Blind Justice: Fairness with Encrypted Sensitive Attributes

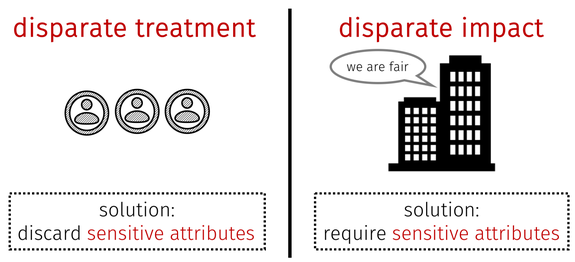
Niki Kilbertus[1,2], Adrià Gascón[3,4], Matt J. Kusner[3,4], Michael Veale[5], Krishna P. Gummadi[6], Adrian Weller[1,3]

[1]University of Cambridge, [2]MPI for Intelligent Systems, [3]The Alan Turing Institute, [4]Warwick University, [5]University College London, [6]MPI for Software Systems

Motivation and Notation

Dilemma in Fair Learning

U.S. law requires decisions in credit, education, employment, and housing do not cause:



To enforce fairness, **sensitive attributes must be examined**. Yet, users may feel **uncomfortable in revealing these attributes** or modelers may be **legally restricted in utilizing them** [1, 2].

Notation

- the users, i.e., individuals using a service
- the modeler providing a service, e.g., bank, insurance company, etc.
- the regulator, e.g., governmental institution, non-profit, etc.
- $\mathbf{x}$ are the non-sensitive **features**, e.g., GPA, salary, etc.
- y is the (non-sensitive) **label**, e.g., paid back loan, recidivism, etc.
- $\mathbf{z}$ are the **sensitive attributes**, e.g., gender, race, etc.
- θ are model parameters
- $s_{\mathbb{F}}(\theta)$ is a signature of a model

Secure Multi-Party Computation (MPC)

MPC allows two (or more) parties holding secret values to evaluate an agreed-upon function without learning anything besides the outcome and what can be inferred from it [3].

Remark: Here, privacy and secrecy constraints are separate from setup-dependent attacks, like model extraction or inversion (see differential privacy).

Challenges

fixed-point arithmetic may lead to under- and overflow

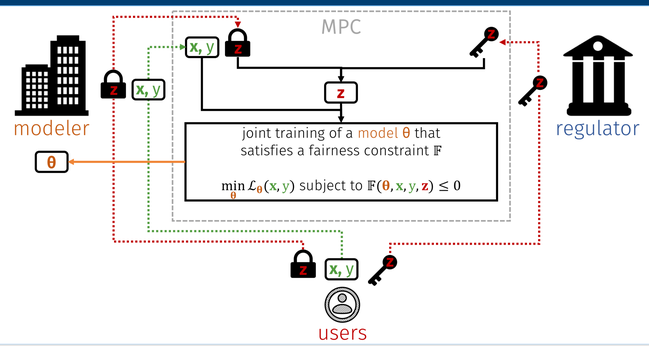
approximate non-linearities may lead to loss of accuracy

Theoretical Guarantees

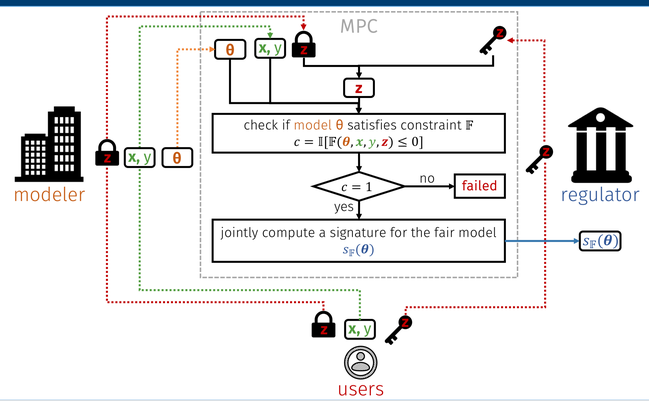
Proposition. For non-colluding modeler and regulator, our protocols implements the functionality of each setting 1), 2), 3) satisfying cryptographic privacy of sensitive user data and model secrecy in the presence of a semi-honest adversary.

Remark: Certification and verification are sub-processes of model training.

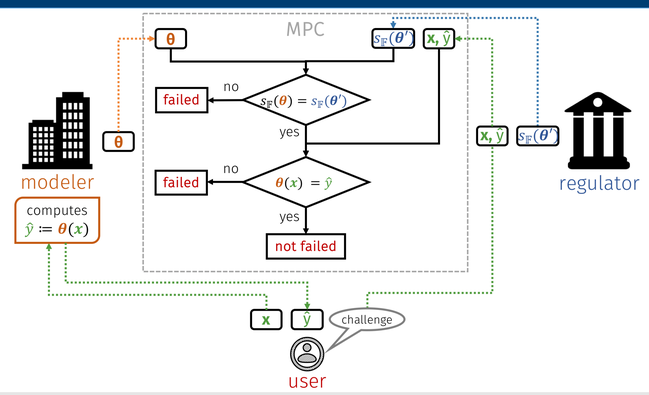
1) Fair Model Training



2) Model Certification



3) Decision Verification



References

[1] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. *Fairness through awareness*. ITCS, 2012.

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

[3] Mohassel, P., Zhang, Y. *SecureML: A system for scalable privacy-preserving machine learning*. S&P, 2017.

[4] Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P. *Fairness Constraints: Mechanisms for Fair Classification*. AISTATS, 2017.

Experiments

Fairness Notion: p%-Rule (Acceptance Rate Parity)

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

with the linear proxy [4]

$$\mathbb{F}(\theta) := \frac{1}{n}|(Z-\overline{Z})^T X\theta| - c \leq 0$$

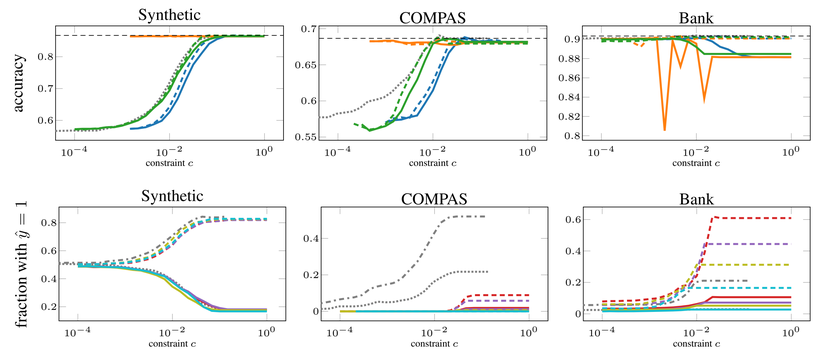
sensitive attributes; mean of sensitive attributes; non-sensitive features; constraint value

Fixed-Point-Friendly Optimization Techniques

baseline:	non-private SLSQP [4]
Lagrangian multipliers:	$\min_\theta \sum_{i=1}^n \ell_\theta(x_i, y_i) + \lambda \max\{\mathbb{F}(\theta, x_i, y_i, z_i), 0\}$
projected gradients:	$\theta \leftarrow \pi_{\mathbb{F}}(\theta - \eta\nabla_\theta)$
interior point log barrier:	$\min_\theta \sum_{i=1}^n \ell_\theta(x_i, y_i) - \mu\log(-\mathbb{F}(\theta, x_i, y_i, z_i))$

Results

Accuracy and p%-Rule



Datasets and Feasibility

	Adult	Bank	COMPAS	German	SQF
n training examples	2^{14}	2^{15}	2^{12}	2^{9}	2^{16}
d features	51	62	7	24	23
p sensitive attributes	1	1	7	1	1
certification	802 ms	827 ms	288 ms	250 ms	765 ms
training (online time, 10 epochs)	43 min	51 min	7 min	1 min	111 min

Conclusion

Addressing concerns in privacy, fairness, and accountability, our proposal helps empower regulators to provide better oversight, modelers to develop fair and secret models, and users to retain control over sensitive data.

arXiv:1806.03281

nk470@cam.ac.uk, agascon@turing.ac.uk, mkusner@turing.ac.uk, m.veale@ucl.ac.uk, gummadi@mpi-sws.org, aw665@cam.ac.uk