CS861: Theoretical Foundations of Machine Learning Lecture 5 - 09/15/2023
University of Wi... Fall 2023

Lec... Function and VC Dimension

Lecturer: Kirthe... *Scribed by: Chenghui Zheng, Yixuan Zhang*

Disclaimer: *These n... ected 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 first bound the Radamacher complexity using the growth function. Then, we will introduce the VC dimension and provide some examples.

1 Bounding Rademacher Complexity Using the Growth Function

First, we will prove Massart's lemma, which upper bounds the empirical Radamacher complexity.

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

$$\widehat{\mathrm{Rad}}(S, \mathcal{H}) \leq \frac{1}{n}\left(\max_{v\in\mathcal{L}(S,\mathcal{H})} \|v\|_2\right)\sqrt{2\log(|\mathcal{L}(S,\mathcal{H})|)},$$

where $\|v\|_2^2 = \sum_{i=1}^n v_i^2$.

Proof First, observe that we can write

$$\widehat{\mathrm{Rad}}(S, \mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(h(x_i), y_i)\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\max_{v\in\mathcal{L}(S,\mathcal{H})}\sum_{i=1}^n \sigma_i v_i\right]. \tag{1}$$

Next, let $s > 0$, whose value we will specify later.

$$\begin{aligned}
\mathbb{E}_\sigma\left[\max_{v\in\mathcal{L}(S,\mathcal{H})}\sum_{i=1}^n \sigma_i v_i\right] &= \frac{1}{s}\mathbb{E}_\sigma\left[\max_{v\in\mathcal{L}(S,\mathcal{H})} s\sum_{i=1}^n \sigma_i v_i\right]\\
&= \frac{1}{s}\mathbb{E}_\sigma\left[\log\left(\exp\left(\max_{v\in\mathcal{L}(S,\mathcal{H})} s\sum_{i=1}^n \sigma_i v_i\right)\right)\right]\\
&\leq \frac{1}{s}\log\left(\mathbb{E}_\sigma\left[\exp\left(\max_{v\in\mathcal{L}(S,\mathcal{H})} s\sum_{i=1}^n \sigma_i v_i\right)\right]\right) \qquad \text{by Jensen's Inequality}\\
&\leq \frac{1}{s}\log\left(\mathbb{E}_\sigma\left[\sum_{v\in\mathcal{L}(S,\mathcal{H})}\exp\left(s\sum_{i=1}^n \sigma_i v_i\right)\right]\right)\\
&\leq \frac{1}{s}\log\left(\sum_{v\in\mathcal{L}(S,\mathcal{H})}\mathbb{E}_\sigma\left[\exp\left(s\sum_{i=1}^n \sigma_i v_i\right)\right]\right)\\
&\overset{(i)}{\leq} \frac{1}{s}\log\left(\sum_{v\in\mathcal{L}(S,\mathcal{H})}\exp\left(\frac{s^2}{2}\sum_{i=1}^n v_i^2\right)\right)
\end{aligned}$$

$$\le \frac{1}{s}\log\left(|\mathcal{L}(S,\mathcal{H})| \max_{v\in\mathcal{L}(S,\mathcal{H})} \exp\left(\frac{s^2}{2}\sum_{i=1}^n v_i^2\right)\right)$$

$$|) + \frac{s}{2}\max_{v\in\mathcal{L}(S,\mathcal{H})} \|v\|_2^2. \tag{2}$$

The inequality (i) h $\dots, \sigma_n)$ and σ_i is 1-subgaussian. Then,

$$\mathbb{E}_\sigma \prod_{i=1}^n \mathbb{E}_\sigma\left[\exp\left((sv_i)\sigma_i\right)\right] \le \prod_{i=1}^n \exp\left(\frac{s^2 v_i^2}{2}\right).$$

Equation (2) holds for all s, we can choose

$$s = \sqrt{\frac{2\log|\mathcal{L}(S,\mathcal{H})|}{\max_{v\in\mathcal{L}(S,\mathcal{H})}\|v\|_2^2}}. \tag{3}$$

Equation (1), (2) and (3) imply

$$\mathrm{Rad}_n(S,\mathcal{H}) \le \frac{1}{n}\left(\max_{v\in\mathcal{L}(S,\mathcal{H})}\hat{}\;\|v\|\right)\sqrt{2\log(|\mathcal{L}(S,\mathcal{H}|)}.$$

□

Corollary 1. *$\forall S$ such that $|S| = n$, we have*

$$\widehat{\mathrm{Rad}}(S,\mathcal{H}) \le \sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}}.$$

Moreover,

$$\mathrm{Rad}_n(\mathcal{H}) \le \sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}}.$$

Proof $\|v\|_2 \le \sqrt{n}$ and $|\mathcal{L}(S,\mathcal{H})| \le g(n,\mathcal{H})$ by definition of $g(n,\mathcal{H})$. The second statement follows by taking the expectation over S of the LHS of the first statement. □

To motivate the ensuing discussion about the VC dimension, recall that with probability at least $1 - 2e^{-2n\epsilon^2}$

$$R(\hat{h}) \le \inf_{h\in\mathcal{H}} R(h) + c_1\mathrm{Rad}_n(\mathcal{H}) + c_2\epsilon.$$

Then, with fixed n, δ, where $\epsilon \in O(\sqrt{\frac{1}{n}\log(\frac{1}{\delta})})$. From the previous lecture, we obtained $g(n,\mathcal{H}) \le 2^n$. However, when $g(n,\mathcal{H}) = 2^n$, $\mathrm{Rad}_n(\mathcal{H})$ will never goes to 0. At the very least, we hope to have: $g(n,\mathcal{H}) \in o(2^n)$, but ideally we would like to have $g(n,\mathcal{H}) \in \mathrm{poly}(n)$ so that $\sqrt{\frac{\log(g(n,\mathcal{H}))}{n}} \lesssim \sqrt{\frac{\log(n)}{n}}$.

2 VC dimension

In this section, we begin with the definition of Shattering.

Definition 1. *Let $S^X = \{x_1, \dots, x_n\} \in X^n$ be a set of n points in X. We say that S^X is shattered by a hypothesis class $\mathcal{H}$ if $\mathcal{H}$ "can realize any label on S^X". That is*

$$H(S^X)| = 2^n,$$

where $H(S^X) = \{[h(x_1)$

Then, we give two e under the same hypothesis class $\mathcal{H}$, which is the two sided threshold classifiers:

$$\forall a \in \mathbb{R}\} \cup \{h_a(x) = \mathbb{1}_{\{x<a\}} \mid \forall a \in \mathbb{R}\}.$$

Example 1. Consider $S^X = \{x_1, x_2\}$ and we can assume $x_1 < x_2$ without loss of generality. Therefore, we can try different classifiers in $\mathcal{H}$ to achieve different labels.

- When we use $h_a(x) = \mathbb{1}_{\{x \geq x_1 - 1\}}$, the label is $[1, 1]$.
- When we use $h_a(x) = \mathbb{1}_{\{x \geq \frac{x_1+x_2}{2}\}}$, the label is $[0, 1]$.
- When we use $h_a(x) = \mathbb{1}_{\{x \geq x_2 + 1\}}$, the label is $[0, 0]$.
- When we use $h_a(x) = \mathbb{1}_{\{x < \frac{x_1+x_2}{2}\}}$, the label is $[1, 0]$.

Then, $|H(S^X)| = 2^2$ and we can say S^X is shattered by $\mathcal{H}$

Example 2. Consider $S^X = \{x_1, x_2, x_3\}$ and we can assume $x_1 < x_2 < x_3$ without loss of generality. We can do the similar thing as Example 1

- When we use $h_a(x) = \mathbb{1}_{\{x \geq x_1 - 1\}}$, the label is $[1, 1, 1]$.
- When we use $h_a(x) = \mathbb{1}_{\{x \geq \frac{x_1+x_2}{2}\}}$, the label is $[0, 1, 1]$.
- When we use $h_a(x) = \mathbb{1}_{\{x \geq \frac{x_2+x_3}{2}\}}$, the label is $[0, 0, 1]$.
- When we use $h_a(x) = \mathbb{1}_{\{x \geq x_3 + 1\}}$, the label is $[0, 0, 0]$.
- When we use $h_a(x) = \mathbb{1}_{\{x < \frac{x_1+x_2}{2}\}}$, the label is $[1, 0, 0]$.
- When we use $h_a(x) = \mathbb{1}_{\{x < \frac{x_2+x_3}{2}\}}$, the label is $[1, 1, 0]$.

However, the label $[0, 1, 0]$ and $[1, 0, 1]$ can't be achieved by any $h \in \mathcal{H}$. Then, $|H(S^X)| = 6 < 2^3$ and we can say S^X can't be shattered by $\mathcal{H}$.

After introducing the shattering, we are ready to give the definition of VC-dimension. Here we use $d_{\mathcal{H}}$ to denote VC-dimension of $\mathcal{H}$ and we will use d when $\mathcal{H}$ is clear from contest.

Definition 2. *The VC-dimension $d_{\mathcal{H}}$ of a hypothesis class $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.*

Below we introduce three examples of VC-dimension.

Example 3. Two-sided threshold classifiers

By Example 1, we can obtain $d \geq 2$. By Example 2, we have $d < 3$. Therefore, we can conclude that $d = 2$.

Example 4. One-sided threshold classifiers

The hypothesis class $\mathcal{H}$ is defined as

$$\mathcal{H} = \{h_a(x) = \mathbb{1}_{\{x \geq a\}} \mid \forall a \in \mathbb{R}\}.$$

Similarly, we can show $d = 1$ by showing $d \geq 1$ and $d < 2$.

1. Consider $S^X = \{x_1\}$.

   - When we use [illegible] e label is [1].
   - When we use [illegible] e label is [0].

   Then, $|H(S^X)| =$ [illegible] s shattered by $\mathcal{H}$, which implies $d \geq 1$

2. Consider $S^X = \{x$ [illegible] me $x_1 < x_2$ without loss of generality.

   - When we use [illegible] e label is [1, 1].
   - When we use [illegible] he label is [0, 1].
   - When we use $h_a(x) = \mathbb{1}_{\{x \geq x_2+1\}}$, the label is [0, 0].

   However, the label [1, 0] can't be achieved by any $h \in \mathcal{H}$. Then, $|H(S^X)| = 3 < 2^2$ and we can say S^X can't be shattered by $\mathcal{H}$, which implies $d < 2$.

Example 5. Two-dimensional linear classifiers. Firstly, we consider three data points located at 2-dimensional space, which have the triangle shape. By Figure 1, we can say the dataset generated by three data distributed as Figure 1 can be shatter by $\mathcal{H}$, which implies $d \geq 3$.

Furthermore, the distribution of 4 points in 2-dimensional space can only have 4 different cases. By Figure 2, we give an counterexample for each of 4 cases to show all the dataset contained 4 data can't be shattered by $\mathcal{H}$, which implies $d < 4$.

Therefore, we have $d = 3$.

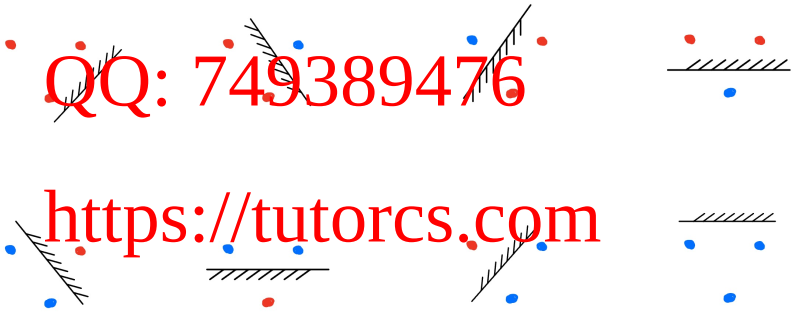

Figure 1: 8 different labels generated by linear classifier under 3 data in 2-dimensional space.

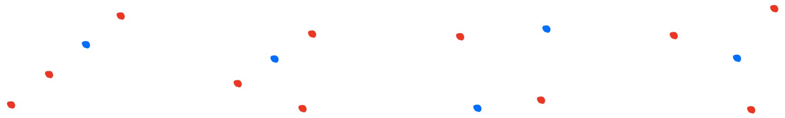

Figure 2: Unattainable labels by linear classifier under 4 different cases of 4 data in 2-dimensional space.

Example 6. K-dimensional linear classifiers. We directly give the result without proof here. $d = K + 1$. The proof of this result will appear on the next homework.