Neural Networks 65 (2015) 115–125

Contents lists available at ScienceDirect

Neural Networks

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

Local Rademacher Complexity: Sharper risk bounds with and without unlabeled samples

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

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

[b] DIBRIS - University of Genova, Via Opera Pia 13, I-16145 Genova, Italy

ARTICLE INFO

Article history:
Received 23 October 2014
Accepted 4 February 2015
Available online 16 February 2015

Keywords:
Statistical learning theory
Performance estimation
Local Rademacher Complexity
Unlabeled samples

ABSTRACT

We derive in this paper a new Local Rademacher Complexity risk bound on the generalization ability of a model, which is able to take advantage of the availability of unlabeled samples. Moreover, this new bound improves state-of-the-art results even when no unlabeled samples are available.

1. Introduction

A learning process can be described as the selection of a hypothesis in a fixed set, based on empirical observations (Vapnik, 1998). Its asymptotic analysis, through a bound on the generalization error, has been thoroughly investigated in the past (Talagrand, 1987; Vapnik, 1998). However, as the number of samples is limited in practice, finite sample analysis with global measures of the complexity of the hypothesis set was proposed, and represented a fundamental advance in the field (Bartlett & Mendelson, 2003; Bousquet & Elisseeff, 2002; Koltchinskii, 2006; McAllester & Akinbiyi, 2013; Valiant, 2013; Vapnik, 1998). A further refinement has consisted in exploiting local measures of complexity, which take in account only those models that well approximate the available data (Bartlett, Bousquet, & Mendelson, 2002, 2005; Blanchard & Massart, 2006; Koltchinskii, 2006; Lever, Laviolette, & Shawe-Taylor, 2013). Recently, some attempts to further improve these results have been made (Audibert & Tsybakov, 2007; Srebro, Sridharan, & Tewari, 2010; Steinwart & Scovel, 2007): unfortunately, these approaches require additional assumptions that, in general, are not satisfied or cannot be justified by inferring them from the data. Alternative paths have been explored like, for example, exploiting additional a-priori information (Parrado-Hernández, Ambroladze, Shawe-Taylor, & Sun, 2012). Recently, the use of unlabeled samples has been proposed for improving the tightness of Global Rademacher Complexity based bounds (Anguita, Ghio, Oneto, & Ridella, 2011). Such results are appealing since unlabeled samples are commonly available in many real world applications, as also confirmed by the success of learning procedures able to exploit them (Chapelle, Schölkopf, & Zien, 2006).

In this paper, we extend the recent results on the use of unlabeled samples in global complexity measures to the case of local ones and derive sharper Local Rademacher Complexity risk bounds on the generalization ability of a model. For this purpose, two steps are completed. First, we propose a proof for the Local Rademacher Complexity bound, simplified with respect to the milestone result of (Bartlett et al., 2005) through the exploitation of the well-known bounded difference inequality (McDiarmid, 1989). Such simplification enables us to apply results on concentration inequalities of self-bounding functions (Boucheron, Lugosi, & Massart, 2013), and to obtain a sharper Local Rademacher Complexity risk bound. The latter improves the state-of-the-art results both when unlabeled samples are used and the dataset is entirely composed of labeled samples.

2. The learning framework

We consider the conventional learning problem (Vapnik, 1998): based on a random observation of $X \in \mathcal{X}$, one has to estimate $Y \in \mathcal{Y}$ by choosing a suitable hypothesis $h : \mathcal{X} \to \hat{\mathcal{Y}}$, where $h \in \mathcal{H}$. A learning algorithm selects h by exploiting a set of labeled samples $\mathcal{D}_{n_l} : \left\{\left(X_1^l, Y_1^l\right), \ldots, \left(X_{n_l}^l, Y_{n_l}^l\right)\right\}$ and, eventually, a set of unlabeled

* Corresponding author.
E-mail addresses: Luca.Oneto@unige.it (L. Oneto), Alessandro.Ghio@unige.it (A. Ghio), Sandro.Ridella@unige.it (S. Ridella), Davide.Anguita@unige.it (D. Anguita).

http://dx.doi.org/10.1016/j.neunet.2015.02.006
0893-6080/

ones $\mathcal{D}_{n_u} : \left\{\left(X_1^u\right), \ldots, \left(X_{n_u}^u\right)\right\}$. $\mathcal{D}_{n_l}$ and $\mathcal{D}_{n_u}$ consist of a sequence of independent samples distributed according to μ over $\mathcal{X} \times \mathcal{Y}$. The generalization error

$$L(h) = \mathbb{E}_\mu \ell(h(X), Y), \tag{1}$$

associated to a hypothesis $\mathcal{H}$, is defined through a loss function $\ell(h(X), Y) : \hat{\mathcal{Y}} \times \mathcal{Y} \to [0, 1]$. As μ is unknown, $L(h)$ cannot be explicitly computed, thus we have to resort to its empirical estimator, namely the empirical error

$$\hat{L}_{n_l}(h) = \frac{1}{n_l} \sum_{i=1}^{n_l} \ell\left(h\left(X_i^l\right), Y_i^l\right). \tag{2}$$

Note that $\hat{L}_{n_l}(h)$ is a biased estimator, since the data used for selecting the model and for computing the empirical error coincide. We estimate this bias by studying the discrepancy between the generalization error and the empirical error. For this purpose we exploit powerful statistical tools like concentration inequalities and the Local Rademacher Complexity.

2.1. Definitions

In the seminal work of Bartlett et al. (2005), a bound, defined over the space of functions, is provided. In this work, we generalize this result to a more general supervised learning framework. For this purpose, we switch from the space of functions $\mathcal{H}$ to the space of loss functions.

Definition 2.1. Given a space of functions $\mathcal{H}$ with its associated loss function $\ell(h(X), Y)$, the space of loss functions $\mathcal{L}$ is defined as:

$$\mathcal{L} = \left\{\ell(h(X), Y) \middle| h \in \mathcal{H}\right\}. \tag{3}$$

Let us also consider the corresponding star-shaped space of function.

Definition 2.2. Given the space of loss functions $\mathcal{L}$, its star-shaped version is:

$$\mathcal{L}^s = \left\{\alpha\ell \middle| \alpha \in [0, 1],\ \ell \in \mathcal{L}\right\}. \tag{4}$$

Then, the generalization error and the empirical error can be rewritten in terms of the space of loss functions:

$$L(h) \equiv L(\ell) = \mathbb{E}_\mu \ell(h(X), Y), \tag{5}$$

$$\hat{L}_{n_l}(h) \equiv \hat{L}_{n_l}(\ell) = \frac{1}{n_l} \sum_{i=1}^{n_l} \ell\left(h\left(X_i^l\right), Y_i^l\right). \tag{6}$$

Moreover we can define, respectively, the expected square error and the empirical square error:

$$L(\ell^2) = \mathbb{E}_\mu \left[\ell(h(X), Y)\right]^2, \tag{7}$$

$$\hat{L}_{n_l}(\ell^2) = \frac{1}{n_l} \sum_{i=1}^{n_l} \left[\ell\left(h\left(X_i^l\right), Y_i^l\right)\right]^2. \tag{8}$$

Consequently, the variance of $\ell \in \mathcal{L}$ can be defined as:

$$V^2(\ell) = \mathbb{E}_\mu \left[\ell(h(X), Y) - L(\ell)\right]^2 = L(\ell^2) - [L(\ell)]^2. \tag{9}$$

Note that the following relations hold:

$$V^2(\ell) \le L(\ell^2) \le L(\ell), \qquad L[(\alpha\ell)^2] = \alpha^2 L(\ell^2). \tag{10}$$

Since we do not know in advance which $h \in \mathcal{H}$ will be chosen during the learning phase, in order to estimate $L(\ell)$ we have to study the behavior of the difference between the generalization error and the empirical error.

Definition 2.3. Given $\mathcal{L}$, the Uniform Deviation of the loss $\hat{U}_{n_l}(\mathcal{L})$ and square loss $\hat{U}^2_{n_l}(\mathcal{L})$ are:

$$\begin{aligned} \hat{U}_{n_l}(\mathcal{L}) &= \sup_{\ell \in \mathcal{L}} \left[L(\ell) - \hat{L}_{n_l}(\ell)\right], \\ \hat{U}^2_{n_l}(\mathcal{L}) &= \sup_{\ell \in \mathcal{L}} \left[\hat{L}_{n_l}(\ell^2) - L(\ell^2)\right], \end{aligned} \tag{11}$$

while their deterministic counterparts are:

$$U_{n_l}(\mathcal{L}) = \mathbb{E}_\mu \hat{U}_{n_l}(\mathcal{L}), \qquad U^2_{n_l}(\mathcal{L}) = \mathbb{E}_\mu \hat{U}^2_{n_l}(\mathcal{L}). \tag{12}$$

The Uniform Deviation is not computable, but we can upper bound its value through some computable quantity. One possibility is to use the Rademacher Complexity.

Definition 2.4. The Rademacher Complexity of the loss and of the square loss are:

$$\hat{R}_{n_l}(\mathcal{L}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\ell \in \mathcal{L}} \frac{2}{n_l} \sum_{i=1}^{n_l} \sigma_i \ell\left(h\left(X_i^l\right), Y_i^l\right), \tag{13}$$

$$\hat{R}^2_{n_l}(\mathcal{L}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\ell \in \mathcal{L}} \frac{2}{n_l} \sum_{i=1}^{n_l} \sigma_i \left[\ell\left(h\left(X_i^l\right), Y_i^l\right)\right]^2, \tag{14}$$

where $\sigma_1, \ldots, \sigma_{n_l}$ are n_l $\{\pm 1\}$-valued independent Rademacher random variables for which $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Their deterministic counterparts are:

$$R_{n_l}(\mathcal{L}) = \mathbb{E}_\mu \hat{R}_{n_l}(\mathcal{L}), \qquad R^2_{n_l}(\mathcal{L}) = \mathbb{E}_\mu \hat{R}^2_{n_l}(\mathcal{L}). \tag{15}$$

In Appendix A, some propaedeutic properties of the Uniform Deviation and Rademacher Complexity are recalled, which will be useful for deriving the main results of this work.

Finally, we will also make use of the notion of sub-root function (Bartlett et al., 2005).

Definition 2.5. A function is a sub-root function if and only if:

(I) $\psi(r)$ is positive,

(II) $\psi(r)$ is non-decreasing,

(III) $\psi(r)/\sqrt{r}$ is non-increasing,

with $r > 0$.

Its properties are reported in Appendix B.

3. Local Rademacher complexity error bound

In this section, we propose a proof of the Local Rademacher Complexity bound on the generalization error of a model (Bartlett et al., 2005; Koltchinskii, 2006), which is simplified with respect to the original proof in literature and allows us also to obtain optimal constants.

In order to improve the readability of the paper, an outline of the main steps of the proof is presented. As a first step, Theorem 3.1 shows that it is possible to bound the generalization error of a function chosen in $\mathcal{H}$, through an assumption over the Expected Uniform Deviation of a normalized and slightly enlarged version (see Lemma 3.2) of $\mathcal{H}$. As a second step, Theorem 3.3 shows how to relate the Expected Uniform Deviation and the Expected Rademacher Complexity through the use of a sub-root function. The fixed point of this sub-root function is used to bound the generalization error of a function chosen in $\mathcal{H}$. As a third step, Lemma 3.4 shows that, instead of using any sub-root function, we can directly use the Expected Rademacher Complexity of a local space of functions,