Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 1-2 — Week 8

Instructor: S. Minsker *Scribe: S. Minsker*

Recall the key definitions (shattering number, growth function and VC dimension) from the previous lecture. Next, we will show a surprising result stating that the growth function $m^{\mathcal{C}}(n)$ can grown either exponentially or polynomially, and that there is no possible mode "in between". Let us introduce the following convenient notation:

$$\binom{n}{\le k} \overset{\text{def}}{=} \sum_{j=0}^{k} \binom{n}{j}.$$

Lemma 1. The following inequality holds: $\binom{n}{\le k} \le \left(\frac{ne}{k}\right)^k$.

Proof. Observe that $\binom{n}{j} \le (\frac{n}{k})^j \frac{k^j}{j!}$ (exercise: fill in the details using the definition pf the binomial coefficient), hence

$$\sum_{j=1}^{k} \binom{n}{k} \le \sum_{j=0}^{k} \left(\frac{n}{k}\right)^j \frac{k^j}{j!} \le \left(\frac{n}{k}\right)^k \sum_{j=0}^{\infty} \frac{k^j}{j!} = \left(\frac{ne}{k}\right)^k.$$

□

Theorem 1. Assume that the class $\mathcal{C}$ has finite VC dimension V. Then

$$m^{\mathcal{C}}(n) \le \binom{n}{\le V} \le \left(\frac{ne}{V}\right)^V.$$

Before we prove it, let us state an important corollary.

Corollary 1. Assume that the class $\mathcal{C}$ has finite VC dimension V. Then there exists an absolute constant $K > 0$ such that

$$\mathbb{E}\|P_n - P\|_{\mathcal{C}} \le K\sqrt{\frac{V \log n}{n}}$$

Indeed, it follows immediately from Theorem 6 in the last week's notes. The proof of Theorem 1 itself can be deduced from the following lemma.

Lemma 2 (Shelah-Sauer-Perles-Vapnik-Chervonenkis)**.** Let F be a subset of S with $\text{card}(F) = n$, and assume that $\Delta^{\mathcal{C}}(F) > \binom{n}{\le k-1}$. Then $\exists F' \subset F$, $\text{card}(F') = k$ and F' is shattered by $\mathcal{C}$.

Before looking at the proof, let us assume that the lemma is valid. Then the statement of Theorem 1 can be deduced as follows: assume that $m^{\mathcal{C}}(n) > \binom{n}{\leq V}$. Then there exist a subset $\{x_1, ..., x_{V+1}\}$ of cardinality $V+1$ that is shattered by $\mathcal{C}$, yielding a contradiction.

Proof of the lemma. Given $F = \{x_1, ..., x_n\}$, let $T = \{(I_C(x_1), ..., I_C(x_n), C \in \mathcal{C}\} \subseteq \{0,1\}^n$. If $\Delta^{\mathcal{C}}(F) > \binom{n}{\leq k-1}$, then $\mathrm{card}(T) = \Delta^{\mathcal{C}}(F) > \binom{n}{\leq k-1}$.

For $J \subset \{1, \dots, n\}$, define $\Pi_J : \mathbb{R}^n \longmapsto \mathbb{R}^{|J|}$, $\Pi_J(t) = (t_{i_1}, \cdots, t_{i_{|J|}})$ to be a projection on the coordinates indexed by J. The claim of the lemma is then equivalent to showing that there exist J, $\mathrm{card}(J) = k$ and $\mathrm{card}(\Pi_J T) = 2^k$. With some abuse of notation, we will say that J is shattered if $\Pi_J T = \{0,1\}^{\mathrm{card}(J)}$. To this end, we will construct a sequence of sets $T = T_0 \longmapsto T_1 \longmapsto ... \longmapsto T_n$ such that

(a) $\mathrm{card}(T_i) = \mathrm{card}(T_{i+1}),\ i = 0, ..., n-1$;

(b) If J is shattered by T_{i+1}, it is also shattered by T_i.

The map $\tau_1 : T_0 \mapsto T_1$ is constructed as follows: $t = (t_1, t_2, ..., t_n) \xrightarrow{\tau_1} t' = (t_1', ..., t_n')$ such that

1. $t_2' = t_2, ..., t_n' = t_n$;
2. if $t_1 = 0 \Rightarrow t_1' = 0$;
3. $t_1 = 1$ and $(0, t_2, ..., t_n) \in T \Rightarrow t_1' = 1$;
4. $t_1 = 1$ and $(0, t_2, ..., t_n) \notin T \Rightarrow t_1' = 0$.

It is easy to check that τ_1 satisfies required properties. Similarly for $i = 2, ..., n-1$, $\tau_i : T_{i-1} \mapsto T_i$ is defined via

1. $t_j' = t_j, j \neq i$;
2. $t_i = 0 \Rightarrow t_i' = 0$;
3. $t_i = 1$ and $(t_1, \dots, 0, \dots, t_n) \in T \Rightarrow t_i' = t_i$ (0 is at the i-th position);
4. $t_i = 1$ and $(t_1, \dots, 0, \dots, t_n) \notin T \Rightarrow t_1' = 0$ (0 is at the i-th position).

Let's look at the structure of the set T_n: assume that $t \in T_n$. If $s \leq t$ (meaning that $s_j \leq t_j, j = 1, ..., n$), then $s \in T_n$. Indeed, $t \in T_n$ and $t_i = 1$ for some $i \Longrightarrow (t_1, ...0, ..t_n) \in T_n$. Since $\mathrm{card}(T_n) > \begin{pmatrix} n \\ \leq k-1 \end{pmatrix}$, then clearly there exists $J, \mathrm{card}(J) = k$ such that J is shattered by T_n (because the total number of points t with $\mathrm{card}\{j : t_j = 1\} < k$ is $\binom{n}{k-1}$).

Let $t_\star \in T_n$ be such that $\mathrm{card}\{j : t_{\star,j} = 1\} \geq k$. Then by the properties of T_n the set $J = \{j : t_{\star,j} = 1\}$ is shattered. $\square$

Next, we will prove several useful facts that allow one to construct rich families of sets with finite VC dimension.

1.1 More on VC dimension

The following result is perhaps the best justification of associating VC index with the term "dimension."

Theorem 2 (R.M.Dudley)**.** Let $L := \{f : S \to \mathbb{R}\}$ be a linear subspace of functions with finite dimension $\dim(L) = d < \infty$. Define $\mathcal{C} := \{\{x : f(x) > 0\}, f \in L\}$. Then $\mathcal{C}$ has VC-dimension d.

Exercise. The same statement holds for $\mathcal{C} := \{\{x : f(x) \geq 0\}, f \in L\}$.

Example 1. Let L be a subspace of linear functions $f(x) := \langle w, x\rangle + b$, where $w \in \mathbb{R}^d, b \in \mathbb{R}$. Then $dim(L) = d + 1$. In this case, $\mathcal{C}$ is a collection of all half-spaces $\mathbb{R}^d$ and VC-dimension of $\mathcal{C}$ $V(\mathcal{C}) = d + 1$.

Proof. (a) First, we'll show that no set of $d+1$ points can be shattered. Let $\{x_1, ..., x_{d+1}\} \subset S$, and define a linear map $T : L \to \mathbb{R}^{d+1}$ via

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

Then $\operatorname{Im} T$ is at most d-dimensional subspace of $\mathbb{R}^{d+1}$. Hence there exists $w \in \mathbb{R}^{d+1}, w \neq 0$ such that $w \perp \operatorname{Im} T$. Without loss of generality, $\exists j$ such that $w_j < 0$, where $w = (w_1, ..., w_{d+1})$. Let us define

$$\mathbb{A}_- := \{1 \leq j \leq d + 1 : w_j < 0\},$$
$$\mathbb{A}_+ := \{1 \leq j \leq d + 1 : w_j \geq 0\}.$$

Let's assume that $\{x_1, ..., x_{d+1}\}$ is shattered by $\mathcal{C}$. Then $\exists f \in L$ such that $f(x_j) > 0$ whenever $j \in \mathbb{A}_-$. Since $w \perp (f(x_1), ..., f(x_{d+1}))$, $\sum_{j=1}^n w_j f(x_j) = 0$, but

$$\sum_{j=1}^n w_j f(x_j) = \sum_{j\in\mathbb{A}_-}^n w_j f(x_j) + \sum_{j\in\mathbb{A}_+}^n w_j f(x_j) < 0,$$

because $\sum_{j\in\mathbb{A}_-}^n w_j f(x_j) < 0$ (we chose w in a way that there exist at least a single j such that and $w_j < 0$) and $\sum_{j\in\mathbb{A}_+}^n w_j f(x_j) \leq 0$. We obtain a contradiction, therefore no set of $d + 1$ points can be shattered.

(b) It remains to find $\{x_1, ..., x_d\}$ that are shattered by $\mathcal{C}$. Let L' be the dual space of L, the space of linear functionals on L. Let $\phi_1, ..., \phi_d$ be the basis of L, so that

$$\forall f \in L,\ \exists \alpha_1(f), ..., \alpha_d(f) : f = \sum_{j=1}^d \alpha_j(f)\phi_j.$$

By linearity, for any $l \in L'$

$$l(f) = \sum_{j=1}^d \alpha_j(f) l(\phi_j) = \sum_{j=1}^d \alpha_j(f) l_j,$$

where $l_j = l(\phi_j)$. Define the inner product of two linear functionals l_1 and l_2 via

$$\langle l_1, l_2 \rangle = \sum_{j=1}^{d} l_1(\phi_j) l_2(\phi_j).$$

Let's consider a particular linear functional - the point evaluation function $\delta_x(f) := f(x)$ and $\phi_j = l_j(\delta_x)$. So we can rewrite $f(x) = \sum_{j=1}^{n} \alpha_j(f)\phi_j(x)$ as

$$f(x) = \delta_x(f) = \sum_{j=1}^{n} \alpha_j(f) l_j(\delta_x),$$

meaning that $l_j(\delta_x) = \phi_j(x)$.

Assume there exists a linear functional $l \in L'$ such that it is orthogonal to the linear span of point-evaluation functionals $\{\delta_x,\ x \in S\}$. In other words

$$\forall x \in S, \langle l, \delta_x \rangle = 0 \Leftrightarrow \forall x \in S, \sum_{j=1}^{d} l_j \phi_j(x) = 0.$$

But since $\phi_1, ..., \phi_d$ is a basis, $l_1 = \ldots = l_d = 0$.

This way we have shown that linear span of point-evaluation functions is equal to L'. So, $\exists \{x_1, ..., x_d\}$ such that $\{\delta_{x_1}, ..., \delta_{x_d}\}$ are linearly independent. Hence,

$$\text{lin.span}\{(\delta_{x_1}(f), ..., \delta_{x_d}(f)), f \in L\} = \mathbb{R}^d,$$

which easily implies that $\mathcal{C}$ shatters these $\{x_1, ..., x_d\}$. $\square$