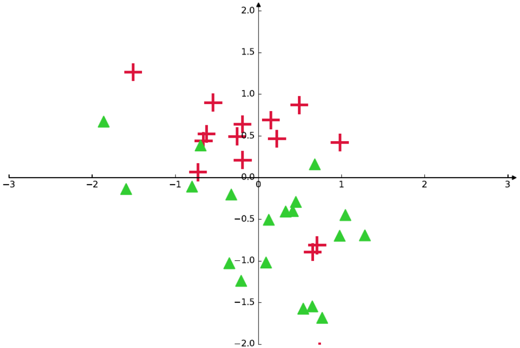
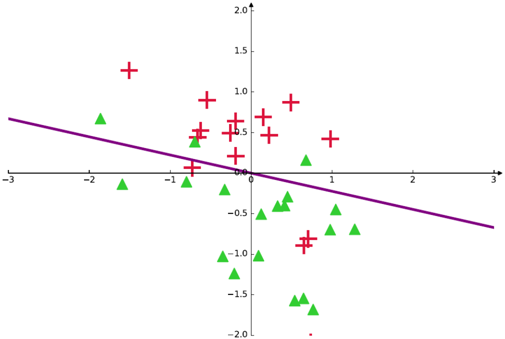
Differential Privacy has Bounded Impact on Fairness

Paul Mangold
(Joint work with Michaël Perrot, Aurélien Bellet and Marc Tommasi)

CMAP, École Polytechnique

Journées MAS
August 28th, 2024


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−3
−2
−1
0
1
2
3



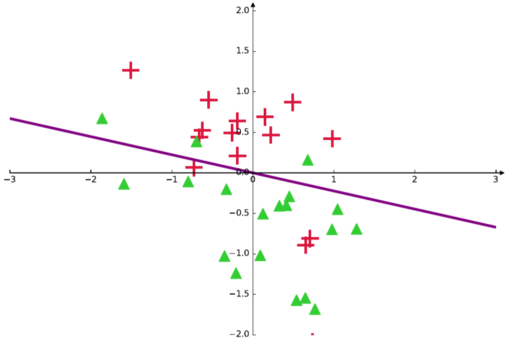
The resulting model:

- is (quite) accurate
- contains info on data

Privacy Issues?

Membership Inference:

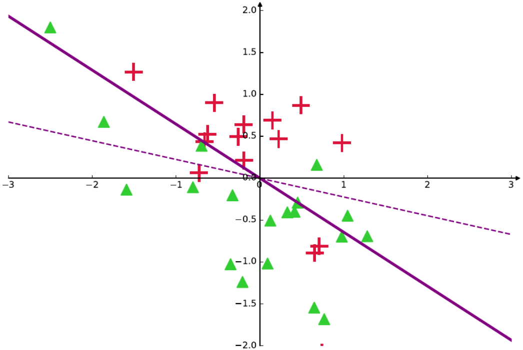
> *"determine whether a given record was part of a model's training dataset"*



Privacy Issues?

Membership Inference:

> *"determine whether a given record was part of a model's training dataset"*



Guaranteeing Privacy

Perturb the predictor with a Gaussian noise b:

$$h_w(x) = w_0 + w_1 \cdot x_1 + \cdots + w_p \cdot x_p$$

Guaranteeing Privacy

Perturb the predictor with a Gaussian noise b:

$$h_{w+b}(x) = w_0 + b_0 + (w_1 + b_1) \cdot x_1 + \cdots + (w_p + b_p) \cdot x_p$$

Guaranteeing Privacy

Perturb the predictor with a Gaussian noise b:

$$h_{w+b}(x) = w_0 + b_0 + (w_1 + b_1) \cdot x_1 + \cdots + (w_p + b_p) \cdot x_p$$

✓ noise gives plausible deniability $\rightarrow$ better privacy

✗ noisy predictions $\rightarrow$ lower accuracy

How Strong is the Protection?

$\mathcal{A} : D \mapsto w$ is (ϵ, δ)-differentially private[1]

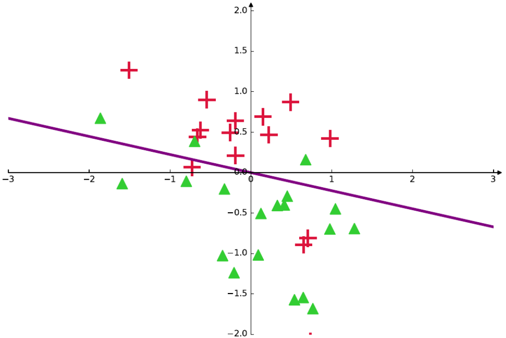
$$\mathbb{P}(\mathcal{A}(D) \in \mathcal{S}) \leq \exp(\epsilon)\mathbb{P}(\mathcal{A}(D') \in \mathcal{S}) + \delta$$

for all datasets D, D' that differ on one element, and any set $\mathcal{S}$

Rule of thumb: $\epsilon \leq 1$, $\delta = o(1/|D|)$

[1] Cynthia Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. 2006.

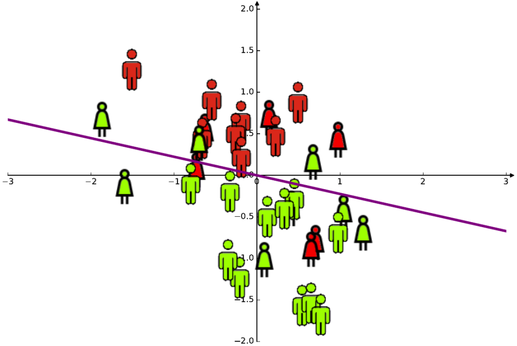
How About Fairness?



Group Fairness:

different groups can be treated differently

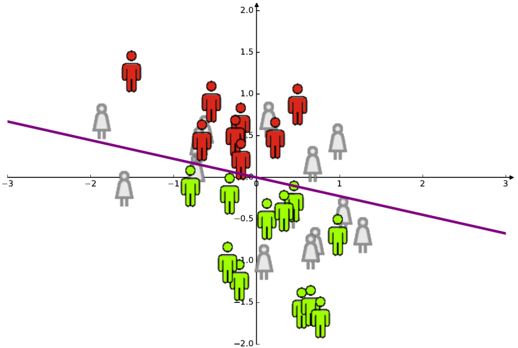
How About Fairness?



Group Fairness:

different groups can be treated differently

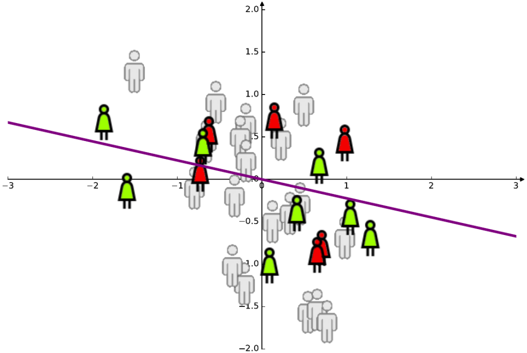
How About Fairness?



Group Fairness:

different groups can be treated differently

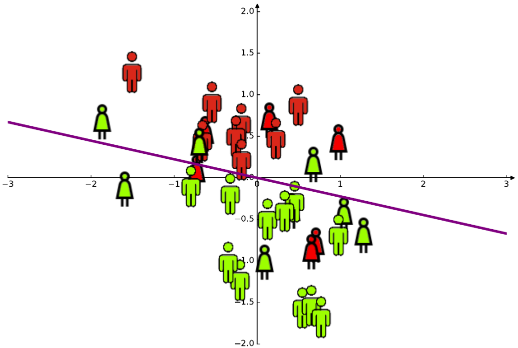
How About Fairness?



Group Fairness:

different groups can be treated differently

How About Fairness?



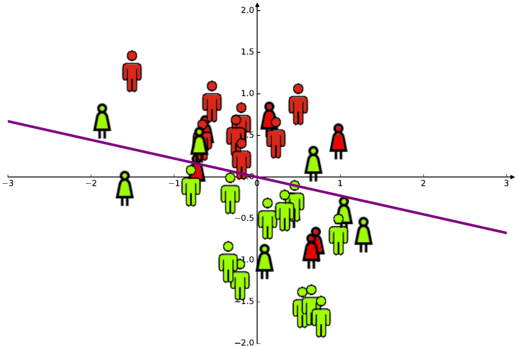
Group Fairness:

different groups can be treated differently

Note: perturbing the model can have disparate impact[2]

[2] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. "Differential Privacy Has Disparate Impact on Model Accuracy". In: *NeurIPS*. 2019.

Modelling the Problem
with a sensitive group $\mathcal{S}$



Take: $\mathcal{X} \times \mathcal{S} \to \{0, 1\}$

Goal: learn $h : \mathcal{X} \to \mathbb{R}$

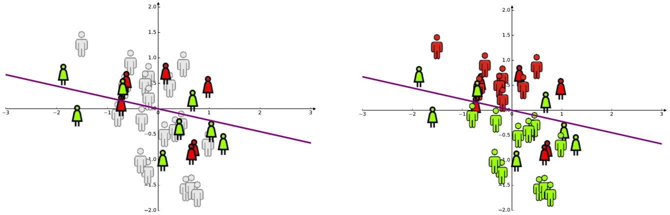
$\to$ classify $x \in \mathcal{X}$ as

$$\hat{y} = \text{sign}(h(x))$$

Measuring Group Fairness

Example: Demographic Parity[3]

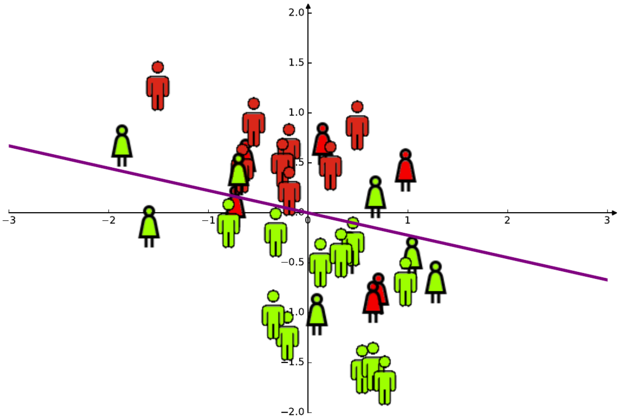
$$F_k(h) = \boxed{\mathbb{P}(h(X) > 0 | S = k)} - \boxed{\mathbb{P}(h(X) > 0)}$$



[3] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. "Building Classifiers with Independency Constraints". In: *2009 IEEE International Conference on Data Mining Workshops*. 2009.

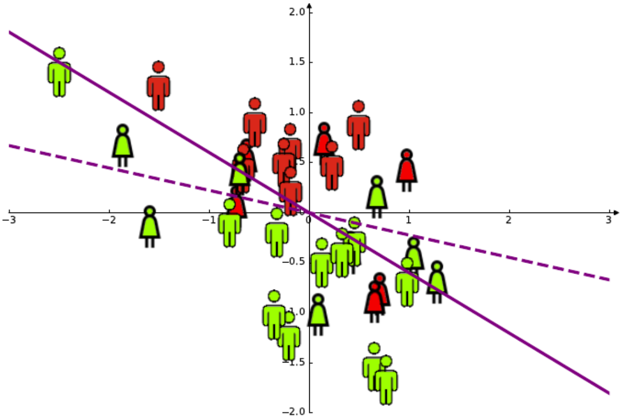
Fairness and Privacy

How much can fairness be affected by privacy?



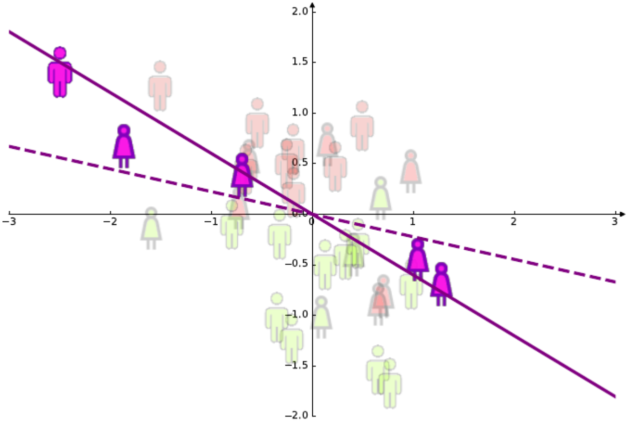
Fairness and Privacy

How much can fairness be affected by privacy?



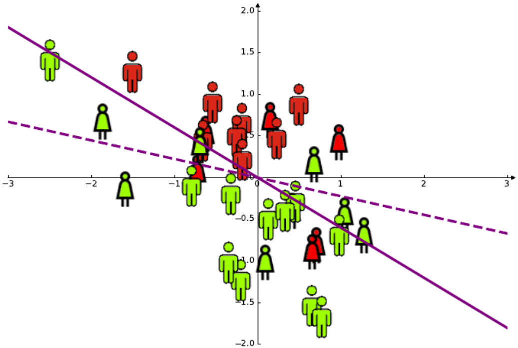
Fairness and Privacy

How much can fairness be affected by privacy?



Fairness and Privacy

How much can fairness be affected by privacy?



Key assumption:

confidence margin is lipschitz

$$|h(x) - h(x')| \leq L_{x,y} \|h - h'\|$$

for $x, y \in \mathcal{X} \times \{0, 1\}$

Bound on Difference of Fairness

Difference of Fairness

$$|F_k(h) - F_k(h')| \leq \chi_k(h) \, \|h - h'\|$$

Where $\chi_k(h) = \mathbb{E}\left(\frac{L_{X,Y}}{|h(X)|} \;\middle|\; S = k \right) + \mathbb{E}\left(\frac{L_{X,Y}}{|h(X)|} \right)$

Loss of Fairness due to Privacy is Bounded

Take $h = h_{\text{priv}}$ and $h' = h_\star$:

$$|F_k(h^{\text{priv}}) - F_k(h_\star)| \le O\left(\chi_k(h^{\text{priv}})\frac{\sqrt{p}}{n\epsilon}\right)$$

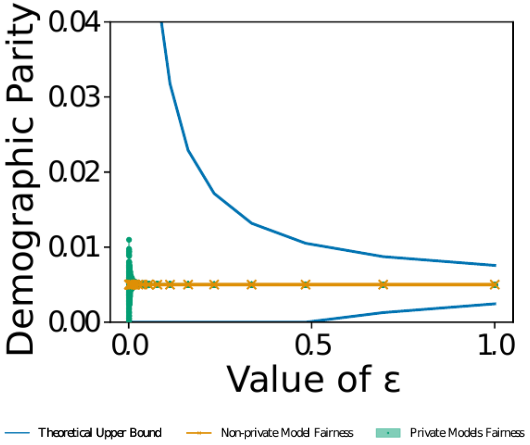
Since from DP literature (assuming strongly convex loss)[4]

$$\|h_{\text{priv}} - h_\star\| \le O\left(\frac{\sqrt{p}}{n\epsilon}\right) \qquad \text{w.h.p.}$$

$\Rightarrow$ No need to know optimal model $h_\star$!

[4] Raef Bassily, Adam Smith, and Abhradeep Thakurta. "Private ERM: Efficient Algorithms and Tight Error Bounds". In: *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. 2014.

Numerical Illustration
Not super tight, but meaningful!



- folktables dataset
- $n = 182,339$ records
- $p = 40$ features
- Green $=$ real private models

Summary

Fairness of private models:

- is "close" to the one of non-private model
- is influenced by confidence margin of the model

More results: for other group fairness measures, multi-class problems...

Open questions: use fairness-promoting methods, broader study of large-margin classifiers...

Thank you! :)

Questions?

See the Paper:

Paul Mangold et al. "Differential Privacy Has Bounded Impact on Fairness in Classification". In: *ICML*. 2023