Machine Learning Theory
Tübingen University, WS 2016/2017
Lecture 12

Tolstikhin Ilya

Abstract

In this lecture we derive risk bounds for kernel methods. We will start by showing that Soft Margin kernel SVM corresponds to minimizing empirical *hinge loss* over the ball in RKHS. Combining this observation with the Rademacher analysis, we derive risk bounds for various kernel methods by bounding Rademacher Complexity of the ball in RKHS.

1 Balls of RKHS

Recall the soft-margin SVM. Our input space $\mathcal{X}$ is $\mathbb{R}^d$ and we are solving a binary classification problem with $\mathcal{Y} = \{-1, +1\}$. We saw that this algorithm corresponds to solving

$$\min_{\alpha_1,\dots,\alpha_n \in \mathbb{R}} \frac{1}{2} \left\| \sum_{i=1}^n \alpha_i X_i \right\|^2_{\mathbb{R}^d} + \frac{C}{n} \sum_{i=1}^n \left(1 - Y_i \sum_{j=1}^n \alpha_j \langle X_i, X_j \rangle_{\mathbb{R}^d} \right)_+ . \tag{1}$$

More generally, we could also apply this algorithm in the feature space, corresponding to any reproducing kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In other words, we may first map all our inputs to the RKHS with $X \in \mathcal{X} \mapsto k(X, \cdot) \in \mathcal{H}_k$ and apply the above algorithm there. In order to do so, all we need to do is to replace all the inner products $\langle X_i, X_j \rangle_{\mathbb{R}^d}$ with the values of kernel $k(X_i, X_j)$, as this is now the inner product between points in the feature space:

$$\langle k(X_i, \cdot), k(X_j, \cdot) \rangle_{\mathcal{H}_k} = k(X_i, X_j)$$

by reproducing property. We arrive at the following kernelized problem:

$$\min_{\alpha_1,\dots,\alpha_n \in \mathbb{R}} \frac{1}{2} \left\| \sum_{i=1}^n \alpha_i k(X_i, \cdot) \right\|^2_{\mathcal{H}_k} + \frac{C}{n} \sum_{i=1}^n \left(1 - Y_i \sum_{j=1}^n \alpha_j k(X_i, X_j) \right)_+ . \tag{2}$$

Let's denote $f_{\boldsymbol{\alpha}} = \sum_{i=1}^n \alpha_i k(X_i, \cdot)$. We can notice that the same problem can be equivalently written in the following way:

$$\min_{R \in \mathbb{R}} \left\{ \min_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^n : \\ \|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}_k} = R}} \frac{1}{2} R^2 + \frac{C}{n} \sum_{i=1}^n \left(1 - Y_i \cdot f_{\boldsymbol{\alpha}}(X_i) \right)_+ \right\} . \tag{3}$$

At this point we only said that instead of optimizing over all $f_{\boldsymbol{\alpha}}$ simultaneously, we'll split these functions into the subsets $\mathcal{C}_R$ (based on their norms) and reduce our problem to two steps: (a)

optimize the objective over every subset $\mathcal{C}_R$, (b) choose the subset $\mathcal{C}_{R^*}$ with the minimal overall value of the objective. Now we can rewrite the same problem equivalently in the following form:

$$\min_{R\in\mathbb{R}}\left\{\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n:\\ \|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}_k}\leq R}}\frac{1}{2}R^2+\frac{C}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}(X_i)\right)_+\right\}. \tag{4}$$

and this step is, perhaps, less obvious than the previous one. In order to show that (4) is equivalent to (3), it is enough to notice the following fact. Fix any $R\in\mathbb{R}$, say R', and denote the solution of the corresponding inner optimization problem in (4) with $f_{\boldsymbol{\alpha}}^{R'}$. In other words,

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n:\\ \|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}_k}\leq R'}}\frac{1}{2}(R')^2+\frac{C}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}(X_i)\right)_+=\frac{1}{2}(R')^2+\frac{C}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}^{R'}(X_i)\right)_+.$$

Assume $\|f_{\boldsymbol{\alpha}}^{R'}\|_{\mathcal{H}_k}=R''<R'$. Now take $R=R''$ and notice that

$$\begin{aligned}
&\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n:\\ \|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}_k}\leq R''}}\frac{1}{2}(R'')^2+\frac{C}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}(X_i)\right)_+\\
&\qquad\leq\frac{1}{2}(R'')^2+\frac{C}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}^{R'}(X_i)\right)_+<\frac{1}{2}(R')^2+\frac{C}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}^{R'}(X_i)\right)_+\\
&\qquad\qquad=\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n:\\ \|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}_k}\leq R'}}\frac{1}{2}(R')^2+\frac{C}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}(X_i)\right)_+.
\end{aligned}$$

This shows that R' is definitely not a solution to the outer optimization problem.

We have just proved that if R_C solves the outer optimization problem in (4), then, if $f_{\boldsymbol{\alpha}}^{R_C}$ is the solution of the corresponding inner problem

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n:\\ \|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}_k}\leq R_C}}\frac{1}{2}(R_C)^2+\frac{C}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}(X_i)\right)_+,$$

it necessarily satisfies $\|f_{\boldsymbol{\alpha}}^{R_C}\|_{\mathcal{H}_k}=R_C$.

Concluding, we proved that there is a real number R_C (here we make explicit the fact that, actually, R_C depends on the regularization coefficient C) such that the unconstrained optimization problem (2) is equivalent to the following constrained optimization problem:

$$\begin{aligned}
&\min_{\alpha\in\mathbb{R}^n}\frac{1}{n}\sum_{i=1}^{n}\left(1-Y_i\cdot f_{\boldsymbol{\alpha}}(X_i)\right)_+\\
&\qquad\text{such that}\quad\|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}_k}\leq R_C.
\end{aligned} \tag{5}$$

Thus, we see that running soft-margin kernel SVM with regularizer C results in minimizing empirical risk over the ball in RKHS of certain radius R_C defined by C. Couple of remarks are in place here. First, we see that SVM minimizes empirical risk corresponding not to the binary loss function, but to the *hinge loss* function $\ell(f(X),Y)=\max\{0,1-Yf(X)\}$. Notice that hinge loss actually upper bounds the binary loss $\max\{0,1-Yf(X)\}\geq 1\{f(X)Y\leq 0\}$. Also, hinge

loss is *convex* in contrast to the binary loss. This illustrates a nice property of SVM that the corresponding optimization problem is convex, i.e. can be solved efficiently.

Another interesting thing to look at is the relation between R_C and C. Looking at (2), we see that the larger the C, the less we care about minimizing the norm of $f_{\boldsymbol{\alpha}}$. In extreme case when $C \to \infty$, we ignore the norm of $f_{\boldsymbol{\alpha}}$ and only concentrate at minimizing empirical hinge loss. On the other hand, as we decrease C more and more, the term $\|f_{\boldsymbol{\alpha}}\|^2_{\mathcal{H}_k}$ starts dominating and it becomes more important to make sure this term is small. We may conclude that larger C correspond to balls of larger radius R_C. In other words, choosing regularizer C in the soft-margin SVM corresponds to performing a *model selection*, i.e. choosing the ball over which we are going to minimize the hinge loss.

2 Risk bounds for kernel methods

We saw already that certain kernel methods result in searching functions of the form

$$\sum_{i=1}^{n} \alpha_i k(X_i, \cdot)$$

in balls of RKHS corresponding to the reproducing kernel k. I.e., these methods use elements of the following class of functions:

$$\mathcal{F}_R := \left\{ f = \sum_{i=1}^{n} \alpha_i k(X_i, \cdot) \in \mathcal{H}_k \colon n \in \mathbb{N},\ \alpha_1, \dots, \alpha_n \in \mathbb{R},\ X_1, \dots, X_n \in \mathcal{X},\ \|f\|_{\mathcal{H}_k} \le R \right\}.$$

In this section we will derive simple risk bounds for these methods, based on the Rademacher analysis.

First we state the main result of this lecture.

Theorem 1. *Consider a binary classification problem with $\mathcal{Y} = \{-1, +1\}$, any input space $\mathcal{X}$, and unknown probability P over $\mathcal{X} \times \mathcal{Y}$. Let k be a reproducing kernel over $\mathcal{X}$ satisfying $k(x, x) \le B$ for any $x \in \mathcal{X}$. Take any $\delta \in (0, 1)$, any $\gamma, R > 0$ and a function*

$$\varphi(z) = \begin{cases} 1, & z < 0; \\ 0, & z \ge \gamma; \\ 1 - z/\gamma, & z \in [0, \gamma]. \end{cases}$$

Then, with probability at least $1 - \delta$ (over the random i.i.d. training set $S_n = \{(X_i, Y_i)\}_{i=1}^n$), for any $f \in \mathcal{F}_R$ it holds that

$$\mathbb{P}_{(X,Y)\sim P}\big(Y \ne \operatorname{sgn} f(X)\big) \le \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i f(X_i)) + \frac{2R}{\gamma}\sqrt{\frac{B}{n}} + 2\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Before proving the result, let's discuss it.

First of all, we see that two last terms in the upper bound tend to zero as $n \to \infty$. Thus, as soon as the first term (empirical φ-risk) is small, we get a nice bound. At this point it may be useful to refresh the discussion of the margin bound for AdaBoost, as a similar argument may be applied here. Note that the φ-risk penalizes mistakes *and correct but low confidence answers*, i.e. answers with small margin $0 < Y_i f(X_i) < \gamma$. We can decrease the first term by letting $\gamma \to 0$, but this would explode last two terms.

Also, note that φ with $\gamma \to 0$ corresponds to the binary loss. It is not surprising that in this case the upper bound explodes to infinity. Indeed, from the VC theory we know that the problem is learnable if and only if the VC dimension of the hypotheses class is finite. Meanwhile, in Theorem 1 we are using classifiers based on elements in the balls of RKHS. From our previous discussion of soft-margin kernel SVM we know that it also corresponds to performing linear classification in the feature space (after mapping $\mathcal{X}$ to $\mathcal{H}_k$). Now, VC dimension of linear classifiers in $\mathbb{R}^d$ is $d+1$. If we use a Gaussian kernel, RKHS is infinite dimensional, and clearly VC theory tells us that the problem is not learnable. This is why the upper bound with $\gamma = 0$ *should* actually explode, otherwise it would lead to a contradiction.

For certain fixed $\gamma > 0$, in order to make sure the upper bound is small, apart from letting $n \to \infty$ we also need to assume that the first term is small. This corresponds to saying that f should make a lot of correct confident answers with large margin. Whether or not there is such an f in $\mathcal{F}_R$ is generally unknown. However, we may assume this, i.e. , we may have some expert knowledge telling us that the kernel k is good enough and R is big enough so that $\mathcal{F}_R$ actually contains an element nicely separating our training set. In this case (if this data assumption of ours is correct) we will arrive at a tight upper bound.

Risk bound for soft-margin kernel SVM? Finally, we see that Theorem 1 is not directly applicable to the Soft-Margin kernel SVM (5), because the hinge-loss is unbounded, while we require φ to be bounded in $[0,1]$. Nevertheless, there is a workaround. First of all, it will be clear from the proof that we can replace φ in the statement of theorem with *any* L-Lipschitz function bounded in $[0,M]$ interval, such that $\varphi(z) \geq 1\{z \leq 0\}$. This will result in $2/\gamma$ being replaced with $2L$ and extra factors M in the last two terms. Now, we see that this new updated result is applicable for hinge loss $\varphi(z) = (1-z)_+$ with $L = 1$, as soon as we guarantee that our distribution P is such that $\varphi(Yf(X)) = (1-f(X)Y)_+ \leq M$ with probability 1 (for $(X,Y) \sim P$) for all $f \in \mathcal{F}_R$. This is equivalent to asking $Yf(X) \geq 1-M$, which, in turn, can be guaranteed if $|f(X)| \leq |1-M|$ with probability 1. For a linear kernel k and $\mathcal{X} = \mathbb{R}^d$, as we know, $f(X) = \langle X, w\rangle_{\mathbb{R}^d}$. So, one case when Theorem 1 may be applied for a soft-margin SVM is when $\mathcal{X} = \mathbb{R}^d$, $k(x,x') = \langle x, x'\rangle_{\mathbb{R}^d}$, and P is such that P-almost surely $\|X\|_{\mathbb{R}^d} \leq |1-M|/R$ and $\|w\|_{\mathbb{R}^d} \leq R$.

3 Proof of Theorem 1

First we will repeat proof of Theorem 5 from Lecture 10 and write that with probability at least $1-\delta$ for all $f \in \mathcal{F}_R$ it holds that

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

where the second term corresponds to the empirical Rademacher complexity of $\mathcal{F}_R$. It remains to prove the following Lemma:

Lemma 2.

$$\hat{R}_n(\mathcal{F}_R) := \mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}_R}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right] \leq \frac{R}{n}\sqrt{\sum_{i=1}^n k(X_i,X_i)}.$$

Proof. Note that $\mathcal{F}_R \subseteq \{f \in \mathcal{H}_k \colon \|f\|_{\mathcal{H}_k} \leq R\}$. Using this we write:

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_R} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right] &\leq \mathbb{E}_\epsilon \left[\sup_{\substack{f \in \mathcal{H}_k: \\ \|f\|_{\mathcal{H}_k} \leq R}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right] \\
&= \mathbb{E}_\epsilon \left[\sup_{\substack{f \in \mathcal{H}_k: \\ \|f\|_{\mathcal{H}_k} \leq R}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \langle f, k(X_i, \cdot) \rangle_{\mathcal{H}_k} \right] \\
&= \mathbb{E}_\epsilon \left[\sup_{\substack{f \in \mathcal{H}_k: \\ \|f\|_{\mathcal{H}_k} \leq R}} \left\langle f, \frac{1}{n} \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \right\rangle_{\mathcal{H}_k} \right] \\
&= \frac{R}{n} \cdot \mathbb{E}_\epsilon \left[\left\| \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \right\|_{\mathcal{H}_k} \right],
\end{aligned}$$

where we used the reproducing property and a simple geometry. Jensen's inequality says that for any concave function $g \colon \mathbb{R} \to \mathbb{R}$ it holds that $\mathbb{E}_X[g(X)] \leq g(\mathbb{E}[X])$. We may write

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_R} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right] &\leq \frac{R}{n} \cdot \mathbb{E}_\epsilon \left[\left\| \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \right\|_{\mathcal{H}_k} \right] \\
&= \frac{R}{n} \cdot \mathbb{E}_\epsilon \left[\sqrt{\left(\left\| \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \right\|_{\mathcal{H}_k} \right)^2} \right] \\
&\leq \frac{R}{n} \cdot \sqrt{\mathbb{E}_\epsilon \left[\left(\left\| \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \right\|_{\mathcal{H}_k} \right)^2 \right]}.
\end{aligned}$$

Now,

$$\left(\left\| \sum_{i=1}^n \epsilon_i k(X_i, \cdot) \right\|_{\mathcal{H}_k} \right)^2 = \sum_{i=1}^n (\epsilon_i)^2 k(X_i, X_i) + \sum_{i \neq j}^n \epsilon_i \epsilon_j k(X_i, X_j).$$

Noticing that $(\epsilon_i)^2 = 1$ and $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ we conclude the proof. □