Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 21 — October 9

Instructor: Stas Minsker *Scribe: Mose Wintner*

21.1 VC Dimension

Theorem 1. Let $L = \{f : S \to \mathbb{R}\}$ be a linear subspace of functions with $\dim L = d$. Define

$$\mathcal{C} = \{\{x : f(x) > 0\} : f \in L\}.$$

Then $\mathcal{C}$ has VC dimension d.

Remark 1. Same holds for $\mathcal{C}' = \{\{f \geq 0\},\, f \in L\}$.

Example 1. Let L be a subspace of linear functions, $f(x) = \langle w, x\rangle + b$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$. Then $\dim(L) = d + 1$ (standard basis of functions plus the constant function 1) and $\mathcal{C} = \{\text{all half-spaces on } \mathbb{R}^d\}$.

Proof of theorem: First we'll show that no set of $d + 1$ points can be shattered. Let $\{x_1, \ldots, x_{d+1}\} \subset S$ be arbitrary. Define $T : L \to \mathbb{R}^{d+1}$ as

$$T(f) = (f(x_1), \ldots, f(x_{d+1})) \in \mathbb{R}^{d+1}.$$

The range of T is at most a d dimensional vector space since $\dim(L) = d$. Therefore there exists nonzero $w \in \mathbb{R}^{d+1}$ such that $w \perp R(T)$. Assume that $\{x_1, \ldots, x_{d+1}\}$ is shattered by $\mathcal{C}$. Then without loss of generality there exists j such that $w_j < 0$, where $w = (w_1, \ldots, w_{d+1})$. Define

$$\begin{aligned}\mathcal{A}_- &= \{1 \leq j \leq d+1, \text{ s.t. } w_j < 0\} \neq \emptyset\\ \mathcal{A}_+ &= \{1 \leq j \leq d+1, \text{ s.t. } w_j \geq 0\}\end{aligned}$$

By assumption, there exists $f \in L$ such that $f(x_j) > 0$ if and only if $j \in \mathcal{A}_-$. Since $w \perp (f(x_1), \ldots, f(x_{d+1}))$,

$$0 = \sum w_j f(x_j) = \sum_{j \in \mathcal{A}_-} w_j f(x_j) + \sum_{j \in \mathcal{A}_+} w_j f(x_j).$$

The first term is negative, and the second is ≤ 0, contradiction.

Now we must find $\{x_1, \ldots, x_d\}$ shattered by $\mathcal{C}$. Let L' be the dual space of L, and let $(\phi_1, \ldots, \phi_d)$ be a basis of L. Since L is finite-dimensional, for each $f \in L$ we have

$$f = \sum_{j=1}^{d} \alpha_j(f)\phi_j.$$

If $\ell \in L'$, then

$$\ell(f) = \sum \alpha_j(f)\ell(\phi_j).$$

For $\ell = (\ell_1, \ldots, \ell_d)$, $\ell' = (\ell_1', \ldots, \ell_d')$, we have $\langle \ell, \ell' \rangle = \sum \ell_j \ell_j'$. Let $\delta_x(f) := f(x)$ be the evaluation functional. Since $f(x) = \sum \alpha_j(f)\phi_j(x)$, we must have that $\phi_j(x) = \ell_j(\delta_x)$. Assume that $\langle \ell, \delta_x \rangle = 0$ for all $x \in S$ and some $\ell \in L'$. Then $\sum \ell_j \phi_j(x) = 0$ for all x. Since the ϕ_j form a basis, we must have that $\ell = 0$. Therefore the evaluation funtionals span L'. Therefore there exist $\{x_1, \ldots, x_d\}$ such that the evaluation functionals at those points are linearly independent. Therefore the set

$$\{\langle \delta_{x_1}(f), \ldots, \delta_{x_d}(f)\rangle : f \in L\} = \mathbb{R}^d,$$

so that $\mathcal{C}$ shatters $\{x_1, \ldots, x_d\}$.

Example 2. The collection of polygons (intersections of subspaces) in $\mathbb{R}^d$ with $k \geq 1$ faces has finite VC dimension. For instance, triangles in $\mathbb{R}^2$ have finite VC dimension (prof. thinks it's 7–nontrivial).

Let $\mathcal{C}$ be a VC class of sets. We have shown that

$$\mathbb{E}||P_n - P||_{\mathcal{C}} = \mathbb{E}\sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum I\{X_j \in C\} - P(C) \right| \leq K\sqrt{\frac{\log m^{\mathcal{C}}(n)}{n}} \leq K'\sqrt{\frac{V(\mathcal{C})\log n}{n}}.$$

We can use a chaining argument to obtain a sharper bound. Note that

$$\mathbb{E}||P_n - P||_{\mathcal{C}} \leq 2\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{C \in \mathcal{C}} \left| \frac{1}{n}\varepsilon_j I\{X_j \in C\} \right|$$

by the symmetrization inequality. We can write this

$$\frac{2}{\sqrt{n}}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{C \in \mathcal{C}} \left| \varepsilon_j \frac{I\{X_j \in C\}}{\sqrt{n}} \right|.$$

Define

$$T := \{t = (I\{x_1 \in C\}, \ldots, I\{X_n \in C\}), C \in \mathcal{C}\} \subset \mathbb{R}^n.$$

Then

$$X(t) := \frac{1}{\sqrt{n}} \sum \varepsilon_j t_j \in SG\left(\frac{||t||_2^2}{n}\right)$$

and $X(t) - X(s) \in SG(||t-s||_2^2/n)$. Note that

$$\frac{||t-s||^2}{n} = \frac{1}{n}\sum (I\{X_j \in C_1\} - I\{X_j \in C_2\})^2,$$

which is a random distance. We want to use this to get a nonrandom bound for the covering number.