ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Improving the Union Bound: a Distribution Dependent Approach

Luca Oneto, Sandro Ridella, Davide Anguita

DIBRIS - University of Genoa - Via Opera Pia 11a, 16145, Genoa, Italy

Abstract. Statistical Learning Theory deals with the problem of estimating the performance of a learning procedure. Any learning procedure implies making choices and this choices imply a risk. When the number of choices is finite, the state-of-the-art tool for evaluating the total risk of all the choice made is the Union Bound. The problem of the Union Bound is that it is very loose in practice if no a-priori information is available. In fact, the Union Bound considers all choices equally plausible while, as a matter of fact, a learning procedure targets just particular choices disregarding the others. In this work we will show that it is possible to improve the Union Bound based results using a distribution dependent weighting strategy of the true risks associated to each choice. Then we will prove that our proposal outperforms or, in the worst case, it degenerate in the Union Bound.

1 Introduction

Statistical Learning Theory (SLT) [1, 2] deals with the problem of understanding and estimating the performance of a statistical learning procedure. Although asymptotic analysis is a crucial first step in this direction, finite sample error bounds are of more value as they allow the design of model selection procedures [3]. These error bounds typically have the following form: with high probability, the generalization error of the selected hypothesis, chosen in a space of possible ones, is bounded by an empirical estimate of error plus a penalty term which depends on the size of the hypothesis space and the number of samples available. The latter terms basically take into account that the learning procedure made a choice between a set of possible options based on the available data. Every data dependent choice implies a risk and the penalty term is exactly the measure of this risk. When the number of choices is finite, namely the hypothesis space is composed by an arbitrary finite number of hypothesis, the state-of-the-art tool for evaluating their total risk is the Union Bound (UB) also called Bonferroni Bound [1, 4]. The Union Bound is an ubiquitous building block in SLT [3]: in the Vapnik-Chervonenkis theory, in the Rademacher Complexity theory, in the Algorithmic Stability theory, in the Compression Bound, in the PAC-Bayes theory, and even in the Differential Privacy theory.

Our proposal, takes inspiration from several works in the field for reaching a better result. The first idea, which is also a driver of the Shell Bound, is that, during any learning procedure the hypotheses with high error will be never taken into account and consequently we should not pay the risk for those hypotheses [5]. The second idea is that, since we do not know the true error of the hypotheses but just its empirical one, we should discard those hypotheses for which the estimated confidence intervals do not overlap [6] with the ones of the hypothesis of minimal training error. The third idea is that, since there is no supporting

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

theory for discarding the hypothesis with non-overlapping confidence intervals, we should weight differently the risk associated to each hypothesis based on their true error analogously to what is done in the field of multiple hypotheses testing [7]. The fourth idea is that other researchers have shown that a distribution dependent weighting strategy can be performed without the actual knowledge of the distribution [8]. By combining all these ideas we will be able to derive our proposal and show that improves both on the Union Bound and on the Shell Bound.

2 Preliminaries

Let us consider, for simplicity, the classical binary classification framework. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{-1, +1\}$ be the set of binary output labels. Let $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$ $\forall i \in \{1, \cdots, n\}$, be a sequence of $n \in \mathbb{N}^*$ samples drawn independently from an unknown probability distribution μ over $\mathcal{X} \times \mathcal{Y}$. Let us consider an hypothesis $h : \mathcal{X} \to \mathcal{Y}$ chosen from a finite set $\mathcal{H}$ of possible hypotheses of cardinality $m \in \mathbb{N}^*$ such that $\mathcal{H} = \{h_i : i \in \mathcal{I}\}$ where $\mathcal{I} = \{1, \cdots, m\}$. The error of h in approximating $\mathbb{P}\{Y|X\}$ is measured by a prescribed bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. The generalization error of h is defined as $\mathtt{R}(h) = \mathbb{E}\{\ell(h(X), Y)\} \in [0, 1]$. Since the probability measure μ is usually unknown, the generalization error cannot be computed, however we can compute the empirical error $\hat{\mathtt{R}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i) \in [0, 1]$. If the choice of $h \in \mathcal{H}$ does not depend on $\mathcal{D}_n$, namely if we want to bound the generalization error of a single hypothesis in the hypothesis space chosen before seeing the data, it is possible to prove that

$$\mathbb{P}\left\{\mathtt{R}(h) \geq \mathtt{L}(\hat{\mathtt{R}}(h), \delta)\right\} \geq 1 - \delta, \quad \mathbb{P}\left\{\mathtt{R}(h) \leq \mathtt{U}(\hat{\mathtt{R}}(h), \delta)\right\} \geq 1 - \delta,$$

where $\delta \in (0, 1)$ while $\mathtt{L}$ and $\mathtt{U}$ are respectively lower and upper bounds of the generalization error (e.g. [9]).

In general the choice of $h \in \mathcal{H}$ does depend on $\mathcal{D}_n$: in this case we have to estimate the risk due to this data dependent choice. As an example, common practice for choosing $h \in \mathcal{H}$ based on $\mathcal{D}_n$ is to choose the hypothesis with minimum empirical error $i^* : \arg\min_{i \in \mathcal{I}} \hat{\mathtt{R}}(h_i)$, and this approach is called Empirical Risk Minimization, but others possibilities exist such as the Structural Risk Minimization, or the penalized (regularized) Empirical Risk Minimization [10].

In order to guarantee a prescribed confidence level, or risk, of the chosen hypothesis, the UB can be applied.

Theorem 1. *Let $q_i = q(h_i) \in (0, 1)$ and $p_i = p(h_i) \in (0, 1)$ be some weight associated to h_i with $i \in \mathcal{I}$ before seeing the data and such that $\sum_{i \in \mathcal{I}} q_i + p_i = 1$, then the following bounds hold*

$$\mathbb{P}\left\{\mathtt{L}\left(\hat{\mathtt{R}}(h_i), \delta q(h_i)\right) \leq \mathtt{R}(h_i) \leq \mathtt{U}\left(\hat{\mathtt{R}}(h_i), \delta p(h_i)\right) \ \ \forall i \in \mathcal{I}\right\} \geq 1 - \delta.$$

Theorem 1 introduces a weight for each risk associated to each choice. Weighting more the risk associated to useful choices leads to tighter bounds on the generalization error of hypotheses that will be selected by the algorithm (hypotheses characterized by small empirical error) and looser estimates over the others (hypotheses characterized by high empirical error). Unfortunately, this approach

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

is mainly theoretical since the weights must be chosen before seeing the data and consequently we cannot set them without an *a priori* knowledge about the problem. Since Theorem 1 does not propose any solution for the choice of these weights these are set to be the same for each of the choices $q(h_i) = p(h_i) = 1/2m$ $\forall i \in \mathcal{I}$.

3 Our Proposal

In this work, we propose a Distribution Dependent Weighted UB (DDWUB) where the weights depend on some parameters of the distribution which generated them. In particular, we define a set of functions $f_i^p : \mathbb{R}^m \to \mathbb{R}$ and $f_i^q : \mathbb{R}^m \to \mathbb{R}$ with $i \in \mathcal{I}$ such that

$$q(h_i) = f_i^q(\mathtt{R}_1, \cdots, \mathtt{R}_m), p(h_i) = f_i^p(\mathtt{R}_1, \cdots, \mathtt{R}_m) \in (0,1),\ \forall i \in \mathcal{I},$$
$$\textstyle\sum_{i\in\mathcal{I}} f_i^q(\mathtt{R}_1, \cdots, \mathtt{R}_m) + f_i^p(\mathtt{R}_1, \cdots, \mathtt{R}_m) = 1.$$

Note that f_i^q, f_i^p with $i \in \mathcal{I}$ are quite general and data independent and for this reason they can be inserted in Theorem 1.

Since in our case, the scope is to find tighter upper bounds of the empirical minimizer, we study the case when, for $\mathtt{L}$ and $\mathtt{U}$, the proposal of [9] is exploited, and we set

$$f_i^q(\mathtt{R}_1, \cdots, \mathtt{R}_m) = 1/2m, f_i^p(\mathtt{R}_1, \cdots, \mathtt{R}_m) = 1/2e^{-\gamma \max[\theta, \mathtt{R}_i]} / \textstyle\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, \mathtt{R}_j]},\ \forall i \in \mathcal{I},$$

with particular values of $\gamma \in [0, \infty)$ and $\theta \in [0, 1]$. The choiche of the weights takes inspiration from the work of [8] which proposed, in the context of the PAC-Bayes theory, a distribution dependent method for assigning an a-priori distribution over a set of hypotheses in order to give an higher probability to the hypothesis with small generalization error. This method has been shown to possess interesting theoretical properties and to be also quite effective in practical applications [3]. In this setting it is possible to state our DDWUB.

Corollary 1. *If* $\forall i \in \mathcal{I}$ $f_i(r_1, \cdots, r_m) = e^{-\gamma \max[\theta, r_i]} / \sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]}$ *with* $\gamma \in [0, \infty)$ *and* $\theta \in [0, 1]$*, then the following bound holds*[1]

$$\mathbb{P}\left\{ \max\left[0, \hat{\mathtt{R}}_i - \sqrt{\frac{\log\left(\frac{2m}{\delta}\right)}{2n}}\right] \leq \mathtt{R}_i \leq \min\left[1, \hat{\mathtt{R}}_i + \sqrt{\frac{\log\left(\frac{2}{\delta f_i(\mathtt{R}_1, \cdots, \mathtt{R}_m)}\right)}{2n}}\right]\ \ \forall i \in \mathcal{I} \right\} \geq 1 - \delta.$$

Corollary 1 is a direct consequence of Theorem 1 and [9]. In Corollary 1, γ acts as a weighting factor. The larger is γ the larger are the weights of the risks associated to hypotheses with small empirical error and the smaller are the weights of the risks associated to hypotheses with large empirical error. For $\gamma \to \infty$ we have that $p_{i^*} \to 1$ and $p_i \to 0$ $\forall i \in \mathcal{I} \setminus i^*$. The smaller is γ the less is the difference between the weights of the risks. For $\gamma \to 0$ we have that $p_i = 1/m$ $\forall i \in \mathcal{I}$. In Corollary 1, θ, instead, acts as a protection against the fact that the empirical error is measured over a finite number of samples and,

[1] We replace, for brevity, $\mathtt{R}(h_i)$ and $\hat{\mathtt{R}}(h_i)$ with $\mathtt{R}_i$ and $\hat{\mathtt{R}}_i$.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

if the sample size is small, hypotheses with a small difference in the empirical error are indistinguishable. In other words, the weights depend on unknown parameters of the data generating distribution, then we will have to estimate them and since the number of sample is finite these estimates will not allow us to distinguish hypotheses which show similar empirical error. For this reasons, θ allows to give the same the weight to the risks associated to hypotheses with small empirical error. As we will see in this section, the values of γ and θ must be set in a particular way in order to be sure that DDWUB improves over the UB. In particular: Lemma 1 shows that in order to upper bound the generalization error of the empirical risk minimizer based on DDWUB of Corollary 1 we have to solve an optimization problem, Theorem 2 show that for particular values of γ the solution is unique and can be found by simply search for the fixed point of a simple function, and Theorem 3 show that for particular values of θ it is possible to prove that DDWUB is tighter than, or in the worst case as tight as, the UB. The proof the theorems can be found in Appendix A.

Lemma 1. *Under the same conditions of Corollary 1 if* $i^* = \arg\min_{i\in\mathcal{I}} \hat{\mathtt{R}}_i$*, then we can state that following bound holds*

$$\begin{aligned}
&\mathtt{R}_{i^*} \leq \max_{r_1,\cdots,r_m} \; r_{i^*} \\
&s.t.\; r_i \geq \max\left[0, \hat{\mathtt{R}}_i - \sqrt{\log\left(2m/\delta\right)/2n}\right],\; \forall i \in \mathcal{I} \\
&\qquad r_i \leq \min\left[1, \hat{\mathtt{R}}_i + \sqrt{\log\left(2\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]}/\delta e^{-\gamma \max[\theta, r_i]}\right)/2n}\right],\; \forall i \in \mathcal{I}.
\end{aligned}$$

Theorem 2. *Under the same conditions of Lemma 1 if* $\gamma \leq 2\sqrt{n}$*, the solution of the optimization problem of Lemma 1 exists, it is unique, and it is the fixed point* $r^*_{i^*}$ *of the following function of* r_{i^*}

$$r_{i^*} = \min\left[1, \hat{\mathtt{R}}_{i^*} + \sqrt{\log\left(2\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]}/\delta e^{-\gamma \max[\theta, r_{i^*}]}\right)/2n}\right]$$

where $r_i = \max\left[0, \hat{\mathtt{R}}_i - \sqrt{\log\left(2m/\delta\right)/2n}\right]\; \forall i \in \mathcal{I} \setminus i^*$.

Note that, in order to find the fixed point defined in Theorem 2 a simple bisection method can be applied.

Theorem 3. *Under the same conditions of Theorem 2 if*

$$\theta = \min\left[1, \min\left[\mathtt{R}_1, \cdots, \mathtt{R}_m\right] + 2\sqrt{\log\left(2m/\delta\right)/2n}\right]. \tag{1}$$

then $r^*_{i^*} \leq \min\left[1, \hat{\mathtt{R}}_{i^*} + \sqrt{\log\left(2m/\delta\right)/2n}\right]$.

By finding the $r^*_{i^*}$ for all possible values of θ and then by selecting the largest one which satisfies Eq. (1) we have the results of our DDWUB.

3.1 Example

Before presenting DDWUB in the general setting we would like to show an application of DDWUB in the simplified setting. Let us consider the case when

$$\hat{\mathtt{R}}_1 = \hat{\mathtt{R}}_2 = 0,\; \hat{\mathtt{R}}_3 = \hat{\mathtt{R}}_4 = \cdots = \hat{\mathtt{R}}_m = \nu, \quad \nu \in \{1/n, 2/n, \cdots, 1\}.$$

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Let us set $\gamma = 2\sqrt{n}$ (see Theorem 2) and note that, in order to upper bound the function with the smallest empirical error (i.e. the one corresponding to $\hat{\mathtt{R}}_1$) we have that DDWUB states that, $r_1 = \sqrt{\ln\left(2\sum_{i=1}^{m} e^{-2\sqrt{n}\max[\theta, r_i]}/\delta e^{-2\sqrt{n}\max[\theta, r_1]}\right)/2n}$, $r_2 = 0$, $r_3 = r_4 = \cdots = \nu - \sqrt{\ln(2m/\delta)/2n}$, $\theta = \min\left[1, \min[r_1, \cdots, r_m] + 2\sqrt{\log(2m/\delta)/2n}\right]$. Note that $\min[r_1, \cdots, r_m] = 0 \rightarrow \theta = \sqrt{\ln(2m/\delta)/2n}$. Thanks to the theory of DDWUB we can state that $r_1^* \leq \theta$. Let us note that if $m < \frac{\delta e^{2n\nu^2}}{2}$, then $r_3 = \cdots = r_m > \theta$. Than we can easily state that

$$\lim_{n\to\infty} \frac{2\sum_{i=1}^{m} e^{-2\sqrt{n}\max[\theta, r_i]}}{\delta e^{-2\sqrt{n}\max[\theta, r_1]}} = \frac{4}{\delta},$$

which means that all the hypothesis in the space with $\hat{\mathtt{R}} \neq 0$, if $m < \delta e^{2n\nu^2}/2$ are not taken into account, asymptotically, in estimating the upper bound of the hypothesis with the smaller error with DDWUB.

4 Discussion

In this work, for an arbitrary finite hypothesis space, we consider the hypothesis of minimal training error, we give a fully empirical new upper bound on the generalization error of this hypothesis and we show that our proposal is always tighter than the one based on the UB and, in the worst case, it degenerates in the UB. Our approach applies to finite hypotheses spaces and surely more sophisticated techniques, such as the Local Rademacher Complexity [3], can be employed and can sometimes result in tighter bounds. However, insight into finite classes remains quite useful [11]. Finite class analysis can be exploited for as a pedagogical tool. Finite class analysis can teach new directions in which to look for the development and evolution of more sophisticated bounds. Finite class analysis can be useful for model selection purposes (e.g. selecting the most suitable hypothesis space, or set of hyperparameters, or algorithm). Finite class analysis can be useful when the models are represented with limited number of bits because of the constants involved in the bounds. In the future, we will investigate how to generalize the approach, investigate more its performance, and exploit it to improve results in SLT which exploit the UB.

References

[1] V. N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[2] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer, 2001.

[3] L. Oneto. *Model Selection and Error Estimation in a Nutshell*. Springer, 2020.

[4] J. Langford. *Quantitatively tight sample complexity bounds*. Carnegie Mellon Thesis, 2002.

[5] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[6] O. Maron and A. W. Moore. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 11(1-5):193–225, 1997.

[7] K. Roeder and L. Wasserman. Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24(4):398, 2009.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

[8] O. Catoni. *PAC-Bayesian Supervised Classification*. Institute of Mathematical Statistics, 2007.
[9] W Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.
[10] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
[11] J. Langford and D. McAllester. Computable shell decomposition bounds. *Journal of Machine Learning Research*, 5(May):529–547, 2004.

A Proofs

Proof of Lemma 1. The proof is a direct consequence of Theorem 1. □

Proof of Theorem 2. Note that, under the assumptions of the lemma, $\forall r_1, \cdots, r_m \in [0,1]$, and $\forall i \in \mathcal{I} \setminus i^*$ and $\forall r'_k, r''_k \in [0,1]$ such that $r'_k < r''_k$

$$\sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}\setminus k} e^{-\gamma \max[\theta, r_j]} + e^{-\gamma \max[\theta, r'_k]}}{\delta e^{-\gamma \max[\theta, r_i]}}\right)}{2n}} - \sqrt{\frac{\log\left(\frac{2\sum_{j\in\mathcal{I}\setminus k} e^{-\gamma \max[\theta, r_j]} + e^{-\gamma \max[\theta, r''_k]}}{\delta e^{-\gamma \max[\theta, r_i]}}\right)}{2n}} \geq 0.$$

Moreover $0 < \min\left[1, \hat{\mathtt{R}}_i + \sqrt{\log\left(2\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]}/\delta e^{-\gamma \max[\theta, r_{i^*}]}\right)/2n}\right] \leq 1$ and if $\gamma \leq 2\sqrt{n}$ and $p_{i^*} = e^{-\gamma r_{i^*}}/\sum_{j\in\mathcal{I}} e^{-\gamma r_j}$ then

$$\partial \sqrt{\log\left(2\sum_{j\in\mathcal{I}} e^{-\gamma r_j}/\delta e^{-\gamma r_{i^*}}\right)/2n}/\partial r_{i^*} = \gamma p_{i^*}(1-p_{i^*})/4np_{i^*}\sqrt{\ln\left(\frac{2}{\delta p_{i^*}}\right)/2n} < \frac{(1-p_{i^*})}{2\sqrt{\ln\left(2/\delta p_{i^*}\right)/2}} < 1.$$

Consequently the statement of the lemma is proved. □

Proof of Theorem 3. Let us define $\vartheta = \min\left[1, \hat{\mathtt{R}}_{i^*} + \sqrt{\log(2m/\delta)/2n}\right]$. Let us suppose that $r_i \leq \vartheta$, $\forall i \in \mathcal{I} \setminus i^*$. If we set $r^*_{i^*} = \vartheta$, we have, thanks to the hypothesis of the theorem, that $e^{-\gamma \max[\theta, r_i]}/\sum_{j\in\mathcal{I}} e^{-\gamma \max[\theta, r_j]} = 1/m$, $\forall i \in \mathcal{I}$, then $r^*_{i^*}$ is a fixed point and since it is unique by Theorem 2 we have that $r^*_{i^*} \leq \vartheta$. Let us suppose now that $r_i \leq \vartheta$, $\forall i \in \mathcal{I} \setminus \{i^*, k\}$, $r_k > \vartheta$, $k \in \mathcal{I} \setminus i^*$. Thank to the hypothesis of the theorem we can state that

$$\frac{e^{-\gamma \max[\theta, r_{i^*}]}}{\sum_{j\in\mathcal{I}\setminus\{i^*,k\}} e^{-\gamma \max[\theta, r_j]} + e^{-\gamma \max[\theta, r_{i^*}]} + e^{-\gamma \max[\theta, r_k]}} > \frac{e^{-\gamma \max[\theta, r_{i^*}]}}{\sum_{j\in\mathcal{I}\setminus i^*} e^{-\gamma\theta} + e^{-\gamma \max[\theta, r_{i^*}]}},$$

and, consequently, we can also state that

$$r_{i^*} \leq \min\left[1, \hat{\mathtt{R}}_{i^*} + \sqrt{\log\left(2\sum_{j\in\mathcal{I}\setminus i^*} e^{-\gamma\theta} + e^{-\gamma \max[\theta, r_{i^*}]}/\delta e^{-\gamma \max[\theta, r_{i^*}]}\right)/2n}\right].$$

As a consequence, by exploiting the same reasoning exploited before, $r^*_{i^*} \leq \vartheta$. By induction and by noting that

$$\min\left[\mathtt{R}_1, \cdots, \mathtt{R}_m\right] \geq \max\left[0, \min\left[\hat{\mathtt{R}}_1, \cdots, \hat{\mathtt{R}}_m\right] - \sqrt{\log(2m/\delta)/2n}\right],$$

we can state that $\hat{\mathtt{R}}_{i^*} \leq \min\left[\mathtt{R}_1, \cdots, \mathtt{R}_m\right] + \sqrt{\log(2m/\delta)/2n}$, and consequently the statement of the theorem is proved. □