CSCI699: Theory of Machine Learning Fall 2021

Lecture 5: VC Dimension

Instructor: Vatsal Sharan *Scribe: Berk Tinaz*

Today

- We will talk about the VC dimension.

Recap: In previous classes, we showed that for hypothesis classes $\mathcal{H}$ with finite size $|\mathcal{H}|$,

- $\mathcal{H}$ is PAC learnable with $n_{\mathcal{H}}(\epsilon, \delta) = \mathcal{O}\left(\frac{\log(|\mathcal{H}|/\delta)}{\epsilon}\right)$ samples (with the realizability assumption).
- $\mathcal{H}$ is agnostic-PAC learnable with $n_{\mathcal{H}}(\epsilon, \delta) = \mathcal{O}\left(\frac{\log(|\mathcal{H}|/\delta)}{\epsilon^2}\right)$ samples.

What about when size of $\mathcal{H}$ is infinite?

- Discretization: One way to handle infinite size is by discretizing it.
  - Think about a linear classifier in $\mathbb{R}^d$. For a 32-bit system $\implies$ $(2^{32})^d$ possible classifiers which is large but finite!
- VC dimension gives a nice way to handle ∞ classes.

1 Vapnik-Chervonenkis (VC) Dimension

Shattering

Definition 1 (Restriction & Shattering)**.** *The **restriction** of a hypothesis class $\mathcal{H}$ to a set of examples $C = \{c_1, \ldots, c_n\} \in \mathcal{X}$ is a subset of $\{0,1\}^{|C|}$, given by $\mathcal{H}_C = \{(h(c_1), \ldots, h(c_n)), \forall h \in \mathcal{H}\}$. We say that $\mathcal{H}$ **shatters** C if $|\mathcal{H}_C| = 2^{|C|}$.*

Basically, shattering indicates that all possible labelings are realized when $\mathcal{H}$ labels the set C.

Corollary 2 (of No Free-lunch Theorem)**.** *Let $\mathcal{H}$ be a hypothesis class and assume there exists a set $C \subseteq \mathcal{X}$ of size $2n$ such that $\mathcal{H}$ shatters C. Then, $\exists$ a distribution D over $\mathcal{X} \times \{0,1\}$ and a predictor $h^* \in \mathcal{H}$ such that $R(h^*) = 0$, but for any learning algorithm A, $\mathbb{P}_{S \sim D^n}[R(A(S)) \geq 1/8] \geq 1/7$.*

In short, if $\mathcal{H}$ shatters a set of size $2n$ then one cannot learn with just n examples. Can we do something if C is such that $|\mathcal{H}_C| \ll 2^{|C|}$?

Idea: For any distribution supported on C, the real hypothesis space under consideration is actually $\mathcal{H}_C$. Moreover, because of the construction, $\mathcal{H}_C$ is finite. Therefore, if $|\mathcal{H}_C|$ is small, then maybe one can learn.

Definition 3 (VC Dimension)**.** *The VC dimension of a hypothesis class* $\mathcal{H}$*, denoted by* $\text{VCdim}(\mathcal{H})$ *is the size of the largest set* $C \subseteq \mathcal{X}$ *that can be shattered by* $\mathcal{H}$*. If* $\mathcal{H}$ *can shatter sets of arbitrary size, then* $\text{VCdim}(\mathcal{H}) = \infty$*.*

How to show $\text{VCdim}(\mathcal{H}) = d$?

1. Check if there exists some set C of size d that can be shattered by $\mathcal{H}$.

2. Check that no set of size $d+1$ is shattered by $\mathcal{H}$.

Examples

- Example 1 (Threshold functions): Let $x = [0,1]$, $\mathcal{H} = \{h_\delta(x) = \mathbb{1}(x \geq \delta), \delta \in [0,1]\}$. $\mathcal{H}$ are set of thresholds in $\mathbb{R}$.

  Claim: $\text{VCdim}(\mathcal{H}) = 1$.

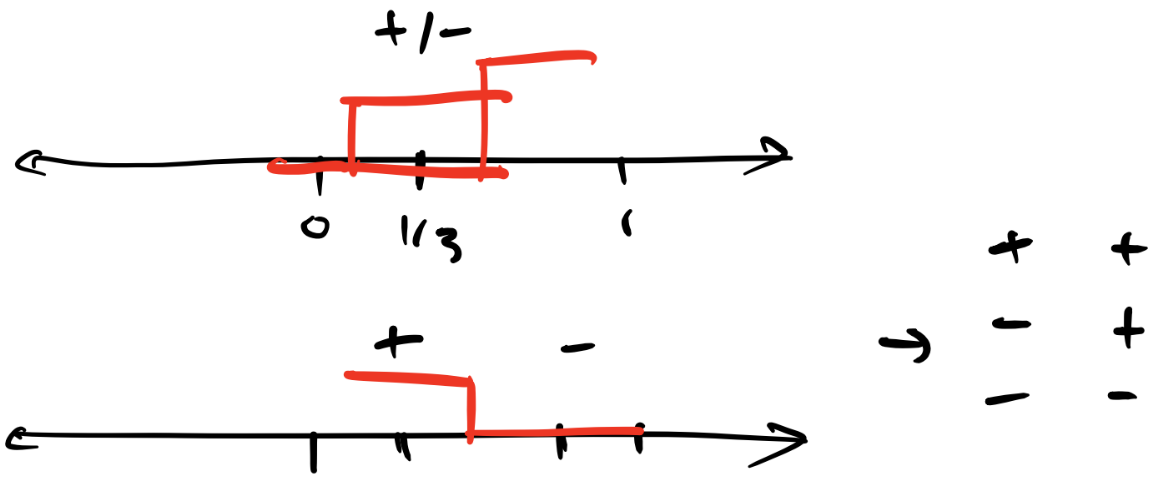

Figure 1: Setup in the first row is used to show that $\mathcal{H}$ can shatter a set C of size 1. Setup in the second row is used to show that $\text{VCdim}(\mathcal{H}) < 2$. To realize all possible labelings for $|C| = 2$, one requires reverse thresholds.

To verify the claim, we will use the 2-step approach depicted above. As a first step, we will check if there is a set C of size 1 that can be shattered by $\mathcal{H}$. Select any point x, e.g. $x = 1/3$. We can see that for $\delta \leq 1/3$, $h_\delta(x) = 1$ and similarly for $\delta > 1/3$, $h_\delta(x) = 0$. Hence, all possible labeling are realized for $|C| = 1$. For visualization, refer to first row of Figure 1.

Now we have to check that $\mathcal{H}$ can't shatter any set C with $|C| = 2$. To see this, pick two points $x_1 = a$ and $x_2 = b$ such that $a, b \in [0,1]$ and without loss of generality (w.l.o.g.) assume $a < b$. Then, for $\delta \leq a$ we have $h_\delta(a) = h_\delta(b) = 1$. For, $a < \delta \leq b$ we have $h_\delta(a) = 0, h_\delta(b) = 1$. Finally, for $b < \delta$ we have $h_\delta(a) = h_\delta(b) = 0$. Notice that with this hypothesis class $\mathcal{H}$, we can't get the labeling $h_\delta(a) = 1, h_\delta(b) = 0$ for any $\delta \in [0,1]$ (which requires a reverse threshold as can be seen in second row of Figure 1). Therefore, $\mathcal{H}$ does not shatter C with $|C| = 2$. Hence, claim is proven.

If we also allow reverse thresholds, i.e. $\mathbb{1}(x < \delta)$, then one can show that $\text{VCdim}(\mathcal{H}) = 2$.

- Example 2 (Axis-aligned rectangles): Let $\mathcal{X} = \mathbb{R}^2$ and define,

$$\mathcal{H}_{a_1,a_2,b_1,b_2}(x_1, x_2) = \mathbb{1}\left(a_1 \le x_1 \le b_1 \;\&\; a_2 \le x_2 \le b_2\right)$$

  Claim: $\text{VCdim}(\mathcal{H}) = 4$.

  Similar to the previous example, let us first show that there is a set C of size 4 that can be shattered by $\mathcal{H}$. Consider the points in the first row of Figure 2 (points organized in diamond shape). Notice that we can enclose any subset of these points with a rectangle. Therefore, all labelings can be realized with $\mathcal{H}$.

  To see that $\mathcal{H}$ cannot shatter any set C with size 5, consider the case in the second row of Figure 2. Pick 5 distinct points and label left-most point c_1, right-most point c_2, bottom-most point c_3, and top-most point c_4. Last point c_5 can be anywhere in the tightest rectangle fitted to the first 4 points. We would like to label first 4 points 1 but the last point 0. For the first 4 points to be labeled 1, $\mathcal{H}$ must enclose them with the rectangle. However, due to construction, c_5 must also be in that rectangle which means it can't be labeled 0. Therefore, desired labeling cannot be realized. Hence, $\text{VCdim}(\mathcal{H}) < 5$ which proves the claim.

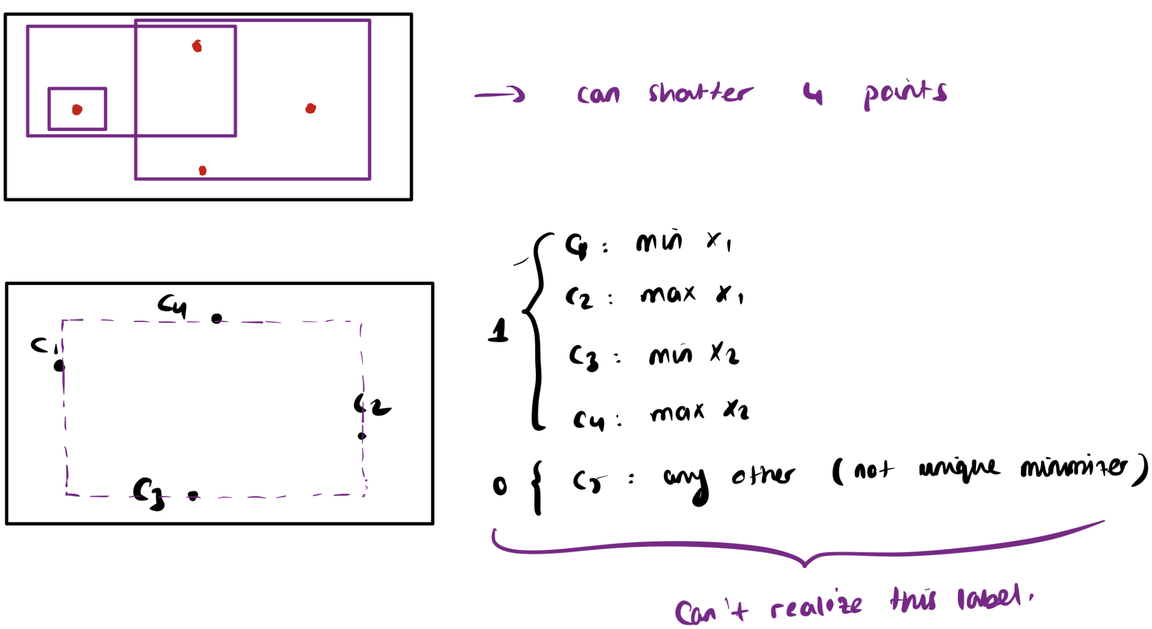

Figure 2: Setup in the first row is used to show that $\mathcal{H}$ can shatter a set C of size 4. Setup in the second row is used to show that $\text{VCdim}(\mathcal{H}) < 5$.

- Example 3 (Finite classes): For any finite hypothesis class $\mathcal{H}$, we have $\text{VCdim}(\mathcal{H}) \le \log(|\mathcal{H}|)$. This is because for any set C, $|\mathcal{H}_C| \le |\mathcal{H}|$. Therefore, if $2^{|C|} > |\mathcal{H}|$, then we cannot shatter C.

2 VC Theorem

Theorem 4 (VC Theorem)**.** *Let $\mathcal{H}$ be a hypothesis class with* $\text{VCdim}(\mathcal{H}) = d < \infty$*. Then there is an absolute constant $c > 0$ such that $\mathcal{H}$ has uniform convergence property with,*

$$n_{\mathcal{H}}^{VC}(\epsilon, \delta) = c \cdot \frac{d \cdot log(d/\epsilon) + log(1/\epsilon)}{\epsilon^2}$$

Corollary 5. *$\mathcal{H}$ is agnostic-PAC learnable with $\mathcal{O}\left(\frac{d \cdot log(d/\epsilon) + log(1/\epsilon)}{\epsilon^2}\right)$ samples.*

Note:

(1) It is also possible to show that $n_{\mathcal{H}}^{VC}(\epsilon, \delta) \leq c \cdot \frac{d + log(1/\delta)}{\epsilon^2}$. For $d = log(|\mathcal{H}|)$, this bound reduces to $c \cdot \frac{log(|\mathcal{H}|/\delta)}{\epsilon^2}$ which boils down to the $\mathcal{O}\left(\frac{log(|\mathcal{H}|/\delta)}{\epsilon^2}\right)$ sample complexity that we derived earlier for agnostic-PAC learning.

(2) The result above is for binary classification with 0/1 loss. Later, we will see how to generalize to other losses beyond 0/1 loss.

Proof Outline:

1) For any set $C \subseteq \mathcal{X}$, effective size of restriction of $\mathcal{H}$ on C ($\mathcal{H}_C$) is approximately $|C|^d$ ($|\mathcal{H}_C| \approx |C|^d$).

2) We want small "effective size" which will be good when we are using union bound to get VC result.

Step 1: Polynomial growth of $\mathcal{H}_C$

Definition 6 (Growth function)**.** *The growth function of $\mathcal{H}$, $T_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as*

$$T_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}, |C| = n} |\mathcal{H}_C|$$

If $\text{VCdim}(\mathcal{H}) = d$, then $T_{\mathcal{H}}(n) = 2^n, \forall n \leq d$. Sauer's Lemma gives a good upper bound $\forall n > d$.

Lemma 7 (Sauer's Lemma)**.** $\forall n, \text{VCdim}(\mathcal{H}) = d$,

$$T_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

For $n > d + 1$, this implies:

$$T_{\mathcal{H}}(n) \leq \left(\frac{n \cdot e}{d}\right)^d \quad \textit{\textbf{(exponential to polynomial regime)}}$$

Proof. (of Step 1). We will instead show a stronger inequality. For any $C = \{c_1, \ldots, c_n\}$ & any $\mathcal{H}$,

$$|\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}| \tag{1}$$

This is sufficient since if VCdim($\mathcal{H}$) $= d$, $\mathcal{H}$ cannot shatter any set B of size $|B| > d$. There are $\binom{n}{i}$ subsets of size i, hence, we will get our bound.

We will prove (1) by induction.

Base Step ($n = 1$): We have either,

1) $|\mathcal{H}_C| = 2^0 = 1$. Then, $LHS = RHS$ in (1) since one labeling shatters $\{\emptyset\}$.
2) $|\mathcal{H}_C| = 2^1 = 2$. Then, again $LHS = RHS$ as two labelings shatter $\{\{\emptyset\}, \{c_1\}\}$.

Induction Step: Assume that (1) holds for all sets of size $k < n$. Let $C = \{c_1, \ldots, c_n\}$ & $C' = \{c_2, \ldots, c_n\}$. Define,

$$Y_0 = \{(y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathcal{H}_C \textit{ or } (1, y_2, \ldots, y_n) \in \mathcal{H}_C\}$$
$$Y_1 = \{(y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathcal{H}_C \textit{ and } (1, y_2, \ldots, y_n) \in \mathcal{H}_C\}$$

Claim: $|\mathcal{H}_C| = |Y_0| + |Y_1|$. This is true because, $(y_2, \ldots, y_n)$ is counted once in Y_0, but counted again in Y_1 if can be shattered.

By induction we get,

$$|Y_0| \leq \left|\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}\right| = |\{B \subseteq C : c_1 \notin B \textit{ and } \mathcal{H} \text{ shatters } B\}|$$

for Y_1, define $\mathcal{H}' \subseteq \mathcal{H}$ to be:

$$\mathcal{H}' = \left\{h \in \mathcal{H}, \exists h' \in \mathcal{H} \textit{ such that } ((1 - h'(c_1), h'(c_2), \ldots, h'(c_n)) = (h(c_1), h(c_2), \ldots, h(c_n))\right\}$$

$\mathcal{H}'$ is the set of hypothesis for which that hypothesis that agrees everywhere in C except c_1 is also in $\mathcal{H}$.

Note:

1) If $\mathcal{H}'$ shatters $B \subseteq C'$ then it also shatters $B \cup \{c_1\}$.
2) $Y_1 = \mathcal{H}'_{C'}$

Then,

$$\begin{aligned}
|Y_1| = |\mathcal{H}'_{C'}| &\leq \left|\{B \subseteq C' : \mathcal{H}' \textit{ shatters } B\}\right| \quad \text{(By induction hypothesis (1))} \\
&= \left|\left\{B \subseteq C' : \mathcal{H}' \textit{ shatters } B \cup \{c_1\}\right\}\right| \\
&= \left|\left\{B \subseteq C : c_1 \in B \textit{ and } \mathcal{H}' \textit{ shatters } B\right\}\right| \\
&\leq |\{B \subseteq C : c_1 \in B \textit{ and } \mathcal{H} \textit{ shatters } B\}|
\end{aligned}$$

From previous claim:

$$
\begin{aligned}
|\mathcal{H}_C| &= |Y_0| + |Y_1| \\
&\leq |\{B \subseteq C : c_1 \notin B \textit{ and } \mathcal{H} \textit{ shatters } B\}| + |\{B \subseteq C : c_1 \in B \textit{ and } \mathcal{H} \textit{ shatters } B\}| \\
&= |\{B \subseteq C : \mathcal{H} \textit{ shatters } B\}|
\end{aligned}
$$

which completes our proof. □

Step 2: Symmetrization

Lemma 8. *For a class $\mathcal{H}$ with growth function $T_{\mathcal{H}}$,*

$$\mathbb{E}_{S\sim D^n}\left[\sup_{h\in\mathcal{H}} |R(h) - \hat{R}_S(h)|\right] \leq \sqrt{\frac{2\cdot\log(2\cdot T_{\mathcal{H}}(2n))}{n}}$$

Once we have this, we can use Markov's inequality to get a high probability statement such as:

$$\Pr\left[\sup_{h\in\mathcal{H}} |R(h) - \hat{R}_S(h)| > t\right] \leq \frac{\sqrt{\frac{2\cdot\log(2\cdot T_{\mathcal{H}}(2n))}{n}}}{t}$$

But instead, we will use McDiarmid's inequality to get an even better bound.