CS861: Theoretical Foundations of Machine Learning Lecture 5 - 09/13/2023
University of Wisconsin–Madison, Fall 2024

Lecture 05: VC Dimension

Lecturer: Kirthevasan Kandasamy *Scribed by: Travis Pence, Alejandro Calle-Saldarriaga*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this lecture, we will continue our introduction to VC dimension by defining the growth function, shattering, and the VC dimension itself. The VC dimension is then used in a bound about Rademacher complexity.

1 Growth Function

From last lecture, we introduced the definitions

$$\begin{aligned}\text{Let } A &= \{x_1, \ldots, x_n\},\\ \mathcal{H}(A) &\triangleq [h(x_1), \ldots, h(x_n)]|h \in \mathcal{H}\}\\ \text{Let } S &= \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}\\ \mathcal{L}(S, \mathcal{H}) &\triangleq \{[\ell(h(x_1), y_1), \ldots, \ell(h(x_n), y_n)]|h \in \mathcal{H}\}\end{aligned}$$

The following example was then stated

Example 1.
Let $\mathcal{H} = \{h_a(x) = \mathbb{1}_{x \geq a}|h \in \mathcal{H}\}$ be all one-sided threshold classifiers. Write out $\mathcal{H}(A)$ for:

$$\begin{aligned}A_1 &= \{x_1 = -1, x_2 = 1\}\\ A_2 &= \{x_1 = 0, x_2 = 0\}\end{aligned}$$

Write out $\mathcal{L}(S, \mathcal{H})$ for:

$$\begin{aligned}S_1 &= \{(x_1 = -1, y_1 = 0), (x_2 = 1, y_2 = 1)\}\\ S_2 &= \{(x_1 = 0, y_1 = 0), (x_2 = 0, y_2 = 1)\}\end{aligned}$$

Note that the x values in S_i match those in A_i. At the beginning of this lecture, we gave the answers, which are

$$\begin{aligned}\mathcal{H}(A_1) &= \{[1, 1], [0, 1], [0, 0]\}\\ \mathcal{H}(A_2) &= \{[1, 1], [0, 0]\}\\ \mathcal{L}(S_1, \mathcal{H}) &= \{[1, 0], [0, 0], [0, 1]\}\\ \mathcal{L}(S_2, \mathcal{H}) &= \{[0, 1], [1, 0]\}\end{aligned}$$

If you look at the above answers, you will notice that the cardinality of $\mathcal{H}(A_i)$ and $L(S_i, \mathcal{H})$ are the same. This is not a coincidence, and is stated below. From now on, $\mathcal{L}(S, \mathcal{H})$ uses the $0-1$ loss.

Claim 1. *For any $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$, let $S_{\mathcal{X}} = \{x_1, \dots, x_n\}$. Then,*

$$|\mathcal{L}(S, \mathcal{H})| = |\mathcal{H}(S_{\mathcal{X}})|.$$

While we did not prove this rigorously, we gave a proof sketch. There exists a bijection between $\mathcal{L}(S, \mathcal{H})$ and $\mathcal{H}(S_{\mathcal{X}})$, which would imply the result. The intuition behind the bijection is that each element in $\mathcal{H}(A)$ will compare differently with the $y's$, and so will map to a unique element in $\mathcal{L}(S, \mathcal{H})$. Now, we define the growth function.

Definition 1. *Let $n \in \mathbb{N}$ and $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ be a hypothesis class. Then, the* ***growth function*** *is*

$$g(n, \mathcal{H}) := \max_{A \subseteq \mathcal{X}, |A|=n} |\mathcal{H}(A)| = \max_{S \subseteq \mathcal{X} \times \mathcal{Y}, |S|=n} |\mathcal{L}(S, \mathcal{H})|.$$

To interpret this definition, the growth function is equal to the maximum number of label vectors that can be realized on a set of n points in $\mathcal{X}$, or loss vectors with a dataset of n points in $\mathcal{X} \times \mathcal{Y}$, via hypothesis in $\mathcal{X}$. Note that $g(n, \mathcal{H}) \leq 2^n$ as we are using $0-1$ loss. Below is an example of the growth function calculation.

Example 2. Consider as our hypothesis class two-sided threshold classifiers, that is

$$\mathcal{H} = \{h_a(x) = \mathbb{1}_{\{x \geq a\}} | a \in \mathbb{R}\} \cup \{h_a(x) = \mathbb{1}_{\{x \leq a\}} | a \in \mathbb{R}\}$$

From Example 1, we have that $g(1, \mathcal{H}) = 2$ for the one-sided classifiers. As we are dealing with two-sided classifiers, a superset of the one-sided, the growth function of the two-sided must be at least the growth function of the one-sided. But the growth function for the one-sided is the maximum, so we have that $g(1, \mathcal{H}) = 2$.

$g(2, \mathcal{H}) = 4 = 2^2$ as for $x_1 \neq x_2$, we have

$$\mathcal{H}(\{x_1, x_2\}) = \{[0, 0], [0, 1], [1, 1], [1, 0]\}.$$

$g(3, \mathcal{H}) = 6 < 2^3$ as for $x_1 \neq x_2 \neq x_3$, we have

$$\mathcal{H}(\{x_1, x_2, x_3\}) = \{[1, 1, 1], [1, 1, 0], [1, 0, 0], [0, 0, 0], [0, 0, 1], [0, 1, 1]\}$$

Loosely speaking, $g(n, \mathcal{H}) = 2^n$ as long as $n \leq$ degrees of freedom of $\mathcal{H}$, which can be observed in the two examples above. For one-sided classifiers, we only have one degree of freedom, the choice of a. For two-sided classifiers, we have two degrees of freedom, the choice of a and the direction of the inequality. Recall that our purpose for studying the VC dimension is to bound the Rademacher complexity via the growth function. First however, we must bound the empirical Rademacher complexity.

Lemma 1 (Massart's Lemma)**.** *Let $S = \{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$, and $\mathcal{H}$ be a hypothesis class. Then,*

$$\widehat{\text{Rad}}(S, \mathcal{H}) \leq \sqrt{\frac{2 \log(|\mathcal{L}(S, \mathcal{H})|)}{n}}.$$

Proof

The idea of this proof is to use the maximal inequality we showed last class, so expanding the LHS

$$\begin{aligned}\widehat{\text{Rad}}(S, \mathcal{H}) &= \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i R(h(x_i), y_i)\right] \\ &= \frac{1}{n} \mathbb{E}_{\sigma}\left[\max_{v \in \mathcal{L}(S, \mathcal{H})} \sum_{i=1}^{n} \sigma_i v_i\right]\end{aligned}$$

In the second equality, instead of looking for a supremum over $\mathcal{H}$, we can look at the max over the set of possible realizations of the loss function for this $\mathcal{H}$ with dataset S. Now we have some random variables, $Z_v = \sum_{i=1}^{n} \sigma_i v_i$, that we want to use the maximal inequality for. Note that we must show that they are sub-Gaussian in order to use the result. Noting that $\sigma_i v_i \in [-1, 1]$ so then $\sigma_i v_i$ is 1-sub-Gaussian. As $\{\sigma_i v_i\}_{i\in[n]}$ are independent, we have their sum $\sum_{i=1}^{n} \sigma_i v_i$ is $\sqrt{n}$-sub-Gaussian. Therefore, by the maximal inequality

$$\mathbb{E}_\sigma \left[\max_{v \in \mathcal{L}(S,\mathcal{H})} \sum_{i=1}^{n} \sigma_i v_i \right] \leq \sqrt{n}\sqrt{2\log(|\mathcal{L}(S,\mathcal{H})|)}$$

So,

$$\begin{aligned} \widehat{\text{Rad}}(S,\mathcal{H}) &\leq \frac{1}{n}\sqrt{n}\sqrt{2\log(|\mathcal{L}(S,\mathcal{H})|)} \\ &= \sqrt{\frac{2\log(|\mathcal{L}(S,\mathcal{H})|)}{n}} \end{aligned}$$

□

We can use this result to bound the Rademacher complexity.

Corollary 1. $Rad_n(\mathcal{H}) \leq \sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}}$

Proof By the definition of the growth function, $\mathcal{L}(S,\mathcal{H}) \leq g(n,\mathcal{H}), \forall S$ such that $|S| = n$. Also, as $Rad_n(\mathcal{H}) = \mathbb{E}_{S\sim P_{\mathcal{X}\times\mathcal{Y}}}[\widehat{\text{Rad}}(S,\mathcal{H})]$, we are done by applying expectation at both sides of the inequality. □

From our previous theorem, with probability $1-\delta$ we have

$$\begin{aligned} R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h) &\leq C_1 Rad_n(\mathcal{H}) + O\left(\sqrt{\frac{1}{n}\log\left(\frac{1}{\delta}\right)}\right) \\ &\leq C_1\sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}} + O\left(\sqrt{\frac{1}{n}\log\left(\frac{1}{\delta}\right)}\right) \end{aligned} \tag{1}$$

We want the LHS of Equation 1 to go to zero, i.e., the ERM $\hat{h}$ to be very close to the best hypothesis as $n \to \infty$. We also know that $g(n,\mathcal{H}) \leq 2^n$, so at the very least, $g(n,\mathcal{H}) \in o(2^n)$, but ideally, we would want something like $g(n,\mathcal{H}) \in O(1)$. We can (and will) show that for finite VC classes, $g(n,\mathcal{H}) \in O(poly(n))$

2 Shattering and VC dimension

Now we introduce the definitions of shattering and VC dimension, which will help us with PAC bounds.

Definition 2. *Let $A = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be a set of n points in $\mathcal{X}$. We say that A is **shattered** by a hypothesis class $\mathcal{H}$ if $\mathcal{H}$ can realize any label on A. That is, $|\mathcal{H}(A)| = 2^n$*

Since we are working with binary classification, a hypothesis class $\mathcal{H}$ shattering a set A means that an hypotheses from $\mathcal{H}$ are expressive enough so that they can give all possible labels (0 or 1) to each $x_i \in A$. A couple examples:

Example 3. Let $A = \{x_1, x_2\}$. First, let us consider the hypothesis class of one-sided threshold classifiers, that is, $\mathcal{H}_1 = \{h_a(x) = \mathbb{1}_{\{x\geq a\}} | \forall a \in \mathbb{R}\}$. Without loss of generality, let $x_1 < x_2$. We can propose different values of a to get the following labels:

- If $a < x_1$, we get labels $[1,1]$.
- If $x_1 < a < x_2$, we get labels $[0,1]$.
- If $a > x_2$, we get labels $[0,0]$

Note that we cannot realize $[1,0]$ using hypotheses in $\mathcal{H}_1$, so it cannot shatter A.

Now, consider the hypothesis class of two-sided threshold classifiers, that is,

$$\mathcal{H}_2 = \{h_a(x) = \mathbb{1}_{\{x \geq a\}} | \forall a \in \mathbb{R}\} \cup \{h_a(x) = \mathbb{1}_{\{x < a\}} | \forall a \in \mathbb{R}\}$$

Note that $\mathcal{H}_1 \subset \mathcal{H}_2$, so $\mathcal{H}_2$ can realize all the labels realized by $\mathcal{H}_1$. Now, to get the label $[1,0]$, consider $h_a(x) = \mathbb{1}_{\{x<a\}}$ with $x_1 < a < x_2$. So we can realize every possible pair of labels, meaning that $\mathcal{H}_2$ shatters A.

Example 4. Now, let $A = \{x_1, x_2, x_3\}$. Let uss see if we can shatter A with $\mathcal{H}_2$. Without loss of generality, let $x_1 < x_2 < x_3$. Consider all possible values of a and both types of hypotheses to see what labels we can get

- First, consider hypotheses of the type $h_a(x) = \mathbb{1}_{\{x \geq a\}}$:
  - If $a < x_1$, we get label $[1,1,1]$
  - If $x_1 < a < x_2$, we get $[0,1,1]$.
  - If $x_2 < a < x_3$, we get $[0,0,1]$
  - If $a > x_3$ we get $[0,0,0]$
- Now consider hypotheses of type $h_a(x) = \mathbb{1}_{\{x<a\}}$:
  - If $a < x_1$, we get label $[0,0,0]$
  - If $x_1 < a < x_2$, we get $[1,0,0]$.
  - If $x_2 < a < x_3$, we get $[1,1,0]$
  - If $a > x_3$ we get $[1,1,1]$

Note that there are some repeated labels, and we can not get the full $2^3 = 8$ possibilities for labels, since labels $[0,1,0]$ and $[1,0,1]$ are not realized. Therefore, we get that $\mathcal{H}_2(A) = 6 < 2^3$ so A cannot be shattered by $\mathcal{H}_2$

Now we are ready to introduce the definition of the Vapnik–Chervonenkis dimension.

Definition 3. *The* ***VC dimension*** *of an hypothesis class $\mathcal{H}$, called $d_{\mathcal{H}}$, is the size of the largest set shattered by $\mathcal{H}$.*

In a nutshell, if a hypothesis class has higher VC dimension, it means it is more expressive since it can shatter bigger sets (or for our interests, datasets).

Example 5. Let us recall $\mathcal{H}_2$, the set of two-sided threshold classifiers. From Example 3, we know that this hypothesis class can shatter two points, so $d_{\mathcal{H}} \geq 2$. But from Example 4, we know it cannot shatter sets of three points, so the VC dimension of this hypothesis class is $d_{\mathcal{H}} = 2$

Example 6. Consider $2D$ linear classifiers. The linear classifiers can shatter 3 points. For an illustration of this, see Figure 1. But can it shatter a set $A = \{x_1, x_2, x_3, x_4\}$ of 4 points? Different possible configurations of unattainable labels by $2D$ linear classifiers are given in Figure 2

So the hypothesis class of linear classifiers in $2D$ can not shatter a set of 4 points, but it can shatter 3 points, so $d_{\mathcal{H}} = 3$

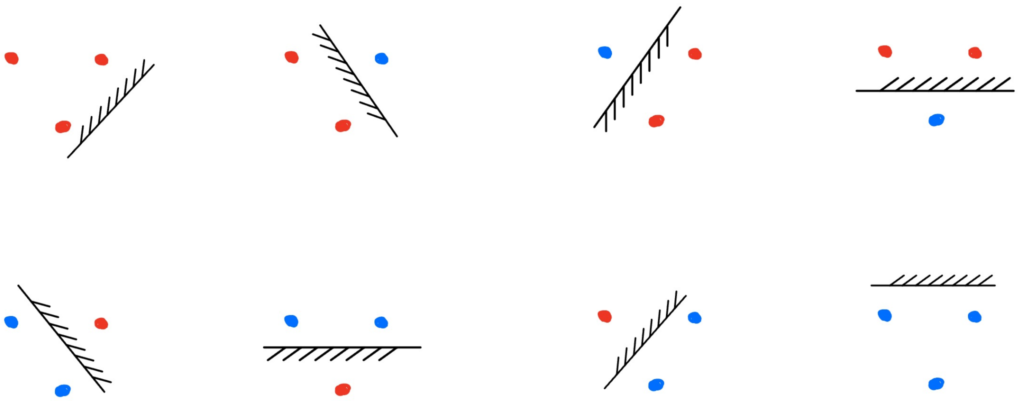

Figure 1: Different possible configurations of labels on three points, and corresponding linear classifier who can realize that configuration of labels

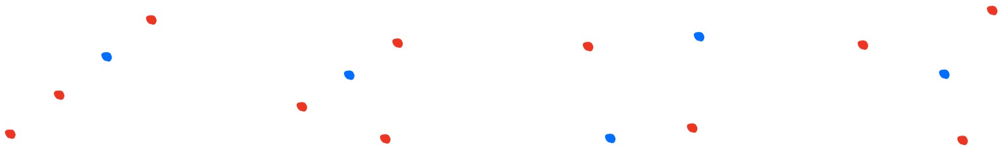

Figure 2: Different configurations of 4 points in $2D$ space. Red dots correspond to label 1, blue dots correspond to label 0.

3 Sauer's Lemma and a PAC Bound

Now we present a result that does the heavy lifting when we bound VC dimensions.

Theorem 7. *If the VC dimension of an hypothesis class is d, then*

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

The proof will be presented next class. This result means that

$$g(n, \mathcal{H}) = \begin{cases} = 2^n & \text{if } n \leq d \\ \leq \sum_{i=0}^{d} \binom{n}{i} & \text{if } n > d \end{cases}$$

because $\binom{n}{i} = 0$ if $n < i$. Note that the term on the bottom is of order $O(n^d)$. Let us bound then the term $\sum_{i=0}^{d} \binom{n}{i}$ to get a better sense of how fast does it grow.

$$
\begin{aligned}
\sum_{i=0}^{d} \binom{n}{i} &= \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^d \\
&\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i \qquad \text{(as } d < n \text{ and } i \leq d\text{)} \\
&= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \qquad \text{(binomial expansion)} \\
&\leq \left(\frac{n}{d}\right)^d e^d \qquad \text{(as } \left(1 + \frac{x}{n}\right) \leq e^x\text{)} \\
&= \left(\frac{en}{d}\right)^d
\end{aligned}
$$

That is, when $n > d$, the growth function grows polynomially on n. We can combine this result with Equation 1 to get a PAC bound in terms of VC dimension, instead of the growth function.

Theorem 8 (PAC bound for finite VC dimension)**.** *Let $\mathcal{H}$ be a hypothesis class with finite VC dimension d. Let $\hat{h}$ be obtained via ERM using n i.i.d. samples, where $n \geq d$. Let $\varepsilon > 0$. Then, with probability at least $1 - 2e^{-2n\varepsilon^2}$, we have*

$$
R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h) \leq O\left(\sqrt{\frac{d \log\left(\frac{n}{d}\right)}{n}}\right) + 2\varepsilon
$$

This is exactly what we have been working towards the last few lectures! We have bounded the estimation error of our ERM procedure using VC dimensions.

Once we prove Sauer's Lemma, we are done with PAC learning. But we will revisit VC dimension when we prove lower bounds involving risks.

Our work these lectures has been focused on binary classification. The ideas here can be extended in some directions, like multi-class classification (Mohri et al., 2012, Chapter 9) or regression (Mohri et al., 2012, Chapter 11). There are technical details to be considered when extending the ideas presented in the last few lectures to those scenarios, but the core ideas stay the same.

References

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

Acknowledgements

These notes are based on scribed lecture 5 and 6 from Fall 2023 by Chenghui Zheng, Yixuan Zhang, Justin Kiefel, Joseph Salzer.