Machine Learning Theory
Tübingen University, WS 2016/2017
Lecture 3

Tolstikhin Ilya

Abstract

In this lecture we will prove the VC-bound, which provides a high-probability excess risk bound for the ERM algorithm when performing a binary classification over classes of finite VC dimension. This result generalizes the agnostic bound for finite classes, discussed in the previous lecture. Most of the material follows the exposition of Bousquet et al. (2004). I also invite the interested students to think about questions marked with blue. You won't get extra points for them, but you will certainly get a better understanding of the material.

1 Let's recall the setting and some basic facts

We have an input space $\mathcal{X}$ and an output space $\mathcal{Y} := \{0, 1\}$. There is an unknown probability distribution P over $\mathcal{X} \times \mathcal{Y}$. We receive a training sample $S_n := \{(X_i, Y_i)\}_{i=1}^n$ of n i.i.d. input-output pairs from P. We fix a set of classifiers $\mathcal{H}$. We denote the expected risk for any $h \in \mathcal{H}$ as

$$L(h) := \mathbb{P}_{(X,Y)\sim P}\{h(X) \neq Y\}$$

and the empirical risk as

$$L_n(h) := \frac{1}{n}\sum_{i=1}^{n} 1\{h(X_i) \neq Y_i\}.$$

We introduce the Empirical Risk Minimization (ERM) algorithm $\hat{h}_n := \hat{h}_n(S_n, \mathcal{H})$:

$$L_n(\hat{H}_n) = \inf_{g\in\mathcal{H}} L_n(g).$$

We will require the following concentration inequality, introduced in the second lecture:

Theorem 1 (Hoeffding's inequality)**.** *Let $\xi_1, \ldots, \xi_n$ be independent random variables such that $\xi_i \in [a_i, b_i]$, $a_i, b_i \in \mathbb{R}$, for $i = 1, \ldots, n$ with probability one. Denote $Z_n := \sum_{i=1}^n \xi_i$. Then for any $\varepsilon > 0$ it holds that:*

$$\mathbb{P}\{Z_n - \mathbb{E}[Z_n] \geq \varepsilon\} \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Same inequality holds for $\mathbb{P}\{\mathbb{E}[Z_n] - Z_n \geq \varepsilon\}$. Moreover,

$$\mathbb{P}\{|Z_n - \mathbb{E}[Z_n]| \geq \varepsilon\} \leq 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Show that the third inequality of theorem follows simply from the first two ones. The union bound is our favourite trick!

2 Agnostic bound for finite classes

Let's shortly recall the agnostic excess risk bound for finite classes, introduced in the second lecture. We will provide a slightly modified proof leading to minor changes in the constant factors:

Theorem 2. *Assume $\mathcal{H} = \{h_1, \ldots, h_N\}$. Then for any $\delta > 0$ with probability larger than $1 - \delta$ the following holds:*

$$L(\hat{h}_n) \leq \min_{i=1,\ldots,N} L(h_i) + 2\sqrt{\frac{\log N + \log \frac{2}{\delta}}{2n}}. \tag{Th1}$$

Proof. For our further discussion it will be useful to recall the idea behind a proof. Assuming h^* is the minimizer of the expected risk over $\mathcal{H}$ we may write:

$$\begin{aligned}
&L(\hat{h}_n) - L(h^*) \\
&= L(\hat{h}_n) - L_n(\hat{h}_n) + L_n(\hat{h}_n) - L_n(h^*) + L_n(h^*) - L(h^*) \\
&\leq L(\hat{h}_n) - L_n(\hat{h}_n) + L_n(h^*) - L(h^*) \qquad (*) \\
&\leq \sup_{h\in\mathcal{H}} \big(L(h) - L_n(h)\big) + \sup_{h\in\mathcal{H}} \big(L_n(h) - L(h)\big) \\
&\leq 2 \sup_{h\in\mathcal{H}} \big|L(h) - L_n(h)\big|. \qquad (1)
\end{aligned}$$

Next we write:

$$\begin{aligned}
\mathbb{P}\{\sup_{h\in\mathcal{H}} \big|L(h) - L_n(h)\big| \geq \epsilon\} &= \mathbb{P}\left\{\cup_{i=1}^N \big|L(h_i) - L_n(h_i)\big| \geq \epsilon\right\} \qquad (2) \\
&\leq \sum_{i=1}^N \mathbb{P}\left\{\big|L(h_i) - L_n(h_i)\big| \geq \epsilon\right\}, \qquad (3)
\end{aligned}$$

where we used the union bound in the last line. We may now apply Hoeffding's inequality of Theorem 1 and get:

$$\mathbb{P}\{\sup_{h\in\mathcal{H}} \big|L(h) - L_n(h)\big| \geq \epsilon\} \leq \sum_{i=1}^N 2e^{-\frac{2\epsilon^2}{n/n^2}} = 2Ne^{-2n\epsilon^2}.$$

We want the r.h.s. of the previous inequality to be smaller than δ. In other words, we want to find $\tilde{\epsilon}$ such that:

$$\delta = 2Ne^{-2n\tilde{\epsilon}^2}.$$

Solving the equation for $\tilde{\epsilon}$ we get:

$$\tilde{\epsilon} = \sqrt{\frac{\log(2N/\delta)}{2n}}.$$

Note that for this choice of ϵ we have

$$\mathbb{P}\{\sup_{h\in\mathcal{H}} \big|L(h) - L_n(h)\big| \geq \tilde{\epsilon}\} \leq \delta,$$

or equivalently

$$\mathbb{P}\{\sup_{h\in\mathcal{H}} \big|L(h) - L_n(h)\big| < \tilde{\epsilon}\} \geq 1 - \delta.$$

In other words, with probability larger than $1 - \delta$ we have

$$\sup_{h\in\mathcal{H}} \big|L(h) - L_n(h)\big| \leq \sqrt{\frac{\log(2N/\delta)}{2n}}.$$

Inserting this bound back to (1) we conclude the proof. □

Try to slightly improve this result. You may replace $2\sqrt{\frac{\log N+\log\frac{2}{\delta}}{2n}}$ in the upper bound with $\sqrt{\frac{\log N+\log\frac{2}{\delta}}{2n}}+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$. For this get back to (*) and do something smarter. Notice that h^* does not depend on S_n — so why upper bounding last two terms with supremum?

3 One step further: infinite classes $\mathcal{H}$, VC-bound

The main goal of this lecture is to drop the assumption of Theorem 2 that the class $\mathcal{H}$ is finite. Now we assume that $\mathcal{H}$ may be infinite. Actually, there can be *uncountably many* classifiers in $\mathcal{H}$ (just think about linear classifiers in $\mathbb{R}^d$ or simply about thresholds in one dimension).

3.1 Spoiler

Before even introducing all the necessary definitions, let us start with the statement of theorem, which we are going to prove.

Theorem 3 (VC-bound)**.** *For any δ with probability larger than $1-\delta$ it holds that:*

$$L(\hat{h}_n) \leq \inf_{g\in\mathcal{H}} L(g) + 2\sqrt{2}\sqrt{\frac{\log S_{\mathcal{H}}(2n)+\log\frac{4}{\delta}}{n}}.$$

Compare this bound to (Th1). It looks almost the same, but N is replaced with $S_{\mathcal{H}}(2n)$— a quantity known as the growth function, which will be introduced later in the proof. For now it is instructive to note the similarity between these two results: perhaps, it means that we can proceed with the same (or almost the same) proof, where, magically, N events appearing on lines (2)–(3) will be eventually replaced with $S_{\mathcal{H}}(2n)$ events?

It turns out that this is indeed the case! In the following we present the proof of Theorem 7.

3.2 Debugging the proof

Can we still repeat the proof of Theorem 2? Let's assume for now that there is $h^* \in \mathcal{H}$ such that $L(h^*) = \inf_{g\in\mathcal{H}} L(g)$ (Show that generally this is not true). It turns out that we can still repeat the first steps, but we can no more apply the union bound. Indeed, the union bound $\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i)$ holds at most for countable set of events A_i. In our case, as we already mentioned, we may end up with uncountably many events. In summary, we can not apply step (2)–(3) any more.

Let's try to find a workaround. What is actually causing the problem? Note that $L_n(h)$ in lines between (1) and (2) still takes only finitely many values as h runs through the $\mathcal{H}$ (prove this yourself!). If we had only $L_n(h)$ appearing inside of probability sign in (2) we could still enumerate all the different values of $L_n(h)$ and get back to finitely many events and proceed with all the previous steps. The real problem is the $L(h)$ term, which also appears in the events of (2). In principle, $L(h)$ can take any value between 0 and 1 for $h \in \mathcal{H}$ (prove this yourself!). This is the reason we may end up with uncountably many events.

Fortunately, the following nontrivial inequality helps us to get rid of the adversarial $L(h)$ term:

Lemma 4 (Symmetrization inequality)**.** *Assume $S'_n := \{(X'_i, Y'_i))\}_{i=1}^n$ is an independent "copy" of S_n, that is $S_n \cup S'_n$ forms a sequence of $2n$ i.i.d. input-output pairs distributed according to P. Denote $L'_n(h) := \frac{1}{n}\sum_{i=1}^n 1\{h(X'_i) \neq Y'_i\}$. Then for any $\epsilon > 0$, such that $n\epsilon^2 \geq 2$, it holds that:*

$$\mathbb{P}_{S_n}\left\{\sup_{h\in\mathcal{H}}\big(L(h)-L_n(h)\big) \geq \epsilon\right\} \leq 2\cdot\mathbb{P}_{S_n\cup S'_n}\left\{\sup_{h\in\mathcal{H}}\big(L'_n(h)-L_n(h)\big) \geq \epsilon/2\right\}.$$

Inequality also holds for $\sup_{h\in\mathcal{H}}\big(L_n(h)-L(h)\big)$.

3.3 Modifying the proof: getting rid of $L(h)$

Now, let us return to the beginning and try to apply this result:

$$\begin{aligned}
&L(\hat{h}_n)-L(h^*)\\
&=L(\hat{h}_n)-L_n(\hat{h}_n)+L_n(\hat{h}_n)-L_n(h^*)+L_n(h^*)-L(h^*)\\
&\le L(\hat{h}_n)-L_n(\hat{h}_n)+L_n(h^*)-L(h^*)\\
&\le \sup_{h\in\mathcal{H}}\big(L(h)-L_n(h)\big)+\sup_{h\in\mathcal{H}}\big(L_n(h)-L(h)\big).
\end{aligned}$$

As we already now, if for two events A and B it holds that $A\subseteq B$ then necessarily $\mathbb{P}(A)\le\mathbb{P}(B)$. This gives us

$$\mathbb{P}\{L(\hat{h}_n)-L(h^*)\ge\epsilon\}\le\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\big(L(h)-L_n(h)\big)+\sup_{h\in\mathcal{H}}\big(L_n(h)-L(h)\big)\ge\epsilon\right\}. \tag{4}$$

Also note that by the same reason for any random variables a and b we have

$$\mathbb{P}\{a+b\ge\epsilon\}\le\mathbb{P}\{a\ge\epsilon/2\cup b\ge\epsilon/2\}\le\mathbb{P}\{a\ge\epsilon/2\}+\mathbb{P}\{b\ge\epsilon/2\}$$

Applying this to (4) and using Lemma 4 we get:

$$\begin{aligned}
&\mathbb{P}\{L(\hat{h}_n)-L(h^*)\ge\epsilon\}\\
&\le\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\big(L(h)-L_n(h)\big)\ge\epsilon/2\right\}+\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\big(L_n(h)-L(h)\big)\ge\epsilon/2\right\}\\
&\le 4\cdot\mathbb{P}_{S_n\cup S'_n}\left\{\sup_{h\in\mathcal{H}}\big(L'_n(h)-L_n(h)\big)\ge\epsilon/2\right\}.
\end{aligned} \tag{5}$$

At this point note that no matter what h is, $L'_n(h)-L_n(h)$ can take only finitely many values (prove this yourself!). The value of $L'_n(h)-L_n(h)$ depends only on the *projection* of $\mathcal{H}$ on the double sample $S_n\cup S'_n$, where for any sample $S_m:=\{(X_j,Y_j)\}_{j=1}^m$ we define a projection in the following way:

$$\mathcal{H}_{S_m}:=\Big\{\big(1\{h(X_1)\ne Y_1,h(X_2)\ne Y_2,\dots,h(X_m)\ne Y_m\big),h\in\mathcal{H}\Big\}\subseteq\{0,1\}^m.$$

Note that $\mathcal{H}_{S_n\cup S'_n}$ is a subset of the $\{0,1\}^{2n}$ and thus its cardinality $\mathrm{card}(\mathcal{H}_{S_n\cup S'_n})$ is upper bounded by 2^{2n}. We may write

$$\mathbb{P}\{L(\hat{h}_n)-L(h^*)\ge\epsilon\}\le 4\cdot\mathbb{P}_{S_n\cup S'_n}\left\{\sup_{v\in\mathcal{H}_{S_n\cup S'_n}}\big(L'_n(v)-L_n(v)\big)\ge\epsilon/2\right\},$$

where we have overloaded notations $L'_n(v)$ and $L_n(v)$ in a natural way. All in all, it seems like we may now proceed with the original (2)–(3) steps to bound the r.h.s. of the previous inequality, since sup is now over the finite set. This is indeed what we did during the lecture, but the thing is, this step is not quite correct. Notice that the union bound assumes that events A_i are *fixed*. In our case, there are finitely many events $A_v:=\{L'_n(v)-L_n(v)\ge\epsilon\}$ indexed by v, but they all depend on the random samples S_n and S'_n, so the union bound (at least in its usual form) can not be applied.

3.4 Another neat trick: Rademacher symmetrization

Instead, we will proceed with a trick commonly known as the *Rademacher symmetrization*. Next lines are taken from Section 12.4 of Devroye et al. (1996).

Introduce random variables $\sigma_1, \ldots, \sigma_n$ which are all independent (also independent from S_n and S'_n) and take values -1 and $+1$ with probabilities 0.5. Rewrite (5) in the following way:

$$\mathbb{P}\{L(\hat{h}_n) - L(h^*) \geq \epsilon\} \leq 4 \cdot \mathbb{P}_{S_n \cup S'_n} \left\{ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \big(1\{h(X'_i) \neq Y'_i\} - 1\{h(X_i) \neq Y_i\} \big) \geq \epsilon/2 \right\}$$

and notice that distribution of

$$\frac{1}{n} \sum_{i=1}^{n} \big(1\{h(X'_i) \neq Y'_i\} - 1\{h(X_i) \neq Y_i\} \big)$$

is the same as distribution of

$$\frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(1\{h(X'_i) \neq Y'_i\} - 1\{h(X_i) \neq Y_i\} \big)$$

(proof this yourself!). We may thus write

$$\begin{aligned}
\mathbb{P}\{L(\hat{h}_n) - L(h^*) \geq \epsilon\} &\leq 4 \cdot \mathbb{P}_{S_n \cup S'_n} \left\{ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \big(1\{h(X'_i) \neq Y'_i\} - 1\{h(X_i) \neq Y_i\} \big) \geq \epsilon/2 \right\} \\
&= 4 \cdot \mathbb{P}_{\sigma, S_n \cup S'_n} \left\{ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(1\{h(X'_i) \neq Y'_i\} - 1\{h(X_i) \neq Y_i\} \big) \geq \epsilon/2 \right\}.
\end{aligned}$$

Next we use the tower rule of expectation, which can be written for any event A and any random variable Z as $\mathbb{P}(A) = \mathbb{E}_Z[\mathbb{P}(A|Z)]$. This gives us

$$\mathbb{P}\{L(\hat{h}_n) - L(h^*) \geq \epsilon\} \leq 4\mathbb{E}_{S_n \cup S'_n} \left[\mathbb{P}_\sigma \left\{ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(1\{h(X'_i) \neq Y'_i\} - 1\{h(X_i) \neq Y_i\} \big) \geq \frac{\epsilon}{2} \middle| S_n \cup S'_n \right\} \right].$$

It is left to bound the conditional probability appearing inside of expected value. Using our definition of the projection we may rewrite

$$\begin{aligned}
&\mathbb{P}_\sigma \left\{ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(1\{h(X'_i) \neq Y'_i\} - 1\{h(X_i) \neq Y_i\} \big) \geq \frac{\epsilon}{2} \middle| S_n \cup S'_n \right\} \\
&= \mathbb{P}_\sigma \left\{ \sup_{v \in \mathcal{H}_{S_n \cup S'_n}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(v'_i - v_i \big) \geq \frac{\epsilon}{2} \middle| S_n \cup S'_n \right\},
\end{aligned}$$

where we once again (perhaps confusingly) used v_i and v'_i to denote indicators $1\{h_v(X_i) \neq Y_i\}$ and $1\{h_v(X'_i) \neq Y'_i\}$, where $h_v \in \mathcal{H}$ is any classifier with projection equal to v. Notice that, because we conditioned on S_n and S'_n, these sets are now "fixed", and thus the projection $\mathcal{H}_{S_n \cup S'_n}$ is now not random any more, but instead just some fixed subset of $\{0,1\}^{2n}$. We may now safely use our initial (2)–(3) trick (union bound) and write

$$\begin{aligned}
&\mathbb{P}_\sigma \left\{ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(1\{h(X'_i) \neq Y'_i\} - 1\{h(X_i) \neq Y_i\} \big) \geq \frac{\epsilon}{2} \middle| S_n \cup S'_n \right\} \\
&\leq \sum_{v \in \mathcal{H}_{S_n \cup S'_n}} \mathbb{P}_\sigma \left\{ \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(v'_i - v_i \big) \geq \frac{\epsilon}{2} \middle| S_n \cup S'_n \right\}.
\end{aligned}$$

Individual probabilities may be again bounded using Hoeffding's inequality (prove it yourself!):

$$\mathbb{P}_\sigma \left\{ \frac{1}{n} \sum_{i=1}^{n} \sigma_i (v_i' - v_i) \geq \frac{\epsilon}{2} \middle| S_n \cup S_n' \right\} \leq e^{-\frac{2\epsilon^2/4}{n4/n^2}} = e^{-n\epsilon^2/8}.$$

3.5 VC combinatorics

Putting all the bits together we finally get:

$$\mathbb{P}\{L(\hat{h}_n) - L(h^*) \geq \epsilon\} \leq 4e^{-n\epsilon^2/8} \mathbb{E}_{S_n \cup S_n'} \left[\text{card}(\mathcal{H}_{S_n \cup S_n'}) \right].$$

Again, making the upper bound equal to δ and solving for ϵ we get that for any $\delta > 0$ with probability larger than $1 - \delta$ it holds that:

$$L(\hat{h}_n) \leq \inf_{g \in \mathcal{H}} L(g) + 2\sqrt{2} \sqrt{\frac{\log E_{\mathcal{H}}(2n) + \log \frac{4}{\delta}}{n}},$$

where we denoted

$$E_{\mathcal{H}}(n) := \mathbb{E}_{S_n} \left[\text{card}(\mathcal{H}_{S_n}) \right].$$

The quantity $E_{\mathcal{H}}(n)$ is known as the *VC entropy*. Obviously, the VC entropy can be upper bounded in the following (perhaps, extremely crude) way:

$$E_{\mathcal{H}}(n) \leq S_{\mathcal{H}}(n) := \sup_{S:\ \text{card}(S)=n} \text{card}(\mathcal{H}_S).$$

All we did is replaced the average (expectation) with the maximum value. The quantity $S_{\mathcal{H}}(n)$ is commonly known as the *growth function*. We showed that with probability larger than $1 - \delta$ it also holds that:

$$L(\hat{h}_n) \leq \inf_{g \in \mathcal{H}} L(g) + 2\sqrt{2} \sqrt{\frac{\log S_{\mathcal{H}}(2n) + \log \frac{4}{\delta}}{n}}.$$

This concludes the proof of Theorem 7.

But are we satisfied with this result? The good thing about Theorem 2 is that as the sample size n grows to infinity the last term on the r.h.s. of (Th1) decreases to zero, showing that the performance of ERM achieves the best possible one. Does Theorem 7 have the same behaviour?

Of course, the answer depends on the growth function $S_{\mathcal{H}}(2n)$, which is defined *purely* by the geometry of $\mathcal{H}$. As we already mentioned, the trivial upper bound gives $S_{\mathcal{H}}(2n) \leq 2^{2n}$. However, if we insert it in the VC-bound we end up with $2\sqrt{2} \cdot \sqrt{2}$, which does not tend to zero. An important question is: how should $\mathcal{H}$ look like so that $\log S_{\mathcal{H}}(2n)/n \to 0$ as $n \to \infty$?

The answer to this question is hidden in the following definition:

Definition 5 (VC dimesnion)**.** *The VC dimension of the class $\mathcal{H}$ is the largest n such that*

$$S_{\mathcal{H}}(n) = 2^n.$$

If there is no such an n we say that $\mathcal{H}$ has infinite VC dimension.

The following fact[1] establishes the polynomial growth of $S_{\mathcal{H}}(n)$ for classes $\mathcal{H}$ of finite VC dimension:

[1] There is a curious history behind this lemma. It was (apparently) simultaneously proved by several groups around late 60s – early 70th, including Vapnik and Chervonenkis, Sauer, and Shelah and Perles. A wonderful overview of this fact can be found in Leon Bottou's slides available online here: `http://leon.bottou.org/_media/papers/vapnik-symposium-2011.pdf`.

Lemma 6 (Vapnik, Chervonenkis, Sauer, Shelah)**.** *Let $\mathcal{H}$ be a class of VC dimension $d < \infty$. Then for all n it holds that*

$$S_{\mathcal{H}}(n) \leq \sum_{i=1}^{d} \binom{n}{i},$$

and for all $n \geq d$ it holds that:

$$S_{\mathcal{H}}(n) \leq \left(\frac{e \cdot n}{d}\right)^{d}.$$

We may finally state the following bound, which behaves exactly like the one of original Theorem 2:

Theorem 7 (VC-bound)**.** *Assume $\mathcal{H}$ has a VC dimension $d < \infty$. For any δ with probability larger than $1 - \delta$ it holds that:*

$$L(\hat{h}_n) \leq \inf_{g \in \mathcal{H}} L(g) + 2\sqrt{2}\sqrt{\frac{d \log \frac{2en}{d} + \log \frac{4}{\delta}}{n}}.$$

References

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, 2004. URL `http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/pdfs/pdf2819.pdf`.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.