VC-Dimension, Covering, and Packing

John Duchi: Notes for Statistics 300b

March 2, 2017

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 $\mathcal{C} \subset 2^{\mathcal{X}}$, 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 $\mathcal{C}$ realizes the labeling $y \in \{-1, 1\}^n$ for $\{x_i\}$ if the containment (1) holds. The collection $\mathcal{C}$ has VC-dimension $\mathsf{VC}(\mathcal{C}) = d$ if the largest set of points $x_1, \ldots, x_n$ it shatters is of size $n = d$. That is,

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

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

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

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

Then an equivalent definition to the VC-dimension is that

$$\mathsf{VC}(\mathcal{C}) := \sup_n \left\{n : \sup_{x_1, \ldots, x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) = 2^n\right\}.$$

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 $\mathcal{C}$ cannot shatter any set of n points, the number of labelings it can realize on the points is at most $n^{\mathsf{VC}(\mathcal{C})} \ll 2^n$. This is the content of the Sauer-Shelah lemma.

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

$$\sup_{x_1, \ldots, x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \le \sum_{k=0}^{\mathsf{VC}(\mathcal{C})} \binom{n}{k}.$$

Proof Our proof follows an argument by Martin Wainwright. Define $\Psi_k(n) := \sum_{i=0}^k \binom{n}{i}$ and

$$\Phi_k(n) := \sup_{\mathcal{C}:\mathsf{VC}(\mathcal{C})\le k} \sup_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n)$$

in which case the assertion is equivalent to $\Phi_k(n) \le \Psi_k(n)$ for all k, n. We prove this by induction on the sum $n + k$.

For the base case of the induction, in which $n = 0$ or $k = 0$, the result is trivial—both $\Phi_k(n) = \Psi_k(n) = 0$. Taking $n = 1$, $k = 1$, we have certainly that $\Psi_1(1) = 2$ and that there are at most two labelings of a set with 1 element, so $\Phi_1(1) = 2$.

Now, assume that we know the result holds for all pairs (n, k) with $n + k < m$ for some $m \in \mathbb{N}$. Let $n + k = m$ and let $\mathsf{VC}(\mathcal{C}) = k$ for some collection of sets $\mathcal{C}$. Now, for $i \in \{1, \ldots, n\}$ and a set $A = \{x_1, \ldots, x_n\}$, let $A' = A \setminus \{x_1\} = \{x_2, \ldots, x_n\}$, and let $\mathcal{C}' \subset \mathcal{C}$ label A' in as many ways as possible, i.e.

$$\mathcal{C}' = \operatorname*{argmax}_{\mathcal{C}_0 \subset \mathcal{C}} \Delta_{n-1}(\mathcal{C}_0, x_2, \ldots, x_n).$$

We claim that

$$\Delta_n(\mathcal{C}, A) = \Delta_{n-1}(\mathcal{C}', A') + \Delta_{n-1}(\mathcal{C} \setminus \mathcal{C}', A').$$

Indeed, consider a binary labeling $y \in \{-1, 1\}^n$ of $x_1, \ldots, x_n$ that $\mathcal{C}$ realizes (recall definition (1)). Then either its latter $n - 1$ components are realized by $\mathcal{C}'$, or (by the maximality of $\mathcal{C}'$, they are a duplicate labeling and are realized by a unique set in $\mathcal{C} \setminus \mathcal{C}'$.

Now, of course, we have $\mathsf{VC}(\mathcal{C}') \le k$, so that $\Delta_{n-1}(\mathcal{C}', A') \le \Phi_k(n - 1) \le \Psi_k(n - 1)$ by the induction hypothesis. We claim that $\mathsf{VC}(\mathcal{C} \setminus \mathcal{C}') \le k - 1$. Indeed, if $\mathcal{C} \setminus \mathcal{C}'$ shatters a set $B \subset A'$ then $\mathcal{C}$ must shatter $B \cup \{x_1\}$, and so we must have $\mathrm{card}(B) \le k - 1$ because $\mathsf{VC}(\mathcal{C}) = k$, and $\Delta_{n-1}(\mathcal{C} \setminus \mathcal{C}', A') \le \Phi_{k-1}(n - 1) \le \Psi_{k-1}(n - 1)$, again by the induction hypothesis. Then we have

$$\Psi_k(n-1) + \Psi_{k-1}(n-1) = \sum_{i=0}^{k} \binom{n-1}{i} + \sum_{i=0}^{k-1} \binom{n-1}{i} = \sum_{i=0}^{k} \binom{n}{i},$$

which gives the result. □

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 Θ with metric d as

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

and

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

where we recall an ϵ-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 Θ with metric d,*

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

For a probability distribution P, we recall the definition of $L_r(P)$ norms on functions $f : \mathcal{X} \to \mathbb{R}$ 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 $\mathsf{VC}(\mathcal{C}) < \infty$. Then there exist universal constants $C, K < \infty$ such that for all $0 \le \epsilon < 1$*

$$N(\mathcal{C}, L_r(P), \epsilon) \le C \cdot \mathsf{VC}(\mathcal{C}) K^{\mathsf{VC}(\mathcal{C})} \left(\frac{1}{\epsilon} \right)^{r\mathsf{VC}(\mathcal{C})}.$$

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 $\mathsf{VC}(\mathcal{C}) = d < \infty$. Then for any $\tau > 0$, there exist universal constants $C, K < \infty$ (which may depend on τ) such that for all $0 \le \epsilon \le 1$*

$$N(\mathcal{C}, L_r(P), \epsilon) \le C \cdot K^{d \log d} \left(\frac{1}{\epsilon} \right)^{rd + \tau}.$$

Proof We provide the proof in three parts. First, we let $C_1, \ldots, C_N$ be a maximal $\delta = \epsilon^r$-packing in the $L_r(P)$ norm, so that for $X \sim P$ we have

$$\mathbb{E}[|1_{C_i}(X) - 1_{C_j}(X)|^r] = \mathbb{E}[|1_{C_i}(X) - 1_{C_j}(X)|] > \delta = \epsilon^r.$$

It is thus clear that $N(\mathcal{C}, \epsilon^r, L_1(P)) \ge N(\mathcal{C}, \epsilon, L_r(P))$, so we may thus focus on the L_1 case with the δ-packing. By Lemma 3.1, we thus have $N(\mathcal{C}, \delta, L_1(P)) \le N$.

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 $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P$, 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

$$P(\mathcal{E}^c) \le \sum_{i<j} P(C_i \cap \{X_1, \ldots, X_n\} = C_j \cap \{X_1, \ldots, X_n\}) < \sum_{i<j} (1 - \delta)^n = \binom{N}{2} (1 - \delta)^n, \tag{2}$$

so that the probability $P(\mathcal{E}) \geq 1 - \binom{N}{2}(1-\delta)^n$.

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

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

So the probabilistic method implies that at least some such set exists, i.e. that $\Delta_n(\mathcal{C}, x_1, \ldots, x_n) \geq N$ 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

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

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

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

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

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

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

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

Evidently taking $C \gg d2^{-\tau}$ gives the desired contradiction. We obtain the theorem when we replace δ with ϵ^r. □

References

[1] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York, 1996.