An Equivalence
Between Private Classification and Online Prediction

Reading Groups Presentation

Master 2 Data Science
Université de Lille

Presenters

Zakaria BOULKHIR
Omar IKEN
Gabriel LOISEAU

Authors

Mark Bun
Roi Livni
Shay Moran

October 21, 2021

Université de Lille

Overview

Introduction

Preliminaries

Classification in Statistical Learning

H - a class of $X \to \{\pm 1\}$ hypotheses

- ✓ In the paper, examples drawn from a distribution D on $X \times \{\pm 1\}$
- ✓ Input: training sample $(x_1, y_1) \ldots (x_n, y_n) \sim \mathcal{D}^n$
- ✓ Output: hypothesis $h : X \to \{\pm 1\}$

Online Prediction

Real-time predictions on sequentially arriving data

$H - \mathrm{a}$ known class of $X \to \{\pm 1\}$ experts

Input: a sequence of examples $(x_1 y_1), \ldots, (x_T, y_T) \in X \times \{\pm 1\}$

At each round t :

- ✓ Observe x_t
- ✓ Predict $\hat{y}_t$
- ✓ Suffer loss $1[\hat{y}_t \neq y_t]$

Definition (Goal. Minimize Regret:)

$$\sum_t 1[\hat{y}_t \neq y_t] - \inf_{h^* \in H} \sum_t 1\left[h^*(x_t) \neq y_t\right]$$

PAC Learnability

Definition (PAC Learnability)

$\mathcal{H}$ is PAC learnable if $\exists$ a hypothesis h with vanishing expected excess loss with respect to any input distribution $\mathcal{D}$.

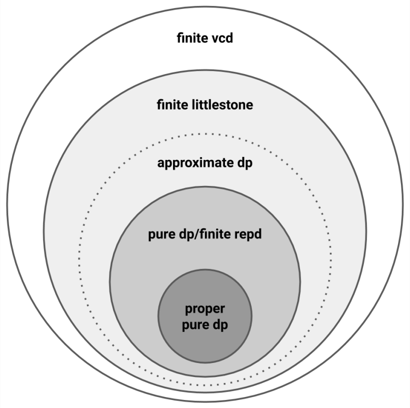

Figure 1: https://differentialprivacy.org/private-pac/

Differential Privacy

Definition (Setting)

- ✓ $\epsilon, \delta \in [0,1]$ - privacy parameters
- ✓ $x =_{\epsilon,\delta} y$ denotes the following statement: $x \leq e^{\epsilon} \cdot y + \delta$ and $y \leq e^{\epsilon} \cdot x + \delta$

Definition ((ϵ, δ)-indistinguishable)

Two distributions P, Q are (ϵ, δ)-indistinguishable if for every event E :

$$\mathrm{P_t}(E) = {}_{\epsilon,\delta} Q(E)$$

- ✓ A-a (randomized) algorithm
- ✓ Input: sample $S = (x_1, y_1), \ldots, (x_n, y_n)$
- ✓ Output: a randomized hypothesis $A(S)$
- ✓ $A(S)$ is a distribution over hypotheses

Definition (Differentially Private Algorithm)

A is (ϵ, δ)-DP if $\forall$ pair of neighboring samples S', S'' : $A\left(S'\right)$ and $A\left(S''\right)$ are (ε, δ)-indistinguishable.

Stability : robustness of output under small changes of input

Differential Private PAC Learnability

Definition (DP-PAC Learnability)

A class H is *DP*-PAC learnable if it is PAC learnable by an (ϵ, δ)-DP algorithm s.t.

- ✓ $\epsilon =$ constant
- ✓ $\delta \ll$ poly $\left(\frac{1}{n}\right)$

Global Stability

Definition (Global Stability)

Let $n \in \mathbb{N}$ be a sample size and $\eta > 0$ be a global stability parameter. An algorithm A is (n, η)-globally-stable with respect to a distribution $\mathcal{D}$ if there exists an hypothesis h such that:

$$Pr_{S \sim \mathcal{D}^n}[A(S) = h] \geq \eta$$

Definition ("Local" DP Stability)

$$Pr_{S,S' \sim \mathcal{D}^n}[A(S) =_\gamma A(S')] \geq \Gamma$$

Proofs

Proof Overview

Theorem (Main result)

For a class $\mathcal{H} \subseteq \{\pm 1\}^X$, the following statements are equivalent:

- ✓ *$\mathcal{H}$ is online learnable.*
- ✓ *$\mathcal{H}$ is approximate differentially-privately PAC learnable.*

Authors proved that any online learner is a Differentially-Private Learner.

The converse statement is already proved by Alon et al.

Prove in two steps:

1. Globally-Stable Learning $\implies$ Differentially-Private Learning
2. Online Learning $\implies$ Globally-Stable Learning

Step 1: Globally-Stablity implies DP Learning

Aim: Construct a differentially-private learner !

Idea: Combine the **Stable Histograms algorithm** with the **Generic Private Learner** to convert any globally-stable learning algorithm into a differentially-private one.

Process

If $\mathcal{A}$ is a globally stable learner with respect to $\mathcal{D}$, we obtain a differentially private learner using roughly m/η samples from that distribution as follows:

1. Run $\mathcal{A}$ on k independent samples, non-privately producing a list of k hypotheses.
2. Apply a differential-private "Stable Histograms" algorithm to this list.
3. Global stability of $\mathcal{A}$ guarantees that with high probability, this contains some hypothesis h with small population loss.
4. Apply generic differentially-private learner (based on exponential mechanism) on a fresh set of examples to identify such a accurate hypothesis from the list.

Definition (Littlestone dimension)

The **Littlestone dimension** of $\mathcal{H}$, denoted $Ldim(\mathcal{H})$, is the depth of largest complete tree that is shattered by $\mathcal{H}$. It captures mistake and regret bounds in online learning.

A mistake tree is a binary decision tree whose internal nodes are labeled by elements of X.

$$\mathcal{H} \text{ is online learnable } \iff \mathcal{H} \text{ has a finite Littlestone dimension } d < +\infty$$

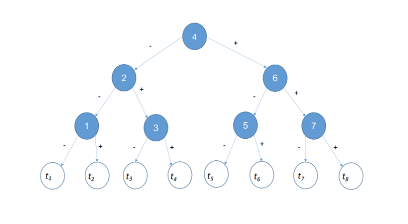

Figure 2: Example of mistake tree

In the realizable case: Any algorithm might make at least $Ldim(\mathcal{H})$ mistakes.

Now suffices to prove that: Finite Littlestone Dimension $\implies$ Global stability

Step 2: Finite Littlestone Dimension implies Globally-Stability

Let $\mathcal{H}$ be a concept class $s.t$ $d = Ldim(\mathcal{H}) < +\infty$

Aim: Design a globally-stable learning algorithm $\mathcal{A}$ for $\mathcal{H}$!

Littlestone showed that the minimum mistake bound achievable by any online learner is exactly $Ldim(\mathcal{H})$

The simplest setting in which learnability is captured by the Littlestone dimension is called the mistake-bound model.

In the realizable case $\rightsquigarrow$ some expert has excess zero !

Mistake bound model

1. The competitor pick $h^* \in \mathcal{H}$
2. The learner receives instance $x_t \in X$ and predict $\hat{y}_t \in \{0, 1\}$.
3. The competitor shows $y_t = h^*(x_t)$.
4. The leaner makes a mistake if $y_t \neq \hat{h}_t$.
5. SO what is the bound on the number of mistakes ??

$$d = \text{Mistake bound} = \text{Littlestone dimension} = \mathcal{O}(\sqrt{dT})$$

Step 2: Finite Littlestone Dimension implies Globally-Stability

Particular case: $Ldim(\mathcal{H}) = 1$

Let $\{(x_1, y_1), \dots (x_n, y_n)\}$ i.i.d. samples. Run the online learner sequentially $\rightsquigarrow$ outputs the predictor h^{n+1}.

- ✓ Case 1. If the algorithm does not make mistakes.
  - ✓ Given a sample, this hypothesis is already good.
  - ✓ Does not depend on the sample $\implies$ globally stable.
- ✓ Case 2. If the algorithm makes only one mistake.
  - ✓ If the sample is consistent with h^*.
  - ✓ If $h^{(n+1)} \neq h \implies \exists x, h^{(n+1)}(x) \neq h^*(x)$
  - ✓ Then the algorithm makes 2 mistakes on $\{S, (s, h^*(x)))\}$

Same reasoning in case $Ldim(\mathcal{H}) = 2$.

Applications

Build new Private Learners

Every class $\mathcal{H}$ that is online learnable (i.e. has finite Littlestone dimension) is also privately PAC learnable

- ✓ We can therefore construct using this method Private Algorithms using Online learnable concept classes

Finite classes have finite Littlestone dimension, and are known to be privately learnable

- ✓ Example of finite concept classes : Conjunction of Boolean literals

About infinite classes with finite Littlestone dimension, these can be privately learnable

- ✓ Example of infinite classes : Affine subspaces of $\mathcal{F}^n$ $\{1_V \mid V \subseteq \mathcal{F}^n, \dim V = d\}$

Private learnability in terms of the Littlestone dimension has new consequences for boosting the privacy and accuracy guarantees of differentially-private learners.

✓ (ϵ, δ)-DP algorithms can have their privacy parameters improved to $(p\epsilon, p\delta)$-DP if we pre-process it by randomly sub-sampling a $1/p$ of the input dataset (sample complexity)

But it can be difficult to amplify a weak value of δ

✓ Lemma : the existence of a $(0.1, O(1/n^2))$-DP learner for a given class implies the existence of a $(0.1, exp(-\Omega(n)))$-DP learner for that class.

The construction of such algorithm is left as an open question in the paper

Paper Overview

Paper Overview

What's its Type ?

Argumentative Research Paper : Demonstration of a concept where the converse statement has already been done.

What's the Problem ?

Convert any Private Classifier into an Online Predictor and Vice-Versa

What's the Claim/Contribution ?

The Proof of this equivalence

Is it well Supported ?

Well supported, recall of every definition needed (DP, OP, PAC...)

Is it well Written ?

Well written clear approach, step by step with definitions

What do you like ?

Every notation is well defined, we can find examples of the possible applications of this demonstration

What's its Limitations ?

Applications are currently only limited to open-questions.

References

Mark Bun, Roi Livni and Shay Moran. An Equivalence Between Private Classification and Online Prediction.
https://arxiv.org/abs/2003.00563