.04	69.8	0.08	88.5	0.04	66.1	0.08	83.0	0.04	66.3	0.08	82.8	0.04	66.2	0.07	82.1	0.04
	0	1	1	1	69.7	0.08	88.1	0.03	69.9	0.07	89.1	0.03	69.8	0.08	88.3	0.03	66.1	0.09	82.9	0.07	66.3	0.08	82.8	0.06	66.2	0.08	82.1	0.06
	1	0	0	0	78.4	0.09	82.3	0.09	78.6	0.07	83.2	0.09	78.5	0.08	82.5	0.09	74.6	0.12	77.3	0.10	74.8	0.11	77.2	0.09	74.7	0.10	76.5	0.09
	1	0	0	1	81.7	0.13	83.1	0.08	81.9	0.11	84.0	0.07	81.8	0.12	83.3	0.08	77.6	0.13	78.1	0.09	77.8	0.12	78.0	0.09	77.7	0.11	77.3	0.08
	1	0	1	0	73.7	0.02	80.7	0.01	73.9	0.02	81.6	0.01	73.8	0.02	80.9	0.01	70.1	0.03	75.9	0.01	70.3	0.03	75.8	0.01	70.2	0.03	75.1	0.01
	1	0	1	1	76.8	0.03	81.5	0.01	77.0	0.03	82.4	0.01	76.9	0.03	81.7	0.01	73.1	0.05	76.7	0.01	73.3	0.05	76.6	0.01	73.2	0.04	75.9	0.01
	1	1	0	0	73.0	0.14	89.1	0.09	73.2	0.12	90.1	0.08	73.1	0.13	89.3	0.09	69.3	0.17	83.7	0.09	69.5	0.15	83.6	0.08	69.4	0.14	82.8	0.08
	1	1	0	1	72.8	0.15	88.9	0.10	73.0	0.13	89.9	0.09	72.9	0.14	89.1	0.10	69.3	0.15	83.7	0.10	69.5	0.14	83.6	0.09	69.4	0.13	82.9	0.09
	1	1	1	0	68.0	0.06	87.4	0.01	68.2	0.05	88.3	0.01	68.1	0.05	87.6	0.01	64.7	0.06	82.3	0.01	64.9	0.05	82.1	0.01	64.8	0.05	81.4	0.01
	1	1	1	1	68.0	0.06	87.4	0.01	68.1	0.05	88.3	0.01	68.1	0.06	87.6	0.01	64.6	0.06	82.1	0.01	64.8	0.06	82.0	0.01	64.7	0.05	81.3	0.01
R	0	0	0	0	80.3	0.08	84.2	0.07	80.5	0.07	85.1	0.06	80.4	0.08	84.4	0.07	80.2	0.09	84.2	0.08	80.4	0.08	85.1	0.07	80.3	0.09	84.4	0.08
	0	0	0	1	83.2	0.09	85.3	0.09	83.4	0.08	86.2	0.08	83.3	0.09	85.5	0.09	83.2	0.10	84.9	0.08	83.4	0.09	85.8	0.07	83.3	0.10	85.1	0.08
	0	0	1	0	75.3	0.02	82.6	0.01	75.5	0.02	83.5	0.01	75.4	0.02	82.8	0.01	75.5	0.04	82.4	0.03	75.7	0.04	83.3	0.03	75.6	0.04	82.6	0.03
	0	0	1	1	78.4	0.03	83.4	0.03	78.6	0.03	84.3	0.02	78.5	0.03	83.6	0.03	78.5	0.05	83.5	0.02	78.7	0.04	84.4	0.02	78.6	0.05	83.7	0.02
	0	1	0	0	67.4	0.13	91.8	0.10	67.6	0.11	92.8	0.08	67.5	0.13	92.0	0.10	67.3	0.12	91.7	0.08	67.5	0.11	92.7	0.07	67.4	0.12	92.0	0.08
	0	1	0	1	67.2	0.13	91.8	0.08	67.4	0.12	92.8	0.07	67.3	0.13	92.1	0.08	67.4	0.13	91.8	0.09	67.5	0.11	92.8	0.08	67.4	0.13	92.0	0.09
	0	1	1	0	62.5	0.05	90.0	0.03	62.7	0.05	90.9	0.03	62.6	0.05	90.2	0.03	62.4	0.07	90.1	0.02	62.6	0.06	91.0	0.02	62.5	0.07	90.3	0.02
	0	1	1	1	62.6	0.06	90.4	0.03	62.7	0.05	91.3	0.03	62.6	0.06	90.6	0.03	62.4	0.07	90.0	0.03	62.5	0.07	91.0	0.03	62.4	0.07	90.2	0.03
	1	0	0	0	78.5	0.07	83.2	0.04	78.7	0.06	84.1	0.04	78.6	0.07	83.4	0.04	78.4	0.08	83.3	0.06	78.6	0.07	84.2	0.05	78.5	0.08	83.5	0.06
	1	0	0	1	81.8	0.09	84.1	0.06	82.0	0.08	85.0	0.05	81.9	0.09	84.3	0.06	81.7	0.09	84.4	0.07	81.9	0.08	85.3	0.06	81.8	0.09	84.6	0.07
	1	0	1	0	73.7	0.02	81.6	0.01	73.9	0.02	82.5	0.01	73.8	0.02	81.8	0.01	73.7	0.01	81.5	0.01	73.9	0.01	82.4	0.01	73.8	0.01	81.7	0.01
	1	0	1	1	77.1	0.01	82.5	0.01	77.3	0.01	83.4	0.01	77.2	0.01	82.7	0.01	77.0	0.02	82.4	0.01	77.2	0.01	83.2	0.01	77.1	0.02	82.5	0.01
	1	1	0	0	65.8	0.12	90.8	0.06	66.0	0.11	91.8	0.05	65.9	0.12	91.0	0.06	65.5	0.12	90.8	0.05	65.7	0.11	91.8	0.05	65.6	0.12	91.0	0.05
	1	1	0	1	65.8	0.11	90.7	0.05	66.0	0.10	91.7	0.04	65.9	0.11	91.0	0.05	65.7	0.12	90.8	0.07	65.8	0.11	91.7	0.07	65.7	0.12	91.0	0.07
	1	1	1	0	61.2	0.06	89.3	0.01	61.3	0.05	90.3	0.01	61.2	0.06	89.5	0.01	60.8	0.05	89.2	0.01	61.0	0.05	90.1	0.01	60.9	0.05	89.4	0.01
	1	1	1	1	60.8	0.06	89.2	0.01	61.0	0.05	90.2	0.01	60.9	0.06	89.4	0.01	60.9	0.04	89.0	0.01	61.1	0.04	89.9	0.01	61.0	0.04	89.2	0.01
G+R	0	0	0	0	80.2	0.16	84.6	0.14	80.4	0.14	85.3	0.14	80.3	0.15	84.9	0.14	80.2	0.16	84.8	0.14	80.4	0.14	85.5	0.14	80.3	0.15	85.1	0.14
	0	0	0	1	83.1	0.18	85.7	0.16	83.4	0.16	86.4	0.16	83.3	0.17	86.0	0.16	83.3	0.18	85.5	0.16	83.5	0.15	86.2	0.16	83.4	0.16	85.8	0.16
	0	0	1	0	75.2	0.05	83.2	0.04	75.3	0.04	83.9	0.04	75.3	0.05	83.5	0.04	75.3	0.05	83.1	0.05	75.5	0.04	83.8	0.05	75.4	0.05	83.4	0.05
	0	0	1	1	78.5	0.05	83.9	0.05	78.7	0.04	84.6	0.05	78.6	0.05	84.2	0.05	78.6	0.06	84.1	0.04	78.8	0.05	84.7	0.04	78.7	0.06	84.3	0.04
	0	1	0	0	64.0	0.23	91.5	0.15	64.2	0.20	92.2	0.15	64.1	0.22	91.8	0.15	64.2	0.24	91.4	0.16	64.3	0.21	92.2	0.16	64.3	0.22	91.7	0.16
	0	1	0	1	63.9	0.24	91.7	0.16	64.0	0.21	92.4	0.16	64.0	0.23	92.0	0.16	64.1	0.23	91.5	0.15	64.3	0.20	92.2	0.15	64.2	0.22	91.8	0.15
	0	1	1	0	59.3	0.14	89.8	0.05	59.4	0.12	90.6	0.05	59.3	0.13	90.1	0.05	59.2	0.13	90.1	0.05	59.4	0.11	90.8	0.05	59.3	0.12	90.4	0.05
	0	1	1	1	59.2	0.13	90.0	0.05	59.4	0.11	90.8	0.05	59.3	0.12	90.3	0.05	59.4	0.13	89.9	0.05	59.5	0.11	90.7	0.05	59.5	0.12	90.3	0.05
	1	0	0	0	78.5	0.13	83.9	0.12	78.7	0.11	84.6	0.12	78.6	0.12	84.2	0.12	78.4	0.13	83.8	0.09	78.6	0.12	84.5	0.09	78.5	0.13	84.1	0.09
	1	0	0	1	81.9	0.14	84.8	0.11	82.1	0.12	85.5	0.11	82.0	0.13	85.1	0.11	81.7	0.15	84.7	0.11	81.9	0.13	85.4	0.11	81.8	0.14	85.0	0.11
	1	0	1	0	73.7	0.01	82.3	0.01	73.9	0.01	83.0	0.01	73.8	0.01	82.6	0.01	73.6	0.02	82.5	0.01	73.8	0.02	83.1	0.01	73.7	0.02	82.8	0.01
	1	0	1	1	76.8	0.04	83.1	0.01	76.9	0.04	83.8	0.01	76.8	0.04	83.4	0.01	76.8	0.04	83.2	0.01	77.0	0.04	83.8	0.01	76.9	0.04	83.4	0.01
	1	1	0	0	62.5	0.21	90.8	0.12	62.6	0.19	91.5	0.12	62.5	0.20	91.1	0.12	62.5	0.21	90.6	0.11	62.6	0.18	91.4	0.11	62.6	0.20	91.0	0.11
	1	1	0	1	62.3	0.21	90.8	0.11	62.5	0.18	91.5	0.11	62.4	0.20	91.1	0.11	62.5	0.22	90.7	0.11	62.6	0.19	91.4	0.11	62.6	0.20	91.0	0.11
	1	1	1	0	57.7	0.10	89.1	0.02	57.9	0.08	89.9	0.02	57.8	0.09	89.5	0.02	57.8	0.11	89.2	0.01	58.0	0.10	89.9	0.01	57.9	0.11	89.5	0.01
	1	1	1	1	57.7	0.10	89.0	0.01	57.9	0.09	89.8	0.01	57.8	0.10	89.3	0.01	57.7	0.10	89.0	0.01	57.8	0.08	89.8	0.01	57.7	0.09	89.3	0.01

Table 9: Complete results set.

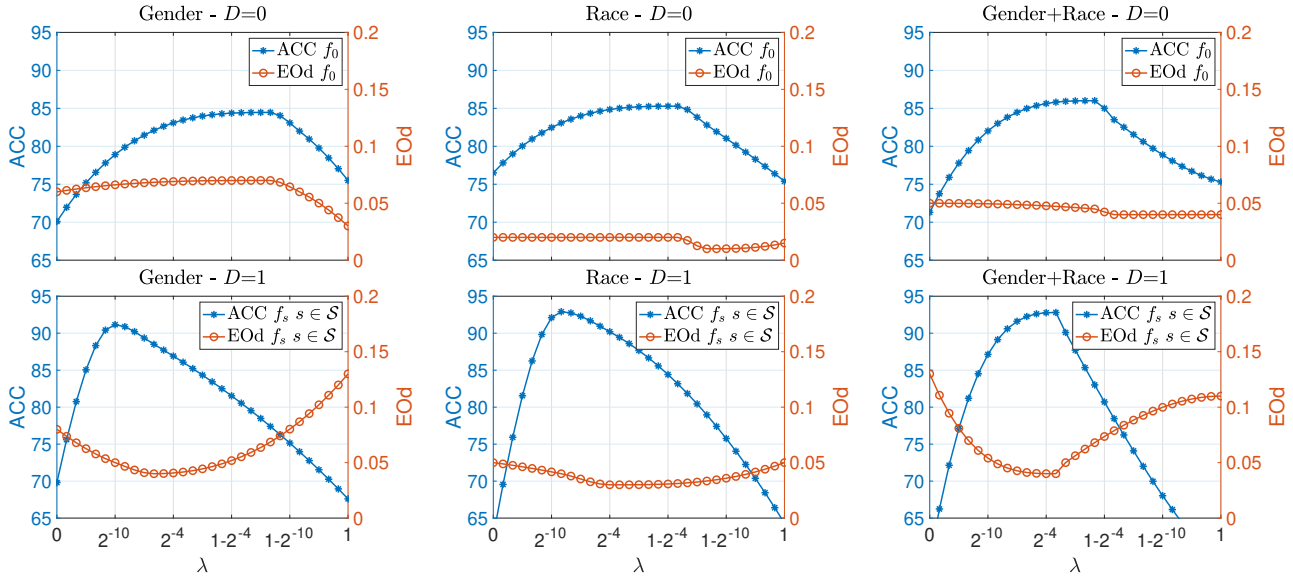

Figure 2: Adult Dataset: ACC and EOd of MTL, when we fix θ and ρ to be the best values found during the validation procedure and we vary λ with P=0, F=1, and S=0.

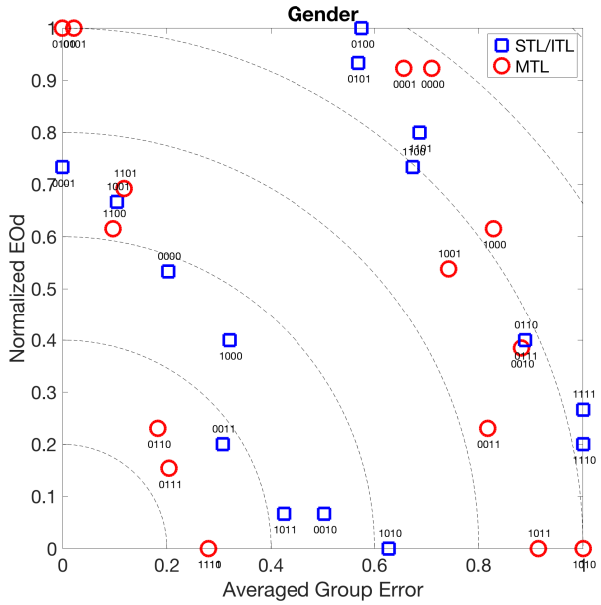

Figure 3: Adult dataset: complete results set for Gender (text close to the symbols in plot are P, D, F, and S).

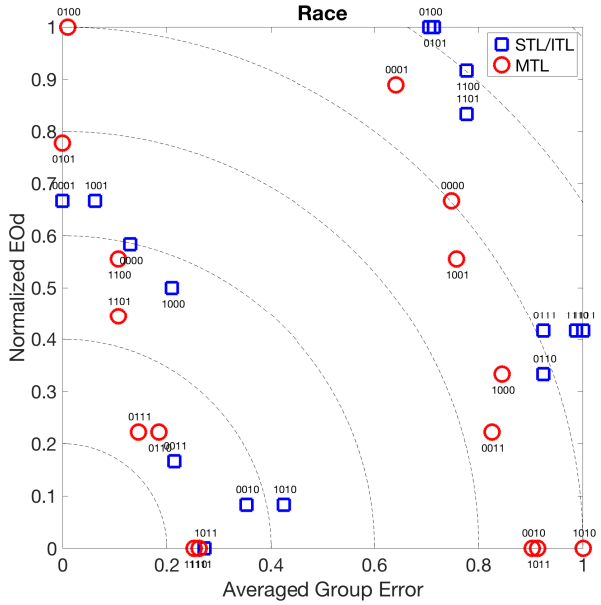

Figure 4: Adult dataset: complete results set for Race (text close to the symbols in plot are P, D, F, and S).

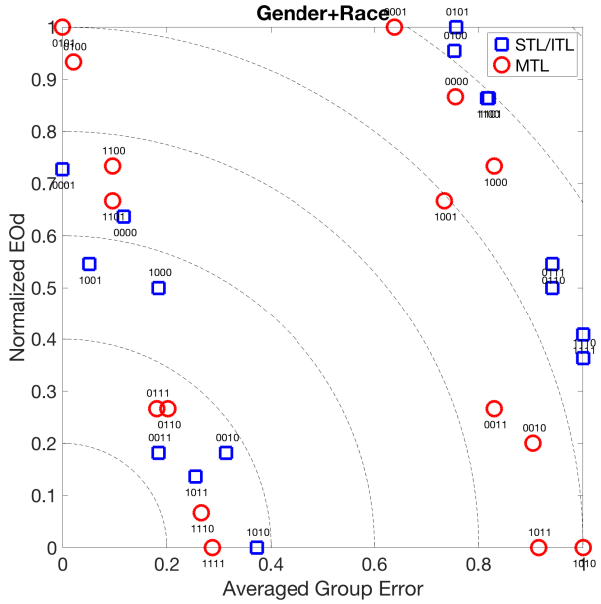

Figure 5: Adult dataset: complete results set for Gender+Race (text close to the symbols in plot are P, D, F, and S).