36-710: Advanced Statistical Theory **Fall 2018**

Lecture 25: November 28

Lecturer: Alessandro Rinaldo *Scribe: Theresa Gebert*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

In the last lecture we covered VC theory. In this lecture, we cover refinements of the last results and covering numbers, as well as some results for suprema of empirical processes.

25.1 Refinements and Covering Numbers

One example of an application of what we saw last class is the following: uniformly over $A \in \mathcal{A}$,

$$P(A) \leq P_n(A) + \sqrt{P(A)\frac{\log S_A(2n) + \log(4/\delta)}{n}} + \frac{4\log S_A(2n) + \log(4/\delta)}{n}$$

with probability $\geq 1 - \delta$. This is a Bernstein-like inequality with $P(A)$ playing the role of σ^2. We can also show that

$$E(\sup_A |P_n(A) - P(A)|) \leq \sqrt{\frac{2\log S_A(2n)}{n}} \approx \sqrt{\frac{\nu \log n}{n}}.$$

Useful extensions include the VC inequality for relative deviations:

$$P\left(sup_{A\in\mathcal{A}} \frac{P(A) - P_n(A)}{\sqrt{P(A)}} \geq \lambda\right) \leq 4S_A(2n)e^{-\frac{\lambda n^2}{4}}$$

as well as

$$P\left(sup_{A\in\mathcal{A}} \frac{P(A) - P_n(A)}{\sqrt{P_n(A)}}\right) \leq 4S_A(2n)e^{-\frac{\lambda n^2}{4}}.$$

Recall that $P(A) = E(1_{\{X\in A\}})$ and $\sqrt{P(A)}$ is an upper bound on $\sqrt{\text{Var}(1_{\{X\in A\}})} \leq \sqrt{P(A)(1-P(A))}$.

Proof: This is just a sketch of the proof. Use the fact that for any $a, b, c > 0$, $a \leq b + c\sqrt{a} \implies a \leq b + c^2 + \sqrt{b}c$. ■

So far we have been working with sets, but many of these concepts also extend to functions.

Proposition 25.1 *Let $\mathcal{F}$ be a collection of functions on $\mathcal{X}$ (e.g. $\mathbb{R}^d$) such that $0 \leq f(x) \leq 1$, for all $f \in \mathcal{F}$ and for all $x \in \mathcal{X}$. (The extension to function classes such that $-B \leq f(x) \leq B$ for some B is easy, so we can work with the simpler case.) Then*

$$\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \leq \sup_{f\in\mathcal{F}, t>0} \frac{1}{n}\left|\sum_{i=1}^n 1_{\{f(x)\geq t\}} - P(f(x) \geq t)\right|.$$

Let $\mathcal{G} = \{x \to 1_{\{f(x)\geq t\}}, f \in \mathcal{F}, t > 0\}$, then

$$\sup_{g\in\mathcal{G}} |P_n(g) - P(g)| = \|P_n - P\|_g.$$

Notice that $\mathcal{G}$ is a class of binary functions (which is what we have been studying so far). Then the right-hand side is the supremum of an empirical process indexed by binary functions.

If we let $A_{f,t} = \{x \in \mathcal{X} : f(x) > t\}$ for $f \in \mathcal{F}, t > 0$, so that

$$\mathcal{A} = \{A_{f,t}, f \in \mathcal{F}, t > 0\}$$

then

$$\|P_n - P\|_g = \sup_{A \in \mathcal{A}} |P_n(A) - P(A)|.$$

Assuming that the class $\mathcal{A}$ has finite VC dimension.

Definition 25.2 *A class $\mathcal{F}$ has finite VC dimension when $\mathcal{A}$ has finite VC dimension.*

This is a convenient assumption. Now we will actually prove the proposition.

Proof: Use the fact that

$$E(f(X)) = \int_0^1 P(f(x) \geq t)dt$$

and also the fact that if $x \geq 0$,

$$x = \int_0^\infty 1_{\{x>t\}}dt$$

so

$$f(x_1) = \int_0^\infty 1_{\{f(x_1)>t\}}dt$$

and all we need to do is plug this in, use the fact that the absolute value of the integral is less than the integral of the absolute value, and we are done. ■

However, a finite VC dimension is a strong assumption and not necessarily easy to verify. See the book [2]. How do we deal with VC theory for functions? Rather than assuming that $\mathcal{F}$ has finite VC dimension, it is more convenient to use covering arguments and assumptions. Let us review some of the concepts we saw previously in the course for this.

Definition 25.3 *An ϵ-covering of $\mathcal{F}$ with respect to a metric d on $\mathcal{F} \times \mathcal{F}$ is the smallest number $N = N(\epsilon)$ such that there exists $\{f_1, \ldots, f_n\}$ (a finite collection of functions) such that for all $f \in \mathcal{F}$, there exists $f_i = f_i(f)$ such that $d(f, f_i) \leq \epsilon$.*

For i.i.d. $(X_1, \ldots, X_n) \sim P$ on $\mathcal{X}$, let

$$d_{1,P_n}(f, g) = \frac{1}{n} \sum_{i=1}^n |f(x_1) - g(x_1)|$$

where P_n is the empirical measure. This is a random metric (since it depends on random variables) and is technically a "pseudo"-metric. The random ℓ_1 distance over $\mathcal{F}$ is

$$\|f - g\|_{\Delta,P} = \int |f(x) - g(x)|dP(x).$$

Recognize that $E(d_{1,P_n}(f, g))$ is the quantity above.

Theorem 25.4

$$P(\|P_n - P\|_{\mathcal{F}} \geq \epsilon) \leq E\left(N_{1,P_{2n}}\left(\mathcal{F}, \frac{\epsilon}{8}\right)\right) e^{-\frac{n\epsilon^2}{32}}$$

where $N_{1,P_{2n}}(\mathcal{F}, \frac{\epsilon}{8})$ is the $\epsilon/8$-covering number of $\mathcal{F}$ with respect to the random distance $d_{1,P_{2n}}$.

Since the covering number is with respect to a random distance, this makes it a random covering number, hence taking an expectation makes sense. The $2n$ comes from symmetrization. For practical purposes, this covering number is still a pain to calculate.

Most of the time we upper bound $E(d_{1,P_{2n}}(\mathcal{F}, \frac{\epsilon}{8}))$ by $N_\infty(\mathcal{F}(\epsilon/8))$, which is the $\epsilon/8$-covering number of $\mathcal{F}$ with respect to $d_\infty(f,g) = \sup_x |f(x) - g(x)|$. (This is on the homework assignment.)

The theorem below can be found as Theorem 9.4 of [3].

Theorem 25.5 *Let $\mathcal{F}$ be a class of functions on $\mathbb{R}^d$ with $0 \leq f \leq B$ with VC dimension ν. Then for all $p \geq 1$,*

$$N_{L_p}(\mathcal{F}, \epsilon) \leq 3\left(\frac{2eB^p}{\epsilon^p} \log \frac{3eB^p}{\epsilon^p}\right)^\nu$$

where $0 < \epsilon \leq B/4$ and $N_{L_p}(\mathcal{F}, \epsilon)$ is the L_p covering number with respect to any probability distribution on $(\mathbb{R}^d, \mathcal{B}^d)$. Specifically,

$$\|f - g\|_{L_p} = \left(\int (f-g)^p dp\right)^{1/p}.$$

Is it known whether this can be extended to the case where $p = \infty$? Now we must go beyond VC dimension; we need to prove bounds on covering numbers. If functions are highly irregular, the class will be bigger and so the covering number will be larger. The smoother the function, the smaller the covering number.

Always remember that these are can opener assumptions: they provide nice theoretical results but are not useful in practice.

25.2 Suprema of Empirical Processes

This is Bosquet's version of a Talagrand inequality for the supremum of an empirical process [4].

Theorem 25.6 *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ onto $[0,1]$. Then*

$$P\left(\|P_n - P\|_{\mathcal{F}} \geq E(\|P_n - P\|_{\mathcal{F}}) + \sqrt{\frac{2}{n}t\sigma^2(\mathcal{F}) + 2E(\|P_n - P\|_{\mathcal{F}})} + \frac{t}{3n}\right) \leq e^{-t}$$

and similarly

$$P\left(\|P_n - P\|_{\mathcal{F}} \leq E(\|P_n - P\|_{\mathcal{F}}) - \sqrt{\frac{2}{n}t\sigma^2(\mathcal{F}) + 2E(\|P_n - P\|_{\mathcal{F}})} + \frac{t}{n}\right) \leq e^{-t}$$

where

$$\sigma^2(\mathcal{F}) = \sup_{f \in \mathcal{F}} Var(f(x)).$$

We would need to bound $E(\|P_n - P\|_{\mathcal{F}})$ in order to use this inequality in practice. Rademacher complexity would be one way to do it. This inequality is actually straightforward to prove, in the sense that it is just calculus. (Who knows how Talagrand surmised it would all turn out like this.)

25.2.1 Supremum of Sub-Gaussian Processes

It is a hard task to compute the expected value of $sup_{\theta\in\mathbb{T}}X_\theta$, where $\{X_\theta, \theta \in \mathbb{T}\}$ is a stochastic process. Assume $\mathbb{T}$ is some random metric space that is well-behaved. Oftentimes we can express $X_\theta = \langle \theta^T, \epsilon\rangle$, assuming an inner product structure, where ϵ is a random variable.

Example 25.7 *Let $\mathbb{T} = \mathcal{F}(x_1^n) = \{f(x_1), \ldots, f(x_n), f \in \mathcal{F}\}$. The empirical Rademacher complexity is*

$$R_n(\mathcal{F}(x_1^n)) = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) = \sup_{\theta\in\mathbb{T}} \epsilon^T\theta.$$

There are many more examples we have seen throughout class. More generally, we can talk about $\mathbb{T} \subseteq \mathbb{R}^n$, and we can talk about $R_n(\mathbb{T}) = \sup_{\theta\in\mathbb{T}} \epsilon^T\theta$ as the Rademacher complexity of $\mathbb{T}$. Similarly,

$$G_n(\mathbb{T}) = \sup_{\theta\in\mathbb{T}} \omega^T\theta$$

where ω are i.i.d. random Gaussians. Next time, we will discuss the Gaussian Complexity of a set (which arises in nonparametric regression).

References

[1] M. J. Wainwright. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. *Cambridge University Press.*

[2] L. Devroye & L. Györfi & G. Lugosi. (1996). A Probabilistic Theory of Pattern Recognition. *Springer.* 10.1007/978-1-4612-0711-5.

[3] Distribution-Free Theory of Nonparametric Regression.

[4] O. Bousquet. (2003). Concentration inequalities for sub-additive functions using the entropy method. *Progress in Probability.* 213?248.