

Responsible Machine Learning

Nicolas Vayatis

Lecture - On Privacy (part I)

What privacy means in data analysis

- Consider a database and a user who makes queries on the database and receives answers.
- Suppose information about Zorro can be found in the database.
- Protecting the privacy of Zorro means the user should not learn anything new about Zorro she does not already know.
- If the user may learn something about him then it should be some general characteristic of the whole population.

The flaws of privacy-preserving data analysis

But... what if the purpose of the user is to segment the population wrt to credit risk or health?

- Then, in order not to unveil the risk status of Zorro, the user should:
 - either not know Zorro belongs to the database!
 - or she should not have access to the features driving the classifier or risk score!
- Two strategies arise:
 - Anonymization
 - Summary statistics

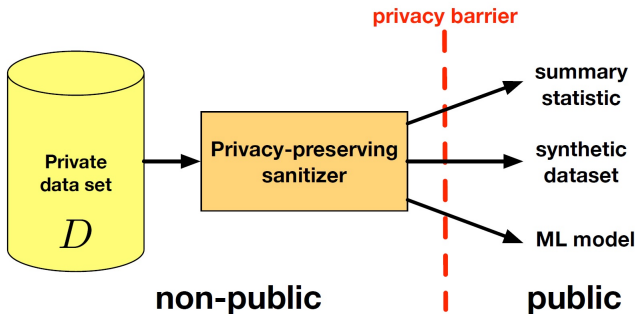
Are those two strategies safe? Well...

Reported cases of privacy leaks

- Data leakage in 2020 (at Q3)
 - 2,935 publicly reported breaches
 - 36 billion records exposed
 - Among which: Facebook, Instagram, Microsoft, TikTok, Google Cloud Server, etc.
- Data breaches with anonymized data by linkage between different but overlapping databases
 - AOL search data leak (2006)
 - Netflix prize (2007-2009)

Ref. Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In IEEE Symposium on Security and Privacy.

Property of Sanitizer



Aggregate information computable

Individual information protected
(robust to side-information)

Privacy in Machine Learning

Machine Learning under attacks

- Membership inference attack: infer data from training set (e.g. patient from hospital)
- Feature inference attack: infer sensitive properties of group of individuals in the training set
- Model extraction attack: infer the prediction model in order to twist decisions

Membership inference attacks

- ML algorithms are prone to membership inference and variants
 - An attack is made to determine whether a subject belongs to a training data set.
 - If successful, it becomes possible to infer individual information: e.g. participating to a clinical study can thus unveil the fact that the patient was treated in a certain hospital for a given disease.
- Being prone to membership inference attacks increases the risk for ML algorithms outcome to be classified as personal data under the GDPR.

Shokri et al. (2017): Membership inference attacks against machine learning models

Hu et al. (2021): Membership Inference Attacks on Machine Learning: A Survey

Privacy vs. Accuracy vs. Sample size

- If sample size is small, one cannot achieve both privacy and accuracy
- To achieve accuracy, need many features which will eventually identify the individual if the data set is small

N.B.: large/small sample size should be discussed wrt dimension

(Regularized) Empirical Risk Minimization

- Mother of global Machine Learning procedures: Optimization of a risk functional formed by the sum of a data-fitting term and a penalty (regularizer):

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \text{pen}(f) \right\}$$

- In shallow learning: most algorithms boil down to an optimization problem with explicit penalty
- In the case of deep learning: no explicit regularization ($\text{pen}(f) = 0$) but regularization operates through SGD and operators linking successive layers of computation

Private Regularized Empirical Risk Minimization

1. Data perturbation

- Same procedure, perturbed data:

$$\hat{f}^D \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(\tilde{Y}_i, f(\tilde{X}_i)) + \lambda \text{pen}(f) \right\}$$

- Example: *k-anonymity* (Sweeney, 2002)

Private Regularized Empirical Risk Minimization

2. Output perturbation

- Same procedure, change decision rule: $\hat{f}^O = T(\hat{f})$ where

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \text{pen}(f) \right\}$$

- Example: **Global Sensitivity Method** also referred to as Laplace or Gaussian mechanisms (Dwork et al., 2006)

Private Regularized Empirical Risk Minimization

3. Risk perturbation

- Same procedure, change risk criterion:

$$\hat{f}^R \in \arg \min_{f \in \tilde{\mathcal{F}}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \widetilde{\text{pen}}(f) \right\}$$

- Example: **Private SVM** with finite feature maps (Rubinstein et al., 2009)

Private Regularized Empirical Risk Minimization

4. Algorithm perturbation

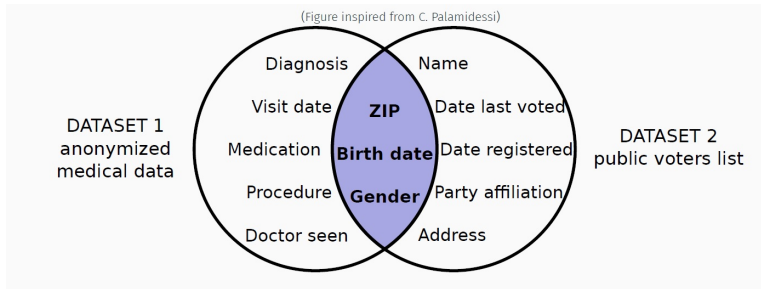
- Same procedure, change algorithm:

$$\hat{f}^A \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \text{pen}(f) \right\}$$

- Example: **Private SGD** (Abadi et al. (2016), Song et al. (2013))

Anonymization: The limits of simple ideas

Anonymization is not safe due to linkage



Indeed: 87% of the US population can be identified based on ZIP/BD/Gender!

Anonymization is not safe due to linkage on nonsensitive data!

Let U and V be feature vectors of nonsensitive data, S a feature vector of sensitive data. Assume:

- Database #1 contains (private) data (ID, U, S)
- Database #1a contains (public) anonymized data (U, S)
- Database #2 contains public data (ID, U, V)

Then:

- If U is unique, then ID may be linked to S from DB #1a and DB #2
- The larger the dimension of U (and/or the smaller the sample size), the more likely U will be unique

Composition - example

Assume Alice belongs to both tables and we know she is 26, lives in zip code 13012 area...

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 1: A 4-anonymous table.

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

Figure 2: A 6-anonymous table.

This example highlights the risks of independent k -anonymization among databases...

Summary statistics are not safe!

Two types of threats:

- 1 Differential attacks by querying the data set

Example: average performance of a group of people before and after a new member joins...

- 2 Membership inference attacks

Contingency tables or test statistics can actually lead to recover the identity of an individual if the data set is not too large.

Example: Intensive research in the field of Genome-wide association studies (GWAS) [Homer et al. (2008), Wang et al. (2009), Sei and Ohsuga (2021)]

Randomized response

A toy example Setup

- Consider a database with n bits $x = (x_1, \dots, x_n) \in \{0, 1\}^n$
- Assume an external source wants to read the database, and the response Y_i for any i is picked at random:

$$Y_i = Z \cdot x_i + (1 - Z) \cdot (1 - x_i)$$

where Z is a Bernoulli random variable with parameter p .

- How much privacy does this mechanism guarantee?
- Check this ratio: for any i and any $y \in \{0, 1\}$, and if $p > 1/2$

$$\frac{1 - p}{p} \leq \frac{\mathbb{P}(Y_i = y \mid x_i = 1)}{\mathbb{P}(Y_i = y \mid x_i = 0)} \leq \frac{p}{1 - p}$$

- This is referred to as *plausible deniability*

Counting query

- Now consider a unitary predicate ϕ applying to a single bit such that $\phi(x_i) \in \{0, 1\}$
- Querying the full database, we can infer the fraction $\Phi(x)$ of bits which satisfy the predicate ϕ

$$\Phi(x) = \frac{1}{n} \sum_{i=1}^n \phi(x_i)$$

- Examples of such queries are: histograms, CDF, marginals, contingency tables...
- Under random response with plausible deniability (that is $p \simeq 1/2$), it can be shown that any statistic $\Phi(x)$ can be estimated with a mean square error at precision of the order $O(n/\varepsilon^2)$.

Hint: set $p = e^\varepsilon / (1 + e^\varepsilon)$ with $o(\varepsilon) = 1$

A useful inequality

Hoeffding's lemma

Proposition.

Consider Z a random variable such that:

- $\mathbb{E}(Z) = 0$
- $Z \in [a, b]$ almost surely

Then, for any $s > 0$, we have:

$$\mathbb{E}(e^{sZ}) \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Interpretation: the Laplace transform of bounded random variables exhibits subgaussian behavior.

Hoeffding's inequality

Proposition.

Consider Z_1, \dots, Z_n IID over $[a, b]$ and $\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i$. We then have, for any $t > 0$

$$\mathbb{P}\{\bar{Z}_n - \mathbb{E}(Z_1) > t\} \leq \exp(-2nt^2/(b-a)^2)$$

and

$$\mathbb{P}\{\bar{Z}_n - \mathbb{E}(Z_1) < -t\} \leq \exp(-2nt^2/(b-a)^2)$$

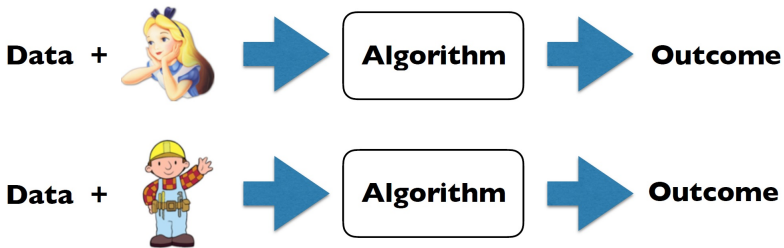
Consequence: This bound implies the strong law of large numbers for bounded random variables (by Borel-Cantelli lemma)

Proof technique: Chernoff's bounding method

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}(Z_1) > t\right) \leq \inf_{s>0} \exp\left(-nst + n \log \mathbb{E}(e^{s(Z_1 - \mathbb{E}(Z_1))})\right)$$

Differential privacy

Differential Privacy



Participation of a person does not change outcome

Definition of Differential Privacy (DP)

Dwork, McSherry, Nissim, Smith (2006)

- Consider $A(x)$ where A is a *randomized* algorithm operating on a data set x
- Let x' be a data set which differs from x by one data point.
- We consider that the randomized algorithm will satisfy differential privacy at level ϵ (privacy loss) if the following loglikelihood ratio is uniformly bounded over x ; x' and B :

$$\sup_B \sup_{x, x'} \left| \log \left(\frac{\mathbb{P}(A(x) \in B)}{\mathbb{P}(A(x') \in B)} \right) \right| \leq \epsilon$$

Approximate Differential Privacy

- We say the randomized algorithm A is (ϵ, δ) -DP if for all x, x' such that x and x' differ by one element and any measurable set B , we have:

$$\mathbb{P}(A(x) \in B) \leq e^\epsilon \mathbb{P}(A(x') \in B) + \delta$$

- Properties of DP:
 - Preservation of DP under post-processing
 - Simple composition is additive in ϵ and δ .
 - Group of k individuals privacy is preserved at $k \times \epsilon$

The Laplace mechanism

Definitions

- A Laplace distribution $Lap(0, b)$ with scale parameter b has density $p(u) = (1/(2b)) \exp(-|u|/b)$
- The global sensitivity of a query function is defined as $\Delta(f) = \sup_{x, x'} \|f(x) - f(x')\|$ where x, x' differ by one element
- The Laplace mechanism is an algorithm applying to the database x as follows:

$$A(x, f(\cdot), \varepsilon) = f(x) + (Y_1, \dots, Y_k)$$

where f is a vector-valued (in \mathbb{R}^k) query function and the Y_i 's are IID Laplace random variables with scale parameter $b = \Delta(f)/\varepsilon$

Properties of the Laplace mechanism

- The Laplace mechanism preserves $(\varepsilon, 0)$ -DP
- The accuracy of the Laplace mechanism can be monitored by the following bound: set $y = A(x, f(\cdot), \varepsilon)$ and any $\delta \in (0, 1]$

$$\mathbb{P} \left(\|f(x) - y\|_{\infty} \geq \ln \left(\frac{k}{\delta} \right) \left(\frac{\Delta(f)}{\varepsilon} \right) \right) \leq \delta$$

Strategies to make Machine Learning private

(Regularized) Empirical Risk Minimization

- Mother of global Machine Learning procedures: Optimization of a risk functional formed by the sum of a data-fitting term and a penalty (regularizer):

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \text{pen}(f) \right\}$$

- In shallow learning: most algorithms boil down to an optimization problem with explicit penalty
- In the case of deep learning: no explicit regularization ($\text{pen}(f) = 0$) but regularization operates through SGD and operators linking successive layers of computation

Private Empirical Risk Minimization

1. Data perturbation

- Same procedure, perturbed data:

$$\hat{f}^D \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(\tilde{Y}_i, f(\tilde{X}_i)) + \lambda \text{pen}(f) \right\}$$

- Example: *k-anonymity* (Sweeney, 2002)
 - Define a set of attributes as quasi-identifiers
 - Suppress/generalize attributes and/or add dummy records to make every record in the dataset indistinguishable from at least $k - 1$ other records with respect to quasi-identifiers

k-anonymity example

Name	Birth date	Zip code	Gender	Diagnosis	...
Ewen Jordan	1993-09-15	13741	M	Asthma	...
Lea Yang	1999-11-07	13440	F	Type-1 diabetes	...
William Weld	1945-07-31	02110	M	Cancer	...
Clarice Mueller	1950-03-13	02061	F	Cancer	...

Name	Birth date	Zip code	Gender	Diagnosis	...
	1993-09-15	13741	M	Asthma	...
	1999-11-07	13440	F	Type-1 diabetes	...
	1945-07-31	02110	M	Cancer	...
	1950-03-13	02061	F	Cancer	...

	Quasi identifiers			Sensitive attribute	
Name	Age	Zip code	Gender	Diagnosis	...
	20-30	13***		Asthma	...
	20-30	13***		Type-1 diabetes	...
	70-80	02***		Cancer	...
	70-80	02***		Cancer	...

Question: pros/cons?

Private Empirical Risk Minimization

2. Output perturbation

- Same procedure, change decision rule: $\hat{f}^O = T(\hat{f})$ where

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \text{pen}(f) \right\}$$

- Example: **Global Sensitivity Method** also referred to as Laplace or Gaussian mechanisms (Dwork et al., 2006)

Other example: linear SVM case

- Consider the following inference principle:

$$\hat{w} \in \arg \min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i w^T X_i) + \frac{\lambda}{2} \|w\|^2 \right\}$$

with ℓ convex

- Pseudocode for private version

Algorithm 1 Private linear SVM with output perturbation

Input: training data $\{(X_i, Y_i) : i = 1, \dots, n\}$, privacy parameter ϵ , amount of regularization λ

Solve raw optimization problem to get \hat{w}

Draw $Z = z$ according to $\mathbb{P}\{Z = z\} \propto e^{-\epsilon \|z\|}$

return Compute $\tilde{w} = \hat{w} + \frac{z}{n\lambda}$

Private Empirical Risk Minimization

3. Risk perturbation

- Same procedure, change risk criterion:

$$\hat{f}^R \in \arg \min_{f \in \tilde{\mathcal{F}}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \widetilde{\text{pen}}(f) \right\}$$

- Example: **Private SVM** with finite feature maps (Rubinstein et al., 2009)

Private SVM - second version

- Main ingredients:
 - Random and finite feature map and induced kernel
 - Dual optimization solver
 - Laplace mechanism

- Pseudocode

Algorithm 2 Private linear SVM with objective perturbation

Input: training data $\{(X_i, Y_i) : i = 1, \dots, n\}$, convex loss ℓ , parameter ε , amount of regularization λ , finite feature map $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^F$ and induced kernel

Solve dual optimization problem to get $\tilde{\alpha}$ based on induced kernel

Compute $\tilde{w} = \sum_{i=1}^n \tilde{\alpha}_i Y_i \Phi(X_i)$

Draw IID sample $Z = z$ from Laplace distribution $(0, \lambda)$

return Compute $\tilde{w}^{\mathbb{R}} = \tilde{w} + z$

Discussion and further topics related to privacy

Some names on differential privacy

- Cynthia Dwork (Harvard) - 2014 book on "The Algorithmic Foundations of Differential Privacy" (with Aaron Roth)
- Helen Nissenbaum (Cornell Tech)
- Catuscia Palamidessi (INRIA, France) - book and Master course about Foundations of Privacy
- Kamalika Chaudhuri* (UCSD) - NIPS 2017 tutorial
- Aurélien Bellet* (INRIA, France) - Master course on Privacy Preserving Machine Learning

*more ML flavor in their research

Check workshop series at the Simons Foundation on "Data Privacy: Foundations and Applications" - Jan. 15 – May 17, 2019

Further topics

- Regulatory - How to account for privacy?
- Implementation - Where to place sanitizers along a pipeline?
How to deal with privacy during the data exploration stage?
Can deep learning preserve the privacy of all its parameters and still generalize?
- Under constraints - How to optimize privacy budget along several stages ?

preprocessing/ training/ cross-validation/ testing/
hyperparameter calibration
- Resilience to attacks

Next lecture: Part 2 on privacy

- Alternative definitions and properties of formal privacy
- Private Empirical Risk Minimization with algorithm perturbation (approach #4)
- Relationship between membership attacks and Differential Privacy, privacy auditing
- Federated learning with DP