Report - Fair Learning with Private Demographic Data

Ramanathi Manjunath

180050083

Under Supervision of

Prof. Harish Guruprasad Ramaswamy

Computer Science and Engineering

Indian Institute of Technology, Bombay

2021

Abstract

Sensitive attributes such as race are rarely available to models in real world as their collection is often restricted by laws and regulations. But the Opposing notion is that Government still wants to achieve non-discrimination on all genders. The paper "Fair Learning with Private Demographic Data" by Mozannar et al, tries to solves this issue by using derived private attributes and learn models from them.

The method follows a two step process beginning with problem formulation - pre-processing learning with normal models and post processing correction to minimize the discrimination while not severely effecting the error. We then try to visualize the results using a set of employee details with gender as sensitive attribute.

Table of Contents

Chapter 1

Introduction

As algorithmic systems driven by machine learning start to play an increasingly important role in society, concerns arise over their compliance with laws, regulations and societal norms. To ensure non-discrimination in learning tasks, knowledge of the sensitive attributes is essential, however, laws and regulation often prohibit access and use of this sensitive data. Hence we want our system to be non-discriminatory while maintaining the privacy of our sensitive attributes.

One potential workaround to this problem, ignoring legal feasibility, is to allow the individuals to release their data in a locally differentially private manner and then try to learn from this privatized data a non-discriminatory predictor. This allows us to guarantee that our decisions are fair while maintaining a degree of individual privacy to each user

Chapter 2

Methodology

2.1 Problem Formulation

Let $\hat{Y}$ be binary predictor of target $Y \in \{0, 1\}$ is a function of non-sensitive attributes $X \in \mathbf{X}$ and sensitive (or protected) attribute $A \in \mathbf{A}$. Our focus here is on statistical notions of group-wise non-discrimination and their abstraction is defined as follows let $\varepsilon 1$, $\varepsilon 2$ be two probability events defined with respect to (X, Y, $\hat{Y}$)

$$\mathrm{P}\ (\varepsilon 1 \mid \varepsilon 2,\ \mathrm{A} = \mathrm{a}) = \mathrm{P}(\varepsilon 1 \mid \varepsilon 2,\ \mathrm{A} = \mathrm{a'})\ \forall a, a' \in \mathbf{A}$$

For $\varepsilon 1 = \{\ \hat{Y} = 1\ \}$ and $\varepsilon 2 = \{\ \mathrm{Y} = \mathrm{y}\ \}$, we have
Equalized odds (EO) :

$$\text{if } \forall a \in \mathbf{A}\ ,\ \mathrm{P}(\ \hat{Y} = 1|\mathrm{A} = \mathrm{a},\ \mathrm{Y} = \mathrm{y}) = \mathrm{P}(\ \hat{Y} = 1|\mathrm{Y} = \mathrm{y})\ \forall y \in \{0, 1, \},$$

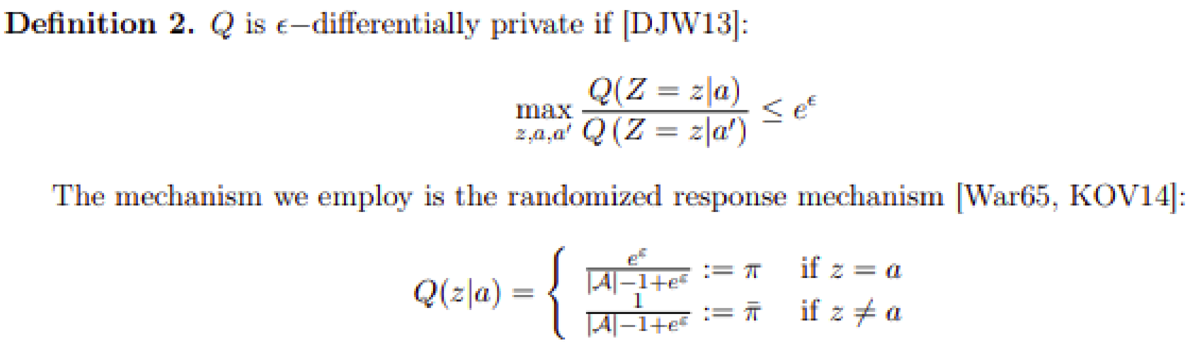
Define $\gamma_{y,a}$ ($\hat{Y}$) = P($\hat{Y}$= 1|Y = y, A = a), then $\hat{Y}$ satisfies α-EO with respect to A if:

$$\max_{y \in \{0,1\}, a \in A}\ \Gamma_{ya} := \gamma_{y,a}(\ \hat{Y}\ )\ \text{-}\ \gamma_{y,0}(\ \hat{Y}\ ) \leq \alpha$$

Local differential privacy : (LDP) guarantees that the entity holding the data does not know for certain the protected attribute of any data point, which in turn makes sure that any algorithm built on this data is deferentially private

we assume that we have access to n samples S of the form $(x_i, y_i, z_i), i = \{1, 2, ..., n\}$

Z is sampled from Q(.|A), defined as shown below independently from X and



Definition 2. Q is $\epsilon-$differentially private if [DJW13]:

$$\max_{z,a,a'} \frac{Q(Z=z|a)}{Q(Z=z|a')} \leq e^{\epsilon}$$

The mechanism we employ is the randomized response mechanism [War65, KOV14]:

$$Q(z|a) = \begin{cases} \frac{e^{\varepsilon}}{|\mathcal{A}|-1+e^{\varepsilon}} := \pi & \text{if } z = a \\ \frac{1}{|\mathcal{A}|-1+e^{\varepsilon}} := \bar{\pi} & \text{if } z \neq a \end{cases}$$

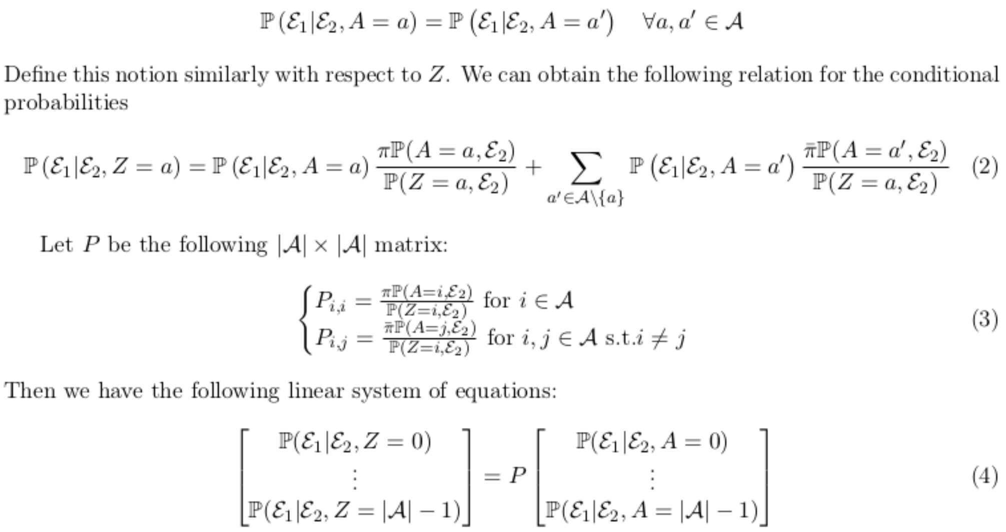
Y . We call Z the *privatized protected attribute*

let $q_{y,a}$($\hat{Y}$) = P(Ŷ = 1|Y = y, Z = a), note that $\hat{Y}$ satisfies α -EO with respect to Z if:

$$\max_{y\in\{0,1\},a\in Z} \mid q_{y,a}(\hat{Y}) \text{ - } q_{y,0}(\hat{Y}) \mid \leq \alpha$$

2.2 Proposition

Auditing for discrimination : Consider any exact non-discrimination notion among equalized odds, demographic parity, accuracy parity, or equality of false discovery/omission rates. Let $\hat{Y}$:= h(X) be a binary predictor, then $\hat{Y}$ is non-discriminatory with respect to A if and only if it is non-discriminatory with respect to Z.



$$\mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A=a) = \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A=a') \quad \forall a, a' \in \mathcal{A}$$

Define this notion similarly with respect to Z. We can obtain the following relation for the conditional probabilities

$$\mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, Z=a) = \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A=a)\frac{\pi\mathbb{P}(A=a,\mathcal{E}_2)}{\mathbb{P}(Z=a,\mathcal{E}_2)} + \sum_{a'\in\mathcal{A}\backslash\{a\}} \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A=a')\frac{\bar{\pi}\mathbb{P}(A=a',\mathcal{E}_2)}{\mathbb{P}(Z=a,\mathcal{E}_2)} \quad (2)$$

Let P be the following $|\mathcal{A}| \times |\mathcal{A}|$ matrix:

$$\begin{cases} P_{i,i} = \frac{\pi\mathbb{P}(A=i,\mathcal{E}_2)}{\mathbb{P}(Z=i,\mathcal{E}_2)} \text{ for } i \in \mathcal{A} \\ P_{i,j} = \frac{\bar{\pi}\mathbb{P}(A=j,\mathcal{E}_2)}{\mathbb{P}(Z=i,\mathcal{E}_2)} \text{ for } i,j \in \mathcal{A} \text{ s.t.} i \neq j \end{cases} \quad (3)$$

Then we have the following linear system of equations:

$$\begin{bmatrix} \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, Z=0) \\ \vdots \\ \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, Z=|\mathcal{A}|-1) \end{bmatrix} = P \begin{bmatrix} \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A=0) \\ \vdots \\ \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A=|\mathcal{A}|-1) \end{bmatrix} \quad (4)$$

2.3 Learning Fair Predictors

let H be a hypothesis class of functions that depend only on X
Let ΔH be the set of all distributions over H, and denote such a randomized predictor by Q $\in \Delta$ H
The goal is to learn a predictor that approximates the performance of the optimal non-discriminatory distribution

$$\mathrm{Y}^{\star} = \mathrm{argmin}_{Q \in \Delta H} P(Q(X) \neq Y)$$

$$\text{s.t. } \gamma_{y,a} \text{ (Q)} = \gamma_{y,0} \text{ (Q) } \forall \text{y} \in \{0,1\},\ \forall a \in \mathbf{A}$$

Our approach is to adapt the two-step procedure :

- The first step is to learn an approximately non-discriminatory predictor $\hat{Y}$ = Q(X) with respect to Z on S1(half of the dataset)
- The aim of the second step is to produce a final predictor $\tilde{Y}$ that corrects for this discrimination, without increasing its error by much to get $\tilde{Y}$ = f ($\hat{Y}$, Z)

2.3.1 Step I

$$\hat{Y} = \mathrm{argmin}_{Q \in \Delta H} err^{S1}(Q(X))$$

$$\text{s.t. } \mathrm{max}_{y \in \{0,1\}, a \in Z} \mid q^{S1}_{y,a}(\text{ Q }) \text{ - } q^{S1}_{y,0}(\text{ Q }) \mid \leq \alpha_n$$

where, we use the shorthand err(Q) = P(Q(X) $\neq$ Y) and S1 to be first half of dataset defined as before Let J = Y × A, K = Y × A \ { 0 } × {-, +} and define $\gamma(Q) \in R^{|J|}$ with $\gamma(Q)_{(y,a)} = \gamma_{y,a}(Q)$, with the matrix M $\in R^{|K| \times |J|}$
Define the Lagrangian:

$$L(Q, \boldsymbol{\lambda}) = \mathrm{err}(Q) + \boldsymbol{\lambda}^{\top}(M\boldsymbol{\gamma}(Q) - \alpha\mathbf{1})$$

We constrain the norm of λ with B $\in$ R^{+} and consider the following two dual problems (a saddle constraint problem):

$$\min_{Q\in\Delta_{\mathcal{H}}} \max_{\boldsymbol{\lambda}\in\mathbb{R}_{+}^{|\mathcal{K}|}, ||\boldsymbol{\lambda}||_1\leq B} L(Q,\boldsymbol{\lambda}) \qquad\qquad \max_{\boldsymbol{\lambda}\in\mathbb{R}_{+}^{|\mathcal{K}|}, ||\boldsymbol{\lambda}||_1\leq B} \min_{Q\in\Delta_{\mathcal{H}}} L(Q,\boldsymbol{\lambda})$$

Algorithm 1: Exp. gradient reduction for fair classification [ABD+18]

```
Input: training data $(X_i, Y_i, Z_i)_{i=1}^{n/2}$, bound $B$, learning rate $\eta$, rounds $T$
$\boldsymbol{\theta}_1 \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{K}|}$
for $t = 1, 2, \cdots, T$ do
    $\boldsymbol{\lambda}_{t,k} \leftarrow B\frac{\exp(\theta_{t,k})}{1+\sum_{k'}\exp(\theta_{t,k})} \forall k \in \mathcal{K}$
    $h_t \leftarrow \mathrm{BEST}_h(\boldsymbol{\lambda}_t)$
    $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \eta(M\boldsymbol{\gamma}^S(h_t) - \alpha_n\mathbf{1})$
end
$\hat{Y} \leftarrow \frac{1}{T}\sum_{t=1}^{T} h_t, \hat{\boldsymbol{\lambda}} \leftarrow \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\lambda}_t$
Return $(\hat{Y}, \hat{\boldsymbol{\lambda}})$
```

The auditor follows the exponentiated gradient algorithm and the learner picks it's best response to the auditor. The approach is fully described in Algorithm described as above.

2.3.2 Step II

The derived second step predictor $\tilde{Y}$ is fully characterized by 2|A| probabilities P($\tilde{Y}$ = 1| $\hat{Y}$ = $\hat{y}$, Z = a) := p$_{\hat{y},z}$.

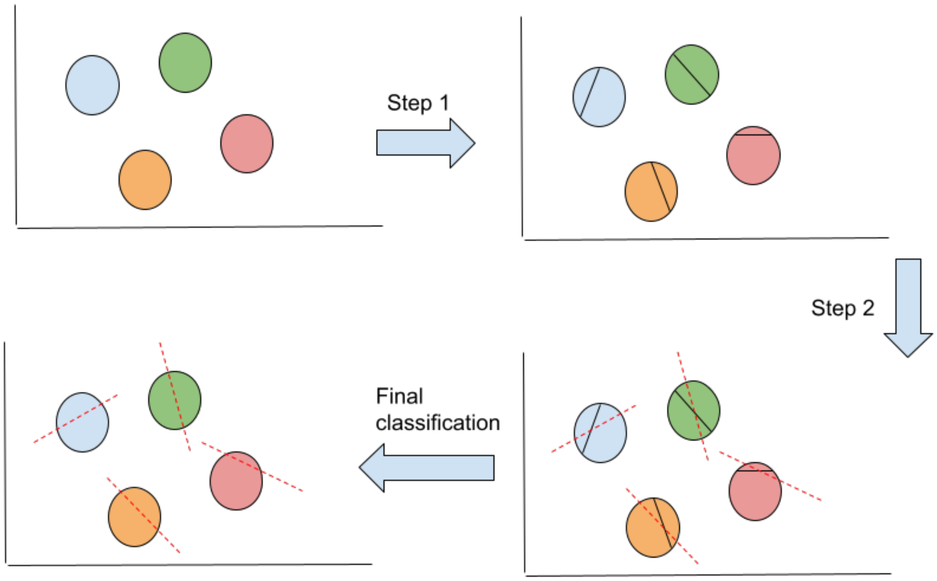
$$\begin{aligned}\mathbb{P}(\widetilde{Y} = 1|Y = y, A = a) = \mathbb{P}(\widetilde{Y} = 1|\hat{Y} = 0, A = a)\mathbb{P}(\hat{Y} = 0|Y = y, A = a)\\ + \mathbb{P}(\widetilde{Y} = 1|\hat{Y} = 1, A = a)\mathbb{P}(\hat{Y} = 1|Y = y, A = a)\end{aligned}$$

Constrained linear program for $\tilde{Y}$:

$$\widetilde{Y} = \arg\min_{p_{\cdot,\cdot}} \sum_{\hat{y},a} \left(\widetilde{\mathbb{P}}^{S_2}(\hat{Y} = \hat{y}, Z = a, Y = 0) - \widetilde{\mathbb{P}}^{S_2}(\hat{Y} = \hat{y}, Z = a, Y = 1)\right) \cdot \widetilde{p}_{\hat{y},a}$$

$$\begin{aligned} s.t. \quad & \Big|\widetilde{p}_{0,a}\widetilde{\mathbb{P}}^{S_2}(\hat{Y} = 0|Y = y, A = a) + \widetilde{p}_{1,a}\widetilde{\mathbb{P}}^{S_2}(\hat{Y} = 1|Y = y, A = a) \\ & - \widetilde{p}_{0,0}\widetilde{\mathbb{P}}^{S_2}(\hat{Y} = 0|Y = y, A = 0) - \widetilde{p}_{1,0}\widetilde{\mathbb{P}}^{S_2}(\hat{Y} = 1|Y = y, A = 0)\Big| \leq \widetilde{\alpha}_n, \ \forall y, a \\ & 0 \leq p_{\hat{y},a} \leq 1 \quad \forall \hat{y} \in \{0, 1\}, \forall a \in \mathcal{A}\end{aligned}$$

2.3.3 Abstraction

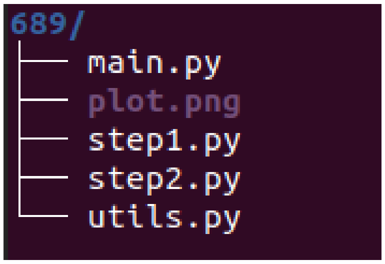


In above figure, different colors represent different classes (sensitive) and consider a binary classification. Step 1 tries to classify by normal cost function minimization methods and step 2 adjusts those classification to reduce discrimination using other half of data set without much increasing the error. Point to remember, this two step classification applies to group-wise fairness rather than individual fairness.

Chapter 3

Implementation

My implementation is based on original source code fairlearn_private_data. Because of various constraints I could not complete step 2 - Post Processing part of the implementation. So, to re-generate the desired results, I imported Post processing class. The file structure is as follows,



```
689/
├── main.py
├── plot.png
├── step1.py
├── step2.py
└── utils.py
```

The important packages used are **shap** to get dataset mentioned in paper, **sklearn** to access helper functions and **fairlearn** package for exponentiated gradient algorithm. Dataset is divided into train data, validation data and test data. test data always being one fourth and validation being one tenth of remianing in only step 1 and half of remaining in two step process.

File	**Purpose**
main.py	This is the main file and running this script gives desired graph
step1.py	contains a class about implementing Preprocessing
step2.py	contains a class about implementing PostProcessing
utils.py	Functions that are not present in any packages but are regularly used are placed here

Chapter 4

Results and Conclusion

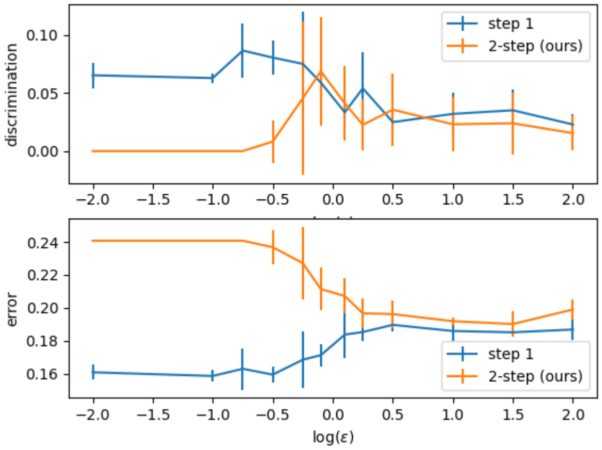
4.1 Results

Our final predictor $\tilde{Y}$ has a discrimination guarantee that is independent of the model complexity, however this comes at a cost of a privacy penalty entering the error bound as shown below. This creates a new set of trade-offs that do not appear in the absence of the privacy constraint, fairness and error start to trade-off more severely with increasing levels of privacy.

$$\mathrm{err}(\widetilde{Y}) \leq_{\delta/2} \mathrm{err}(\widetilde{Y}^{*}) + 4|\mathcal{A}|C\sqrt{\frac{\log(32|\mathcal{A}|/\delta)}{2n}}$$

$$\mathrm{disc}(\widetilde{Y}) \leq_{\delta/2} \sqrt{\frac{\log(\frac{64}{\delta})}{2n}}\,\frac{8|\mathcal{A}|C^{2}}{\min_{ya}\mathbf{P}_{ya}^{2}}$$

where $\tilde{Y}^{\star}$ being an optimal non-discriminatory derived predictor from $\hat{Y}$. The results of the above implementation are shown below. The plots are discrimination and error against level of privacy ϵ. Blue plot being when the classification is done only using step 1 and orange plot being when the discrimination is mitigated after step 1 i.e. two step method.



We observe that $\tilde{Y}$ achieves lower discrimination than $\hat{Y}$ across the different privacy levels. This comes at a cost of lower accuracy, which improves at lower privacy regimes (large epsilon). The predictor of step 1 only begins to suffer on error when the privacy level is low enough as the fairness constraint is void at high levels of privacy (small epsilon).

4.2 Conclusion

We studied learning non-discriminatory predictors when the protected attributes are privatized or noisy. We observed that, in the population limit, non-discrimination against noisy attributes is equivalent to that against original attributes. We then characterized the amount of difficulty, in sample complexity, that privacy adds to testing non-discrimination. Using this relationship, we proposed how to carefully adapt existing nondiscriminatory learners to work with privatized protected attributes.