1 Introduction

In this note, we sketch a few properties of covering numbers, VC-dimension, and provide a few pointers to more general resources for more detailed treatment of the results.

To define Vapnik-Chervonenkis dimension (VC-dimension), we begin by recalling the notion of shattering a set of points. Give a set of points \( x_1, \ldots, x_n \in \mathcal{X} \), we call a vector \( y \in \{-1, 1\}^n \) a labeling of the set \( \{x_i\} \). Then a collection of sets \( C \subset \mathcal{X} \), where \( C \in \mathcal{C} \) are subsets of \( \mathcal{X} \), shatters \( \{x_i\} \) if for each labeling \( y_1, \ldots, y_n \) of the points \( x_i \), there is a set \( C \in \mathcal{C} \) such that

\[
x_i \in C \text{ for } i \text{ s.t. } y_i = 1, \quad x_i \notin C \text{ otherwise.} \tag{1}
\]

In general, we say \( C \) realizes the labeling \( y \in \{-1, 1\}^n \) for \( \{x_i\} \) if the containment (1) holds. The collection \( C \) has VC-dimension \( \text{VC}(C) = d \) if the largest set of points \( x_1, \ldots, x_n \) it shatters is of size \( n = d \). That is,

\[
\text{VC}(C) = \sup \{ n \in \mathbb{N} : \exists x_1, \ldots, x_n \text{ s.t. } C \text{ shatters } \{x_i\}_{i=1}^n \}.
\]

Put another way, if there is no set of points \( x_1, \ldots, x_{n+1} \) that \( C \) shatters, then \( \text{VC}(C) < n + 1 \).

With this in mind, we follow van der Vaart and Wellner [1] and define the shattering number of the points \( x_1, \ldots, x_n \) as

\[
\Delta_n(C, x_1, \ldots, x_n) := \text{card} \{ y \in \{-1, 1\}^n \text{ s.t. } C \text{ realizes } y \text{ for } \{x_1, \ldots, x_n\} \}.
\]

Then an equivalent definition to the VC-dimension is that

\[
\text{VC}(C) := \sup_n \left\{ n : \sup_{x_1, \ldots, x_n} \Delta_n(C, x_1, \ldots, x_n) = 2^n \right\}.
\]

2 Sauer’s lemma

We now state a few results on VC-dimension, providing proofs of simplifications that make clearer what is happening. Interestingly, VC-sets have at most polynomial growth in their shattering numbers—as soon as a VC collection \( C \) cannot shatter any set of \( n \) points, the number of labelings it can realize on the points is at most \( n^{\text{VC}(C)} \ll 2^n \). This is the content of the Sauer-Shelah lemma.

**Lemma 2.1** (Sauer-Shelah lemma). Let \( \text{VC}(C) < \infty \). Then

\[
\sup_{x_1, \ldots, x_n} \Delta_n(C, x_1, \ldots, x_n) \leq \sum_{k=0}^{\text{VC}(C)} \binom{n}{k}.
\]
Lemma 3.1. For any $\epsilon > 0$ and set $\Theta$ with metric $d$,

$$M(\Theta, d, 2\epsilon) \leq N(\Theta, d, \epsilon) \leq M(\Theta, d, \epsilon).$$

3 Covering numbers for VC-classes

VC-classes of sets have finite covering numbers in a very uniform sense, which allows substantial control in concentration inequalities and uniform laws of large numbers. We begin by recalling the definition of the covering $N$ and packing $M$ numbers of a set $\Theta$ with metric $d$ as

$$N(\Theta, d, \epsilon) := \inf \{N : \exists \text{ an } \epsilon\text{-cover } \{\theta^i\}_{i=1}^N \text{ of } \Theta\}$$

and

$$M(\Theta, d, \epsilon) := \sup \{M : \exists \text{ an } \epsilon\text{-packing } \{\theta^i\}_{i=1}^M \text{ of } \Theta\},$$

where we recall an $\epsilon$-packing satisfies $d(\theta^i, \theta^j) > \epsilon$ for all $i, j$. The following lemma is standard.

**Lemma 3.1.** For any $\epsilon > 0$ and set $\Theta$ with metric $d$,

$$M(\Theta, d, 2\epsilon) \leq N(\Theta, d, \epsilon) \leq M(\Theta, d, \epsilon).$$
For a probability distribution \( P \), we recall the definition of \( L_r(P) \) norms on functions \( f : \mathcal{X} \rightarrow \mathbb{R} \) as
\[
\|f\|_{L_r(P)} := \left( \int |f(x)|^r dP(x) \right)^{\frac{1}{r}}.
\]
For a collection of sets \( \mathcal{C} \), we define the \( L_r(P) \) metric between sets \( A, B \subset \mathcal{X} \) by the distance between their indicators, that is,
\[
\|1_A - 1_B\|_{L_r(P)}^r = \int |1(x \in A) - 1(x \in B)|^r dP(x).
\]
We then define the covering numbers of a collection \( \mathcal{C} \) with respect to this metric on sets, denoting them by \( N(\mathcal{C}, L_r(P), \epsilon) \). A classical result is then the following uniform control on covering numbers.

**Theorem 1.** Let \( \mathcal{C} \) be a class of sets with \( VC(\mathcal{C}) < \infty \). Then there exist universal constants \( C, K < \infty \) such that for all \( 0 \leq \epsilon < 1 \)
\[
N(\mathcal{C}, L_r(P), \epsilon) \leq C \cdot VC(\mathcal{C})K^{VC(\mathcal{C})} \left( \frac{1}{\epsilon} \right)^{rVC(\mathcal{C})}.
\]
We do not prove this theorem in its full generality, referring to van der Vaart and Wellner [1, Theorem 2.6.4] for the full proof (note that they use a slightly different definition of VC-dimension than ours, which is shifted by 1).

We can, however, provide the following weaker theorem, which is a simplification of the preceding result, and gives a flavor of the types of results one can demonstrate.

**Theorem 2.** Let \( \mathcal{C} \) be a VC-class with \( VC(\mathcal{C}) = d < \infty \). Then for any \( \tau > 0 \), there exist universal constants \( C, K < \infty \) (which may depend on \( \tau \)) such that for all \( 0 \leq \epsilon \leq 1 \)
\[
N(\mathcal{C}, L_r(P), \epsilon) \leq C \cdot K^{d\log d} \left( \frac{1}{\epsilon} \right)^{rd+\tau}.
\]

**Proof.** We provide the proof in three parts. First, we let \( C_1, \ldots, C_N \) be a maximal \( \delta = \epsilon^\tau \)-packing in the \( L_r(P) \) norm, so that for \( X \sim P \) we have
\[
\mathbb{E}[|1_{C_i}(X) - 1_{C_j}(X)|^r] = \mathbb{E}[|1_{C_i}(X) - 1_{C_j}(X)|] > \delta = \epsilon^\tau.
\]
It is thus clear that \( N(\mathcal{C}, \epsilon^\tau, L_1(P)) \geq N(\mathcal{C}, \epsilon, L_r(P)) \), so we may thus focus on the \( L_1 \) case with the \( \delta \)-packing. By Lemma 3.1, we thus have \( N(\mathcal{C}, \delta, L_1(P)) \leq N(\mathcal{C}, \delta, L_1(P)) \leq N \).

We now note that for \( X \sim P \), we have
\[
P(X \in C_i \text{ and } X \in C_j) < 1 - \delta,
\]
because \( \delta < \mathbb{E}[|1_{C_i}(X) - 1_{C_j}(X)|] = 1 - \mathbb{E}[1_{C_i \cap C_j}(X)] = 1 - P(X \in C_i, X \in C_j) \). By independence, if \( X_1, \ldots, X_n \overset{iid}{\sim} P \), we obtain
\[
P(X_1 \in C_i \cap C_j, \ldots, X_n \in C_i \cap C_j) < (1 - \delta)^n.
\]
Now, let \( \mathcal{E} \) denote the event that each \( C_i \) “picks out” a different subset of \( X_1, \ldots, X_n \), that is, the sets \( C_1 \cap \{X_1, \ldots, X_n\} \) are distinct. Then by a union bound, we have
\[
P(\mathcal{E}^c) \leq \sum_{i<j} P(C_i \cap \{X_1, \ldots, X_n\} = C_j \cap \{X_1, \ldots, X_n\}) < \sum_{i<j} (1 - \delta)^n = \binom{N}{2}(1 - \delta)^n, \quad (2)
\]
so that the probability $P(\mathcal{E}) \geq 1 - \left(\frac{N}{2}\right)(1 - \delta)^n$.

Now we note that if $n = \frac{2 \log N}{\delta}$, then there exists a set of $n$ points from which $C$ can choose at least $N$ distinct subsets. Indeed, by inequality (2), we have

$$P(C \cap \{X_1, \ldots, X_n\} \text{ are distinct}) > 1 - \left(\frac{N}{2}\right)(1 - \delta)^n \geq 1 - N^2 e^{-\delta n} = 1 - N^2 e^{-2 \log N} = 0.$$ 

So the probabilistic method implies that at least some such set exists, i.e. that $\Delta_n(C, x_1, \ldots, x_n) \geq N$ for some set $\{x_i\}_{i=1}^n$ by the definition of the shattering numbers.

Using the Sauer-Shelah lemma 2.1, we find that

$$N \leq \Delta_n(C, x_1, \ldots, x_n) \leq \sum_{k=0}^{\text{VC}(C)} \binom{n}{k} \leq dn^d,$$

where we have used $d = \text{VC}(C)$. Rearranging, we have that the covering number $N$ must satisfy

$$N \leq d \left(\frac{2 \log N}{\delta}\right)^d \quad \text{or} \quad \frac{N}{\log^d N} \leq d \left(\frac{2}{\delta}\right)^d. \quad (3)$$

We now argue that for any $\tau > 0$, choosing a large enough constant $C = C(d)$ and $N \geq C(2/\delta)^{d+\tau}$ contradicts this inequality. Indeed, rewriting the inequality with such an $N$, we have

$$\frac{C}{\log^d (C(\frac{2}{\delta})^{2+\tau})} \leq d \left(\frac{2}{\delta}\right)^{-\tau} \quad \text{or} \quad \frac{C^{\frac{1}{2}}}{\log C + (2 + \tau) \log \frac{2}{\delta}} \leq d^{\frac{1}{2}} \left(\frac{2}{\delta}\right)^{-\frac{\tau}{2}}.$$ 

If this inequality fails for $\delta = 1$ it fails for all $\delta < 1$, so we must have

$$\frac{C^{\frac{1}{2}}}{\log C + (2 + \tau) \log 2} \leq d^{\frac{1}{2}} 2^{-\tau}.$$ 

Evidently taking $C \gg d^{2-\tau}$ gives the desired contradiction. We obtain the theorem when we replace $\delta$ with $\epsilon^\tau$. \qed

References