Pattern Recognition Letters 89 (2017) 31–38

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Differential privacy and generalization: Sharper bounds with applications

Luca Oneto[a,*], Sandro Ridella[b], Davide Anguita[a]

[a] DIBRIS Department, University of Genova, Via Opera Pia 13, I-16145 Genova, Italy
[b] DITEN Department, University of Genova, Via Opera Pia 11A, I-16145 Genova, Italy

ARTICLE INFO

Article history:
Received 31 July 2016
Available online 9 February 2017

MSC:
41A05
41A10
65D05
65D17

Keywords:
Differential privacy
Generalization bound
Chernoff type bound
Bennett type bounds
Gibbs classifier
Randomized classifier
Catoni prior and posterior
Thresholdout
Model selection
Error estimation

ABSTRACT

In this paper we deal with the problem of improving the recent milestone results on the estimation of the generalization capability of a randomized learning algorithm based on Differential Privacy (DP). In particular, we derive new DP based multiplicative Chernoff and Bennett type generalization bounds, which improve over the current state-of-the-art Hoeffding type bound. Then, we prove that a randomized algorithm based on the data generating dependent prior and data dependent posterior Boltzmann distributions of Catoni (2007) [10] is Differentially Private and shows better generalization properties than the Gibbs classifier associated to the same distributions. With this aim, we also exploit a simple example. Finally, we discuss the advantages of using the Thresholdout procedure, one of the main results generated by the DP theory, for Model Selection and Error Estimation purposes, and we derive a new result which exploits our new generalization bounds.

1. Introduction

The problem of learning from data while preserving the privacy of individual observations has a long history and spans over multiple disciplines [21,29,55]. One way to preserve privacy is to corrupt the learning procedure with noise without destroying the information that we want to extract. Differential Privacy (DP) is one of the most powerful tools in this context [16,21]. DP addresses the apparently self-contradictory problem of keeping private the information about an individual observation while learning useful information about a population. In particular, a procedure is DP if and only if its output is almost independent from any of the individual observations. In other words, the probability of a certain output should not change significantly if one individual is present or not, where the probabilities are taken over the noise introduced by the procedure. In the last years, DP has been deeply studied from a theoretical point of view [11,20,22,33–35,38,43,49,51,52,57] and exploited to develop new learning strategies for solving real world problems [7,12,13,24,31,32,50,56].

Another key problem in learning from data is the one of Model Selection (MS) and Error Estimation (EE) which aim at tuning and assessing the performance of the learning procedure [1]. Resampling techniques like hold out, cross validation and bootstrap [1] are often used by practitioners because work well in many situations, but they may lead to severe problems of false discovery [48] and they do not give insight into the learning process. The first seminal work in filling these gaps is the one of Vapnik [54] about the Vapnik–Chervonenkis Dimension, which states the conditions under which a set of hypotheses is learnable. Later these results have been improved with the introduction of the Rademacher Complexity [3] together with its localized counterpart [2]. The theory of Floyd and Warmuth [23], which tightly connects compression to learning, later extended by Langford and McAllester [37], was another step forward in the direction of understanding the learning properties of an algorithm. A breakthrough was made with the Algorithmic Stability [9,47], which states the properties that a learning algorithm should fulfill in order to achieve good generalization performance. Finally, it is well

* Corresponding author.
E-mail address: luca.oneto@unige.it (L. Oneto).

http://dx.doi.org/10.1016/j.patrec.2017.02.006
0167-8655/

known that combining the output of different learning procedures results in much better performance than using any one of them alone, but it is hard to combine them appropriately in order to obtain good performance [10,42] and it is not trivial the assessment of the performance of the resulting learning procedure [4,27,39,53]. The PAC-Bayes theory is one of the most powerful tool in this context. For example in classification problems, it allows to bound the risk of the Gibbs Classifier (GC) and the Bayes Classifier (BC) and has inspired the development of new theoretically grounded weighting strategies such as the one developed by Catoni [10]. In particular he proposed to use a new data dependent weighting strategy which has shown many strong and interesting theoretical properties [45].

DP allowed to reach a milestone result by connecting the field of privacy preserving data analysis and the generalization capability of a learning algorithm. From one side it proved that a learning algorithm which shows DP properties also generalizes [18]. From the other side, if an algorithm is not DP, it allowed to state the conditions under which a hold out set can be reused without risk of false discovery through a DP procedure called Thresholdout [17,19].

In this paper we make an additional step forward in this direction by developing DP based multiplicative Chernoff and Bennett type generalization bounds, reported in Section 2, which improve over the current state-of-the-art Hoeffding type bound. Then, in Section 3 we show that a randomized algorithm based on the work of Catoni [10] has interesting DP properties and shows better generalization performances than the associated GC by also exploiting a simple example. Finally, in Section 4 we discuss the advantages of exploiting the Thresholdout for MS and EE purposes and we derive a novel result which exploits our newly derived generalization bounds reported in Section 2. Section 5 concludes our work.

2. DP and generalization

In order to present our novel results, we first recall some preliminary definitions [18,21,54]. Let us consider an input space $\mathcal{X}$ and an output space $\mathcal{Y} = \{-1, +1\}$, since in this work we will deal with binary classification problems. We indicate with $\mathfrak{P}_{\mathcal{X}}$, $\mathfrak{P}_{\mathcal{Y}}$, and $\mathfrak{P}_{\mathcal{Z}}$ respectively the distributions over $\mathcal{X}$, $\mathcal{Y}$, and the cartesian product between the input and the output space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. From $\mathcal{Z}$ we observe a series of n i.i.d. samples $s = \{z_1, \ldots, z_n\} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $\forall i \in \mathcal{I}_n = \{1, 2, 3, \ldots, n\}$ we have $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, and $z_i \in \mathcal{Z}$. Moreover, $\boldsymbol{Z}$ is a random variable sampled from $\mathcal{Z}$ according to $\mathfrak{P}_{\mathcal{Z}}$ whereas s is a dataset inside the space of all the possible datasets $\mathcal{S} = \mathcal{Z}^n$ and $\mathfrak{P}_{\mathcal{S}}$ is the distribution of probability generated by $\mathfrak{P}_{\mathcal{Z}}$ over $\mathcal{S}$. Analogously to $\boldsymbol{Z}$, $\boldsymbol{S}$ is a random variable sampled from $\mathcal{S}$ according to $\mathfrak{P}_{\mathcal{S}}$. We denote with $\dot{s}$ the neighborhood dataset of s such that $s = \{z_1, \ldots, z_{i-1}, \dot{z}_i, z_{i+1}, \ldots, z_n\}$ where i may assume any value in $\mathcal{I}_n$ and $\dot{z}_i$ i.i.d. with z_i. We denote with $\breve{\mathcal{S}}$ a subset of the space of datasets $\mathcal{S}$: $\breve{\mathcal{S}} \subseteq \mathcal{S}$. Let us define with $f : \mathcal{X} \to [-1, 1]$ a function in a space $\mathcal{F}$ of all the possible functions and $\breve{\mathcal{F}} \subseteq \mathcal{F}$. A randomized algorithm $\mathscr{A} : \mathcal{S} \to \mathcal{F}$ maps a dataset $s \in \mathcal{S}$ in a function $f \in \mathcal{F}$ with nondeterministic rules that can be encapsulated in a distribution $\mathfrak{P}_{\mathscr{A}}$ over $\mathcal{F}$ given $s \in \mathcal{S}$. We also define an operator $\breve{D}$ which maps a function $f \in \mathcal{F}$ into a subset of all the possible datasets $\breve{\mathcal{S}}$. For example, $\breve{D}$ can be seen as an inverse operator of $\mathscr{A}$ which, given an $f \in \mathcal{F}$, tries to retrieve the datasets $\breve{\mathcal{S}}$ that may have generated f. The accuracy of $f \in \mathcal{F}$ in representing $\mathfrak{P}_{\mathcal{Z}}$ is measured with reference to a loss function $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$. Hence, we can define the true risk of f, namely generalization error, as $L(f) = \mathbb{E}_{\boldsymbol{Z}} \ell(f, \boldsymbol{Z})$, together with its variance $V(f) = \mathbb{E}_{\boldsymbol{Z}}[\ell(f, \boldsymbol{Z}) - L(f)]^2$. Since $\mathfrak{P}_{\mathcal{Z}}$ is unknown, $L(f)$ and $V(f)$ cannot be computed. Therefore, we have to resort to their empirical estimators, respectively the empirical error $\widehat{L}_n^s(f) = 1/n \sum_{i=1}^n \ell(f, z_i)$, and the empirical variance $\widehat{V}_n^s(f) = 1/n(n-1) \sum_{i=1}^n \sum_{j=i+1}^n [\ell(f, z_i) - \ell(f, z_j)]^2$ [41].

2.1. State-of-the-art

Before presenting our advances with respect to the state-of-the-art we need to recall the current results that can be retrieved from the literature. In particular we need to recall the definition of DP.

Definitions 1 ([21]). A randomized algorithm $\mathscr{A}$ is (ϵ, δ)-Differentially Private if $\forall \breve{\mathcal{F}} \subseteq \mathcal{F}$ and $\forall s \in \mathcal{S}$ we have that $\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(s) \in \breve{\mathcal{F}}\} \leq e^{\epsilon} \mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{s}) \in \breve{\mathcal{F}}\} + \delta$.

Note that in this work we will only deal with $(\epsilon, 0)$-Differentially Private algorithms that we will denote as ϵ-DP for brevity.

Since we are dealing with ϵ-DP algorithms it is useful to derive the following lemma which gives an alternative simpler and more intuitive definition of ϵ-DP. Basically Lemma 1 says that if the probability of choosing a function does not change too much if the algorithm is fed with a dataset s or with its neighborhood $\dot{s}$ then the algorithm is private. The latter will be used later in the paper.

Lemma 1. *A randomized algorithm $\mathscr{A}$ is ϵ-DP if $\forall f \in \mathcal{F}$ and $\forall s \in \mathcal{S}$ we have that $\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(s) = f\} \leq e^{\epsilon} \mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{s}) = f\}$.*

Proof. In order to prove our statement note that

$$\mathbb{P}_{\mathscr{A}}\left\{\mathscr{A}(s) \in \breve{\mathcal{F}}\right\} = \int_{\breve{\mathcal{F}}} \mathbb{P}_{\mathscr{A}}\{\mathscr{A}(s) = f\} df \tag{1}$$
$$\leq \int_{\breve{\mathcal{F}}} e^{\epsilon} \mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{s}) = f\} df = e^{\epsilon} \mathbb{P}_{\mathscr{A}}\left\{\mathscr{A}(\dot{s}) \in \breve{\mathcal{F}}\right\}.$$

By looking at Definition 1 the statement is proved. □

The milestone result of Dwork et al. [18] shows that an ϵ-DP algorithm generalizes. In particular two main results are derived. The first one is very general and shows that if a function $\breve{D}(f)$ is defined for each element $f \in \mathcal{F}$ and the probability that $\boldsymbol{S} \in \breve{D}(f)$ is small, then the probability remains small if f is chosen based on $\boldsymbol{S}$ and $\mathscr{A}$. In other words the probability that $\boldsymbol{S} \in \breve{D}(\mathscr{A}(\boldsymbol{S}))$ remains small.[1]

Theorem 1 ([18]). *Let $\mathscr{A}$ be an ϵ-DP. Let us suppose that $\mathbb{P}_{\boldsymbol{S}}\{\boldsymbol{S} \in \breve{D}(f)\} \leq \beta$, $\forall f \in \mathcal{F}$. Then, for $\epsilon \leq \sqrt{\ln(1\beta)/2n}$ we have that $\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{\boldsymbol{S} \in \breve{D}(\boldsymbol{F})\} \leq 3\sqrt{\beta}$.*

The second result, which builds upon Theorem 1, shows that the empirical error of a function chosen with an ϵ-DP algorithm is concentrated around its generalization error.

Corollary 1 ([18]). *Let $\mathscr{A}$ be an ϵ-DP, then for any $t > 0$, setting $\epsilon \leq \sqrt{t^2 - \ln(2)/2n}$ ensures that*

$$\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{|L(\boldsymbol{F}) - \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})| \geq t\right\} \leq 3\sqrt{2} e^{-nt^2}.$$

The result of Corollary 1 can be reformulated in a more convenient expression, which is more suited for the subsequent analysis.

Lemma 2. *Let $\mathscr{A}$ be an ϵ-DP, then we can state that*

$$\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{|L(\boldsymbol{F}) - \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})| \geq \epsilon + \sqrt{1/n}\right\} \leq 3e^{-n\epsilon^2}.$$

Proof. Let us consider Corollary 1. By setting $\epsilon = \sqrt{t^2 - \ln(2)/2n}$, which is the most convenient choice since it leads to a tighter bound, we have that $t^2 = \epsilon^2 + \ln(2)/2n$. By noting that $\sqrt{\epsilon^2 + \ln(2)/2n} \leq \epsilon + \sqrt{1/n}$, the statement is proved. □

The limitation of Corollary 1 (or Lemma 2) is the slow convergence rate $O(1/\sqrt{n})$. When the empirical error is small we would like to retrieve a Chernoff type result [14]. Instead, when the variance is small, a Bernstein or Bennet bound would be preferred

[1] From now on with a little abuse of notation we will identify $\boldsymbol{F} = \mathscr{A}(\boldsymbol{S})$.