CS272 - Theoretical Foundations of Learning, Decisions, and Games

Lecture 4: Statistical Learning III

September 9, 2025

Lecturer: Nika Haghtalab *Readings: Ch 6, UML*
Scribe: Youngmin Park

Recall that for an algorithm $\mathcal{A}$ that learns a concept $\mathcal{C}$ in the consistency model learns $\mathcal{C}$ in the PAC model using sample complexity that satisfies

$$m \geq \frac{2}{\epsilon}\left(\log_2(\Pi_{\mathcal{C}}(2m)) + \log_2\left(\frac{1}{\delta}\right)\right) \tag{1}$$

Reasoning about the inequality is challenging as both sides depend on m. For example, when $\Pi_{\mathcal{C}}(m) = 2^m$, no such m satisfies 1.

We now define an important combinatorial notion that will help to bound $\Pi_{\mathcal{C}}(m)$ and significantly simplify 1.

1 Vapnik–Chervonenkis Dimension

Definition 1.1. A set $S \in \mathcal{X}^m$ is shattered by class $\mathcal{C}$ if $|\mathcal{C}[S]| = 2^{|S|} = 2^m$. That is, for any labeling $y_1, \ldots, y_m \in \{0, 1\}$, there is $c \in \mathcal{C}$ such that $c(x_i) = y_i$ for all $x_i \in S$.

Definition 1.2. The Vapnik–Chervonenkis (VC) dimension of $\mathcal{C}$, denoted by $\mathrm{VCDim}(\mathcal{C})$, is the size of the largest S that can be shattered by $\mathcal{C}$.

Note that in order to show $\mathrm{VCDim}(\mathcal{C}) = d$, we have to prove that

- there exists a set $S = \{x_1, \ldots, x_d\}$ that is shattered by $\mathcal{C}$
- there is no set of size $\geq d + 1$ that can be shattered by $\mathcal{C}$.

Next, we compute the VC dimension of concept classes we have seen in previous lectures.

Example 1.3. *(1-dimensional intervals)*

Let $\mathcal{X} = \mathbb{R}$, $\mathcal{C}_1 = \{c_{ab} \colon a, b \in \mathbb{R}\}$. $c_{ab}(x) = \mathbb{1}\{a \leq x \leq b\}$. For $S = \{-1, 1\}$, we can shatter S as

	x_1	x_2
$c_{-1,4}$	1	1
$c_{4,5}$	0	0
$c_{-1,0}$	1	0
$c_{1,3}$	0	1

However, we cannot shatter any set of three points. Without loss of generality, $x_1 \le x_2 \le x_3$. If $c_{ab}(x_1) = 1$ and $c_{ab}(x_3) = 1$, then $c_{ab}(x_2) = 1$, so assigning labels $(1, 0, 1)$ is not possible. Thus, $\mathrm{VCDim}(\mathcal{C}_1) = 2$.

Example 1.4. *(Axis-aligned rectangles)*

Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{C}_2 = \{c_{a_1,b_1,a_2,b_2} \mid a_1, b_1, a_2, b_2 \in \mathbb{R}\}$, $c_{a_1,b_1,a_2,b_2}(x) = \mathbb{1}\{a_1 \le x_1 \le b_1\}\mathbb{1}\{a_2 \le x_2 \le b_2\}$. We can show that $\mathrm{VCDim}(\mathcal{C}_2) \ge 4$ as we can shatter four vertices of a rhombus:

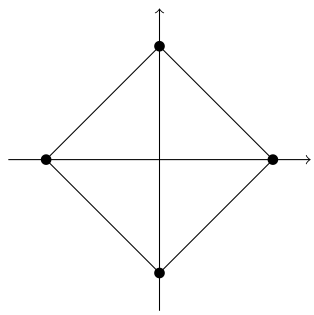

Figure 1: A set of four points that can be shattered

However, no five points can be shattered. Without loss of generality, let x_1 be the leftmost point, x_2 the rightmost point, x_3 the topmost point, and x_4 the bottom most point. Note that while $x_1, \ldots, x_4$ might not be distinct, there is a remaining point x_5 that is not any of these points.

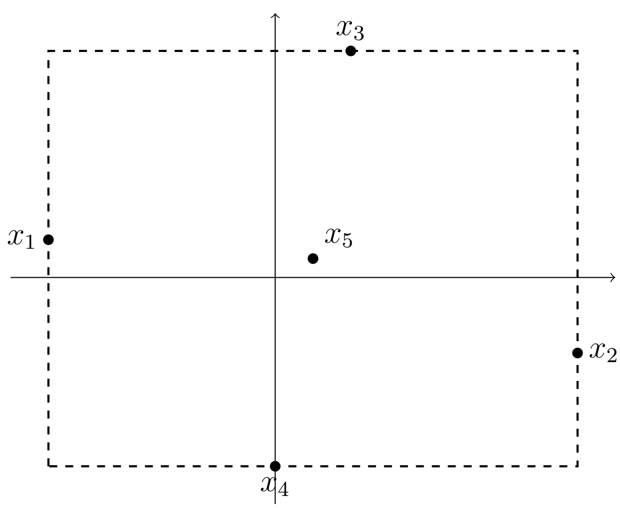

Figure 2: A set of five points cannot be shattered

Note that x_5 is in the bounding box of $\{x_1, \ldots, x_4\}$; the set of extreme points, so the label $(1, 1, 1, 1, 0)$ is not possible. Therefore, the VC dimension of $\mathcal{C}_2$ is 4.

2 Relationship between Growth function and VC dimension

Recall that in previous lectures, we prove that the growth functions for the concept classes above are

$$\Pi_{\mathcal{C}_1}(m) \le \binom{m}{0} + \binom{m}{1} + \binom{m}{2} = O(m^2)$$
$$\Pi_{\mathcal{C}_2}(m) \le (\Pi_{\mathcal{C}_1}(m))^2 = O(m^4).$$

In these examples, $\Pi_{\mathcal{C}}(m)$ is $O(m^d)$, where d is the VC dimension. We can actually bound $\Pi_{\mathcal{C}}(m)$ using VC dimension in general using the following lemma.

Lemma 2.1. *(Sauer-Shelah) A hypothesis class $\mathcal{C}$ with $\mathrm{VCDim}(\mathcal{C}) = d$ satisfies*

$$\Pi_{\mathcal{C}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$$

Proof. The proof involves using the following combinatorial facts:

Fact 2.2. $\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}$

Fact 2.3. $\binom{m}{k} = 0$ *if* $k < 0$ *or* $k > m$.

We will use induction on $m + d$. For ease of notation, let $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$.

Base case 1) $m = 0, d \ge 0$.

$$LHS = \Pi_{\mathcal{C}}(0) = 1$$

and simplifying RHS using Fact 2.3, we have

$$RHS = \Phi_d(0) = \sum_{i=0}^{d} \binom{0}{i} = \binom{0}{0} = 1$$

so it holds for $(0, d)$.

Base case 2) $m \ge 0, d = 0$. Since the VC dimension of $\mathcal{C}$ is 0, $\forall x \in \mathcal{X}$, x cannot be mapped to both 1 and 0 by elements of $\mathcal{C}$, that is, for any $x \in \mathcal{X}$, it has only one label across all $c \in \mathcal{C}$. This implies that LHS is

$$LHS = \Pi_{\mathcal{C}}(m) = 1$$

and RHS is

$$RHS = \Phi_0(m) = \binom{m}{0} = 1$$

so it holds for $(m, 0)$.

Inductive Step. Assume that the lemma holds for any m',d' such that $m' + d' < m + d$. We want to show that $\Pi_{\mathcal{C}}(m) = |\mathcal{C}[S]| \leq \Phi_d(m)$. We construct two hypothesis classes to use our inductive hypothesis. Let $S = \{x_1, \ldots, x_m\}$ be an arbitrary set, and $S' = \{x_1, \ldots, x_{m-1}\}$ be the domain of $\mathcal{C}_1$, $\mathcal{C}_2$. We consider the labeling produced by $h \in \mathcal{C}$, only considering the unique labelings in $\mathcal{C}[S]$. We will define two types for functions in $\mathcal{C}$.

- Pairs: $h, h' \in \mathcal{C}$ are pairs if $\forall i \in [m-1]$, $h(x_i) = h'(x_i)$, and $h(x_m) \neq h'(x_m)$. Define g on $x_1, \ldots, x_{m-1}$ with $g(x_i) = h(x_i) = h'(x_i)$. We add g to $\mathcal{C}_1$ and $\mathcal{C}_2$.
- Singleton: h with no h' that satisfies the pair condition. Define g on $x_1, \ldots, x_{m-1}$ with $g(x_i) = h(x_i)$, and add only to $\mathcal{C}_1$

Fact 2.4. $|\mathcal{C}_1| + |\mathcal{C}_2| = |\mathcal{C}[S]|$

Claim 2.5. $\mathrm{VCDim}(\mathcal{C}_1) \leq \mathrm{VCDim}(\mathcal{C})$

Proof. If a set $T \subseteq \{x_1, \ldots, x_{m-1}\}$ is shattered by $\mathcal{C}_1$, T is also shattered by $\mathcal{C}$. Just from the definition of $\mathcal{C}_1$ as every labeling produced by $\mathcal{C}_1$ on T has a labeling in $\mathcal{C}[S]$ which has the same label on $x_1, \ldots, x_{m-1}$. □

Claim 2.6. $1 + \mathrm{VCDim}(\mathcal{C}_2) \leq \mathrm{VCDim}(\mathcal{C})$

Proof. We claim that if $T \subseteq \{x_1, \ldots, x_{m-1}\}$ is shattered by $\mathcal{C}_2$, then $T \cup \{x_m\}$ is also shattered by $\mathcal{C}$. Every labeling in $\mathcal{C}_2$ corresponds to pairs (h, h'), where $h(x_m) \neq h'(x_m)$, while $h(x_i) = h'(x_i)$ for $x_i \in T$. So $T \cup \{x_m\}$ can be labeled all possible ways. □

By induction, we have

$$
\begin{aligned}
|\mathcal{C}_1| &\leq \Pi_{\mathcal{C}_1}(m-1) \leq \Phi_d(m-1) \\
|\mathcal{C}_2| &\leq \Pi_{\mathcal{C}_2}(m-1) \leq \Phi_{d-1}(m-1)
\end{aligned}
$$

so

$$
\begin{aligned}
\mathcal{C}[S] &= |\mathcal{C}_1| + |\mathcal{C}_2| \\
&\leq \Phi_d(m-1) + \Phi_{d-1}(m-1) \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=1}^{d-1} \binom{m-1}{i-1} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1}
\end{aligned}
$$

$$= \sum_{i=0}^{d} \binom{m}{i} = \Phi_d(m)$$

where in the fourth line, we shift the index by one, and the fifth line holds from 2.3.

□

Using the lemma we can also bound the growth function asymptotically.

Corollary 2.7.

$$\Pi_{\mathcal{C}}(m) \leq \left(\frac{em}{d}\right)^d$$

where e is exponent of natural log.

Proof. Sketch - use Stirling's approximation. □

We can now replace $\Pi_{\mathcal{C}}(2m)$ on the right hand side in 1.

Theorem 2.8. *Any algorithm $\mathcal{A}$ that learns $\mathcal{C}$ in the consistency model learns $\mathcal{C}$ in the PAC model with sample complexity*

$$m_{\epsilon,\delta} = C\left(\frac{1}{\epsilon}\left(d \ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right), \textit{ for some constant } C$$

Proof. Sketch - 1 + 2.7 + rearranging □

In fact, it is possible to have a more efficient sample complexity, albeit not for all algorithms.

Theorem 2.9. *(Hanneke [2016]) There exists an algorithm that learns $\mathcal{C}$ in the PAC model with sample complexity*

$$m_{\epsilon,\delta} = C_1\left(\frac{1}{\epsilon}\left(d + \ln\left(\frac{1}{\delta}\right)\right)\right), \textit{ for some constant } C_1$$

and this bound is tight.

References

Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016. URL `http://jmlr.org/papers/v17/15-389.html`.