Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data

Dylan Slack
UC Irvine
dslack@uci.edu

Sorelle Friedler
Haverford College
sfriedle@haverford.edu

Emile Givental
Haverford College
egivental@haverford.edu

ABSTRACT

In this paper, we advocate for the study of fairness techniques in low data situations. We propose two algorithms *Fairness Warnings* and *Fair-MAML*. The first is a model-agnostic algorithm that provides interpretable boundary conditions for when a fairly trained model *may not* behave fairly on similar but slightly different tasks within a given domain. The second is a fair meta-learning approach to train models that can be trained through gradient descent with the objective of "learning how to learn fairly". This method encodes more general notions of fairness and accuracy into the model so that it can learn new tasks within a domain both quickly and fairly from only a few training points. We demonstrate experimentally the individual utility of each model using relevant baselines for comparison and provide the first experiment to our knowledge of *K-shot fairness*, i.e. training a fair model on a new task with only K data points. Then, we illustrate the usefulness of both algorithms as a combined method for training models from a few data points on new tasks while using Fairness Warnings as interpretable boundary conditions under which the newly trained model may not be fair.

KEYWORDS

machine learning, fairness, meta-learning, covariate shift

1 INTRODUCTION

As machine learning tools become more responsible for decision making in sensitive domains such as credit, employment, and criminal justice, developing methods that are both fair and accurate become critical to the success of such tools. Correspondingly, there has been an increasing amount of academic interest in the field of fair machine learning (for surveys, see [3, 9, 28, 39]). Research on fairness is often concerned with identifying a notion of fairness, developing an approach that mitigates the notion of fairness, and applying the approach to a variety of data sets in a supervised learning setting (see, e.g., [14, 18, 36, 38]).

However, we ask where this leaves fairness-concerned practitioners who are interested in using fair tools for their particular applications but have access to minimal or no training data. In particular, we introduce the following questions:

- When can a practitioner rule out the use of a fair tool trained in a similar but slightly different context (e.g. whether a policy maker in Houston can rule out the use of a fair recidivism tool trained in Philadelphia)?
- How can a practitioner who has access to only a few labeled training points for a particular task still train a fair model?

The most related work to the proposed questions is fairness applied to transfer learning and the covariate shift problem in machine learning. Covariate shift deals with situations where the distribution of data in application differs from the distribution of data in training. Covariate shift is a well studied field, and there are numerous methods that attempt to train supervised learning classifiers that are robust to test distribution shifts with respect to accuracy [6, 24, 31]. Related methods have been developed to address fairness in the covariate shift setting. Kallus et. al. address the problem of systematic bias in data collection and use covariate shift methods to better compute fairness metrics under such conditions [20]. Coston et. al. consider the situation where there are sensitive labels available in only the source or target domain and propose covariate shift methods to solve such problems [10].

Additional work focuses on transferring fair machine learning models across domains. Madras et. al. propose a solution called LAFTR that uses an adversarial approach to create an encoder that can be used to generate fair representations of data sets and demonstrate the utility of the encoder for fair transfer learning [25]. Similarly, Schumman et. al. provide theoretical guarantees surrounding transfer fairness related to equalized odds and opportunity and suggest another adversarial approach aimed at transferring into new domains with different sensitive attributes [29]. Lan and Huan observe that the predictive accuracy of transfer learning across domains can be improved at the cost of fairness [22]. Related to fair transfer learning, Dwork et. al. use a decoupled classifier technique to train a selection of classifiers fairly for each sensitive group in a data set [12].

We argue that our proposed questions are different than the existing work in the following ways. While methods exist that address fairness and covariate shift, such methods do not address the problem of communicating to practitioners and policy makers what domain specific factors might cause a fairly trained model to fail to be fair in practice. Because data sets containing sensitive information can be difficult to obtain, it could be challenging to train a fair machine learning tool using data from only one's specific context. Practitioners might have to rely on data from other geographic locations. Because of demographic or political differences location-to-location, the distribution of data in terms of feature values, sensitive attributes, and labels could be different from one context to another. Thus, determining what changes to the distribution of data might cause a fairly-trained machine learning model to behave unfairly could be useful to practitioners interested in transferring fair models to their particular applications.

Additionally, we note the problem of training fair machine learning models with very little task specific training data is relatively unstudied. Practitioners might have access to minimal training data in one task and sufficient data from other related tasks. This data might be minimal or skewed in terms of which sensitive attribute or label the data belongs to (e.g. only examples of African-Americans

who have been denied a loan) because of data collection issues associated with sensitive data sets like those discussed in Kallus et. al [20]. It could be useful to devise models that are able to achieve satisfactory levels of fairness and accuracy on new tasks with minimal data while being robust enough to handle unfavorable distributions of training data across both labels and sensitive attributes.

In this paper, we propose two different methods to address the proposed problems. First, we discuss the situation where a practitioner has no training data and must decide whether to use a fair machine learning tool trained in another similar but slightly different context. We introduce *Fairness Warnings* — a model agnostic approach that provides interpretable boundary conditions on fairness for when *not* to apply a fair model in a different but related context because the model *may* behave unfairly. Fairness Warnings provide an interpretable model that indicates what distribution shifts to a data set's feature values, labels, and sensitive attributes may cause a fairly trained classifier to act unfairly in terms of a user specified notion of group fairness. While the covariate shift problem setting allows for arbitrary changes to the testing distribution, we only consider mean shifts in this paper. We discuss the limitations imposed by this problem restriction in section 3.1.2. To provide intuition, if Fairness Warnings were trained on a recidivism classifier with respect to the 80% rule of demographic parity [4, 14], the model would provide conditions such as what mean shifts to the features `age` and `priors count` would cause the model to score demographic parity lower than 80%. A practitioner could use this information in combination with their own knowledge about their application to inform their decision surrounding whether *not* to use a fair machine learning tool.

Second, we consider the related situation where a practitioner has access to training data across related tasks in a domain but *minimal* training data for their desired task. We introduce a meta-learning approach, *Fair-MAML*, to address the problem. We empirically demonstrate the ability of Fair-MAML to quickly train models that are both accurate and fair with respect to different notions of fairness. Fair-MAML is based on a meta-learning algorithm called *Model Agnostic Meta Learning* or *MAML* [15] that has shown success in reinforcement learning and image recognition. Fair-MAML is model agnostic in the sense that it is compatible with any model trained through gradient descent. It encourages the learning of more general notions of fairness and accuracy that allow it to achieve strong results on new tasks with only minimal data available. Finally, we connect Fairness-Warnings and Fair-MAML by applying Fairness Warnings as boundary conditions on the fine-tuned fair meta-model.

2 BACKGROUND

2.1 Fairness

We consider a binary fair classification setting with features $X \in \mathbb{R}^n$, labels $Y \in \{0, 1\}$, and sensitive attributes $A \in \{0, 1\}$. Our goal is to train a model that outputs predictions $\hat{Y} \in \{0, 1\}$ such that the predictions are both accurate with respect to Y and fair with respect to the groups defined by A. We consider 1 the "positive" outcome (receiving a loan) and 0 the "negative" outcome (being denied a loan). Within the sensitive attribute, one label is protected and the other unprotected. The protected group might be from a historically disadvantaged group such as women or African-Americans. We will use 0 to denote the protected group and 1 to indicate the unprotected group.

There are three often used ways to define group fairness in this setting. The first, *demographic parity* (or statistical parity [11]), can be formalized as:

$$\frac{P(\hat{Y} = 1|A = 0)}{P(\hat{Y} = 1|A = 1)} \tag{1}$$

This is also known as a lack of disparate impact [4, 14] or discrimination [7]. A value closer to 1 indicates fairness.

The second group fairness definition, *equalized odds* [18], requires that $\hat{Y}$ have equal *true positive rates* and *false positive rates* between groups, where values closer to 1 indicate fairness:

$$\frac{P(\hat{Y} = 1|A = 0, Y = y)}{P(\hat{Y} = 1|A = 1, Y = y)} \; y \in \{0, 1\} \tag{2}$$

This is also known as error rate balance [8] or disparate mistreatment [36]. *Equal opportunity* (or equal true positive rates) introduces relaxed constraints on 2 and requires the equivalence to hold only on the positive outcome in Y. As compared to equalized odds, equal opportunity often allows for increased accuracy [18].

2.2 Meta-Learning

Meta-learning is concerned with training models such that they can be trained on new tasks using only minimal data and few training iterations within a domain. Meta-learning can be phrased as "learning how to learn" because such methods are trained on a range of tasks with the goal of being able to adapt to new tasks more quickly [34]. Metaphorically, this can be likened to finding a base camp (meta-model) from which you can quickly ascend to multiple nearby peaks (optimized per-task models).

In the supervised learning setting, each task $\mathcal{T} = \{\mathcal{D}, \mathcal{L}\}$ where $\mathcal{D}$ is a data set containing pairs (X, Y) and $\mathcal{L}$ is a loss function. We consider a distribution over tasks $P(\mathcal{T})$ which we train the meta-model to adapt to. Supposing the meta-model is a parameterized function f_θ with parameters θ, its optimal parameters are:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \, \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}}(f_\theta) \tag{3}$$

This states that the optimal parameters of the model are those that minimize the loss with respect to both $\mathcal{L}$ and $\mathcal{D}$. Intuitively, the model parameters should be such that they are nearly optimal for a range of tasks. Ideally, this will mean that optimizing for any new task is quick and requires minimal data.

In the meta-learning scenario used in this paper, we train f_θ to learn a new task $\mathcal{T} \sim P(\mathcal{T})$ using K examples drawn from $\mathcal{T}$. Additionally, we assume f_θ can be optimized through gradient descent. During the meta-training procedure, K examples are drawn from $\mathcal{T}$. The model is trained with respect to K and $\mathcal{L}$ and the test performance is evaluated with K new examples. The use of only K training examples for learning a new task is often referred to as K-shot learning and such methods have generally been applied to image recognition and reinforcement learning [35]. Based on the test performance, f_θ is improved. The meta-model f_θ is evaluated at the end of meta-training through a set of tasks that are not included in the meta-training procedure.

3 METHODS

3.1 Fairness Warnings

3.1.1 Framework. Similar to the formalization of LIME in Ribeiro et. al. [27], we define fairness warnings as an interpretable model $g \in G$ where G is a class of interpretable models such as decision trees or logistic regression [30]. Further, g is a function $g : \mathbb{R}^d \rightarrow \{0, 1\}$ where $\mathbb{R}^d$ is a set of distribution shifts applied to the features, labels, and sensitive values of some test data set $\mathcal{D} = \{X, Y, A\}$ under which a fair model is evaluated. We assume f is fair with respect to some notion of group fairness such as equation 1, and the domain of g represents whether the potential shift may result in fair classifications according to that notion of group fairness. Additionally, we assume that group fairness can be evaluated as fair or unfair according to some binary notion of fairness success such as the 80% rule of demographic parity [4, 13, 14]. We assume access to a function $\mathcal{U}_f : \mathcal{D} \rightarrow \{0, 1\}$ that maps between a data set and whether f acts fairly on that data set according to the binary notion of group fairness.

3.1.2 Problem Restrictions. In typical covariate shift settings, the testing distribution can be changed in any number of ways — including being drawn from an entirely different distribution altogether. In this application, we only consider shifts to the mean of the distribution of data that is available for training. Under this assumption there could be more complex changes to the distribution that affect the mean but are not captured by this summary statistic and that may affect fairness. Because we only consider a subset of the possible changes to the testing distribution, Fairness Warnings only indicate what mean shifts may lead a classifier to *not* be fair and do not strongly indicate fairness if no warning is issued. Additionally, it could be the case that Fairness Warnings predict unfairness for certain mean shifts but due to other changes to the testing distribution the classifier actually behaves fairly. Because of these challenges, Fairness Warnings are just that—warnings that there is some evidence that suggests the model may behave unfairly with respect to a notion of group fairness.

3.1.3 SLIM. In practice, we use *Supersparse Linear Integer Models* or *SLIM* as the interpretable model g [33]. SLIM creates a linear perceptron that reduces the magnitudes of the coefficients, removes unnecessary coefficients, and forces the coefficients to be integers. SLIM is a highly interpretable method that is well suited to trading off between model complexity in presentation and accuracy. It has hyperparameters C and ϵ. C controls the marginal accuracy a coefficient must add to stay in the model while ϵ does the same except for the magnitude of the coefficients.

3.1.4 Fairness Warnings Algorithm. In order to train g, we generate some user specified number of perturbed versions of $\mathcal{D}$ using mean shifts. We generate shifts for numerical features by randomly sampling from a Gaussian distribution with the standard deviation of the feature and mean zero. The number sampled is the mean shift across the feature. To perform the shift, we simply add the number to all the values in the feature. We assume categorical features are one-hot encoded and thus only have two binary categorical features in $\{0, 1\}$. We shift each categorical feature by assuming each feature is drawn from a binomial distribution and use the percentage of features labeled 1 as p. We shift the feature vector by drawing a new p from a Gaussian distribution $p \sim \mathcal{N}(p, 1)$ and randomly sample a new vector according to p. If p is less than 0 or greater than 1, we adjust p to 0 or 1 respectively. Doing this a user specified number of times, we create a set of shifted variations $\mathcal{D}'$ of the original $\mathcal{D}$.

For each shifted data set, we generate a fairness label $\mathcal{F}$ using the binary notion of group fairness $\mathcal{U}_f$. We create a data set of mean shifted data sets and their group fairness behavior with respect to f, $\mathcal{Z}_f = (\mathcal{D}', \mathcal{F})$. Finally, we train g on $\mathcal{Z}_f$ using $\mathcal{D}'$ as the features and $\mathcal{F}$ as the labels. Intuitively, we train g so that it learns to predict what mean shifts may result in unfairness. Assuming $shift(;)$ is some function that computes the mean shifting scheme above, the algorithm for generating fairness warnings is given as Algorithm 1.

Require: $\mathcal{D}$: data set
Require: $\mathcal{U}_f$: fairness notion
Require: g: interpretable model
Require: N: number of shifts to perform
  $\mathcal{Z}_f \leftarrow []$
  for $i = 1 : N$ **do**
    $\mathcal{D}' \leftarrow shift(\mathcal{D})$
    $\mathcal{F} \leftarrow \mathcal{U}_f(\mathcal{D}')$
    $\mathcal{Z}_f \leftarrow (\mathcal{D}', \mathcal{F}) \bigcup \mathcal{Z}_f$
  end for
  $g \leftarrow$ Train g with $\mathcal{Z}_f$ using $\mathcal{D}'$ as features and $\mathcal{F}$ as labels
  return g

Algorithm 1: Fairness Warnings

3.2 Fair Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
  randomly initialize θ
  while not done **do**
    Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
    for all $\mathcal{T}_i$ **do**
      Sample K datapoints $\mathcal{D} = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}, \mathbf{a}^{(j)}\}$ from $\mathcal{T}_i$
      Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ using $\mathcal{D}$ and $\mathcal{L}_{\mathcal{T}_i}$
      Compute updated parameters:
      $\theta_i' = \theta - \alpha \nabla_\theta [\mathcal{L}_{\mathcal{T}_i}(f_\theta) + \gamma_{\mathcal{T}_i} \mathcal{R}_{\mathcal{T}_i}(f_\theta)]$
      Sample K *new* datapoints $\mathcal{D}_i' = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}, \mathbf{a}^{(j)}\}$ from $\mathcal{T}_i$ to be used in the meta-update
    end for
    Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} [\mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}) + \gamma_{\mathcal{T}_i} \mathcal{R}_{\mathcal{T}_i}(f_{\theta_i'})]$
    using each $\mathcal{D}_i'$
  end while

Algorithm 2: Fair-MAML

3.2.1 K-shot Fairness. In order to address the problem of learning fairly from minimal data on a new task, we introduce the notion of *K-shot fairness*. Given K training examples, K-shot fairness aims to (1) quickly train a model that is both fair and accurate on a given task. Additionally, because the relationship between fairness and accuracy is often understood as a trade-off [16], an additional aim

is to (2) allow tuning of such a model so that it achieves different balances between accuracy and fairness using just K training points.

The language used in this paper surrounding K-shot learning differs slightly from the language used in typical K-shot learning scenarios such as image recognition. In K-shot image recognition, the goal is to learn how to distinguish between N different image labels using only K training examples of each type. The training set size is then KN examples. Because we assume all the tasks to be binary labeled, all of our tasks are 2-way. In referencing K-shot fairness, we will mean that we are using K training examples *total*—irrespective of class label, with the assumption that all tasks are 2-way.

3.2.2 Fair-MAML Framework. We expand the meta learning framework from section 2.2 such that each task includes a fairness regularization term $\mathcal{R}$ and fairness hyperparameter γ. Additionally, we require that $\mathcal{D}$ have a protected feature A such that $\mathcal{D} = (X, Y, A)$. The goal of $\mathcal{R}$ is to minimize some notion of group fairness and γ dictates the trade off between $\mathcal{R}$ and $\mathcal{L}$. A task is defined as $\mathcal{T} = \{\mathcal{D}, \mathcal{L}, \mathcal{R}, \gamma\}$. We adjust equation 3 such that the optimal parameters are now:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{\mathcal{T}\sim p(\mathcal{T})}[\mathcal{L}_{\mathcal{T}}(f_\theta) + \gamma_{\mathcal{T}}\mathcal{R}_{\mathcal{T}}(f_\theta)] \tag{4}$$

In order to train a fair meta-learning model, we adapt Model-Agnostic Meta-Learning or MAML to our fair meta-learning framework and introduce *Fair-MAML* [15]. MAML is trained by optimizing performance of f_θ across a variety of tasks after one gradient step. MAML is particularly well suited to easy fairness adaption because it works with any model that can be trained with gradient descent. The core assumption of MAML is that some internal representations are better suited to transfer learning. The loss function used by MAML is effectively the loss across a batch of task losses. Thus, the MAML learning configuration encourages learning representations that encode more general features than a traditional learning approach. The MAML algorithm works by first sampling a batch of tasks, computing the updated parameters θ after one gradient step of training on K data points sampled from each task, and finally updating f_θ based on the performance of θ on a new sample of K points.

We modify MAML to Fair-MAML by including a fairness regularization term $\mathcal{R}$ in the task losses. The algorithm for Fair-MAML is given in algorithm 2. By including a regularization term, we hope to encourage MAML to learn generalizable internal representations that strike a desirable balance between accuracy and fairness.

3.3 Fairness Regularizers

A variety of fairness regularizers have been proposed to handle various definitions of group fairness [5, 19, 21]. Because MAML has shown success with the use of deep neural networks [15], we require regularization terms compatible with neural networks. Methods that require the model to be linear are clearly not applicable. In addition, Fair-MAML requires that second derivatives be computed through a Hessian-vector product in order to calculate the meta-loss function which can be computationally intensive and time-consuming. Thus, it is critical that our fairness regularization term be quick to compute in order to allow for reasonable Fair-MAML training times.

We propose two simple regularization terms aimed at achieving demographic parity and equal opportunity that are easy to implement and extremely quick to compute. Let $\mathcal{D}_0$ denote the protected instances in X and Y. The demographic parity regularizer is:

$$\begin{aligned}\mathcal{R}_{dp}(f_\theta, \mathcal{D}) &= 1 - P(\hat{Y} = 1 | A = 0) \\ &\approx 1 - \frac{1}{|\mathcal{D}_0|} \sum_{x \in \mathcal{D}_0} P(f_\theta(x) = 1)\end{aligned} \tag{5}$$

This regularizer incurs a penalty if the probability that the protected group receives positive outcomes is low. Our value assumption here is that we want to adjust the likelihood of the protected class receiving a positive outcome upwards. Namely, we do not reduce the rate at which the unprotected class receives positive outcomes and adjust upwards the rate at which the protected class receives positive outcomes.

Additionally, we consider a regularizer aimed at improving equal opportunity. Let $\mathcal{D}_0^1$ denote the instances within X that are both protected and have the positive outcome in Y.

$$\begin{aligned}\mathcal{R}_{eop}(f_\theta, \mathcal{D}) &= 1 - P(\hat{Y} = 1 | A = 0, Y = 1) \\ &\approx 1 - \frac{1}{|\mathcal{D}_0^1|} \sum_{x \in \mathcal{D}_0^1} P(f_\theta(x) = 1)\end{aligned} \tag{6}$$

We have a similar value assumption using this regularizer as the one for demographic parity. Namely, we adjust the true positive rate of the protected class upwards and do not decrease the true positive rate of the unprotected class.

A nice advantage of these regularizers are that they are easy to implement using common deep learning packages such as PyTorch or TensorFlow [1, 26]. Supposing that Q is a vector of probabilities that f_θ outputs a positive value on X, Y is a vector containing the labels, and A is a vector containing the sensitive attribute, the demographic parity regularizer $\mathcal{R}_{dp}$ and equal opportunity regularizer $\mathcal{R}_{eop}$ can be computed as below, where $\odot$ denotes the Hadamard Product:

$$\begin{aligned}\mathcal{R}_{dp}(f_\theta, \mathcal{D}) &= 1 - \overline{[Q \odot (1 - A)]} \\ \mathcal{R}_{eop}(f_\theta, \mathcal{D}) &= 1 - \overline{[Q \odot Y \odot (1 - A)]}\end{aligned} \tag{7}$$

This formulation is possible because of the proposed binary labels of A and Y from subsection 2.1 where A is a binary vector with 0 indicating a protected value and Y a binary vector with 1 denoting a positive outcome.

4 EXPERIMENTS

We first demonstrate the individual utility of both fairness warnings and Fair-MAML experimentally. We then show their usefulness as a combined method.

4.1 Fairness Warnings

4.1.1 COMPAS Recidivism Experiment Setup. We initially consider applying Fairness Warnings to the COMPAS recidivism data set. The COMPAS recidivism data set consists of data from over 10, 000

criminal defendants from Broward County, Florida. It includes attributes such as the sex, age, race, and priors for the defendants in addition to a categorical variable indicating perceived recidivism risk. We pre-process the data set as described in Angwin et. al. [2]. We create a binary sensitive column for whether the defendant is African-American. We predict the ProPublica collected label of whether the defendant was rearrested within two years.

We trained a neural network as the model, f, to perform Fairness Warnings. We trained two models—one regularized for demographic parity and the other equal opportunity using the regularization terms from equations 5 and 6 respectively. The demographic parity regularized model scored 58% accuracy and 81% demographic parity on a 20% test set. The equal opportunity regularized model scored 54% accuracy and 68% equal opportunity using the same test set. For the demographic parity fairness warnings, we set the fairness warnings demographic parity threshold at 80%. Meaning, if the classifier scored demographic parity above 80%, it was deemed fair. In the equal opportunity setting, we set the threshold to 60%. We generated 2,000 perturbed data sets, 800 of which were classified unfairly according to demographic parity. We set ϵ to $1e-3$ and C to $1e-3$. We found that g was able to classify whether the shifts applied to the perturbed data sets would result in unfair group fairness behavior with 88% accuracy on a 10% test set. Using the same perturbed set, the equal opportunity regularized network was found to be unfair in 550 of the 2,000 perturbed examples. Using the same hyperparameters as before, g was able to classify whether the shifts would result in unfairness with respect to equal opportunity with 86% accuracy. The Fairness Warnings for the COMPAS data set is given in figure 1.

4.1.2 COMPAS Recidivism Experiment Analysis. The COMPAS Fairness Warnings both rely on priors_count and age to determine what mean shifts to the data set may result in unfairness. In the demographic parity warning for instance, if the mean group age applied to f were to increase by 3 years and mean priors were to remain unchanged, the fairness warning would predict unfairness because the score total would be $(0 \cdot 20) + (3 \cdot -2) < -1$. However, in the equal opportunity case, the same shift would not yield unfairness because $(0 \cdot 24) + (3 \cdot -2) \not< -19$. A case that would result in unfairness in the equal opportunity setting would be a decrease in mean priors count by one charge and for age to remain level, i.e. $(-1 \cdot 24) + (0 \cdot -2) < -19$.

Overall, the SLIM implementation of fairness warnings showed good ability to classify whether certain mean shifts applied to the feature values of the COMPAS data set would result in unfairness. Because SLIM is tunable with respect to the importance threshold of features shown in the presentation of the model, the classifier only outputs 2 of a possible 8 feature values in both warnings. The presentation is simple. A practitioner would only have to perform a few arithmetic operations in order to compute the fairness warning outcome.

Additionally, we were able to train a random forest classifiers using 200 estimators from the Scikit-learn implementation which scored 94% and 89% accuracy on the demographic parity and equal opportunity fairness warnings tasks respectively. This suggests that more robust models could serve as much more accurate fairness warnings than SLIM. Presenting a random forest of such size in a digestible way to a user would be difficult. However, the success of the random forest to perform this task indicates that improved interpretable methods that achieve equal levels of interpretability to SLIM but higher levels of accuracy on the fairness warnings task could serve as more desirable fairness warnings.

4.2 Fair-MAML

4.2.1 Synthetic Experiment Setup. We illustrate the usefulness of Fair-MAML as opposed to a regularized pre-trained model in fair few-shot classification through a synthetic example based on Zafar et. al [37]. We generate two Gaussian distributions using the means and covariances from Zafar et. al. The first distribution (1) is set to $p(x) = N([2;2],[5,1;1,5])$ and the second (2) is set to $p(x) = N([-2;-2],[10,1;1,3])$. During training, we simulate a variety of tasks by dividing the class labels along a line with y-intercept of $(0,0)$ and a slope randomly selected on the range $[-5,5]$. All points above the line in terms of their y-coordinate receive a positive outcome while those below are negative. Using the formulation from Zafar et. al., we create a sensitive feature by drawing from a Bernoulli distribution where the probability of the example being in the protected class is: $p(a=0) = p(x'|y=1)/(p(x'|y=1)+p(x'|y=0))$ where $x' = [cos(\phi), -sin(\phi); sin(\phi), cos(\phi))]x$. Here, ϕ controls the correlation between the sensitive attribute and class labels. The lower ϕ, the more correlation and unfairness. We randomly select ϕ from the range $[2,4,8,16]$ to simulate a variety in fairness between tasks.

In order to assess the fine-tuning capacity of Fair-MAML and the pre-trained neural network, we introduced a more difficult fine-tuning task. During training, the two classes were separated clearly by a line. For fine-tuning, we set each of the binary class labels to a distribution. The positive class was set to distribution (1) and the negative class was set to distribution (2). In this scenario, a straight line cannot clearly divide the two classes. We assigned sensitive attributes using the same strategy as above and used a ϕ of 4. Additionally, we only gave 5 *positive-outcome* examples from the *protected class*. We hoped to simulate a situation where a fair classifier is needed on a new task, but there are only a few protected examples in the positive outcome to learn from—simulating the situation where the distribution of fine-tuning task data is biased. An example of such a scenario could be if a practitioner needed to train a new loan tool and had access to only a few examples of African-Americans who received loans.

We randomly generated 100 synthetic tasks that we cached before training. We sampled 5 examples from each task during meta-training, used a meta-batch size of 32 for Fair-MAML, and performed a single epoch of optimization within the internal MAML loop. We trained Fair-MAML for 5,000 meta-iterations. For the pre-trained neural network, we performed a single epoch of optimization for each task. We trained over 5,000 batches of 32 tasks per batch to match the training set size used by Fair-MAML.

The loss used is the cross-entropy loss between the prediction $f(x)$ and the true value using the demographic parity regularizer from equation 5. We use a neural network with two hidden layers consisting of 20 nodes and the ReLU activation function. We used the softmax activation function on the last layer. When training with Fair-MAML, we used $K = 5$ examples and performed one

Predict UNFAIR DEMOGRAPHIC PARITY if SCORE < -1			
Feature	Original Mean	Score (+/- per unit increase/decrease)	Total
priors_count	3.2 priors	20 points / prior	+.....
age	34.5 years	-2 points / year	+.....
ADD POINTS FROM ROWS 1 to 2 (Warning accuracy: 88%)		SCORE	=.....

Predict UNFAIR EQUAL OPPORTUNITY if SCORE < -19			
Feature	Original Mean	Score (+/- per unit increase/decrease)	Total
priors_count	3.2 priors	24 points / prior	+.....
age	34.5 years	-2 points / year	+.....
ADD POINTS FROM ROWS 1 to 2 (Warning accuracy: 86%)		SCORE	=.....

Figure 1: The Fairness Warnings for the COMPAS Recidivism data set for both demographic parity and equal opportunity. The original model is a neural network regularized for the respective notion of fairness. This fairness warning is meant to be read as the expected mean shift *away* from the original mean of the features presented in a practitioner's application. For instance, if age were to decrease 1 year and age were to decrease 3 years, the score would be $(-1 \cdot 20) + (-3 \cdot -2) = -14$ points. -14 points < -1 point, so the warning would predict unfairness. Critically, the fairness warning only makes a claim surrounding unfairness. If the model predicts a score ≥ -1, the model *does not* certify fair behavior.

gradient step update. We set the step size α to 0.3, used the Adam optimizer to update the meta-loss with learning rate β set to $1e-3$. We pre-trained a baseline neural network on the same architecture as Fair-MAML. To one-shot update the pre-trained neural network we experimented with step sizes of $[0.01, 0.1, 0.2, 0.3]$ and ultimately found that 0.3 yielded the best trade offs between accuracy and fairness. Additionally, we tested γ values during training and fine-tuning of $[0, 10]$. We present an example task in figure 2 using 5 fine-tuning points from the positive outcome and protected class. When $\gamma = 0$, Fair-MAML does not incur any fairness regularization, so the model is just MAML. We give comprehensive results over a variety of tasks in the appendix in figure 5.

4.2.2 Synthetic Experiment Analysis. In the new task, there is an unseen configuration of positively labeled points. It was not possible for positively labeled points to fall below $y = 0$ during training. Fair-MAML is able to perform well with respect to both fairness and accuracy on the fine-tuning task when only biased fine-tuning data is available. The pre-trained neural network fails at performing the new task. This example illustrates that Fair-MAML has learned a more useful internal representation for both fairness and accuracy than the pre-trained neural network. Examining the extended results over a variety of randomly selected fine-tuning points in figure 5, Fair-MAML is able to consistently yield both fair and accurate results while the pre-trained neural network is somewhat unstable.

4.2.3 Communities and Crime Experiment. Next we consider an example using the Communities and Crime data set [23]. The Communities and Crime data set includes information relevant to crime (e.g., police per population, income) as well as demographic information (such as race and sex) in different communities across the United States. The goal is to predict the violent crime rate in the community. We convert this data set to a few-shot fairness setting by using each state as a different task. We believe this problem setting is justified because state by state differences ranging from firearm control to weather patterns could affect the generalization ability of a model trained on a selection of states [17, 32]. Because the violent crime rate is a continuous value, we convert it into a binary label based on whether the community is in the top 50% in terms of violent crime rate within a state. Additionally, we add a binary sensitive column that receives a protected label if African-Americans are the highest or second highest population in a community in terms of percentage racial makeup.

The Communities and Crime data set has data from 46 states ranging in number of communities from 1 to 278 communities per state. We only used states with 20 or more communities leaving 30 states. We held out 5 randomly selected states for testing and trained using 25 states. We set $K = 10$ and cached 100 meta-batches of size 8 states for training. For testing, we randomly selected 10 communities from the hold out task that we used for fine-tuning and evaluated on whatever number of communities were left over. The number of evaluation communities is guaranteed to be at least 10 because we only included states with 20 or more communities.

We trained two Fair-MAML models—one with the demographic parity regularizer from equation 1 and another with the equal opportunity regularizer from equation 6. For both models, we used a neural network with two hidden layers of 20 nodes. We trained the model with one gradient step using a step size of $1e-2$ and a meta-learning rate of $1e-4$ using the Adam optimizer. We trained the model for $2,000$ meta-iterations.

In order to assess Fair-MAML, we trained a neural network regularized for fairness using the same architecture and training data. We fine-tuned the neural network for each of the assessment tasks. We used a learning rate of $1e-3$ for training and assessed learning rates of $[1e-4, 1e-3, 1e-2, 1e-1]$ for fine-tuning. We found the

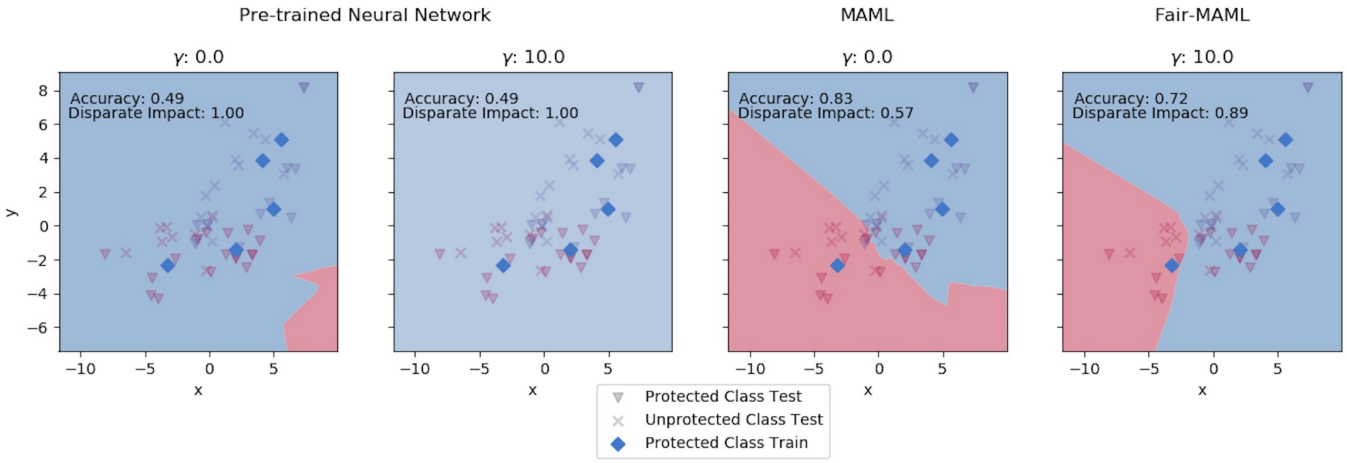

Figure 2: An example decision boundary from the pre-trained neural network, MAML, and Fair-MAML on the synthetic example (note: Fair-MAML is MAML with $\gamma = 0$). Points that are colored the same as the side of the boundary are correct. Only points in the positive outcome and protected class are given for the fine-tuning task. Fair-MAML is able to handle such an imbalance of training points on a previously unseen task while the pre-trained neural network fails—illustrating that Fair-MAML has learned a more useful internal representation for both fairness and accuracy.

fine-tuning rate of $1e-1$ to perform the best trade offs between accuracy and fairness and present results using this learning rate. We varied γ over $[0, 4]$ incremented by 1 for the demographic parity regularizer. We found higher γ's to work better for the equal opportunity regularizer and varied γ from $[0, 40]$ incremented by 10.

Additionally, we trained two LAFTR models on the transfer tasks as comparisons for demographic parity and equalized opportunity. LAFTR is not intended to be compatible with our proposed K-shot fairness experiments because training on fine-tuning tasks with a minimal number of epochs and training points is not expected. However, we find that it is the most relevant fair transfer learning method to use as a baseline. We used the same transfer methodology and hyperparameters as described in Madras et. al. [25] and used a neural network with a hidden layer of 20 nodes as the encoder. We used another neural network with a hidden layer of 20 nodes as the MLP to be trained on the fairly encoded representation. We used the demographic parity and equal opportunity adversarial objectives for the first and second LAFTR model respectively. We trained each encoder for $1,000$ epochs and swept over a range of $\gamma's$: $[0, 0.5, 1.0, 2.0, 4.0]$. We trained with all the data not held out as one of the 5 testing tasks. When training a MLP from the encoder on each of the transfer tasks, we found that LAFTR struggled to produce useful results with only 10 training points from the new task over any number of training epochs. We found that we were able to get reasonable results from LAFTR using 30 fine-tuning points and 100 epochs of optimization—using a minimal number of epochs was unsuccessful. It makes sense that a minimal number of training epochs for the new task is unsuccessful because the MLP trained on the fairly encoded data is trained from scratch. The results are presented in figure 3. Though we do not include the results in presentation, we were able to generate similar results with LAFTR to Fair-MAML using 50 training points from the new task after 100 epochs of optimization.

We observe that Fair-MAML achieves the best trade off between fairness and accuracy both in terms of demographic parity and equal opportunity. In our proposed problem setting, LAFTR was not successful at learning with minimal data and a small number of fine-tuning epochs for the new task. The pre-trained neural network shows some ability to learn the new task using little data and fine-tuning epochs. At low γ's, Fair-MAML is able to achieve higher accuracy than the pre-trained neural network and LAFTR. Crucially, Fair-MAML is able to learn more accurate representations that are also fairer for a range of γ's than both of the baselines. In order to generalize to new states, only 10 communities are needed in order to achieve strong predictive accuracy and fairness using Fair-MAML.

4.3 Fair-MAML with Fairness Warnings

4.3.1 Motivation. We next consider Fairness Warnings applied to Fair-MAML. We argue that Fairness Warnings can serve as a complementary tool to Fair-MAML. Because we expect Fair-MAML to be used in situations with minimal data available, it is possible that testing data given to a fine-tuned Fair-MAML model is unrepresentative of the true distribution of data for a particular task. While in section 4.2.1, we empirically demonstrate that Fair-MAML can still achieve good results when training data is available from one value in a sensitive attribute or label, it still may be useful for practitioners to have indication surrounding situations in which their model may fail to be fair in testing.

4.3.2 Communities and Crime Fairness Warning/Fair-MAML Experiment. We apply fairness warnings to Fair-MAML on the communities and crimes experimental setup from section 4.2.3 using demographic parity as our notion of fairness. We randomly chose an evaluation state to apply Fairness Warnings and left the rest

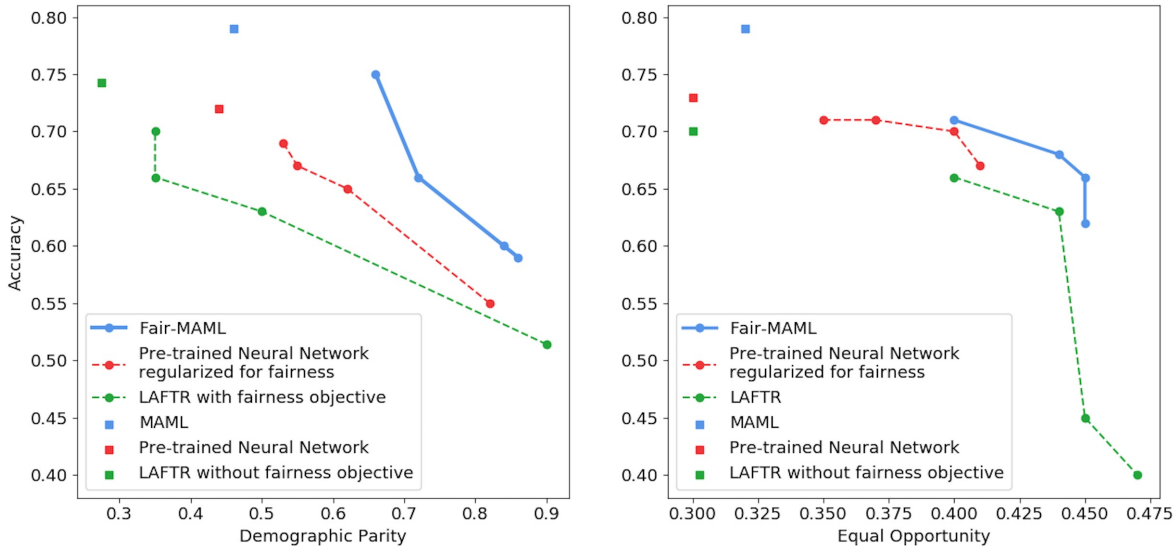

Figure 3: The accuracy/fairness trade off for the communities and crimes example sweeping over a range of γ's. The data presented is the mean across three runs on each γ using 5 randomly selected hold out tasks. The fairness numbers presented are the ratio between the protected and unprotected groups. Higher accuracy and fairness values closer to 1.0 indicate more successful outcomes. The pre-trained neural network and Fair-MAML received 10 fine-tuning points and were optimized for 1 epoch. We did not find useful results using LAFTR with only 10 fine-tuning points or with a minimal number of fine-tuning epochs, so the LAFTR example given here is with 30 fine-tuning points and 100 epochs of optimization. Fair-MAML is able achieve better levels of accuracy and fairness than both the pre-trained network and LAFTR on the transfer tasks using minimal fine-tuning data.

for meta-training. We trained two Fair-MAML models as f in fairness warnings using the demographic parity regularizer for the first model and equal opportunity regularizer for the second model. We used $\gamma = 5$ for the demographic parity Fair-MAML model and $\gamma = 30$ for the equal opportunity Fair-MAML model. We trained for $2,000$ meta-iterations in a 1-step optimization setting, with the update learning rate set to $1e-2$ and the meta learning rate set to $1e-4$. The demographic parity Fair-MAML model scored 87% demographic parity on the test set of the fine-tuning task and accuracy of 69%. The equal opportunity Fair-MAML model scored 64% accuracy and equal opportunity of 63%.

To train Fairness Warnings on the fine-tuning task, we created $2,000$ shifted data sets of the fine-tuning test data. We trained a Fairness Warning for both demographic parity and equal opportunity. We used the 80% rule of demographic parity in the demographic parity warning and a 60% equal opportunity threshold in the equal opportunity warning. We found that $1,034$ or close to 50% of the shifted data sets were classified fairly according to f with respect to demographic parity and that $1,248$ of the shifted data sets were classified fairly according to equal opportunity. We trained SLIM using ϵ of $1e-3$ and C of $1e-5$ for the demographic parity fairness warning. We adjusted C to $1e-3$ for the equal opportunity fairness warning.

SLIM was able to predict whether the mean shifts across the features in the communities and crime data set would result in demographic parity unfairness with 71% accuracy on a 10% test set. A random forest with 200 estimators was able to predict the same task with 88% accuracy. In the equal opportunity setting, SLIM predicted the task with 68% accuracy. A random forest with 200 estimators was able to perform the same task with 77% accuracy. The fairness warnings are presented in figure 4.

4.3.3 Communities and Crime Fairness Warning/Fair-MAML Analysis. The Fairness Warning trained on the fine-tuned Fair-MAML model is able to perform reasonable prediction accuracy and generates informative results. Particularly, it is interesting to consider that the demographic parity fine-tuned model behaves unfairly when the testing data set changes according to features such as number people living under the poverty line, in urban areas, and number of police officer. A similar result is found in the equal opportunity setting with police operating budget. In both the demographic parity and equal opportunity cases, the fairness warnings demonstrate that seemingly small and perhaps innocuous differences between states where Fair-MAML is trained and applied could result in unfair behavior. For instance, the addition of a couple dozen additional police officers across communities in a state in the demographic parity case could lead to the classifier behaving unfairly. The same is true for equal opportunity and a slight increase to the mean police operating budget. As we see in this example, reasonable real world changes to the testing distribution can result in negative changes to the group fairness of the fine-tuned Fair-MAML model. Providing Fairness Warnings to accompany the fine-tune meta model could lend additional guidance to a practitioner and help them better understand if their model will not behave fairly in application.

Predict UNFAIR DEMOGRAPHIC PARITY if SCORE < -3,661,000			
Feature	Original Mean	Score (+/- per unit increase/decrease)	Total
mean people per family	3.1 people	2,000,000 points / person	+.....
number of people living in urban areas	47,700 people	-1 point / person	+.....
number of people living under the poverty line	7,590 people	-5 point / person	+.....
number of sworn full time police officers	77.4 officers	-130,000 points / officer	+.....
ADD POINTS FROM ROWS 1 to 7 (Warning accuracy: 71%)		SCORE	=.....

Predict UNFAIR EQUAL OPPORTUNITY if SCORE < -2			
Feature	Original Mean	Score (+/- per unit increase/decrease)	Total
police operating budget	$3M	-2 points / $1M	+.....
ADD POINTS FROM ROWS 1 to 1 (Warning accuracy: 68%)		SCORE	=.....

Figure 4: The Fairness Warnings for Fair-MAML applied to the communities and crime data set on the fine-tuning task. We consider Fair-MAML trained for both demographic parity and equal opportunity. Unlike in the COMPAS example, the features that the Fairness Warnings use are different though they both relate to aspects of policing.

5 LIMITATIONS AND CONCLUSIONS

In this paper, we introduced Fairness Warnings and Fair-MAML. Fairness Warnings provides an interpretable model that predicts which changes to the testing distribution will cause a model to behave unfairly. Fair-MAML is a method that "learns to learn" fairly and can be used to train a fair model quickly from minimal data. We demonstrate empirically the usefulness of both methods through multiple examples on both synthetic and real data sets.

In this work, we explore Fairness Warnings applied to mean shifts in the testing distribution. It is a relatively straight forward extension to apply Fairness Warnings to other distribution shifts such as changes to the standard deviation. Though we are able to generate Fairness Warnings that show useful results, they ultimately are only applied to summary statistics. Meaning, changes to the distribution that are not captured by such statistics could affect fairness in unpredictable ways. Thus, we only propose fairness warnings as boundary conditions under which the model *may not* be fair. In this regard, receiving a non-unfair score in fairness warnings *does not* guarantee that the model will behave fairly in the new domain. We emphasize the importance of this directionality to any lawmakers or practitioners who would be interested in using Fairness Warnings and advise that they be used only to decide against the use of certain models instead of verify that models will behave fairly. A final limitation to our work is that we assess Fair-MAML when there are many related training tasks to learn from. In reality, there may only be a few related training tasks available. We leave assessing how useful Fair-MAML is on domains with only a few related training tasks to future work.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (2016).

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. *Fairness and Machine Learning*. fairmlbook.org.

[4] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. *ArXiv* abs/1706.02409 (2017).

[6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative Learning Under Covariate Shift. *J. Mach. Learn. Res.* 10 (2009), 2137–2155.

[7] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

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

[9] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. *ArXiv* abs/1810.08810 (2018).

[10] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, Honolulu, HI, USA*.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. ACM, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[12] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*. 119–133.

[13] The U.S. EEOC. 1979. Uniform guidelines on employee selection procedures. (1979).

[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, International Convention Centre, Sydney, Australia, 1126–1135. http://proceedings.mlr.press/v70/finn17a.html

[16] Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton, and Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *ACM Conference on Fairness, Accountability and Transparency (FAT*)*. ACM. http://arxiv.org/abs/1802.04422

[17] Mark E. Hamill, Matthew C. Hernandez, Kent R. Bailey, Martin D. Zielinski, Miguel A. Matos, and Henry J. Schiller. 2019. State Level Firearm Concealed-Carry Legislation and Rates of Homicide and Other Violent Crime. (2019), 1–8.

[18] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 3323–3331. http://dl.acm.org/citation.cfm?id=3157382.3157469

[19] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and Fair Classification. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 2879–2890. http://proceedings.mlr.press/v97/huang19e.html

[20] Nathan Kallus and Angela Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, StockholmsmÃďssan, Stockholm Sweden, 2439–2448. http://proceedings.mlr.press/v80/kallus18a.html

[21] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[22] Chao Lan and Jun Huan. 2017. Discriminatory Transfer. *Workshop on Fairness, Accountability, and Transparency in Machine Learning* (2017).

[23] Moshe Lichman. 2013. UCI machine learning repository. (2013). http://archive.ics.uci.edu/ml/index.php

[24] Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. 2018. Detecting and Correcting for Label Shift with Black Box Predictors. *ICML* (2018).

[25] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. *International Conference on Machine Learning* (2018).

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.

[28] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

[29] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H. Chi. 2019. Transfer of Machine Learning Fairness across Domains. *CoRR* abs/1906.09688 (2019). arXiv:1906.09688 http://arxiv.org/abs/1906.09688

[30] Dylan Slack, Sorelle A. Friedler, Chitradeep Dutta Roy, and Carlos Scheidegger. 2019. Assessing the Local Interpretability of Machine Learning Models. *CoRR* abs/1902.03501 (2019). arXiv:1902.03501 http://arxiv.org/abs/1902.03501

[31] Adarsh Subbaswamy, Peter G. Schulam, and Suchi Saria. 2018. Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport. In *AISTATS*.

[32] Jari Tiihonen, Pirjo Halonen, Laura Tiihonen, Hannu Kautiainen, Markus Storvik, and James Callaway. 2017. The Association of Ambient Temperature and Violent Crime. *Scientific Reports* 7, 1 (2017), 6543. https://doi.org/10.1038/s41598-017-06720-z

[33] Berk Ustun and Cynthia Rudin. 2015. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102 (2015), 349–391.

[34] Joaquin Vanschoren. 2019. *Meta-Learning*. Springer International Publishing, Cham, 35–61. https://doi.org/10.1007/978-3-030-05318-5_2

[35] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In *NIPS*.

[36] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. 1171–1180.

[37] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. *AISTATS* (2017).

[38] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*. 962–970.

[39] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (2015).

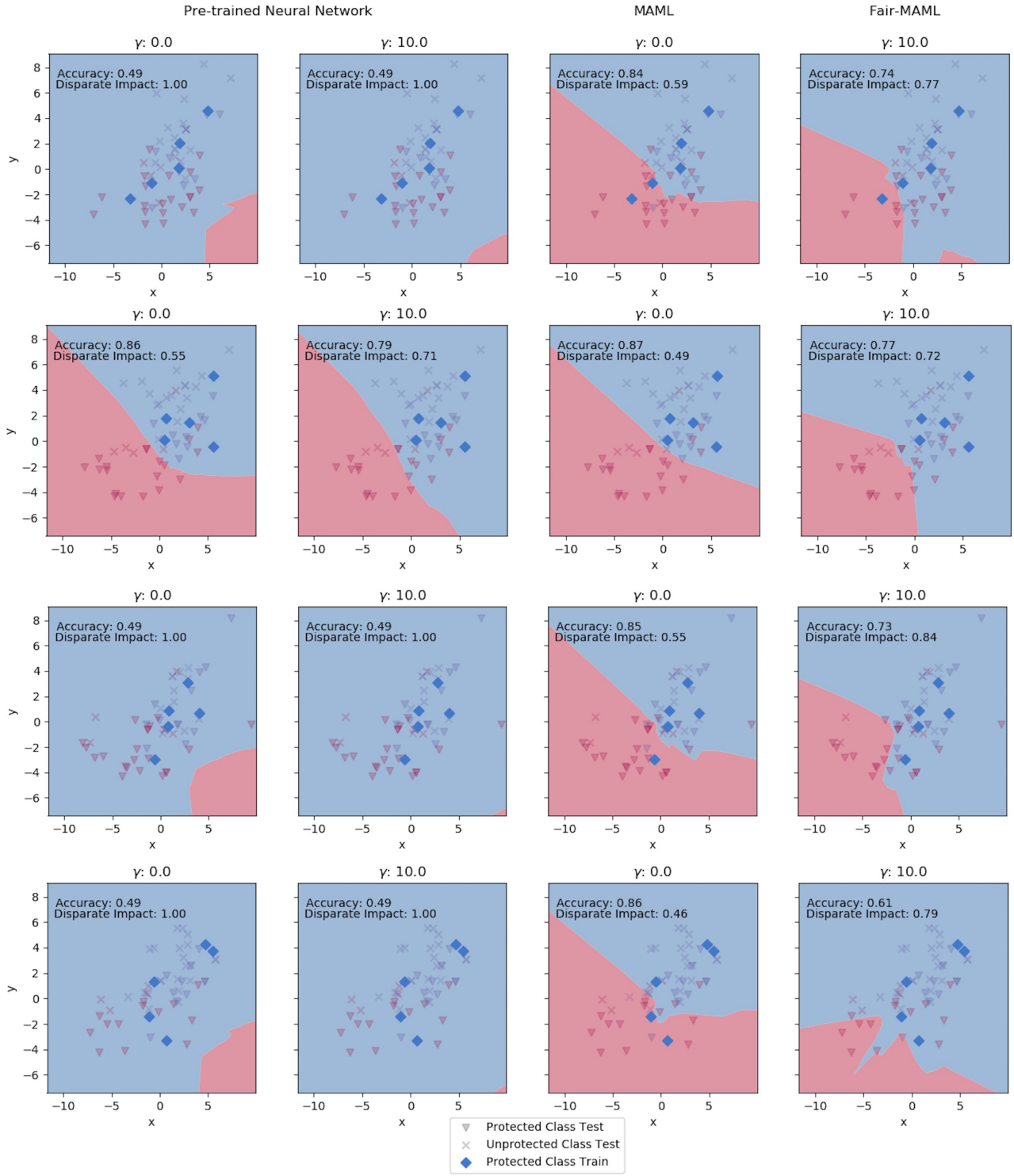

Figure 5: Additional synthetic examples.