Machine Learning Theory
Tübingen University, WS 2016/2017
Lecture 10

Tolstikhin Ilya

Abstract

We will start with several properties of Rademacher Complexity. First we prove the bound on RC for finite function classes, the so-called Massart's lemma. This will allow us to show that VC-bound simply follows from the bound based on Rademacher Complexity. Next, we will show that RC satisfies the *contraction principle*. Also, we will compute RC for convex hulls of functions classes. These two results will allow us to argue regarding AdaBoost algorithm. Blue colour will be used to highlight parts appearing in the upcoming homework assignments.

1 Finite function classes

In the last lecture we saw that Rademacher Complexity (RC) plays an important role in bounding the uniform deviations of expected losses from their empirical counterparts over the classes of predictors:

Theorem 1. *Consider any class of real-valued predictors $\mathcal{H}$ and any loss function ℓ, such that $\ell\big(h(X), Y\big) \in [0,1]$ for all $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, and $h \in \mathcal{H}$. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$ over the training sample S_n we have*

$$\sup_{h \in \mathcal{H}} \big(L(h) - L_n(h)\big) \leq 2 \cdot R_n(\ell \circ \mathcal{H}) + 2\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Just to briefly remind of what RC is, let's write its expression once again:

$$\hat{R}_n(\mathcal{F}) := \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_n f(Z_i) \Big| Z_1, \ldots, Z_n \right], \quad R_n(\mathcal{F}) := \mathbb{E}_{Z_1,\ldots,Z_n \sim P}[\hat{R}_n(\mathcal{F})].$$

In this section we will see how to bound RC in the case when one is working with *finitely many* predictors. The following lemma is commonly known as *Massart's lemma*:

Lemma 2. *Assume $\mathcal{F} = \{f_1, \ldots, f_N\}$ and points $Z_1, \ldots, Z_n \in \mathcal{Z}$ are fixed. Then*

$$\hat{R}_n(\mathcal{F}) \leq \frac{\sqrt{2\log N}}{n} \max_{j=1,\ldots,N} \sqrt{\sum_{i=1}^n f_j^2(Z_i)}\,.$$

In particular, if $f(Z) \in [0,1]$ for all $f \in \mathcal{F}$ and $Z \in \mathcal{Z}$ then

$$\hat{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log N}{n}}.$$

VC bound can be proved using RC To see the power of this result, let's go back to the setting of a binary classification with the binary loss $\ell(y, y') = 1\{y \neq y'\}$ and general (infinite) class of predictors $\mathcal{H}$ of finite VC dimension $\mathrm{VC}(\mathcal{H}) < \infty$. Notice that in this case the empirical RC

$$\hat{R}_n(\ell \circ \mathcal{H}) = \mathbb{E}_\epsilon \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_n \ell\big(h(X_i), Y_i\big) \Big| (X_1, Y_1), \ldots, (X_n, Y_n) \right]$$

depends only on the error vectors induced by the elements of $\mathcal{H}$ on the training set $\{(X_i, Y_i)\}_{i=1}^n$. In other words, the supremum in the above expression can be replaced by the maximum over the subset of $\{0,1\}^n$. Obviously, the cardinality of any such subset is at most 2^n, but that is of course a very loose upper bound. In particular, combining this bound together with Lemma 2 will give an upper bound $\hat{R}_n(\ell \circ \mathcal{H}) \leq 2$, which unfortunately does not decrease to zero with the growing sample size. However, in one of the previous lectures we saw that the cardinality of the binary error vectors induced by a VC class can be bounded using Sauer's lemma, which shows that the logarithm of a maximal possible cardinality is upper bounded by $\mathrm{VC}(\mathcal{H}) \cdot \log(n+1)$. Combining this, together with Lemma 2 and Theorem 1 we recover the VC-bound for binary classification with classes of finite VC dimension.

2 Partially explaining AdaBoost

We will now present two additional results regarding RC. The first one is called *the contraction principle* and allows to bound expressions like $\hat{R}_n(\varphi \circ \mathcal{F})$ in terms of $\hat{R}_n(\mathcal{F})$ for any *Lipschitz* functions φ. We will omit the proof, which is rather technical and long.

Lemma 3. *Assume $\varphi \colon \mathbb{R} \to \mathbb{R}$ is L_φ-Lipschitz, i.e. $|\varphi(x) - \varphi(y)| \leq L_\varphi \cdot |x - y|$ for any $x, y \in \mathbb{R}$, where $L_\varphi > 0$ is a constant. Moreover, assume $\varphi(0) = 0$. Then for any input space $\mathcal{Z}$, any class $\mathcal{F}$ of functions defined over $\mathcal{Z}$ and any given points $Z_1, \ldots, Z_n \in \mathcal{Z}$ it holds that*

$$\hat{R}_n(\varphi \circ \mathcal{F}) := \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_n \varphi\big(f(Z_i)\big) \Big| Z_1, \ldots, Z_n \right] \leq L_\varphi \hat{R}_n(\mathcal{F}).$$

Next, consider a function class $\mathcal{F}$ and introduce the following set of functions, which we call the convex hull:

$$conv(\mathcal{F}) := \left\{ f \colon f(z) = \sum_{i=1}^{N} \alpha_i f_i(z), \quad N \in \mathbb{N}, f_i \in \mathcal{F}, \alpha_i \geq 0, \sum_i \alpha_i = 1 \right\}.$$

The following result shows that, surprisingly, the RC of a function class $conv(\mathcal{F})$ (which is in general much bigger than $\mathcal{F}$) coincides with the RC of $\mathcal{F}$:

Lemma 4.

$$\hat{R}_n\big(conv(\mathcal{F})\big) = \hat{R}_n(\mathcal{F}).$$

Proof.

$$\begin{aligned}
\hat{R}_n\big(conv(\mathcal{F})\big) &= \mathbb{E}_\epsilon \left[\sup_{f \in conv(\mathcal{F})} \frac{1}{n} \sum_{i=1}^{n} \epsilon_n f(Z_i) \right] \\
&= \mathbb{E}_\epsilon \left[\sup_{\substack{N \\ f_1, \ldots, f_N \\ \alpha_1, \ldots, \alpha_N}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_n \sum_{j=1}^{N} \alpha_j f_j(Z_i) \right] = \mathbb{E}_\epsilon \left[\sup_{\substack{N \\ f_1, \ldots, f_N \\ \alpha_1, \ldots, \alpha_N}} \frac{1}{n} \sum_{j=1}^{N} \alpha_j \underbrace{\left(\sum_{i=1}^{n} \epsilon_n f_j(Z_i) \right)}_{c_j} \right].
\end{aligned}$$

Notice that for any $c_1, \ldots, c_N$

$$\max_{\alpha_1,\ldots,\alpha_N} \sum_{i=1}^{N} \alpha_i c_i = \max_{i=1,\ldots,N} c_i.$$

This shows that

$$\hat{R}_n\big(conv(\mathcal{F})\big) = \mathbb{E}_\epsilon \left[\sup_{\substack{N \\ f_1,\ldots,f_N \\ \alpha_1,\ldots,\alpha_N}} \frac{1}{n} \left(\sum_{i=1}^{n} \epsilon_n f_{j^*}(Z_i) \right) \right],$$

where

$$\max_{j=1,\ldots,N} \left(\sum_{i=1}^{n} \epsilon_n f_j(Z_i) \right) = \sum_{i=1}^{n} \epsilon_n f_{j^*}(Z_i).$$

It is now obvious that

$$\hat{R}_n\big(conv(\mathcal{F})\big) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_n f(Z_i) \right] = \hat{R}_n(\mathcal{F}).$$

□

Together these two facts lead to the following curious result:

Theorem 5. *Assume $\mathcal{Y} = \{-1, +1\}$ and consider any set $\mathcal{H}$ of classifiers defined over any input space $\mathcal{X}$. Assume $\mathrm{VC}(\mathcal{H}) < \infty$ and consider a binary loss function $\ell(y, y') = 1\{y \neq y'\}$. Let P be any unknown probability distribution over $\mathcal{X} \times \mathcal{Y}$ and $S_n := \{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. training set. Then for any $\delta \in (0, 1)$ and any $\gamma > 0$ with probability at least $1 - \delta$, for any classifier $\tilde{h}$ of the form $\tilde{h}(x) = \mathrm{sgn} f(x)$, where $f \in conv(\mathcal{H})$, it holds that*

$$L(\tilde{h}) \leq \frac{1}{n} \sum_{i=1}^{n} 1\{f(X_i)Y_i \leq \gamma\} + \frac{2}{\gamma} \sqrt{2 \frac{\mathrm{VC}(\mathcal{H}) \log(n+1)}{n}} + 2 \sqrt{\frac{2 \log(1/\delta)}{n}}.$$

Before we prove this result let's briefly discuss its implications for AdaBoost. We will use AdaBoost together with the base class $\mathcal{H}$ of binary predictors, which has a finite VC-dimension. For simplicity, assume that after completing AdaBoost we will normalize all the weights of the classifiers, i.e. our resulting classifier will be of the form $\tilde{h} = \mathrm{sgn} f_T(X)$, where

$$f_T(X) := \sum_{i=1}^{T} \frac{\alpha_i h_i(X)}{\sum_{j=1}^{T} \alpha_j}, \quad h_j \in \mathcal{H}.$$

This normalization does not change the outputs of the classifier, produced by AdaBoost, as, clearly,

$$\mathrm{sgn} \sum_{i=1}^{T} \frac{\alpha_i h_i(X)}{\sum_{j=1}^{T} \alpha_j} = \mathrm{sgn} \sum_{i=1}^{T} \alpha_i h_i(X).$$

Also, assume that on every t-th step of AdaBoost the week learner manages to find a hypotheses $h_t \in \mathcal{H}$, such that $\epsilon_t \leq 1/2$, where ϵ_t is the weighted empirical binary error of h_t on the t-th round (look into the lecture on AdaBoost for details). This ensures that $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t} \geq 0$. Notice that, no matter how many steps T of AdaBoost we perform, the resulting function f_T belongs to

conv($\mathcal{H}$). This makes it possible to apply Theorem 5 for $\tilde{h}$, no matter what T did we choose, and get that with probability at least $1-\delta$

$$L(\tilde{h}) \leq \frac{1}{n}\sum_{i=1}^{n} 1\{f_T(X_i)Y_i \leq \gamma\} + \frac{2}{\gamma}\sqrt{2\frac{\mathrm{VC}(\mathcal{H})\log(n+1)}{n}} + 2\sqrt{\frac{2\log(1/\delta)}{n}}.$$

The quantity $f_T(X_i)Y_i$ is commonly known as *the margin* of a classifier f_T on the example X_i. As can be seen, this value is negative when we make a mistake and it is positive otherwise. The larger the absolute value of $f_T(X_i)Y_i$ is, the more f_T is confident in making its prediction. In the last upper bound we see that the first term punishes the points where f_T is either making a mistake (negative margin) or where f_T is classifying correctly but with margin smaller than γ (not confidently enough). Taking $\gamma \to 0$ would correspond to the first term approaching the standard binary empirical loss $L_n(\tilde{h})$. However, in this case the second term goes to infinity. On the other hand, increasing γ we decrease the second term and increase the first one. Note also that $f_T(X_i)Y_i \in [-1,1]$, because of the fact that we normalized the weights of our composition, making them sum to one. Thus we should choose $\gamma \in (0,1)$.

There is an interesting aspect of AdaBoost. Empirically it was noticed that, when AdaBoost performs more and more steps (i.e. T increases), the margins $\{Y_i f_T(X_i)\}_{i=1}^n$ of f_T on the training set overall tend to increase. Based on this, we may set γ a bit larger than 0 (say, $\gamma = 0.1$), and hope that the first term in the upper bound will eventually decrease to zero as we make more and more steps T. In the same time, the second term will stay the same, as it does not depend on T. This discussion is rather rough and contains some unclear moments, but overall it gives an intuition of why does AdaBoost perform so well.

Notice that this would not be so clear if we applied a simple VC-bound for AdaBoost, which would give an upper bound of the form $L_n(\tilde{h}) + \sqrt{\frac{T\cdot\mathrm{VC}(\mathcal{H})+\log(1/\delta)}{n}}$, where the second term increases with T.

2.1 Proof

Proof. (of Theorem 5) First of all, note that

$$L(\tilde{h}) = \mathbb{P}\{\tilde{h}(X) \neq Y\} = \mathbb{P}\{\mathrm{sgn} f(X) \neq Y\} = \mathbb{P}\{f(X)Y \leq 0\} = \mathbb{E}_{(X,Y)\sim P} 1\{f(X)Y \leq 0\}.$$

Now, take any L_φ function $\varphi\colon \mathbb{R} \to [0,1]$, such that $\varphi(z) \geq 1\{z \leq 0\}$ for any $z \in \mathbb{R}$ (we will specify the choice of this function later in the proof). In this case, obviously

$$L(\tilde{h}) = \mathbb{E}_{(X,Y)\sim P} 1\{f(X)Y \leq 0\} \leq \mathbb{E}_{(X,Y)\sim P}\varphi\big(f(X)Y\big).$$

Note that we can use Theorem 1 to upper bound $\mathbb{E}_{(X,Y)\sim P}\varphi\big(f(X)Y\big)$. Indeed, we just need to set $\ell(h(X),Y) = \varphi(h(X)\cdot Y)$ and notice that this quantity belongs to $[0,1]$ by construction of φ. Thus, with probability larger than $1-\delta$ we get

$$\mathbb{E}_{(X,Y)\sim P}\varphi\big(f(X)Y\big) \leq \frac{1}{n}\sum_{i=1}^{n}\varphi\big(f(X_i)Y_i\big) + 2\mathbb{E}_{S_n}\mathbb{E}_\epsilon\left[\sup_{f\in \mathit{conv}(\mathcal{H})}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\varphi\big(f(X_i)Y_i\big)\right] + 2\sqrt{\frac{2\log(1/\delta)}{n}}.$$

□

Note that for any $\mathcal{F}$ and any constant C we have

$$\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\big(\varphi\big(f(Z_i)\big) - C\big)\right] = \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\varphi\big(f(Z_i)\big)\right] - \underbrace{\mathbb{E}_\epsilon\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i C\right]}_{=0} = \hat{R}_n(\varphi\circ\mathcal{F}).$$

Together with Lemma 3 (we use the last identity with $C = \varphi(0)$) this implies with probability at least $1 - \delta$

$$L(\tilde{h}) \leq \frac{1}{n}\sum_{i=1}^{n} \varphi\big(f(X_i)Y_i\big) + 2L_\varphi \mathbb{E}_{S_n}\mathbb{E}_\epsilon \left[\sup_{f\in conv(\mathcal{H})} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)Y_i\right] + 2\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Notice that ϵ_i and $\epsilon_i Y_i$ have the same distribution. Indeed,

$$\begin{aligned}
\mathbb{P}\{\epsilon_i Y_i = 1\} &= \mathbb{P}\{Y_i = 1|\epsilon_i = 1\}\mathbb{P}\{\epsilon_i = 1\} + \mathbb{P}\{Y_i = -1|\epsilon_i = -1\}\mathbb{P}\{\epsilon_i = -1\} \\
&= \mathbb{P}\{Y_i = 1\}\mathbb{P}\{\epsilon_i = 1\} + \mathbb{P}\{Y_i = -1\}\mathbb{P}\{\epsilon_i = -1\} \\
&= \frac{1}{2}\left(\mathbb{P}\{Y_i = 1\} + \mathbb{P}\{Y_i = -1\}\right) = \frac{1}{2},
\end{aligned}$$

where we used the fact that ϵ_i is independent of Y_i and the definition of conditional probability. This allows us to conclude that with probability at least $1 - \delta$

$$\begin{aligned}
L(\tilde{h}) &\leq \frac{1}{n}\sum_{i=1}^{n} \varphi\big(f(X_i)Y_i\big) + 2L_\varphi \mathbb{E}_{S_n}\mathbb{E}_\epsilon \left[\sup_{f\in conv(\mathcal{H})} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)Y_i\right] + 2\sqrt{\frac{2\log(1/\delta)}{n}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \varphi\big(f(X_i)Y_i\big) + 2L_\varphi \mathbb{E}_{S_n}\mathbb{E}_\epsilon \left[\sup_{f\in conv(\mathcal{H})} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right] + 2\sqrt{\frac{2\log(1/\delta)}{n}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \varphi\big(f(X_i)Y_i\big) + 2L_\varphi \mathbb{E}_{S_n}\mathbb{E}_\epsilon \left[\sup_{h\in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(X_i)\right] + 2\sqrt{\frac{2\log(1/\delta)}{n}},
\end{aligned}$$

where we used Lemma 4. Finally, we apply the results developed in the first section of this lecture. For this we notice that the cardinalities of the following two sets:

$$\Big\{\big(h(X_1),\dots,h(X_n)\big),\quad h\in\mathcal{H}\Big\},\quad \Big\{\Big(1\big\{h(X_1)\neq Y_1\big\},\dots,1\big\{h(X_n)\neq Y_n\big\}\Big),\quad h\in\mathcal{H}\Big\}$$

are equal. Thus we deduce (with the help of Massart's lemma) that

$$\mathbb{E}_\epsilon\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(X_i)\right] = \hat{R}_n(\mathcal{H}) \leq \sqrt{\frac{2\mathrm{VC}(\mathcal{H})\log(n+1)}{n}}.$$

It is now left to properly choose φ. We will set

$$\varphi(z) := \begin{cases} 1, & z \leq 0; \\ 0, & z \geq \gamma; \\ 1 - z/\gamma, & z \in (0,\gamma). \end{cases}$$

It is easy to check that indeed $\varphi(z) \in [0,1]$ for all $z \in \mathbb{R}$, that it is $(1/\gamma)$-Lipschitz, and that $\varphi(z) \geq 1\{z \leq 0\}$. Overall, we get with probability at least $1 - \delta$

$$L(\tilde{h}) \leq \frac{1}{n}\sum_{i=1}^{n} \varphi\big(f(X_i)Y_i\big) + \frac{2}{\gamma}\sqrt{\frac{2\mathrm{VC}(\mathcal{H})\log(n+1)}{n}} + 2\sqrt{\frac{2\log(1/\delta)}{n}}.$$

To complete the proof (for which we need to deal with the first term of the last expression) we simply notice that

$$\varphi(z) \leq 1\{z \leq \gamma\}.$$