COMS 4773: VC dimension

Daniel Hsu

February 20, 2024

1 VC dimension

Let $\mathcal{F}$ be a collection of $\{-1,1\}$-valued (or $\{0,1\}$-valued) functions on a domain $\mathcal{X}$. We say a set of points in $\mathcal{X}$ is *shattered* by $\mathcal{F}$ if all possible labelings of these points are realized by functions from $\mathcal{F}$. The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{F}$ is the size of the largest set shattered by $\mathcal{F}$ if such a largest set exists; it is ∞ if sets of arbitrarily large size can be shattered by $\mathcal{F}$.

Example: linear threshold functions. Let $\mathsf{LTF}_d = \{x \mapsto \operatorname{sign}(\langle x, w\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ denote the set of linear threshold functions on $\mathcal{X} = \mathbb{R}^d$. We claim that the VC dimension of LTF_d is $d+1$.

To show the VC dimension of LTF_d is at least $d+1$, we need to exhibit $d+1$ points shattered by LTF_d. A choice that works is

$$0, e_1, \dots, e_d,$$

where e_i is the i-th standard basis vector in $\mathbb{R}^d$. Consider any labeling of these points $(y_0, y_1, \dots, y_d) \in \{-1,1\}^{d+1}$. To realize this labeling, we consider the LTF $x \mapsto \operatorname{sign}(\langle x, w\rangle + b)$ where $b = y_0$ and $w = 2(y_1 e_1 + \cdots + y_d e_d)$. Then

$$\begin{aligned} \operatorname{sign}(\langle 0, w\rangle + b) &= \operatorname{sign}(y_0) = y_0, \\ \operatorname{sign}(\langle e_i, w\rangle + b) &= \operatorname{sign}(2y_i + y_0) = \operatorname{sign}(y_i + y_0/2) = y_i \quad \text{for each } i = 1, \dots, d. \end{aligned}$$

To show VC dimension is at most $d+1$, we need to show that no $d+2$ points are shattered by LTF_d. It is a bit easier to think about this in terms of the *homogeneous* linear threshold functions $\mathsf{HLTF}_{d+1} = \{x \mapsto \operatorname{sign}(\langle x, w\rangle) : w \in \mathbb{R}^{d+1}\}$ on $\mathbb{R}^{d+1}$. Let us associate every $x \in \mathbb{R}^d$ with its "lifted" counterpart $(x, 1) \in \mathbb{R}^{d+1}$. The following is easy to show.

Lemma 1. *If some points in $\mathbb{R}^d$ are shattered by LTF_d, then the corresponding lifted points in $\mathbb{R}^{d+1}$ are shattered by HLTF_{d+1}.*

Consider any $d+2$ points in $\mathbb{R}^d$, and consider the lifted points $x_1, \dots, x_{d+2} \in \mathbb{R}^{d+1}$. These points are linearly dependent, so we can write one of them—say, x_{d+2}—as a linear combination of the others: $x_{d+2} = c_1 x_1 + \cdots + c_{d+1} x_{d+1}$. Consider the labeling $(\operatorname{sign}(c_1), \dots, \operatorname{sign}(c_{d+1}), -1)$. Suppose the HLTF $x \mapsto \operatorname{sign}(\langle x, w\rangle)$ realizes the first $d+1$ labels: $\operatorname{sign}(\langle x_i, w\rangle) = \operatorname{sign}(c_i)$. Then

$$\langle x_{d+2}, w\rangle = c_1 \langle x_1, w\rangle + \cdots + c_{d+1} \langle x_{d+1}, w\rangle \geq 0,$$

so $\operatorname{sign}(\langle x_{d+2}, w\rangle) = 1$. So it cannot realize the last label. So the lifted points are not shattered by HLTF_{d+1}, which (by Lemma 1) implies the original points are not shattered by LTF_d.

2 Sauer's lemma

Lemma 2 (Sauer's lemma)**.** *If $\mathcal{F}$ has VC dimension $d < \infty$, then a set of n points can be labeled by $\mathcal{F}$ in at most $\binom{n}{\leq d} := \binom{n}{0} + \cdots + \binom{n}{d}$ ways.*

Sauer's lemma follows from Proposition 1 below. Say a matrix $A \in \{0,1\}^{m \times n}$ has *Property $P_{n,d}$* if every submatrix formed by $k \geq d+1$ of its columns has fewer than 2^k distinct rows.

Proposition 1. *For any $n \geq 1$ and $d \geq 0$, if $A \in \{0,1\}^{m \times n}$ has Property $P_{n,d}$, then A has at most $\binom{n}{\leq d}$ distinct rows.*

Proof. By induction on n and d. The following base cases are easily verified.

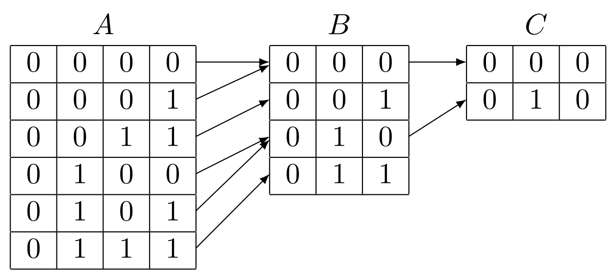
- $n = 1$ and $d = 0$: a matrix with $P_{1,0}$ has at most $1 = \binom{1}{0}$ distinct row.
- $n = 1$ and $d = 1$: a matrix with $P_{1,1}$ has at most $2 = \binom{1}{0} + \binom{1}{1}$ distinct rows.

Now we prove the inductive step. Pick any $n \geq 2$ and $d \geq 1$. Assume, as the (strong) inductive hypothesis, that for any (n', d') with $n' \leq n$, $d' \leq d$, and $n' + d' < n + d$, if a matrix has Property $P_{n',d'}$, then it has at most $\binom{n'}{\leq d'}$ distinct rows.

Consider a matrix A with Property $P_{n,d}$. We use the distinct rows of A to construct two new matrices B and C.

- Let B be the distinct rows of A after removing the n-th column.
- When removing the last column of A, some pairs of distinct rows of A got "collapsed" into the same row of B. For each such pair, put one of the rows in C (but without the n-th column).

Here is an example (with $n = 4$ and $d = 2$).



By construction, the number of distinct rows of A is equal to the number of (distinct) rows of B plus the number of (distinct) rows of C. We make two important observations:

- Matrix B has Property $P_{n-1,d}$. This is because it is obtained from A by removing the last column, and removing a column can only reduce the number of distinct rows.
- Matrix C has Property $P_{n-1,d-1}$. This is because if there was a submatrix of C formed by d columns with 2^d distinct rows, then we could find a submatrix of A formed by $d+1$ columns (one of which is the n-th column) with 2^{d+1} distinct rows, violating Property $P_{n,d}$.

Therefore, invoking the inductive hypothesis, the number of distinct rows of A is at most

$$\binom{n-1}{\leq d} + \binom{n-1}{\leq d-1} = 1 + \sum_{k=1}^{d} \binom{n-1}{k} + \binom{n-1}{k-1} = 1 + \sum_{k=1}^{d} \binom{n}{k} = \binom{n}{\leq d}. \qquad \square$$

Proof of Sauer's lemma. Take any n points, and consider the possible ways to label them by functions in $\mathcal{F}$: this yields a collection of vectors from $\{0,1\}^n$. Organize these vectors as rows in a matrix with n columns; we want to bound the number of distinct rows. Since $\mathcal{F}$ has VC dimension $d < \infty$, this matrix has Property $P_{n,d}$, so Sauer's lemma follows from Proposition 1. $\square$

3 Lower bound in terms of VC dimension

Proposition 2. *Suppose $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ has VC dimension $d < \infty$. Every PAC learner for $\mathcal{H}$ requires a sample size of at least $\Omega(d/\epsilon)$ to guarantee error rate $\leq \epsilon$ with probability at least $3/4$.*

Proof. Let $x_1, \ldots, x_d$ be d points shattered by $\mathcal{H}$, so all possible labelings of these points can be realized by hypotheses from $\mathcal{H}$. Let $Y_1, \ldots, Y_d$ be labels for $x_1, \ldots, x_d$ drawn uniformly at random from $\{-1,1\}^d$ (which corresponds to a random choice of the target hypothesis from $\mathcal{H}$). Let μ be the probability distribution with mass $4\epsilon/(d-1)$ on each of $x_1, \ldots, x_{d-1}$, and mass $1 - 4\epsilon$ on x_d. Suppose S is n points drawn iid from μ, with

$$n \leq \frac{d-1}{16\epsilon}.$$

Let N be the number of points among $x_1, \ldots, x_{d-1}$ that appear in S. Then

$$\mathbb{E}[N] = (d-1)\left(1 - \left(1 - \frac{4\epsilon}{d-1}\right)^n\right) \leq 4\epsilon n,$$

so by Markov's inequality,

$$\Pr(N \geq 8\epsilon n) \leq \frac{1}{2}.$$

So, with probability at least $1/2$, (the labels of) more than half of the points $x_1, \ldots, x_{d-1}$ are not seen by the learner. Without loss of generality, let's say it is $Y_1, \ldots, Y_m$ (with $m > (d-1)/2$) that are not seen by the learner. If H is the hypothesis returned by the learner, then H is independent of $Y_1, \ldots, Y_m$. Let

$$W = |\{i \in [m] : H(x_i) \neq Y_i\}|$$

be the number of mistakes committed by H on these m points. Then W follows the Binomial$(m, 1/2)$ distribution, which has $m/2$ as a median, so

$$\Pr\left(W \geq \frac{m}{2}\right) \geq \frac{1}{2}.$$

So with probability at least $1/2 \times 1/2 = 1/4$ (over both the choice of the target hypothesis and the labeled data provided to the learner), the hypothesis returned by the learner has error rate at least

$$\frac{4\epsilon}{d-1} \cdot \frac{m}{2} > \epsilon.$$

Therefore, there exists a target function $h^* \in \mathcal{H}$ such that, with probability at least $1/4$, the learner returns a hypothesis with error rate $> \epsilon$. $\square$