STAT 620: Asymptotic Statistics **Spring 2022**

Lecture: Apr 5

Lecturer: Xianyang Zhang

This note provides some additional materials on the Vapnik-Chervonenkis (VC) dimension. To learn more, please refer to Chapter 2.6 of van der Vaart and Wellner (1996).[1]

1 Definition

Let $\mathcal{C}$ be a collection of subsets of $\mathcal{X}$. An arbitrary set of n points $x_1^n = \{x_1, \ldots, x_n\}$ contains 2^n subsets. We say that $\mathcal{C}$ picks out a certain subset of x_1^n if the subset is the form $C \cap x_1^n$ for some $C \in \mathcal{C}$. Let $\Delta(\mathcal{C}, x_1^n)$ be the cardinality of the collection of sets $\{C \cap x_1^n : C \in \mathcal{C}\}$, i.e., the number of subsets of x_1^n that can be picked out by $\mathcal{C}$. If $\Delta(\mathcal{C}, x_1^n) = 2^n$, we say that x_1^n is *shattered* by $\mathcal{C}$. The VC-dimension of $\mathcal{C}$, denoted by $VC(\mathcal{C})$, is the largest integer n such that there exists a set of n points $x_1^n \subseteq \mathcal{X}$ with $\Delta(\mathcal{C}, x_1^n) = 2^n$. Equivalently,

$$VC(\mathcal{C}) = \sup_n \left\{ n \in \mathbb{N} : \sup_{x_1^n \subseteq \mathcal{X}} \Delta(\mathcal{C}, x_1^n) = 2^n \right\}.$$

Put another way, if there is no set of points $x_1, \ldots, x_{n+1} \in \mathcal{X}$ that $\mathcal{C}$ shatters, then $VC(\mathcal{C}) < n + 1$. The VC-dimension quantifies the complexity of the collection of sets $\mathcal{C}$.

Remark. Sometimes VC-dimension is defined as the smallest integer n for which no x_1^n is shattered by $\mathcal{C}$. This is the definition used in van der Vaart and Wellner (1996).

Example. We have the following results:

1. If $\mathcal{C} = \{(-\infty, x] : x \in \mathbb{R}\}$, then $VC(\mathcal{C}) = 1$. This is because for two points $x_1 < x_2$, $\mathcal{C}$ can never pick out the set $\{x_2\}$.
2. If $\mathcal{C} = \{(-\infty, x_1] \times \cdots (-\infty, x_d] : x_1, \ldots, x_d \in \mathbb{R}\}$, then $VC(\mathcal{C}) = d$.
3. If $\mathcal{C} = \{\text{all rectangles in } \mathbb{R}^d\}$, $VC(\mathcal{C}) = 2d$.

Let $S_n(\mathcal{C}) = \sup_{x_1^n \subseteq \mathcal{X}} \Delta(\mathcal{C}, x_1^n)$, which is called the shattered coefficient. We have the following properties regarding the shattered coefficient:

1. Let $\mathcal{C}^c = \{C^c : C \in \mathcal{C}\}$. Then $S_n(\mathcal{C}) = S_n(\mathcal{C}^c)$.
2. Let $\mathcal{C}_+ = \{C_1 \cup C_2 : C_1 \in \mathcal{C}_1, C_2 \in \mathcal{C}_2\}$ and $\mathcal{C}_- = \{C_1 \cap C_2 : C_1 \in \mathcal{C}_1, C_2 \in \mathcal{C}_2\}$. Then $S_n(\mathcal{C}_+) \leq S_n(\mathcal{C}_1) \times S_n(\mathcal{C}_2)$ and $S_n(\mathcal{C}_-) \leq S_n(\mathcal{C}_1) \times S_n(\mathcal{C}_2)$.
3. $S_{n+m}(\mathcal{C}) \leq S_n(\mathcal{C}) S_m(\mathcal{C})$ for $n, m \in \mathbb{N}$.
4. $S_n(\mathcal{C}_1 \cup \mathcal{C}_2) \leq S_n(\mathcal{C}_1) + S_n(\mathcal{C}_2)$.

The verification of these results is left as exercises.

[1] van der vaart, A., and Wellner, J. (2000). *Weak convergence and empirical processes: with applications to statistics.* Springer Series in Statistics, New York.

2 Sauer's lemma and covering numbers

We introduce the Sauer's lemma and its proof. It states that as long as $VC(\mathcal{C}) < \infty$, the number of subsets that $\mathcal{C}$ can pick out from x_1^n grows at most polynomially in n.

Lemma. Suppose $VC(\mathcal{C}) < \infty$. Then

$$\sup_{x_1^n \subseteq \mathcal{X}} \Delta(\mathcal{C}, x_1^n) \leq \sum_{k=0}^{VC(\mathcal{C})} \binom{n}{k} \leq (n+1)^{VC(\mathcal{C})}.$$

Proof. To see the second inequality, we note that

$$\sum_{k=0}^{VC(\mathcal{C})} \binom{n}{k} = \sum_{k=0}^{VC(\mathcal{C})} \frac{n!}{(n-k)!k!} \leq \sum_{k=0}^{VC(\mathcal{C})} \frac{n^k}{k!} \leq \sum_{k=0}^{VC(\mathcal{C})} n^k \binom{VC(\mathcal{C})}{k} = (n+1)^{VC(\mathcal{C})},$$

where we have used the fact that $1/k! \leq \binom{VC(\mathcal{C})}{k}$ and the binomial expansion formula.

We shall use the induction argument to prove the first inequality. Let $\Psi_k(n) = \sum_{i=0}^{k} \binom{n}{i}$ and

$$\Phi_k(n) = \sup_{VC(\mathcal{C}) \leq k} \sup_{x_1^n \subseteq \mathcal{X}} \Delta(\mathcal{C}, x_1^n)$$

The assertion is equivalent to

$$\Phi_k(n) \leq \Psi_k(n)$$

for all k, n, which we will prove using induction arguments on the sum $n + k$. When $n = 0$ or $k = 0$, $\Phi_k(n) = \Psi_k(n) = 1$. Taking $n = k = 1$, it is not hard to verify that $\Psi_1(1) = 2$ and $\Phi_1(1) = 2$. Now assume that the results hold for all pairs (n', k') for $n' + k' < m$ with $m \in \mathbb{N}$. Let $n + k = m$ and $VC(\mathcal{C}) = k$ for some collection of sets $\mathcal{C}$. For $i \in \{1, \ldots, n\}$ and a set $x_1^n = \{x_1, \ldots, x_n\}$, define $x_2^n = x_1^n \setminus \{x_1\} = \{x_2, \ldots, x_n\}$.

Let $\mathcal{C}' \subset \mathcal{C}$ be a sub-collection of $\mathcal{C}$ such that it picks out as many subsets of x_2^n as possible. If there exist C_1 and C_2 such that $C_1 \cap x_2^n = C_2 \cap x_2^n$, we keep the one that does not contain x_1 in $\mathcal{C}'$. If all such C_i's contain x_1, we keep all of them in $\mathcal{C}'$. By construction, $\mathcal{C}'$ includes all the sets of $\mathcal{C}$ that do not contain x_1.

We claim that

$$\Delta(\mathcal{C}, x_1^n) = \Delta(\mathcal{C}', x_2^n) + \Delta(\mathcal{C} \setminus \mathcal{C}', x_2^n). \tag{1}$$

Suppose $A \subset x_1^n$. There are three cases:

1. The sets that can pick out $A \setminus \{x_1\}$ (when intersecting with x_2^n) all contain x_1. Then they are all in $\mathcal{C}'$.
2. None of the set that can pick out $A \setminus \{x_1\}$ contains x_1. Then they are all in $\mathcal{C}'$ as well.
3. The sets that can pick out $A \setminus \{x_1\}$ may or may not contain x_1. Then the ones that contain x_1 are in $\mathcal{C} \setminus \mathcal{C}'$ and the rest are in $\mathcal{C}'$.

It is not hard to verify (1) by analyzing each case (**think about why?**).

If we have $VC(\mathcal{C}') \leq k$, then by the induction hypothesis, $\Delta(\mathcal{C}', x_2^n) \leq \Phi_k(n-1) \leq \Psi_k(n-1)$. We claim that $VC(\mathcal{C} \setminus \mathcal{C}') \leq k-1$. To see this, we note that if $\mathcal{C} \setminus \mathcal{C}'$ shatters a set $B \subset x_2^n$, then $\mathcal{C}$ must shatter $B \cup \{x_1\}$ (this is because there exists a set $C \in \mathcal{C}'$ that does not contain x_1 and can pick out the same subset of B). So that the cardinality of B is less than k as $VC(\mathcal{C}) = k$. Therefore, we have $\Delta(\mathcal{C} \setminus \mathcal{C}', x_2^n) \leq \Phi_{k-1}(n-1) \leq \Psi_{k-1}(n-1)$.

We obtain

$$\begin{aligned}\Delta(\mathcal{C}, x_1^n) =&\Delta(\mathcal{C}', x_2^n) + \Delta(\mathcal{C} \setminus \mathcal{C}', x_2^n)\\ \leq&\Psi_k(n-1) + \Psi_{k-1}(n-1)\\ =&\sum_{i=0}^{k} \binom{n-1}{i} + \sum_{i=0}^{k-1} \binom{n-1}{i}\\ =&\sum_{i=0}^{k} \binom{n-1}{i} + \sum_{i=1}^{k} \binom{n-1}{i-1} = \sum_{i=0}^{k} \binom{n}{i}.\end{aligned}$$

To get the last equality, we have used the fact that

$$\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}.$$

This equation can be proved by considering the problem of selecting i balls out of a box of n balls conditioning on whether the first ball is being selected or not.

Remark. A different proof is given in van der Vaart and Wellner (1996). The basic idea is to first prove the result for a collection of sets $\mathcal{C}$ that is hereditary, i.e., $B \in \mathcal{C}$ whenever $B \subset C$ for $C \in \mathcal{C}$. Then it is argued that a general $\mathcal{C}$ can be transformed into a hereditary collection without changing its cardinality and without increasing the number of sets it shatters.

For a collection of sets $\mathcal{C}$ and a probability distribution P on $\mathcal{X}$, we define the $L_r(P)$ metric (with $r > 0$) between sets $A, B \subset \mathcal{X}$ by the distance between their indicator functions, i.e.,

$$\|\mathbf{1}_A - \mathbf{1}_B\|_{L_r(P)} = \left(\int_{\mathcal{X}} |\mathbf{1}_A - \mathbf{1}_B|^r dP(x) \right)^{1/r}.$$

We define the covering numbers of a collection $\mathcal{C}$ with respect to this metric on sets, denoting it by $N(\mathcal{C}, L_r(P), \epsilon)$. A classical result is the following uniform control on covering numbers. See Theorem 2.6.4 of van der Vaart and Wellner (1996).

Theorem. Let $\mathcal{C}$ be a class of sets with $VC(\mathcal{C}) < \infty$. Then there exists a universal constant $K < \infty$ such that for any probability measure P, any $r \geq 1$ and all $0 < \epsilon < 1$,

$$N(\mathcal{C}, L_r(P), \epsilon) \leq K \cdot VC(\mathcal{C}) \cdot (4e)^{VC(\mathcal{C})} \left(\frac{1}{\epsilon}\right)^{rVC(\mathcal{C})}.$$

Example. We show that

$$E\left[\sup_{t \in \mathbb{R}^d} |P_n(X \leq t) - P(X \leq t)|\right] \to 0,$$

where $t = (t_1, \ldots, t_d) \in \mathbb{R}^d$ and $[X \leq t] = [X_1 \leq t_1, \ldots, X_d \leq t_d]$. Set $\mathcal{F} = \{1_{\{X_1 \leq t_1, \ldots, X_d \leq t_d\}} : t =$

$(t_1, \ldots, t_d) \in \mathbb{R}^d\}$. Then $VC(\mathcal{F}) = O(d)$. We thus get

$$
\begin{aligned}
&E\left[\sup_{t\in\mathbb{R}^d} |P_n(X \le t) - P(X \le t)|\right] \\
\le&\frac{2}{\sqrt{n}} E\left[E\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(X_i)\right| \middle| X_1, \ldots, X_n\right]\right] \\
\le&\frac{2c}{\sqrt{n}} E\left[\int_0^\infty \sqrt{\log N(\mathcal{F}, L_2(P_n), \epsilon)}d\epsilon\right] + \frac{c}{\sqrt{n}} \\
\le&\frac{2c\sqrt{d}}{\sqrt{n}}\int_0^1 \sqrt{\log\frac{1}{\epsilon}}d\epsilon + \frac{c}{\sqrt{n}}
\end{aligned}
$$

Definition. The subgraph of a function: $\mathcal{X} \to \mathbb{R}$ is defined as

$$\text{sub} f := \{(x, t) : t < f(x)\} = (\text{epi} f)^c.$$

Note: $\text{sub} f \subseteq \mathcal{X} \times \mathbb{R}$.

Definition. $\mathcal{F}$ is a VC-class (VC-subgraph-class) if $\{\text{sub} f : f \in \mathcal{F}\}$ is a VC-class.

Theorem. For a VC-subgraph-class of functions $\mathcal{F}$ with envelope function F and $r \ge 1$, one has for any probability measure Q with $0 < \|F\|_{Q,r} = (\int F^r dQ)^{1/r} < \infty$,

$$N(\mathcal{F}, L_r(Q), \epsilon\|F\|_{Q,r}) \le K \cdot VC(\mathcal{F}) \cdot 16e^{VC(\mathcal{F})}\left(\frac{1}{\epsilon}\right)^{rVC(\mathcal{F})}.$$