Learning Fair Representations

Richard Zemel ZEMEL@CS.TORONTO.EDU
Yu (Ledell) Wu WUYU@CS.TORONTO.EDU
Kevin Swersky KSWERSKY@CS.TORONTO.EDU
Toniann Pitassi TONI@CS.TORONTO.EDU
University of Toronto, 10 King's College Rd., Toronto, ON M6H 2T1 CANADA

Cynthia Dwork DWORK@MICROSOFT.COM
Microsoft Research, 1065 La Avenida Mountain View, CA. 94043 USA

Abstract

We propose a learning algorithm for fair classification that achieves both group fairness (the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole), and individual fairness (similar individuals should be treated similarly). We formulate fairness as an optimization problem of finding a good representation of the data with two competing goals: to encode the data as well as possible, while simultaneously obfuscating any information about membership in the protected group. We show positive results of our algorithm relative to other known techniques, on three datasets. Moreover, we demonstrate several advantages to our approach. First, our intermediate representation can be used for other classification tasks (i.e., transfer learning is possible); secondly, we take a step toward learning a distance metric which can find important dimensions of the data for classification.

1. Introduction

Information systems are becoming increasingly reliant on statistical inference and learning to render all sorts of decisions, including the setting of insurance rates, the allocation of police, the targeting of advertising, the issuing of bank loans, the provision of health care, and the admission of students. This growing use of automated decision-making has sparked heated debate among philosophers, policy-makers, and lawyers. Critics have voiced concerns with bias and discrimination in decision systems that rely on statistical inference and learning.

Systems trained to make decisions based on historical data will naturally inherit the past biases. These may be ameliorated by attempting to make the automated decision-maker blind to some attributes. This however, is difficult, as many attributes may be correlated with the protected one. The basic aim then is to make fair decisions, i.e., ones that are not unduly biased for or against protected subgroups in the population.

Two important goals of fair classification that have been articulated are: group fairness, and individual fairness. Group fairness, also known as statistical parity, ensures that the overall proportion of members in a protected group receiving positive (negative) classification are identical to the proportion of the population as a whole. While statistical parity is an important property, it may still lead to undesirable outcomes that are blatantly unfair to individuals, such as discriminating in employment while maintaining statistical parity among candidates interviewed by deliberately choosing unqualified members of the protected group to be interviewed in the expectation that they will fail. Individual fairness addresses this by ensuring that any two individuals who are similar with respect to a particular task should be classified similarly.

Only recently have machine learning researchers considered this issue. Several papers, e.g., (Luong et al., 2011; Kamishima et al., 2011), aim to achieve the first goal, group fairness, by adapting standard learning approaches in novel ways, primarily through a form of fairness regularizer, or by re-labeling the training data to achieve statistical parity. In a different line of work, (Dwork et al., 2011) develop an ambitious framework which attempts to achieve both group and individual fairness. In their setup, the goal is to define a

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

probabilistic mapping from individuals to an intermediate representation such that the mapping achieves both. This construction allows the initial mapping, perhaps supervised by an impartial party or regulator concerned with fairness, to produce representations of individuals that can then be used in the second step by multiple vendors to craft classifiers to maximize their own objectives, while maintaining fairness. However, there are several obstacles in their approach. First, a distance metric that defines the similarity between the individuals is assumed to be given. This may be unrealistic in certain settings, and to some extent the problem of establishing fairness in classification (more specifically simultaneously achieving the twin goals) is reduced to the problem of establishing a fair distance function. This was the most challenging aspect of their framework, as was acknowledged in their paper. Secondly, their framework is not formulated as a learning problem, as it forms a mapping for a given set of individuals without any procedure for generalizing to novel unseen data.

Our work builds on this earlier framework in that we try to achieve both group and individual fairness. However, we extend their approach in several important ways. First, we develop a learning approach to solving the fairness problem. Secondly we learn a restricted form of a distance function as well as the intermediate representation, thus making a step toward eliminating the assumption that the distance function is given apriori. Thirdly, we explicitly formulate the problem in a novel way that we feel deserves further study. Namely, we formulate fairness as an optimization problem of finding an intermediate representation of the data that best *encodes* the data (i.e., preserving as much information about the individual's attributes as possible), while simultaneously *obfuscates* aspects of it, removing any information about membership with respect to the protected subgroup. That is, we attempt to learn a set of intermediate representations to satisfy two competing goals: (i) the intermediate representation should encode the data as well as possible; and (ii) the the encoded representation is *sanitized* in the sense that it should be blind to whether or not the individual is from the protected group. We further posit that such an intermediate representation is fundamental to progress in fairness in classification, since it is composable and not ad hoc; once such a representation is established, it can be used in a *black-box* fashion to turn any classification algorithm into a fair classifier, by simply applying the classifer to the sanitized representation of the data.

The remainder of the paper is organized as follows. First we introduce our model formulation, and describe how we use it to learn fair representations. Section 3 reviews relevant work, and Section 4 presents experimental results on some standard datasets, comparing our model to some earlier ones with respect to the fairness and accuracy of the classifications.

2. Our Model

2.1. Overview and notation

The main idea in our model is to map each individual, represented as a data point in a given input space, to a probability distribution in a new representation space. The aim of this new representation is to *lose any information that can identify whether the person belongs to the protected subgroup, while retaining as much other information as possible.* Here we formulate this new representation in terms of a probabilistic mapping to a set of prototypes; note, however, that this is only one of many possible forms of intermediate representation. Finally, we also optimize these representations so that any classification tasks using them are maximally accurate.

To formalize the approach we first introduce some notation and assumptions:

- X denotes the entire data set of individuals. Each $\mathbf{x} \in X$ is a vector of length D where each component of the vector describes some attribute of the person.
- S is a binary random variable representing whether or not a given individual is a member of the protected set; we assume the system has access to this attribute.
- X_0 denotes the training set of individuals.
- $X^+ \subset X$, $X_0^+ \subset X_0$ denotes the subset of individuals (from the whole set and the training set respectively) that are members of the protected set (i.e., $S = 1$), and X^- and X_0^- denotes the subsets that are not members of the protected set, i.e., $S = 0$.
- Z is a multinomial random variable, where each of the K values represents one of the intermediate set of "prototypes". Associated with each prototype is a vector $\mathbf{v}_k$ in the same space as the individuals $\mathbf{x}$.
- Y is the binary random variable representing the classification decision for an individual, and $f : X \to Y$ is the desired classification function.
- d is a distance measure on X, e.g., simple Euclidean distance: $d(\mathbf{x}_n, \mathbf{v}_k) = ||\mathbf{x}_n - \mathbf{v}_k||_2$.

A key property that the learned mapping attempts to ensure is that membership in the protected group is lost. We formulate this using the notion of statistical parity, which requires that the probability that a random element from X^+ maps to a particular prototype is equal to the probability that a random element from X^- maps to the same prototype:

$$P(Z=k|\mathbf{x}^+ \in X^+) = P(Z=k|\mathbf{x}^- \in X^-), \forall k \quad (1)$$

Given the definitions of the prototypes as points in the input space, a set of prototypes induces a natural probabilistic mapping from X to Z via the softmax:

$$P(Z=k|\mathbf{x}) = \exp(-d(\mathbf{x}, \mathbf{v}_k)) / \sum_{j=1}^{K} \exp(-d(\mathbf{x}, \mathbf{v}_j)) \quad (2)$$

The model is thus defined as a discriminative clustering model, where the prototypes act as the clusters. Each input example is stochastically assigned to a prototype, which are in turn used to predict the class for that example. Statistical parity induces an interesting constraint on the prototype assignments, forcing the associated probabilities to be the same in expectation for the protected and unprotected groups.

2.2. Learning fair representations

The goal in our model, which we denote LFR (Learned Fair Representations), is to learn a good prototype set Z such that:

1. the mapping from X_0 to Z satisfies statistical parity;
2. the mapping to Z-space retains information in X (except for membership in the protected set); and
3. the induced mapping from X to Y (by first mapping each $\mathbf{x}$ probabilistically to Z-space, and then mapping Z to Y) is close to f.

Each of these aims corresponds to a term in the objective function we use to learn the representations. In this learning system, there are only two sets of parameters to be learned: the prototype locations $\{\mathbf{v}_k\}$ and the parameters $\{w_k\}$ that govern the mapping from the prototypes to classification decisions y.

For convenience, we use $\mathbf{x}_1, .., \mathbf{x}_N$ to denote N samples of the training set. We also use corresponding indicator variables $s_1, ..., s_N$, to denote whether $\mathbf{x}_n \in X_0^+$, $\forall n \in N$. We use $y_1, .., y_N$ as the outcome for $\mathbf{x}_1, ..., \mathbf{x}_N$ in the training set. We define $M_{n,k}$ as the probability that $\mathbf{x}_n$ maps to $\mathbf{v}_k$, via Eqn. 2:

$$M_{n,k} = P(Z=k|\mathbf{x}_n) \quad \forall n, k \quad (3)$$

Given this setup, the learning system minimizes the following objective:

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y \quad (4)$$

where A_x, A_y, A_z are hyper-parameters governing the trade-off between the system desiderata.

In order to achieve statistical parity, we want to ensure Eqn. 1, which can be estimated using the training data as:

$$M_k^+ = M_k^-, \forall k \quad (5)$$

$$M_k^+ = \mathbb{E}_{\mathbf{x} \in X^+} P(Z=k|\mathbf{x}) = \frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_{n,k} \quad (6)$$

and M_k^- is defined similarly.

Hence the first term in the objective is:

$$L_z = \sum_{k=1}^{K} \left| M_k^+ - M_k^- \right| \quad (7)$$

The second term constrains the mapping to Z to be a good description of X. We quantify the amount of information lost in the new representation using a simple squared-error measure:

$$L_x = \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2 \quad (8)$$

where $\hat{\mathbf{x}}_n$ are the reconstructions of $\mathbf{x}_n$ from Z:

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} M_{n,k} \mathbf{v}_k \quad (9)$$

These first two terms encourage the system to encode all information in the input attributes except for those that can lead to biased decisions.

The final term requires that the prediction of y is as accurate as possible:

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1-y_n) \log(1-\hat{y}_n) \quad (10)$$

Here $\hat{y}_n$ is the prediction for y_n, based on marginalizing over each prototype's prediction for Y, weighted by their respective probabilities $P(Z=k|\mathbf{x}_n)$:

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k \quad (11)$$

We constrain the w_k values to be between 0 and 1. Hence the prototype classification predictions themselves can be viewed as probabilities.

In order to allow different input features to have different levels of impact, we introduce individual weight parameters for each feature dimension, α_i, which act as inverse precision values in the distance function:

$$d(\mathbf{x}_n, \mathbf{v}_k, \alpha) = \sum_{i=1}^{D} \alpha_i (x_{ni} - v_{ki})^2 \tag{12}$$

Finally, we extend the model by using different parameter vectors α^+ and α^- for the protected and unprotected groups respectively. We optimize these parameters jointly with $\{\mathbf{v}_k\}_{k=1}^K, \mathbf{w}$, to minimize the objective; details on the optimization can be found below.

2.3. Explaining the model design choices

The first term in the objective enforces group fairness, as defined by statistical parity. We note however that L_z is not a direct encoding of the aim that the classification decisions are fair. The motivation for this indirect approach derives from our philosophy of a two-step system construction by two parties: an impartial party attempting to enforce fairness, and a vendor attempting to classify individuals. The impartial party builds mapping from individuals to new representations of individuals satisfying statistical parity, and then the vendor will be restricted to mapping the representations to outcomes. These two mappings are composed in order to obtain a fair classification of the individuals.Our learning algorithm attempts to drive L_z to zero. If L_z at test time is small, then $\sum_k |P(Z = k|S = 1) - P(Z = k|S = 0)|$, and it is not hard to show that this implies that $|P(S = 1|Z = k) - P(S = 1)|$, and $|P(S = 0|Z = k) - P(S = 0)|$ are small. Hence the mutual information between Z and S is small, and we have accomplished the goal of obsfucating information about the protected group.

Furthermore we can show that even though the parity constraint does not directly address classification, under the current model formulation the two are closely linked. The key property is that if the parity constraint is met, then the two groups are treated fairly with respect to the classification decisions:

$$\frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_{n,k} = \frac{1}{|X_0^-|} \sum_{n \in X_0^-} M_{n,k} \Rightarrow$$

$$\frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_n \mathbf{w} = \frac{1}{|X_0^-|} \sum_{n \in X_0^-} M_n \mathbf{w} \Rightarrow$$

$$\frac{1}{|X_0^+|} \sum_{n \in X_0^+} y_n^+ = \frac{1}{|X_0^-|} \sum_{n \in X_0^-} y_n^-.$$

This property follows from the linear classification approach.

Another key property of the model is the fact that the mapping to Z is defined for any individual $\mathbf{x} \in X$. This permits generalization to new examples distinct from those in the training set.

Allowing the model to adapt the weights on the input dimensions takes a step towards learning a good distance metric. The use of the same mapping function for all individuals in the group encourages individual fairness, as nearby inputs are mapped to similar representations. Adapting the weights per group allows the model some flexibility in encoding similarities between individuals within a group. The model can thus address the "inversion" problem (Dwork et al., 2011), where different qualities may be deemed important with respect to classification decisions for the two groups. For example, in one community high grades in economics may be a good predictor of success in university (and therefore correlated with admittance), whereas in another community excellence in sports may be a better predictor of success in university. The distance metric can then weight sports and economics grades appropriately for the two sets.

3. Related Work

Previous machine learning research into fair classification can be divided into two general strategies. One involves modifying the labels of the examples, i.e., the $f(X_0)$ values, so that the proportion of positive labels are equal in the protected and unprotected groups. A classifier is then trained with these new labels, assuming that equal-opportunity of positive labeling will generalize to the test set (Pedreschi et al., 2008; Kamiran & Calders, 2009; Luong et al., 2011). We term this a *data-massaging* strategy. The second type of approach, a *regularization* strategy, adds a regularizer to the classification training objective that quantifies the degree of bias or discrimination (Calders & Verwer, 2010; Kamishima et al., 2011). The system is then trained to maximize accuracy while minimizing discrimination.

A good example from the first class is that of (Kamiran & Calders, 2009), where they "massage" the training data labels to remove the discrimination with the least possible changes. The initial step involves ranking the training examples based on the posterior probabilities of positive labels obtained from a Naive-Bayes classifier trained on the original dataset. They then select the set of highest-ranked negatively-labeled items from the protected set and change their labels. The size of this set is chosen to make the proportion of positive labels equal in the two groups; the ranking approach is used to minimize the impact on the system's accuracy in predicting the classification labels. The modified data is then used for learning a classifier for future decisions. They also use Naive-Bayes to learn a classifier based on the modified dataset.

A recent example of a regularization strategy is the work of Kamishima et al (2011); they quantified the degree of prejudice based on mutual information, and added this as a regularizer in a logistic regression model. Our model more closely resembles this regularization approach, as the statistical parity is a regularizer incorporated into the classification objective. Two key differences are that fairness in LFR is defined in terms of the intermediate representation rather than the classification decisions, and that we explicitly attempt to retain information in the input attributes with the exception of membership in the protected group. This enables the intermediate representation to potentially be used for other classification decisions.

A third approach in the literature is the "Fairness Through Awareness" work (Dwork et al., 2011). Here a mapping to an intermediate representation is obtained by optimizing the classification decision criteria while satisfying a Lipschitz condition on individuals, which stipulates that nearby individuals should be mapped similarly. One important difference in our model is that our approach naturally produces out-of-sample representations, whereas this earlier work left open the question of how to utilize this fair mapping for future unseen examples.

On the computational side one notable piece of related work is the information bottleneck approach (Tishby et al., 1999). The aim in information bottleneck is to compress the information in some source variable X while preserving information about another relevant variable Y. The optimization is cast as finding a new representation that simultaneously maximizes the mutual information with Y while minimizing the information about X. Our method similarly attempts to learn a representation that trades off mutual information, and maximizes it with a relevant variable Y. However in our formulation the representation attempts to minimize mutual information with only a portion of the input (S) while maximizing the retained information about the remainder of X.

More broadly there is a large body of work on fairness in social choice theory, game theory, economics, and law. Among the most relevant are theories of fairness and algorithmic approaches to apportionment, e.g., Young's, *Equity*, Moulin's *Fair Division and Collective Welfare*, Roemer's *Equality of Opportunity and Theories of Distributed Justice*, and Rawl's *A Theory of Justice*. Concerns about the impact of classification include: maintaining a fair marketplace, bias, impedence of autonomy and identity formation, and the fear that segmented access to information undermines shared experience and therefore the informational commons considered important to democracy. Recent papers (Dwork & Mulligan, 2012; Zarsky, 2012) articulate these concerns related to classification, and point out that there are no current proposals at present to regulate or build systems addressing these concerns.

Finally, there is a close connection between individual fairness—treating similar people similarly—and *differential privacy* (Dwork et al., 2006). Differential privacy is a definition of privacy designed for privacy-preserving analysis of data; it ensures that the output of any analysis is essentially equally likely to occur on any pair of databases differing only in the data of a single individual. Thus, differential privacy requires that algorithms behave similarly on similar databases, while individual fairness requires that classification outcomes will be similar for similar individuals. Another view of this relationship is that differential privacy involves a constraint with respect to the rows of a data matrix, while fairness involves a constraint with respect to the columns. The analogy is well founded, as techniques from differential privacy may be used to achieve individual fairness (Dwork et al., 2011).

4. Experiments

4.1. Comparisons, datasets, and protocol

For comparison, we implemented the four variations of the models in (Kamiran & Calders, 2009), using Naive-Bayes in both the ranking and classification phase of their algorithm; the variants either use separate or combined classifiers in each phase. We denote this the FNB model, for Fair Naive-Bayes. We also implemented the logistic regression method with a regularizer proposed by (Kamishima et al., 2011), and optimized the setting of the regularization parameter. We denote this the RLR model, for Regularized Logistic Regression. We also trained an un-regularized version of logistic regression, denoted LR, as a baseline.

We defined a performance metric to apply to the validation set in order to determine the best variant and hyper-parameter setting for each method. We chose to focus on two measurements—the accuracy and discrimination in the model output—since both of these were considered in the earlier work on the FNB method (Kamiran & Calders, 2009) and the RLR approach (Kamishima et al., 2011). In addition, in order to evaluate individual fairness, we define a metric that assesses the consistency of the model classifications locally in input space; values close to one indicate that similar inputs are treated similarly.

- **Accuracy**: measures the accuracy of the model classification prediction:

$$yAcc = 1 - \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n| \qquad (13)$$

- **Discrimination**: measures the bias with respect to the sensitive feature S in the classification:

$$yDiscrim = |\frac{\sum_{n:s_n=1}\hat{y}_n}{\sum_{n:s_n=1}1} - \frac{\sum_{n:s_n=0}\hat{y}_n}{\sum_{n:s_n=0}1}| \qquad (14)$$

  This is a form of statistical parity, applied to the classification decisions, measuring the difference in the proportion of positive classifications of individuals in the protected and unprotected groups.

- **Consistency**: compares a model's classification prediction of a given data item $\mathbf{x}$ to its k-nearest neighbors, $kNN(\mathbf{x})$:[1]

$$yNN = 1 - \frac{1}{N}\sum_{n}|\hat{y}_n - \frac{1}{k}\sum_{j\in kNN(\mathbf{x}_n)}\hat{y}_j| \qquad (15)$$

  We applied the kNN function to the full set of examples to obtain the most accurate estimate of each point's nearest neighbors.

Here we present results of the methods on two datasets which are available from the UCI ML-repository (Frank & Asuncion, 2010). The German credit dataset has 1000 instances which classify bank account holders into credit class *Good* or *Bad*. Each person is described by 20 attributes. In our experiments we consider *Age* as the sensitive attribute, following (Kamiran & Calders, 2009). The Adult income dataset has 45,222 instances. The target variable indicates whether or not income is larger than 50K dollars, and the sensitive feature is *Gender*, as in (Kohavi, 1996; Kamishima et al., 2011)). Each datum is described by 14 attributes.

The final, considerably larger dataset we experimented with is derived from the Heritage Health Prize milestone 1 challenge (www.heritagehealthprize.com). It contained 147,473 patients, described using the same 139 features as the winning team, Market Makers. The goal is to predict the number of days a person will spend in the hospital in a given year. To convert this into a binary classification task, we simply predict whether they will spend any days in the hospital that year. We split the patients into two groups based on Age (> 65). For details on the datasets see the Supplementary Material.

All methods were trained in the same way for all datasets. For each variant of FNB, and each setting of the hyper-parameters in the other two methods, we evaluated the performance metrics on the validation set. For our method, LFR, we applied L-BFGS to minimize Eqn. 4. We performed a simple grid search to find a good set of hyper-parameters in Eqn. 4: A_x was 0.01, and we chose A_y, A_z to be the values from the set $\mathcal{S} = \{0.1, 0.5, 1, 5, 10\}$. We also included $A_z = 0$. For RLR, we optimized the regularization parameter $\eta \in \{0, 0.5, 1.0, 1.5, 3.0\}$.

4.2. Results and analysis

A key issue is what measure should be used for model selection; that is, which criteria will be used to evaluate a model's performance on the validation set with a particular setting of hyper-parameters. Here we focus on two measures. In the first the selection was based on minimizing the discrimination criteria *yDisc*, reflecting the primary aim of ensuring fairness. The second selection was based on maximizing the difference between accuracy and discrimination: $Delta = yAcc - yDisc$. In each case we compare the performance of the respective models on a test dataset, examining both the accuracy and discrimination in Y. The results are summarized in Figure 1; LR = Logistic Regression (a baseline method); FNB = Fair Naive Bayes; RLR = Regularized Logistic Regression; and LFR = Learned Fair Representations, our new model.

In these results it is clear that our model is capable of pushing the discrimination to very low values, while maintaining fairly high accuracy. The results are consistent in all three datasets, and across the validation criteria.

In particular, the Fair Naive Bayes method has difficulty in maintaining low values of discrimination at test time. It performs quite well on the Adult dataset, but its performance suffers considerably when the size of the problem increases, as in the Health dataset. The Regularized Logistic Regression has more consistent success in limiting discrimination while preserving accuracy, but still does not match our method's performance overall. This is quite surprising, since our LFR model is not directly minimizing discrimination but instead optimizing a proxy evaluated on the intermediate representations. In addition our model is also trying to preserve information about the data.

We can also compare the models with respect to individual fairness. We use the yNN measure (Eqn. 15) to evaluate the consistency of each model's classification decisions. The results for the models that were selected based on discrimination are shown in Figure 2. For each dataset our model obtained better individual fairness; this is likely due to the optimization criteria

[1] Note that the original version of this equation contained a typo, with a misplaced $1/k$ factor.

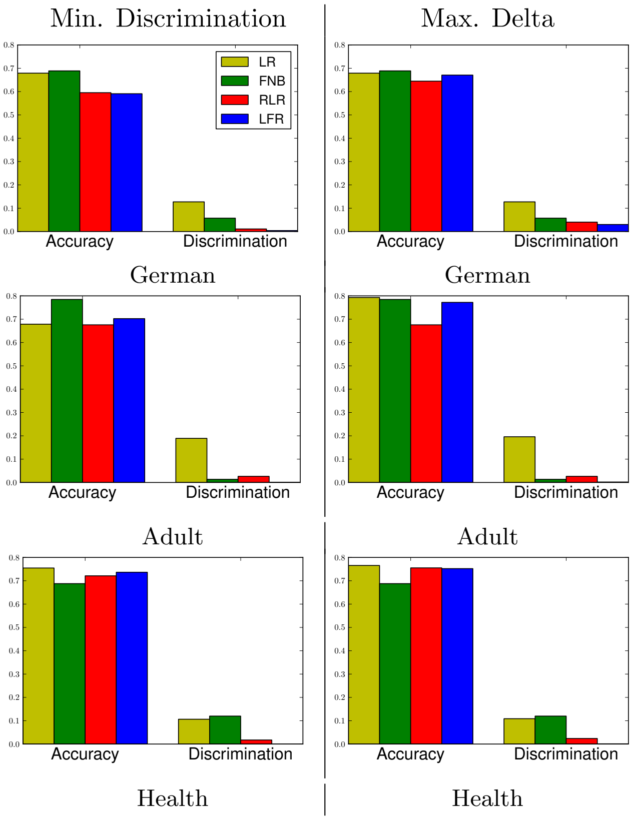

Figure 1. Results on test sets for the three datasets (German, Adult, and Health), for two different model selection criteria: minimizing discrimination and maximizing the difference between accuracy and discrimination.

rewarding Z's preservation of information about X.

For the remainder of this section, we focus on the model that maximizes the difference between accuracy and discrimination. We can gain some insight into our method by analyzing various aspects of the learned representations. First we note that a stated aim of our learning method is to encode as much information as possible about each individual while removing information about membership in the protected set. We can evaluate the degree to which the system succeeded in accomplishing this at test time, in several ways.

First we can measure how much information about S is contained in the latent representation by building a predictor that learns to predict S from Z:

$$\hat{s}_n = \sum_{k=1}^{K} M_{n,k} u_k \tag{16}$$

We optimize this predictor to minimize its difference with the actual s_n, and then evaluate test predictions for S using an $sAcc$ score analogous to $yAcc$:

$$sAcc = 1 - \frac{1}{N} \sum_{n=1}^{N} |s_n - \hat{s}_n| \tag{17}$$

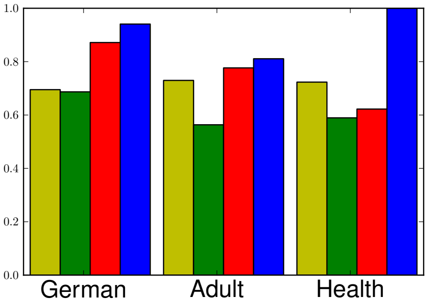

Figure 2. Individual fairness: The plot shows the consistency of each model's classification decisions, based on the yNN measure. Legend as in Figure 1.

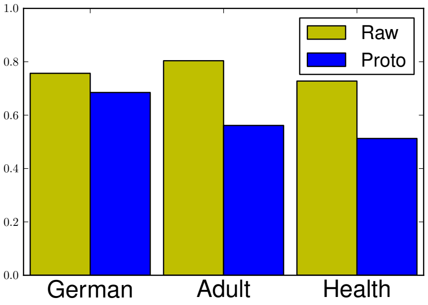

Figure 3. The plot shows the accuracy of predicting the sensitive variable ($sAcc$) for the different datasets. Raw involves predictions directly from all input dimensions except for S, while Proto involves predictions from the learned fair representations.

Note that even though we optimize the predictor in effect to maximize $sAcc$, in contrast to $yAcc$ which we want to be as close to 1 as possible, we want $sAcc$ to be low. We can evaluate how much information about S is removed by the mapping to Z by comparing $sAcc$ to its upper and lower bound. A simple upper bound (based on a linear predictor) predicts $\hat{s}$ from the rest of the input vector X except for S; a lower bound is 0.5, which corresponds to random guessing. Results for the three datasets of predicting S from the raw data (our simple upper bound) versus from the prototype representations in the trained models (as in Eqn. 16) are shown in Figure 3. In all cases the accuracy has moved significantly towards the lower bound of 0.5, showing that the information regarding learned representations has been significantly obfuscated.

This demonstrates that the sensitive information has been obfuscated, but is specific to the method of predicting S. Another evaluation involves looking directly at statistical parity in the representations at test time. An upper bound on how much information about S is contained in the new representations can be gained by finding the maximum amount of bias across the prototypes; in the Adult dataset for example, this is: $\max_k |P(Z = k|S = 1) - P(Z = k|S = 0)| = 0.0027$. Similar results were obtained for the other datasets. As discussed in Section 2.2, this implies that the mutual information between S and Z is very small, and thus Z represents a sanitized version of the input where S has been essentially forgotten.

An additional interesting aspect of the model concerns the learned distance metric. The model is capable of learning different precision parameters along each input dimension, or feature i. When we rank the features based on the magnitude of their difference between the two groups, $|\alpha_i^+ - \alpha_i^-|$, we find that the top-ranked features can reveal something interesting about the datasets. For example, in the Adult dataset, the top-ranked feature represents whether the individual is 'never married'. Here the sensitive feature is gender; the α for female is bigger than α for male on this feature. Similar results were obtained in both datasets. This shows that our algorithm can adapt the distance function to an *inverted* scenario where awareness of the protected group may be relevant to classification.

4.3. Transferring fair representations

We also investigated the ability of our model to learn multiple classifications from a common set of fair representations. This is a form of transfer learning, in which the learned fair representations may be used for several different classifications, such as by different vendors, or various decisions by the same vendor.

In order to examine this, we identified another dimension in the Adult dataset as the second classification variable. We trained an LFR model on a modified version of the dataset, in which this new dimension, Age, was removed from each input example X. Otherwise the setup was exactly as above, with Gender as the sensitive variable S and Income as the target Y. We used a validation set to again find the best setting of the hyper-parameters. After training we used the $\{M_{nk}\}$ values as the representation of each individual n, and trained a linear predictor for the new classification problem, Age.

As a benchmark, we trained a linear predictor (LR) for Age that mapped directly from the input $\mathbf{x}$. The question of interest is how much the learned representations lose in accuracy versus how much fairness is gained, relative to this direct prediction. In this experiment, we found that the transferred representations suffered a small loss in accuracy ($yAcc$ dropped $< 7\%$) while significantly removing bias in the classification ($yDiscrim$ dropped $> 97\%$).

5. Discussion

In this paper, we formulated fairness as an optimization problem of finding an intermediate representation of the data that best *encodes* the data while simultaneously *obfuscates* aspects of it, removing any information about membership with respect to the protected group. Our model maps each individual, represented as a data point in a given input space, to a probability distribution in a new representation space. Classifications can be made based on these new representations. We implemented our algorithm on three data sets and showed positive results compared to other known techniques, and conducted an initial investigation into the ability of our model to learn multiple classifications from a common set of fair representations.

Fairness in classification is a growing concern with tremendous societal importance, pulling in scholars from many diverse areas, such as law, economics, and public policy. For example, a new initiative on big data mining, fairness and privacy has recently been established (http://privacyobservatory.org/current/40-big-data-mining-fairness-and-privacy). Researchers from an algorithmic and machine learning perspective have an important role to play in this new area.

A multitude of interesting and important open problems remain to be solved; here we mention just two. First, all formulations of fairness thus far aim to eliminate all bias. However, in many circumstances, the classification goal does have a direct correlation with membership in the protected group. For example, when predicting who is likely to succeed in university (presumably a key criteria in admission decisions), it might be the case that statistically individuals from a certain population are much more likely to succeed in university than the population at large. One way to nonetheless maintain balance is a quota system, i.e., ensure that the proportions of positive classifications is some value but not necessarily equal between the groups; we can readily handle this case in our framework. Yet it is more desirable to allow some other non-quota control. Formulating a more general framework for fairness that does not force equality or quotas is an important problem. A related problem is understanding how to deconstruct a *given* classifier to determine to what extent it is fair.

Secondly, we would like to develop other forms of intermediate representations beyond prototypes, utilizing multi-dimensional distributed representations, which may offer a richer space for achieving good classifications and transfer to new classifications, while maintaining fairness.

References

Calders, T. and Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21:277–292, 2010.

Dwork, C. and Mulligan, D. Privacy and classification concerns in online behavioral targeting: Mapping objections and sketching solutions. In *Privacy law Scholars Conference*, 2012.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. *In Theory of Cryptograph Conference (TCC)*, 2006.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of Innovations of Theoretical Computer Science*, 2011.

Frank, A. and Asuncion, A. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

Kamiran, F. and Calders, T. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, pp. 1–6, 2009.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *IEEE 11th International Conference on Data Mining*, pp. 643–650, 2011.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

Luong, B., Ruggieri, S., and Turini, F. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM KDD Conference*, pp. 502–510, 2011.

Pedreschi, D., Ruggieri, S., and Turini, F. Discrimination-aware data mining. In *Proceedings of the 14th ACM KDD Conference*, pp. 560–568, 2008.

Tishby, N., Pereira, F.C., and Bialek, W. The Information Bottleneck method. In *The 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999.

Zarsky, T. Automated prediction: Perception, law, and policy. *CACM 15 (9)*, 2012.