Blind Justice: Fairness with Encrypted Sensitive Attributes

Niki Kilbertus 1 2  Adrià Gascón 3 4  Matt Kusner 3 4  Michael Veale 5  Krishna P. Gummadi 6  Adrian Weller 2 3

Abstract

Recent work has explored how to train machine learning models which do not discriminate against any subgroup of the population as determined by sensitive attributes such as gender or race. To avoid disparate treatment, sensitive attributes should not be considered. On the other hand, in order to avoid disparate impact, sensitive attributes must be examined—e.g., in order to learn a fair model, or to check if a given model is fair. We introduce methods from secure multi-party computation which allow us to avoid both. By encrypting sensitive attributes, we show how an outcome-based fair model may be learned, checked, or have its outputs verified and held to account, without users revealing their sensitive attributes.

1. Introduction

The growing field of fair learning seeks to formalize relevant requirements, detection, and mitigation of potential illegal or unfair discrimination against certain subgroups of the population (Schreurs et al., 2008; Calders & Žliobaitė, 2012; Friedler et al., 2016). Most legally-problematic discrimination centers on sensitive attributes (Barocas & Selbst, 2016). Simply not inquiring about sensitive attributes to avoid disparate treatment (Grgić-Hlača et al., 2018), does not protect against disparate impact (Dwork et al., 2012), which occurs when the outcomes of decisions—perhaps unintentionally—disproportionately benefit or hurt particular sensitive groups (Handel et al., 2014). Much recent work has focused avoiding various notions of disparate impact (Feldman et al., 2015; Hardt et al., 2016; Zafar et al., 2017).

In order to check and enforce such requirements, the modeler must have access to the sensitive attributes in the training data (Žliobaitė & Custers, 2016; Lipton et al., 2017)—however, this may be undesirable for several reasons. First, individuals are unlikely to want to entrust sensitive attributes to modelers. Even if a modeler was trusted, the wide provision of sensitive data creates heightened privacy risks in the event of a data breach.

Second, legal barriers such as EU’s General Data Protection Regulation (GDPR), may limit collection and processing of sensitive personal data, which modelers cannot justify with their “legitimate interests” (Veale & Edwards, 2018). Veale & Binns (2017) addressed this problem by involving a highly trusted third party, requiring individuals to disclose their sensitive attributes (risking breaches or hacks) and the modeler to disclose their model (resulting in intellectual property concerns) to this third party.

Contribution. Based on recent methods from secure multi-party computation (MPC), we propose an approach to detect and mitigate disparate impact ensuring that both individuals’ sensitive attributes and the modeler’s model remain private to all other parties. This reflects the notion that decisions should be blind to an individual’s status—depicted in courtrooms by a blindfolded Lady Justice holding balanced scales (Bennett Capers, 2012). In our framework, we assume the existence of a regulator with fairness aims (e.g., a data protection authority or anti-discrimination agency).

Desirable fairness and accountability applications we enable include: 1. Fairness certification: Given a model and a dataset of individuals, check that the model satisfies a given fairness constraint; if yes, generate a certificate. 2. Fair model training: Given a dataset of individuals, learn a model guaranteed and certified to be fair. 3. Decision verification: To prevent a malicious modeler from using a non-certified model in practice (Kroll et al., 2016), we provide for an individual to challenge an outcome and check that it matches the outcome from a certified model.

Our extension of recent theoretical developments in MPC to admit linear constraints may be of independent interest. We demonstrate the real-world efficacy of our methods, and make our code publicly available at https://github.com/nikikilbertus/blind-justice.
2. Fairness and Privacy Requirements

Assumptions and Incentives. We assume three categories of participants: a modeler $M$, a regulator $REG$, and users $U_1 \ldots U_n$. Each user has a vector of binary sensitive attributes $z_i \in \{0,1\}^p$ (e.g., ethnicity or gender), a vector of non-sensitive features $x_i \in \mathbb{R}^n$ (discrete or real), and a non-sensitive recorded outcome $y_i \in \{0,1\}$—the label. We collect user data into matrices $X \in \mathbb{R}^{n \times d}$, $Z \in \{0,1\}^{n \times p}$ and a label vector $y \in \{0,1\}^n$ and refer to non-sensitive data by $D = (X,y)$. The source of societal concern is that sensitive attributes $z_i$ are potentially correlated with $x_i$, $y_i$.

Modeler $M$ wishes to train a model $f_\theta : \mathbb{R}^d \to \{0,1\}$, keeping $\theta$ private. The model $f_\theta$ does not use $z_i$ as input to prevent disparate treatment. For each user $U_i$, $M$ observes or is provided $x_i$, $y_i$. The sensitive information in $z_i$ is required to ensure $f_\theta$ meets a given fairness condition $\mathbb{F}$, but the users want to keep it private from all other parties. The regulator $REG$ aims to ensure that $M$ deploys only models that meet fairness condition $\mathbb{F}$. It has no incentive to collude with $M$. Further, $M$ might be obliged to demonstrate to $REG$ that condition $\mathbb{F}$ is met, before publicly deploying $f_\theta$.

In this work we focus on a variant of balancing acceptance rates across demographic groups, formulated as a constrained optimization problem by Zafar et al. (2017) mimicking the $p\%$-rule: for any binary protected attribute $z \in \{0,1\}$, it aims to achieve

$$\min \left\{ \frac{P(\hat{y} = 1 | z = 1)}{P(\hat{y} = 1 | z = 0)}, \frac{P(\hat{y} = 1 | z = 0)}{P(\hat{y} = 1 | z = 1)} \right\} \geq \frac{p}{100}, \quad (1)$$

We believe that a similar MPC approach could also be used for balancing accuracy, true positive rates, and true negative rates, i.e., all measures which, to our knowledge, have been addressed with efficient, non-private methods.

Fairness Certification. The modeler $M$ works with the regulator $REG$ to obtain a certificate that model $f_\theta$ is fair. To do so, we propose that users send their non-sensitive data $D$ to $REG$; and send private versions of their sensitive data $Z$ to both $M$ and $REG$. Neither $M$ nor $REG$ can read the sensitive data. However, we can design a secure protocol between $M$ and $REG$ to certify if the model is fair. This setup is shown in Figure 1 (Left).

Privacy constraints: (C1) privacy of sensitive user data, (C2) model secrecy, and (C3) minimal disclosure of $D$ to $M$.

Decision Verification. We aim to avoid that a malicious $M$ swaps a successfully certified model $f_\theta$ for a potentially unfair model $f_{\theta'}$ in the real world. Upon receiving a decision $\hat{y}$, a user can challenge it by asking $REG$ for a verification. The verification consists of $M$ and $REG$ jointly verifying that $f_{\theta'}(x) = f_\theta(x)$, where $x$ are the user’s non-sensitive features. While there is no simple technical way to prevent a malicious $M$ from deploying an unfair model, they will get caught if a user challenges a decision that would differ under $f_{\theta'}$. This setup is shown in Figure 1 (Right).

Privacy constraints: (C1) privacy of sensitive user data, and (C2) model secrecy.

Design Choices. We use a regulator for several reasons. While in principle MPC can be carried out without a regulator, using all users as parties, this comes at a significantly greater computational cost and the requirement of all users being online simultaneously. Additionally, given that industry has traditionally resisted privacy-enhancing technologies (Brown, 2014), a coordinating body has been highlighted as practically important in this area (The Royal Society and the British Academy, 2017).

If users are uncomfortable sharing $D$ in the clear, we can keep all of $x_i$, $y_i$, $z_i$ private throughout the three tasks using recent MPC methods, with computational cost increasing only by a factor of 2 (Mohassel & Zhang, 2017). While this extension would sometimes be desirable, it hinders exploratory data analysis by the modeler and validation by the regulator that user-provided data is correct. In this work, privacy or secrecy constraints are separate from other attacks such as model extraction (Tramèr et al., 2016) or inversion (Fredrikson et al., 2015). If relevant, modelers may need to consider these separately.

3. Our MPC Solution

Fair Multi-Party Machine Learning. MPC protocols allow two parties $P_1$ and $P_2$ holding secret values $x_1$ and $x_2$ to evaluate a public function $f$ such that the parties (either both or one of them) learn only $f(x_1, x_2)$. In our setting, $f$ will be model training, verification, or certification and the parties involved are the modeler $M$ and the regulator $REG$.

Because generic solutions do not yet scale to real-world tasks, custom tailored protocols have proved successful, for example in logistic and linear regression (Nikolaenko et al., 2013b; Gascón et al., 2017; Mohassel & Zhang, 2017), neural network training (Mohassel & Zhang, 2017) and evaluation (Juvekar et al., 2018; Liu et al., 2017), matrix factorization (Nikolaenko et al., 2013a), and principal component analysis (Al-Rubaie et al., 2017).
All protocols consist of the two parties jointly evaluating arithmetic and/or Boolean circuits gate by gate. Because floating point arithmetic cannot be represented by practical circuits (see Demmler et al., 2015, Table 4), we are limited to fixed-point arithmetic and need to approximate non-linear functions by (piecewise) linear ones. We use additive secret sharing to keep intermediate values hidden from other parties, i.e., $z$ lives in $Z_q$ with $q = 2^{64}$. To share $z$, a user draws $r$ uniformly at random from $Z_q$, sends $r$ to $M$ and $z-r$ to $R$. None of these values reveals anything about $z$, but $z$ can easily be reconstructed within the MPC computation. In Figure 1, we thus reinterpret the key held by $R$ and the encrypted $z$ by $M$ as their corresponding shares.

For fair training we follow the techniques introduced by Mohassel & Zhang (2017) and extend them to handle linear constraints. This extension may be of independent interest, and has applications for privacy-preserving machine learning beyond fairness. Certification and verification partly correspond to sub-procedures of the fair training task, hence do not add technical difficulties. For certification, we check that $\theta$ satisfies $\mathbb{F}$ in MPC, followed by computing a signature $s(\theta)$. We use cryptographic hash functions such as SHA-256 that can be evaluated quickly in MPC for unencrypted sharing: encrypted data: encrypted operations:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Adult</th>
<th>Bank</th>
<th>COMPAS</th>
<th>German</th>
<th>SQF</th>
</tr>
</thead>
<tbody>
<tr>
<td>n examples</td>
<td>214</td>
<td>215</td>
<td>214</td>
<td>216</td>
<td>216</td>
</tr>
<tr>
<td>d features</td>
<td>51</td>
<td>62</td>
<td>7</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>p sensitive attr.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>certification training</td>
<td>802 ms</td>
<td>827 ms</td>
<td>288 ms</td>
<td>250 ms</td>
<td>765 ms</td>
</tr>
<tr>
<td>Fair Model Training</td>
<td>43 min</td>
<td>51 min</td>
<td>7 min</td>
<td>1 min</td>
<td>111 min</td>
</tr>
</tbody>
</table>

where the absolute value is taken element-wise.

**Technical Challenges.** Zafar et al. (2017) use Sequential Least Squares Programming (SLSQP), which requires solving a sequence of quadratic programs and non-integer divisions by non-constant numbers. These operations are currently infeasible within MPC.

Instead, we run stochastic gradient descent and use Lagrangian multipliers, which performed consistently better than projected gradient descent and interior log barrier (Boyd & Vandenberghe, 2004) in our experiments. Hence, we minimize $\mathcal{L} := \mathcal{L}_{BCE}(X, y, \theta) + \lambda \max\{\mathbb{F}(\theta), 0\}$ for the standard logistic regression loss $\mathcal{L}_{BCE}$ with alternating updates $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}$ and $\lambda \leftarrow \max\{\lambda + \eta_\lambda \nabla_\lambda \mathcal{L}, 0\}$, where $\eta_\theta, \eta_\lambda$ are learning rates. The gradients only require matrix multiplications and a single evaluation of the logistic function, which we approximate by 0 and 1 for $x < -0.5$ and $x > 0.5$ respectively, and linear in between (Mohassel & Zhang, 2017). The largest number representable in fixed-point format with $m$ integer and $n$ fractional bits is roughly $2^{m}+1$. Since we whiten the features $X$ column-wise, we need to be careful whenever we add more than $2^m$ numbers, limiting the minibatch size.

For large $n$, evaluating $\hat{Z}^T X / n$ in the fairness function $\mathbb{F}$ is particularly problematic. To avoid under- and overflows, we perform it in blocks of size $b < 2^m$, divide each block by $b$, and multiply the sum of all blocks by $b/n > 2^{-m}$. Moreover, using powers of two for $n$ and the minibatch size, we can use fast bit shifts to avoid prohibitively expensive divisions.

![Figure 1. Our setup for Fairness certification (Left), Fair model training (Center), and Decision verification (Right).](image-url)
We now show that our approach is feasible for all three tasks on 5 real-world datasets, namely the Adult, Bank, and German datasets from the UCI machine learning repository (Lichman, 2013), the stop, question and frisk 2012 dataset (SQF),1 and the COMPAS dataset (Angwin et al., 2016), see Table 1. Moreover, we also run on synthetic data, generated as described by Zafar et al. (2017), because it allows us to control the correlation between the sensitive attributes and the class labels. We compare to SLSQP as a baseline. Our constraint value covers the range $[10^{-4}, 10^0]$ and a corresponding range for SLSQP.

**Accuracy and Fairness.** The first row of Figure 2 shows test set accuracies over the fairness constraint (smaller is tighter). Like the synthetic dataset, Adult, Bank, and COMPAS exhibit a visible trade-off between accuracy and fairness. The German dataset contains only 512 training examples, which explains the varying and seemingly discrete accuracy. For SQF, accuracy slightly improves as the constraints become active. Further investigation is needed to determine the cause of this behavior. We suspect the constraint to act as a regularizer for SGD. Despite fixed point arithmetic, approximate non-linearities, and SGD, our method retains comparable accuracy across all datasets.

The second row of Figure 2 shows how well disparate impact is mitigated as the fractions of users with positive outcomes in each group gradually approach as we decrease $c$, i.e., increase $p$. The effect is most pronounced for the first three datasets. The COMPAS dataset has 7 sensitive attributes (some with only 10 positive instances) quickly leading to infeasible constraints. Consequently, the $p$%-rule needs careful interpretation when applied for multiple sensitive attributes. A random sensitive attribute is chosen in the second row for COMPAS. Variations in the German dataset are due to its small size. In COMPAS, SQF, and arguably in Bank, the classifier tends to collapse to negative or positive outcomes as $c$ decreases, i.e., we require fractions to be balanced exactly. Our method is competitive in removing disparate impact while retaining high accuracy on all but the challenging COMPAS dataset.

**Runtime.** In Table 1 we show the online running times on a laptop computer. While training takes orders of magnitudes longer than a non-MPC implementation, our approach still remains feasible and realistic. We use the offline precomputation of multiplication triples as described and timed in Mohassel & Zhang (2017, Table 2). Certification consists of checking $F(\theta) > 0$, which is already done for each gradient update during training and only takes negligible computation time, see Table 1. Similarly, the operations required for verification stay well below one second.

## 5. Conclusion

Real world fair learning suffers from a dilemma: in order to enforce fairness, sensitive attributes must be examined; yet in many situations, users may feel uncomfortable in revealing these attributes, or modelers may be legally restricted in collecting and utilizing them. By introducing and extending recent methods from MPC, we have demonstrated that it is practical for certain outcome-based notions of group fairness on real-world datasets to: (i) certify and sign a model as fair; (ii) learn a fair model; and (iii) verify that a fair-certified model has indeed been used; all while maintaining cryptographic privacy of all users’ sensitive attributes. Connecting concerns in privacy, algorithmic fairness and accountability, our proposal reduces the barrier for regulators to provide better oversight, modelers to develop fair and private models, and users to retain control over data they consider highly sensitive. Issues of social fairness are complex. We propose our approach as one tool which may be useful in some settings to mitigate concerns around machine learning and society.

1https://perma.cc/6CSM-N7AQ
Acknowledgments

The authors would like to thank Chris Russell and Philipp Schoppmann for useful discussions and help with the implementation, as well as the anonymous reviewers for helpful comments. AG and MK were supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. MV was supported by EPSRC grant EP/M507970/1. AW acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via the CFI.

References


Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There is software used across the country to predict future criminals and it is biased against blacks. ProPublica, 2016.


Žliobaitė, I. and Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence and Law, 2016.