Risk bounds on statistical learning

Boris Ndjia Njike, Xavier Siebert
Université de Mons, Faculté polytechnique
Département de Mathématique et recherche opérationnelle
e-mail: borisedgar.NDJIANJIKE@umons.ac.be
e-mail: xavier.siebert@umons.ac.be

Abstract

The aim of this paper is to study theoretical risk bounds when using the Empirical Risk Minimization principle for pattern classification problems. We review some recent developments in statistical learning theory, in particular those involving minimal loss strategies. We conclude with a discussion of the practical implications of these results.

Keywords: statistical learning theory, VC dimension, risk bounds, minimax loss.

1 Motivations

The field of machine learning has considerably developed over the last years, from the use of support vector machines (SVM) and its derivatives to the widespread use of deep neural networks. A better theoretical understanding of learning algorithms is important for the understanding of current algorithms as well as for the development of new ones. In this paper we review some recent risk bounds in statistical learning theory, in particular those involving minimal loss strategies.

2 Formalization of the learning problem

In defined in [4], the problem of pattern classification consists in finding a function $\widehat{f}$ among set of hypothesis $\mathcal{H} = \{f : \mathcal{X} \to \{0, 1\}\}$ which minimizes the probability error

$$R(f) = P(Y \neq f(\mathbf{X})) = \int 1_{y \neq f(\mathbf{x})} dP(\mathbf{x}, y) \tag{1}$$

given a i.i.d sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ and an (unknown) probability $P(\mathbf{x}, \mathbf{y})$. Introducing the regression function $\eta(x) = P(Y = 1 \mid X = x)$, the Bayes classifier defined by $f^*(x) = 1_{\eta(x) \geq \frac{1}{2}}$ achieves the minimum risk (1) over all possible measurable functions $f : \mathcal{X} \to \{0, 1\}$, as shown in [2].

3 Risk bounds and strategies in statistical inference

The principle of empirical risk minimization (ERM) consists in replacing (1) by the empirical risk functional

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} 1_{\mathbf{y}_i \neq f(\mathbf{x}_i)}$$

and then approximating the function $f_{\mathcal{H}}$ by the function $\widehat{f}_n$, where $f_{\mathcal{H}}$ and $\widehat{f}_n$ are such that: $f_{\mathcal{H}} = \underset{f\in\mathcal{H}}{\operatorname{argmin}} R(f)$ and $\widehat{f}_n = \underset{f\in\mathcal{H}}{\operatorname{argmin}} R_n(f)$. Risk bounds on the error made by replacing the Bayes classifier f^* with $\widehat{f}_n$ are reviewed in [3] .

The ERM principle implicitely reflects a minimax loss strategy [4], with upper bounds relatively close to the lower bounds on the minimax loss. Indeed, using a loss function ℓ such that: $\ell(f, f^*) = R(f) - R(f^*)$, minimax strategy consists in finding the estimator $\widehat{f}$ that minimizes the supremum (over all P) of expected value of $\ell(\widehat{f}, f^*)$. Two cases can be considered, as detailed in [4] for a set of functions $\mathcal{H}$ with VC dimension V :
In the optimistic case (for set of probability $\mathcal{P}$ for which $R(f^*) = 0$, $\forall P$): for $n > V$,

$$\frac{V+1}{2e(n+1)} \leq \inf_{\widehat{f}} \sup_{P} \mathbb{E}_P(\ell(\widehat{f}, f^*)) \leq \sup_{P} \mathbb{E}_P(\ell(\widehat{f}_n, f^*)) \leq \frac{4}{n} \ln\left(\frac{2en}{V}\right)^V + \frac{16}{n}.$$

In the pessimistic case (for a set of probability $\mathcal{P}$, $\exists P$ such that $R(f^*) \neq 0$), for $n > 2V$:

$$\frac{V}{n}(1-erf(1)) \leq \inf_{\widehat{f}} \sup_{P} \mathbb{E}_P(\ell(\widehat{f}, f^*)) \leq \sup_{P} \mathbb{E}_P(\ell(\widehat{f}_n, f^*)) \leq 4\sqrt{\frac{V\left(\ln\frac{2n}{V}+1\right)+24}{n}} \tag{2}$$

Using geometric and combinatorial quantities related to the class $\mathcal{H}$, it is possible to refine (2) in several ways. First, as in [2], if $n > 2V$ there exists an absolute positive constant c such that:

$$\sqrt{\frac{V}{n}}(1-erf(1)) \leq \inf_{\widehat{f}} \sup_{P} \mathbb{E}_P(\ell(\widehat{f}, f^*)) \leq \sup_{P} \mathbb{E}_P(\ell(\widehat{f}_n, f^*)) \leq c\sqrt{\frac{V}{n}}$$

Second, as in [1], instead of taking P in some arbitrary set $\mathcal{P}$, one can introduce a parameter $h \in [0,1]$, such that $P \in \mathcal{P}(h)$ and $\mathcal{P}(h)$ is the set: $\{P \in \mathcal{P}, \mid 2\eta(x) - 1 \mid \geq h \text{ for all } x \in \mathcal{X}\}$. Then, if we assume that $\mathcal{H}$ has a finite VC-dimension $V \geq 2$, for some absolute positive constant k, if $n \geq V$, one has

$$\inf_{\widehat{f}} \sup_{P\in\mathcal{P}(h)} \mathbb{E}_P(\ell(\widehat{f}, f^*)) \geq k \min\left(\frac{V-1}{nh}, \sqrt{\frac{V-1}{n}}\right).$$

We will show that the latter bounds are in fact a particular case of [5] and discuss the practical implications of these results.

References

[1] P. Massart, E.Nédélec,The Annals of Statistics, Risk bounds for statistical learning, 2326-2366, 2006.

[2] G. Lugosi. Pattern classifcation and learning theory. In Principles of nonparametric learning, pages 1-56. Springer, 2002.

[3] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In Advanced lectures on machine learning, pages 169-207. Springer, 2004.

[4] V. N. Vapnik. Statistical Learning Theory. John Wiley & Sons, 1998.

[5] A. B. Tsybakov. Introduction to nonparametric estimation. Springer Series in Statistics.Springer, New York, 2009.