Machine Learning for Intelligent Systems

Lecture 26: Privacy and Fairness

Reading: Dwork & Roth Chapter 1-2

Some slides thanks to Manish Raghavan and Aaron Roth

Instructors: Nika Haghtalab (this time) and Thorsten Joachims

1

Machine Learning and the Society

Privacy
Fairness
Interpretability
Accountability
Ethics
…

2

Machine Learning and Privacy

Machine Learning seems to be about general statistics of the distribution, not about any one individual.

If we take two large enough sample sets $S \sim D^m$ and $S' \sim D^m$, then effectively we should learn the same thing from S or S'.

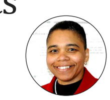
Machine learning is much more about the distribution D or the sample S as a whole, not so much about a specific $x \in S$. So, we should be able to "preserve the privacy of individuals".

Let's formalize what "privacy" means here.

3

Anonymized Data Sets

The trouble with "anonymized data" that other easily available data can "re-identify" the data set.



Latanya Sweeney

Non-anonymized Publicly available data: Voter Registration

Name	ZIP	DoB	Gender

↔

Anonymized Sensitive Data

Gender	DoB	ZIP	Entire Medical Record

At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student Sweeney started hunting for the Governor's hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000

Privacy is not the same as anonymizing the data

date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office.

4

Population level statistics

Only answer queries that are about population as a whole:

November 1

NetID	Prelim Grade
xx123	Hidden
yy123	
aa000	
zz123	

You know your friend aa000 dropped out →

November 2

NetID	Prelim Grade
xx123	Hidden
yy123	
zz123	

What's the class average? 72.75

What's the class average? 82

You can figure out aa00's prelim grade $4 \times 72.75 - 3 \times 82 = 45$.

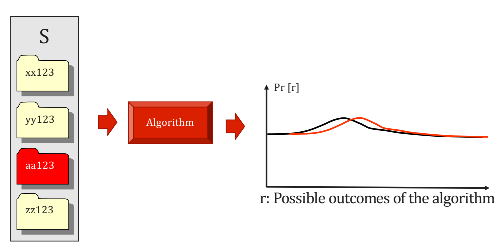
Answering too many queries very accurately reduces privacy.

5

Privacy while Learning

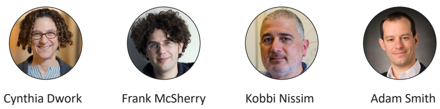
Privacy is about protecting against inferences using your data.

"An analysis of a dataset S is private if the data analyst knows almost no more about Alice after the analysis than he would have known had he conducted the same analysis on an identical database with Alice's data removed."



6

Differential Privacy



S: The data set, where each person's data is one point $x \in S$.

Differential Privacy

An algorithm $\mathcal{L}$ is ϵ-differentially private if for all pairs of datasets S, S' differing in one user's data, and for all outputs r:
$\Pr[\mathcal{L}(S) = r] \leq (1 + \epsilon) \Pr[\mathcal{L}(S') = r]$.

When $\mathcal{L}(\cdot)$ is a learning algorithm, $h = \mathcal{L}(S)$ is a classifier, that can then be applied to any x in the domain X.

Post-processing: If $\mathcal{L}(\cdot)$ is ϵ-differentially private, and f is any function, then $f(\mathcal{L}(\cdot))$ is also ϵ-differentially private.

7

Differential Privacy's Promises

- Differential Privacy and Generalization:
  →If the $h = \mathcal{L}(S)$ doesn't depend heavily on any one sample in x …
  → The algorithm does not overfit to S.
- Differential privacy promises that $h = \mathcal{L}(S)$ doesn't leak information about whose data was in S.
- We can still use differential privacy to find patterns in population:
  →If there is correlations between smoking and lung cancer, we can find it in the data.
  →If x is a smoker $h(x)$ will show high likelihood of getting cancer, and can lead to higher health insurance rate for x.
  → **Still private:** This would have happened even if your data wasn't in the medical dataset.

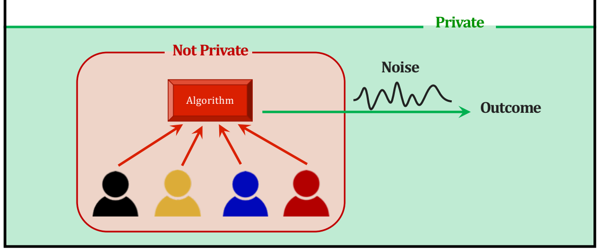
8

The "Centralized" model of Privacy

Implemented at Census, Facebook/Social Science One

The algorithm sees the data fully, but releases information that is differentially private.

Need to trust the algorithm.



9

Privately Releasing Sums

Computing a sum: Add enough noise to obscure participation of a single user in the *aggregated sum*.

"Did you travel during the Thanksgiving break?"

https://tinyurl.com/r5zt4y2

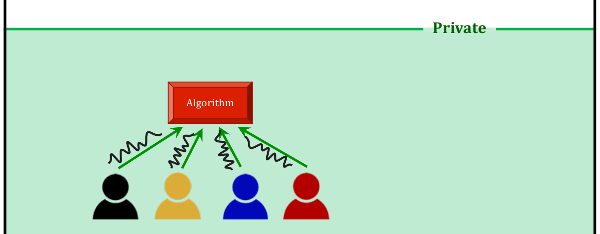
Ensuring ϵ-differential privacy:
1. Compute the *exact* answer p.
2. Perturb that answer: $\hat{p} = p + N(0, \sigma^2)$, $\sigma \approx \frac{1}{\epsilon n}$
3. Release $\hat{p}$

10

The "Distributed" model of Privacy

Implemented on iOS10, Google Chrome

Privacy protected even from the algorithm collecting the data.
- Never hold private data; no breach or subpoena risk.
- Good for when the data could be legal risk or embarrassing.



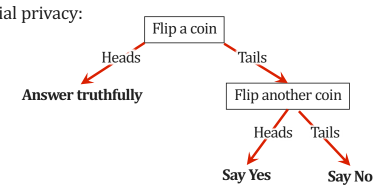
11

Randomized Response

Computing a sum: Each person adds noise to their response.

"Have you ever drunk so much alcohol that you threw up?"

Ensuring 2-differential privacy:



Answer: $p = 2\hat{p} - 0.5$. Where, $\hat{p}$: fraction of people whose response was Yes.

The standard deviation is about $\sigma \approx \frac{1}{\epsilon\sqrt{n}}$.

https://tinyurl.com/tbm7jak

12

Comparison between the two

Distributed setting: randomized response

- Error of $\pm O\left(\frac{1}{\epsilon\sqrt{n}}\right)$, for $\epsilon = 1$ and $n = 500$, error is $\approx \pm 0.044$.
- But very private. Everybody has *plausible deniability.*
- Needs more data: Facebooks and Googles can afford it.

Centralized model:

- Error of $\pm O\left(\frac{1}{\epsilon n}\right)$, for $\epsilon = 1$ and $n = 500$, error is $\approx \pm 0.002$.
- But not that private!
- Needs less data: Smaller stakeholders can also afford it.

13

Private PAC Learning

Many things can be reduced to estimating sums and fractions, e.g., the error of a classifier.

Theorem: Sample Complexity of Private Learning

Let $m \geq O\left(\max\left(\frac{\log|H|}{\epsilon\alpha}, \frac{\log|H|}{\alpha^2}\right)\right)$. For any X, $Y = \{-1, 1\}$, and distribution P on $X \times Y$, with probability 0.99 over i.i.d draws of set S of m samples and

1. Compute the $err_S(h)$ for all $h \in H$.
2. Instead of deterministically picking $h_S = \mathrm{argmin}_{h \in H}\, err_S(h)$, randomly pick one h with prob. that is exponentially decreasing in $err_S(h)$.

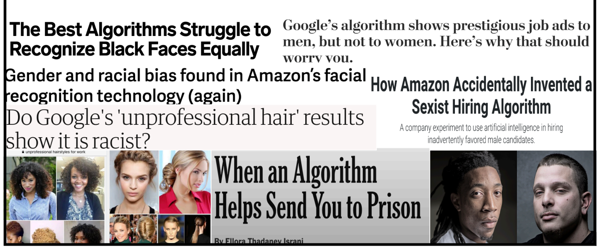
Then $err_P(h_S) \leq \min_{h^* \in H} err_P(h^*) + \alpha$ and the algorithm is ϵ-differentially private.

14

Fairness in Machine Learning

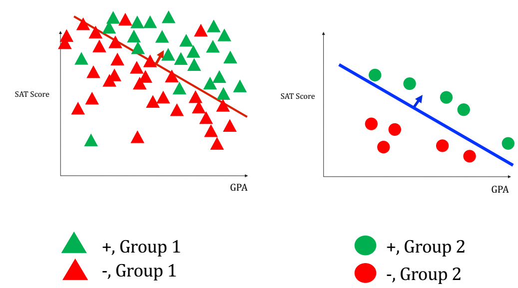
What does it mean to be fair?

- We don't agree on definitions yet. Depends heavily on the context.
- Only starting to understand the tradeoffs between different kinds of fairness and accuracy.



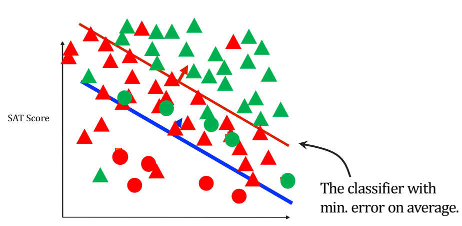
15

Why ML could be unfair?



16

Unfairness as a result of optimization



If we ignore the population and minimize the average error, we fit the majority and choose a classifier that accepts no qualified minority candidates.

17

A Case Study by ProPublica

Each person

- belongs to Positive or Negative class: for re-offending
- Belongs to race 1 or 2.

Risk tool: map people to bins based on prob. re-offending

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

18

Three Notions of Fairness

1. Balanced scores for positive class

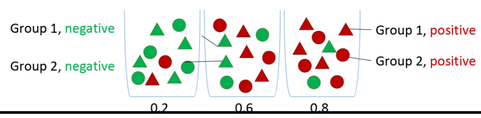
Average score assigned for group 1 **positive class** = Average score assigned for group 2 **positive class**

2. Balanced scores for negative class

Average score assigned for group 1 **negative class** = Average score assigned for group 2 **negative class**

3. Calibration of score within each group

For each group, the same fraction of people in each bin is **positive**.



19

Impossibility of Satisfying all 3

Theorem: Kleinberg, Mullainathan, and Raghavan.

In any instance of risk score assignment, it is impossible* to satisfy all three notions of fairness

*Unless the assignment problem is too trivial: can have perfect prediction or all positive and negative rates are exactly the same in both groups.

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

20

Fairness Challenges

What definitions should we use?

- Depends on the domain and how the outcomes are used by humans later.
- What if data collection was biased to start with?
- What if our decisions skew the data collection further?

....

21