Privacy-Aware Machine Learning: Some progress

Jean-François COUCHOT[1]

[1]Université de Franche-Comté, FEMTO-ST, Besançon, France

JWOC 2024, October 4th 2024

Plan

Introduction to Data Privacy and Differential Privacy (DP)

Local DP to Prototype ICD-10 ML Association Model

Centralized DP-fied Machine Learning to Allow Model Sharing

Conclusion

Outline

Plan

Data Privacy: Legal Framework

Some Regulations

- Universal Declaration of Human Rights[1]: No interference with private life.
- European AI Act[2]: Critical algorithmic decisions by AI only if explainable, safe:
  - ⇝ Evaluation on realistic data.
  - ⇝ Models and outputs: Controlled information leakage.
- GDPR [3]: protective framework for data:
  - ⇝ Reduced constraints on anonymous data.
- e-privacy[4]: Processing of personal data by telephone operators.
  - ⇝ Must be done on the fly (without storage).

Motivation

- Approach for legally compliant analyses.
- Objective: For a defined level of protection, maximization of utility.

---

[1] https://www.un.org/fr/universal-declaration-human-rights/

[2] https://www.europarl.europa.eu/news/fr/press-room/20230609IPR96212/les-deputes-sont-prets-a-negocier-les-regles-pour-une-ia-sure-et-transparente

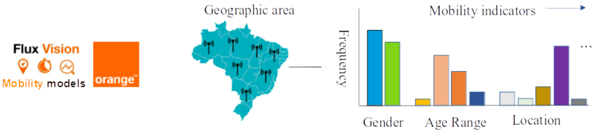
[3] https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

[4] https://www.economie.gouv.fr/files/files/directions_services/cge/e-privacy.pdf

Data Privacy: Overview of Implementations

Syntactic Approaches to Data: k-anonymity[5], l-diversity[6]

- Data grouped into classes of size $\geq k$.
- Easy to implement (but may be attacked with additional knowledge).



Probabilistic Property of Algorithm $\mathcal{M}$: ε-Differential Privacy (ε-DP)[7]

$$\forall D_1, D_2 \text{ (neighboring databases)}, D, O \text{ (output)}, \frac{\Pr(D = D_1 | \mathcal{M}(D) = O)}{\Pr(D = D_2 | \mathcal{M}(D) = O)} \leq e^{\varepsilon} \frac{\Pr(D = D_1)}{\Pr(D = D_2)}.$$

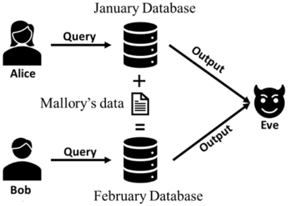
- Publishing $\mathcal{M}(D) = O$: Ability to distinguish D_1 from D_2 is approximately unchanged.
- Practical: Creating randomized mechanisms $\mathcal{M}$ adding controlled noise (see next-slides).

---

[5] Sweeney 2002, "k-Anonymity: A Model for Protecting Privacy".
[6] Machanavajjhala et al. 2006, "l-Diversity: Privacy Beyond k-Anonymity".
[7] Dwork et al. 2006, "Calibrating noise to sensitivity in private data analysis".

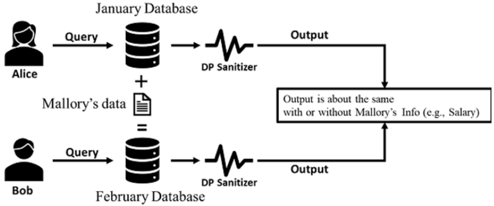
Example of Queries on Neighboring Databases[8]

Without Differential Privacy



- Monthly query: (#employees, average salary).
- Result:
  $\{\text{Jan} : (100, \$55,000), \text{Feb} : (101, \$56,000)\}$.
- Suppl. knowledge: 0 output + Mallory in February.
- $\rightsquigarrow$ Mallory's salary: \$156,000.

With Differential Privacy



- Same queries, same additional knowledge.
- Sanitized results:
  $\{\text{Jan} : (102, \$55,551),$
  $\text{Feb} : (97, \$55,975)\}$.
- Mallory's salary?

[8] Morris Chang et al. 2021, *Privacy-Preserving Machine Learning*.

Key Ideas

Intuition for Two Neighboring Databases D_1 and D_2

- Results (aggregated, statistical, etc.) are close.
- $\Leftrightarrow$ "Probabilities" on $\mathcal{M}(D_1)$ and $\mathcal{M}(D_2)$ are nearly equal (up to ε).

Why Differential Privacy?

- Private data: Desire to have little impact on results.
- $\leadsto$ Difficult to distinguish if a particular individual "participates or not."
- $\leadsto$ Data owner is less concerned about sharing their data.

Plan

Formalization of Differential Privacy[9]

Definition (ε-Differential Privacy (DP))

Let $\varepsilon \in \mathbb{R}^+$. The non-deterministic probabilistic algorithm $\mathcal{M}$ satisfies ε-Differential Privacy if

$$\begin{array}{ll} \forall D_1, D_2 \in \mathbb{N}^{|\mathcal{X}|} \text{ such that } \|D_1 - D_2\|_1 = 1, & (D_1, D_2\text{: neighboring databases}) \\ \forall O \text{ such that } O \in \mathcal{M}(\mathbb{N}^{|\mathcal{X}|}), & (\text{for any output } O \text{ of the algorithm}) \\ \Pr[\mathcal{M}(D_1) = O] \leq e^{\varepsilon} \Pr[\mathcal{M}(D_2) = O] & (\text{if } \varepsilon \text{ is small, } e^{\varepsilon} \approx 1 + \varepsilon) \end{array}$$

Budget of Leakage $\varepsilon \in \mathbb{R}^+$: Allowed Deviation, Permitted Leakage

- $\Pr[\mathcal{M}(D_1) = O] \leq e^{\varepsilon} \Pr[\mathcal{M}(D_2) = O]$: Results are approximately equal (but not necessarily) with or without the data of one person.
- $\varepsilon = 0$: No deviation is allowed (all outputs are equal with or without the data of one person), data is perfectly protected (but less useful).
- Small vs. large ε: It depends on the amount of permitted leakage.

[9] Dwork et al. 2006, “Calibrating noise to sensitivity in private data analysis”.

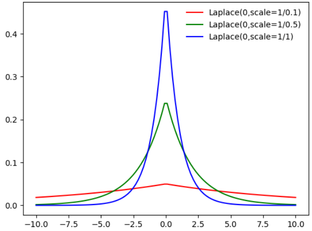
Plan

Query Q_1: Number of Employees in the Database

Objectives, Data, Idea

- Publish the number of employees with an ε-DP mechanism.
- $Q_1(D_{\text{Jan}}) = 100$, $Q_1(D_{\text{Feb}}) = 101$, etc.
- Add Laplace noise centered at 0 depending on ε.

Implementation: Laplace Noise Centered at 0, $\mathcal{M}_L(D) = Q_1(D) + v$, $v \sim Lap(0, \varepsilon^{-1})$

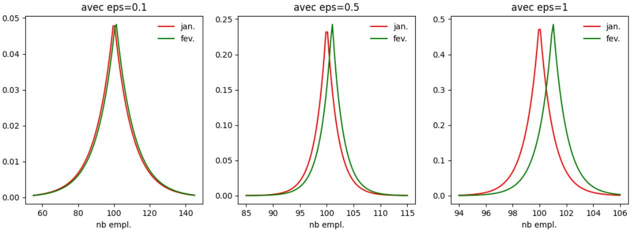


Query Q_1: Number of Employees in the Database

Objectives, Data, Idea

- Publish the number of employees with an ε-DP mechanism.
- $Q_1(D_{\text{Jan}}) = 100$, $Q_1(D_{\text{Feb}}) = 101$, etc.
- Add Laplace noise centered at 0 depending on ε.

Implementation: Laplace Noise Centered at 0, $\mathcal{M}_L(D) = Q_1(D) + v$, $v \sim Lap(0, \varepsilon^{-1})$



Exponential Mechanism[10]

Motivation and Idea

- Directly adding noise to the outputs may result in meaningless outcomes (output of a query is categorical or discrete, e.g.).
- Thanks to a utility function (a score function): One can map any value to a numerical one.

More Formally

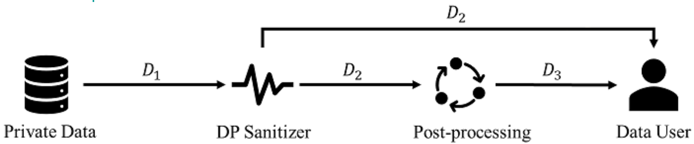
- v in a domain $\mathcal{D}$: The value to be sanitized.
- $\mathcal{R}$: The set of possible output sanitized data.
- $U : \mathcal{D} \times \mathcal{R} \to \mathbb{R}^+$: A score function with sensitivity Δ_U.

The *exponential mechanism* sanitizes v to r with probability proportional to $\exp \frac{\varepsilon U(v,r)}{2\Delta_U}$.

[10] McSherry and Talwar 2007, "Mechanism design via differential privacy".

Robustness to Post-Processing

Intuition for a Database D_1 [11]



Interpretations

- Post-processing if seen as a subsequent algorithm (e.g., removing outliers): Only the DP algorithm needs to be considered carefully.
- Post-processing seen as an attack by an adversary: They can incorporate as much auxiliary information as they want; the privacy guarantee remains valid.

Theorem (Post-Processing of an ε-DP Mechanism)

For any function $f : \mathcal{M}(\mathbb{N}^{|\mathcal{X}|}) \to f(\mathcal{M}(\mathbb{N}^{|\mathcal{X}|}),\ f(\mathcal{M})$ *is also* ε*-DP.*

Direct Application

- Any sanitized real data: Can subsequently be rounded to the nearest integer.

[11] Morris Chang et al. 2021, *Privacy-Preserving Machine Learning*.

Composition of Sequential Leaks

Sequences of Leaks

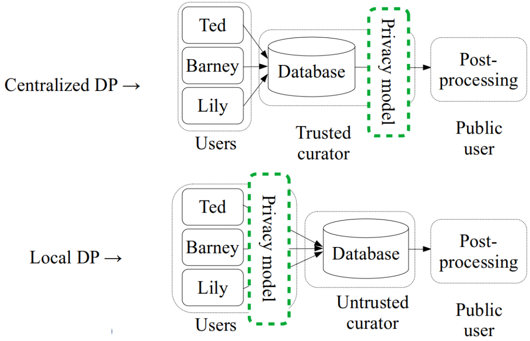
- It is common to query the same database iteratively (e.g., employee count in January, February, etc.).
- Each query corresponds to a data leak, and we want to find the total leakage for a sequence of leaks with ε_1 and ε_2.

Theorem (Sequential Composition of ε-DP Mechanisms)

If $\mathcal{M}_1$ and $\mathcal{M}_2$ operate on non-disjoint sets, $\mathcal{M}_{1,2}$ is $\varepsilon_1 + \varepsilon_2$-DP.

Outline

Centralized vs Local DP in ML



- Centralized DP.
  - Trust required in the Database Curator.
  - Optimal noise per query.
  - Allow to share ML model.
- Local DP
  - Individual noise for all post-processing (e.g., Machine Learning).
  - Unnecessary trust in Data Curator.
  - Allow to develop ML model prototype without accessing to original data.

Definition and Properties

Definition of ε-Local Differential Privacy[12] (ε-LDP, or Local DP)

- $\mathcal{X}$: The set of possible input values.
- $\varepsilon \in \mathbb{R}^+$: Privacy budget.
- $\mathcal{M}$: Non-deterministic probabilistic algorithm respects ε-Local Differential Privacy if

$$\begin{array}{ll} \forall x_1, x_2 \in \mathcal{X} & (x_1 \text{ and } x_2 \text{ are two input data points}) \\ \forall y \text{ s.t. } y \in \mathcal{M}(\mathcal{X}), & (\text{for any output } y \text{ of the algorithm}) \\ \Pr[\mathcal{M}(x_1) = y] \leq e^{\varepsilon} \Pr[\mathcal{M}(x_2) = y] & \end{array}$$

Properties Similar to DP

- Robustness to post-processing.
- Combining two mechanisms ε_1-LDP and ε_2-LDP results in $\varepsilon_1 + \varepsilon_2$-LDP.

[12] Duchi, Jordan, and Wainwright 2013, “Local privacy and statistical minimax rates”.

Outline

Metric-Privacy[13]

Motivation

- (L)DP: It's challenging to determine the origin of a given output.
- 2 data points, far apart $\rightsquigarrow$ may produce the same output.
- Relevance when dealing with a large data space (e.g., centuries, the entire Earth)?
- Introduction of the concept of distance between data points in the probability constraint.

Definition of Metric-Privacy

- $\mathcal{X}$: Set of possible input values, equipped with a Metric d.
- $\mathcal{M}$: Non-deterministic probabilistic algorithm that adheres to metric-privacy if

$$\begin{array}{ll} \forall x_1, x_2 \in \mathcal{X} & \text{(x_1 and x_2 are two input data points)} \\ \forall y \text{ s.t. } y \in \mathcal{M}(\mathcal{X}), & \text{(for any output y of the algorithm)} \\ \Pr[\mathcal{M}(x_1) = y] \leq e^{\varepsilon . d(x_1, x_2)} \Pr[\mathcal{M}(x_2) = y] & \end{array}$$

[13] Chatzikokolakis et al. 2013, "Broadening the scope of differential privacy using metrics".

Plan

Introduction to Data Privacy and Differential Privacy (DP)

Local DP to Prototype ICD-10 ML Association Model

- Introduction to Association of ICD-10 Codes to Sensitive Documents
- NER and Sanitizing Steps of Prototype
- Ready to Build the ICD-10 ML Association Tool

Centralized DP-fied Machine Learning to Allow Model Sharing

Conclusion

Plan

International Classification of Diseases, 10th Revision

ICD-10: Standardized Diagnostic Tool for Recording Health Conditions

- Developed by the World Health Organization (WHO).
- Used globally to classify diseases, injuries, and health conditions.

ICD-10: A Pivotal Role in Healthcare Systems

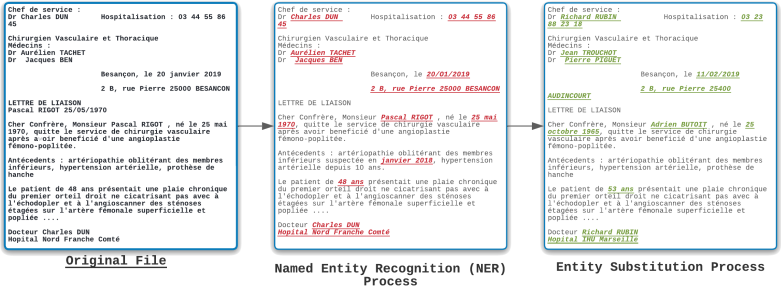
- Patient Records: Codes are used to document diagnoses, procedures, and treatments.
- Healthcare Analytics: They facilitate data analysis for quality improvement, resource allocation, and epidemiological studies.
- Reimbursement: Codes are linked to billing and reimbursement systems.

ICD-10: Coding in Practice

- Manual Coding: Healthcare professionals manually assign ICD-10 codes based on medical records.
- Automated Coding: Natural Language Processing (NLP) is used to automate the coding process $\rightsquigarrow$ Sufficient to have de-identified dataset to build such NLP model.

De-Identification: A Twofold Method

Two Steps



1. Named Entity Recognition (NER) for identifying information (efficiency issue).
2. Sanitizing of detected information (optimization issue: minimizing leakage while preserving utility).

Plan

Named Entity Recognition

Iterative Learning on HNFC Datasets

- Increasingly large, progressively more de-identified datasets.
- Automatically pre-labeled and manually validated.
- Model: Hybrid[14], then deep learning only[15].

NER Results

Method	CamemBERT-ner			MEDINA			FlauBERT-ner			**Hybride**			**Healthinf**			**Dernoncourt**[16]		
Dataset	*HNFC*															i2b2		
Metric	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
PER	89	99	93.8	***98.2***	97.7	*98.2*	91.8	97.6	94.6	96.3	***99.8***	98	97.2	98.9	98	**98.2**	99.1	**98.6**
ORG	7.	21.8	11.1	32.6	24.8	28.1	16.9	34.1	22.6	41.1	*57.3*	47.8	*90*	51	*65.6*	**92.9**	**71.4**	**80.7**
LOC	46	67.2	54.6	98.8	81.1	89.1	75.7	66.3	70.7	88.4	***95.8***	92	***99.4***	94.4	***96.9***	95.9	95.7	95.8
DATE	NA			97.7	86.6	91.9	NA			97.7	86.7	91.9	***99.2***	*95.7*	*97.4*	99	**99.5**	**99.2**
AGE	NA			91.5	66.9	77.3	NA			91.5	66.9	77.3	*98.2*	*91.8*	*95*	**98.9**	**97.6**	**98.2**
TEL	NA			99.5	97.9	98.7	NA			***99.5***	97.9	98.7	99.4	***99.8***	***99.6***	98.7	99.7	99.2
REF	NA			NA			NA			NA			96.1	79.5	87	NA		
QID	NA			NA			NA			NA			77.2	32	45.3	**99.2**	**98.7**	**99**
Mic.-avg.	70.8	51.5	59.6	98.2	91.2	94.5	85.8	86.7	86.3	94.6	94.9	94.7	***98.5***	*96.4*	*97.4*	98.3	**98.5**	**98.4**

[14] Tchouka, Couchot, Coulmeau, et al. 2022, “De-Identification of French Unstructured Clinical Notes for Machine Learning Tasks”.

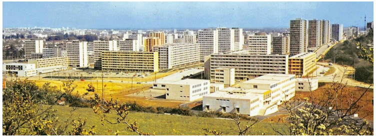
[15] Tchouka, Couchot, and Laiymani 2023, “An Easy-to-Use and Robust Approach for the Differentially Private De-Identification of Clinical Textual Documents”.

[16] Dernoncourt et al. 2016, “De-identification of Patient Notes with Recurrent Neural Networks”.

Sanitizing

Utility of Local DP for Certain Entities?

- $\forall y, x_1, x_2, \Pr(\mathcal{M}(x_1) = y) \leq e^{\varepsilon} \Pr(\mathcal{M}(x_2) = y)$.
- Likely sanitized with the same value:
  - 08/01/42 and 14/03/18 (birth and death dates of St. Hawking).
  - Dijon and Beze (in BFC but epidemiologically $\neq$).



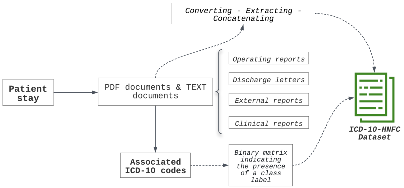
Sanitizing Integrating Metric-Privacy

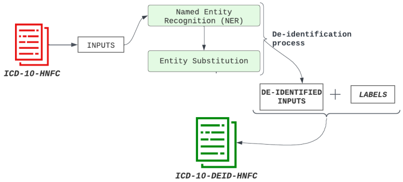
- Theory: $\forall x_1, x_2, y, \Pr(\mathcal{M}(x_1) = y) \leq e^{\varepsilon . d(x_1, x_2)} \Pr(\mathcal{M}(x_2) = y)$.
- Dates: $\mathcal{M}_{date}(x) = x + v$ s.t. $v \sim Lap(\frac{1}{\varepsilon})$.
- Locations: $\Pr(\mathcal{M}_{loc}(x) = o) \propto e^{\varepsilon . d(x, o)}$, s.t. d an epidemiological based distance.

Plan

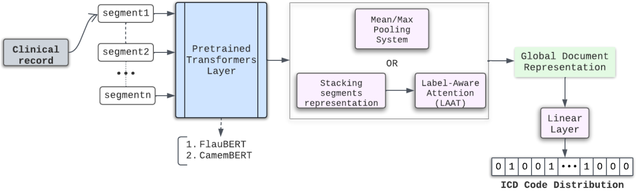
ICD-10 Code Association[18]

Datasets





ICD-10 Code Association Model Architecture, PLM-ICD[17]



[17] Huang, Tsai, and Chen 2022, "PLM-ICD: automatic ICD coding with pretrained language models".

[18] Tchouka, Couchot, Laiymani, Selles, et al. 2023, "Automatic ICD-10 Code Association: A Challenging Task on French Clinical Texts".

ICD-10 Code Association–2

State-of-the-Art[19] Code Association Results

Models	Language	Dataset	Labels	F_1-score
PLM-ICD[20]	*English*	*MIMIC 2*	*5,031*	*0.5*
		MIMIC 3	*8,922*	***0.59***
Dalloux[21]	*French*	*Personnel*	*6,116*	*0.39*
			1,549	*0.52*
PROPOSAL	French	ICD-10-HNFC	6,160	**0.47**
			1,564	**0.55**
Dalloux			6,160	0.27
			1,564	0.35

Impact of De-identification on Results

Dataset	Labels	**Precision**	**Recall**	F_1**-score**
ICD-10-HNFC	**6160**	**0.47**	**0.46**	**0.47**
ICD-10-DEID-HNFC		0.44	0.43	0.44
ICD-10-TAG-HNFC		0.43	0.41	0.42

[19] Tchouka, Couchot, Laiymani, Selles, et al. 2024, “Differentially private de-identifying textual medical document is compliant with challenging NLP analyses: Example of privacy-preserving ICD-10 code association”.

[20] Huang, Tsai, and Chen 2022, “PLM-ICD: automatic ICD coding with pretrained language models”.

[21] Dalloux et al. 2020, “Supervised Learning for the ICD-10 Coding of French Clinical Narratives”.

Plan

Introduction to Data Privacy and Differential Privacy (DP)

Local DP to Prototype ICD-10 ML Association Model

Centralized DP-fied Machine Learning to Allow Model Sharing

- Introduction to DP-fied Machine Learning
- Making GBDT Model computation Differentially Private
- Optimizing a DP GBDT

Conclusion

Plan

ML Models: Vulnerable to Privacy Leakages?

Opacity and Leakage

- Often: ML models are seen as black boxes (opaque, difficult to understand internal workings).
- But is the opacity equivalent to information non-leakage?

Attacks on ML Models

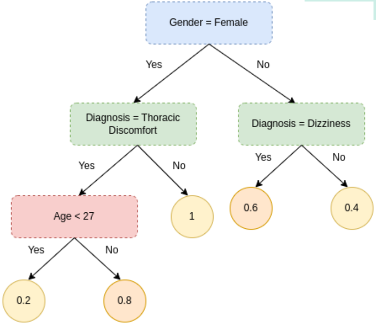
- Membership Inference Attacks[22]: Determine whether a specific data point was used to train a model.
- Property Inference Attacks[23]: Infer sensitive properties about the training data (gender, age distribution of the individuals e.g.).

[22] Shokri et al. 2017, "Membership Inference Attacks Against Machine Learning Models".
[23] Ganju et al. 2018, "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations".

Gradient Boosting Decision Tree (GBDT) Model

An Ensemble Learning Method Based on Decision Trees

- Sequential Learning: A series of models, each correcting errors from previous one.
- Decision Trees as Base Learners: Simple decision trees as building blocks for the complex model.
- Gradient Descent: An optimization algorithm minimizing the loss function at each iteration.
- Two **data-querying and leaking** computations: Internal nodes **splits**, **leaf values** computation



Research Question

- Is it possible to provide a DP version of the model with enough accuracy, and if so, how?

Plan

Introduction to Data Privacy and Differential Privacy (DP)

Local DP to Prototype ICD-10 ML Association Model

Centralized DP-fied Machine Learning to Allow Model Sharing

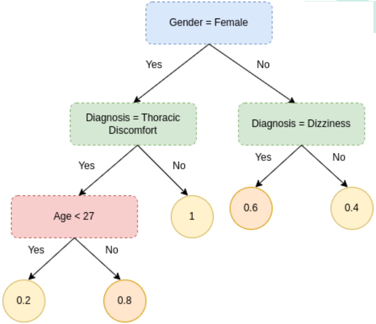
- Introduction to DP-fied Machine Learning
- Making GBDT Model computation Differentially Private
- Optimizing a DP GBDT

Conclusion

DP-BOOST[24]: A DP-fied Instance of GBDT

Method Overview

- When is data queried ? Whilst computing leaf and split nodes.
- How to make this step DP-fied ? Thanks to DP mechanisms.
  - For each numerical leaf value V calculus: Apply Laplace mechanism.
  - Splitting G is choosing between elements according to a metric: Exponential mechanism.



Main contribution

- Bounding sensitivity Δ_G and Δ_V of G and V calculus to avoid useless noise.

[24] Li et al. 2020, “Privacy-preserving gradient boosting decision trees”.

Tree Computation of DP-BOOST

- ε budget allocation: Equally distributed between splitting and leaf calculus.
- Use of reduced Δ_G and Δ_V.

Algorithm 1: TrainSingleTree: Train a differentially private decision tree

Input: I: training data, $Depth_{max}$: maximum depth
Input: ε_t: privacy budget

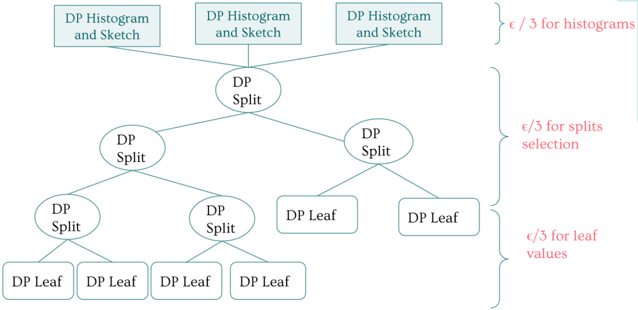
```
ε_leaf ← ε_t / 2                         // privacy budget for leaf nodes
ε_nleaf ← ε_t / (2 Depth_max)            // privacy budget for internal nodes
Perform gradient-based data filtering on dataset I.
for depth = 1 to Depth_max do
    for each node in current depth do
        for each split value i do
            Compute gain G_i according to Equation (3).
            P_i ← exp(ε_nleaf G_i / (2ΔG))
        /* Apply exponential mechanism */
        Choose a value s with probability (P_s / Σ_i P_i).
        Split current node by feature value s.
for each leaf node i do
    Compute leaf value V_i according to Equation (4).
    Perform geometric leaf clipping on V_i.
    /* Apply Laplace mechanism */
    V_i ← V_i + Lap(0, ΔV/ε_nleaf)
```

Output: A ε_t-differentially private decision tree

Other Approaches

DP-XGBoost[25]

- Adding a preprocessing of histogram computation.
- Other arbitrary allocation of ε.



Focus on Other ε Budget Allocation Strategies

- DP-TopDown[26]: Decaying ε budget the deeper the tree goes.
- S-GBDT[27]: splits based on sub-sampled data.

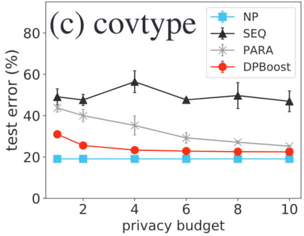
[25] Grislain and Gonzalvez 2021, "DP-XGBoost: Private Machine Learning at Scale".

[26] Wang, Dick, and Balcan 2020, "Scalable and provably accurate algorithms for differentially private distributed decision tree learning".

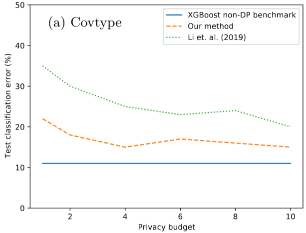
[27] Kirsche et al. 2023, "S-GBDT: Frugal Differentially Private Gradient Boosting Decision Trees".

Classification Results with DP-fied-ML

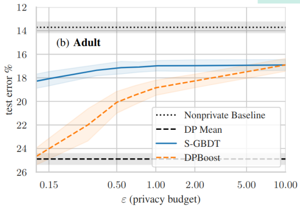
DP-BOOST[28], DP-XGBoost[29], S-GBDT[30]: Comparing Classification Errors



DP-BOOST article



DP-XGBoost article



S-GBDT article

Analysis

- Hard to compare DP-XGBoost and S-GBDT.
- RQ: Is there a way to optimize them?

[28] Li et al. 2020, "Privacy-preserving gradient boosting decision trees".

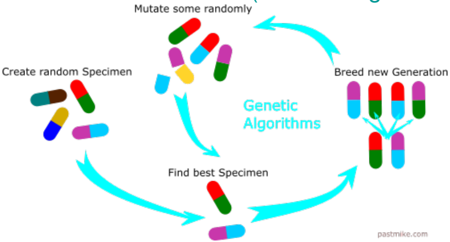
[29] Grislain and Gonzalvez 2021, "DP-XGBoost: Private Machine Learning at Scale".

[30] Kirsche et al. 2023, "S-GBDT: Frugal Differentially Private Gradient Boosting Decision Trees".

Plan

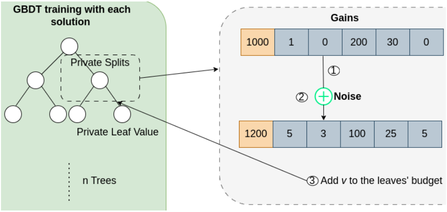
ε Allocation Optimization wrt. Utility Metric

Use Meta-Heuristics (Genetic Algorithm[31]) to Allocate ε Budget



- A vector of ε allocation (ensembles, trees, tree depth).
- Find allocation maximizing a metric, generate new ones, mutate some ones. . .
- Caveat: Requires reading (and thus leaking) data to compute this allocation.

Update of Splitting ε Budget wrt. Other Criterion



- Extreme: Randomized split keep all the ε budget to leafs.
- A part of ε: Used to compute Coefficient of Variation (CV).
- A part of ε: Used to split, wrt. CV, remaining for leafs.
- But, which part for both of them?

[31] https://medium.com/@derya.cortuk/genetic-algorithms-nature-inspired-optimization-for-solving-complex-problems-4dd893a9cb2c

Plan

Conclusion

Contributions on De-Identification[32] for Prototyping ICD-10 codes Association Task

- State-of-the-art NER model for de-identification in the French language.
- Metric privacy based sanitizing approach
- State-of-the-art ICD-10 codes association model in the French language.

Work in Progress in Optimizing DP-GBDT

- Budget allocation: How can we optimize it?

[32] `https://github.com/mlfiab/`