CS861: Theoretical Foundations of Machine Learning Lecture 1 - 09/13/2023
University of Wi[illegible] Fall 2023

Lecture [illegible] Complexity & Growth Function

Lecturer: Kirthe[illegible] *Scribed by: Yixuan Zhang, Elliot Pickens*

Disclaimer: *These n[illegible]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 first introduce a simple example of the Empirical Rademacher Complexity (ERM). Then, we introduce the Rademacher Complexity, which can be applied to derive an upper bound for $\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(\hat{R}_S(h) - R(h))\right]$. After that, we will state a bound for PAC learning. Finally, we will introduce the *growth function*.

1 Rademacher Complexity

Before introducing Rademacher complexity, we first give a simple example to recap the Empirical Rademacher Complexity (ERM).

Figure 1: Two example threshold functions, where the hypothesis is either $h(x) = \mathbb{I}(x \geq a)$ or $h(x) = \mathbb{I}(x \leq a)$

Consider the dataset $S = \{(x_1 = 0, y_1 = 0), (x_2 = 1, y_2 = 1)\}$ and two hypothesis classes:

$$\mathcal{H}_1 = \{h_a(x) = \mathbb{1}_{\{x\geq a\}} \mid \forall a \in \mathbb{R}\} \quad \text{“one-sided threshold”}$$
$$\mathcal{H}_2 = \mathcal{H}_1 \cup \{h'_a(x) = \mathbb{1}_{\{x< a\}} \mid \forall a \in \mathbb{R}\} \quad \text{“two-sided threshold”}$$

In this example, we have two data in our dataset. Therefore, σ is a two-dimensional vector, which can take 4 different possible values: $(1,1)$, $(1,-1)$, $(-1,1)$, $(-1,-1)$. Then, we can calculate the ERM by calculating the supremum under each of the four possible values and taking the expectation. Finally, we obtain

$$\widehat{\text{Rad}}(S,\mathcal{H}_1) = \frac{1}{4}(\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 0) = \frac{3}{8}$$
$$\widehat{\text{Rad}}(S,\mathcal{H}_2) = \frac{1}{4}(1 + \frac{1}{2} + \frac{1}{2} + 0) = \frac{1}{2}.$$

Next, we introduce the definition for Rademacher complexity.

Definition 1. *Given a hypothesis class $\mathcal{H}$ and $n \in \mathbb{N}$, the Rademacher complexity of $\mathcal{H}$ is defined as follows:*

$$\mathrm{Rad}_n(\mathcal{H}) = \mathbb{E}_S\left[\mathrm{Rad}(S, \mathcal{H})\right] = \mathbb{E}_{S,\sigma}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(h(x_i), y_i)\right]$$

Lemma 1. *Given a hypothesis class $\mathcal{H}$ and $n \in \mathbb{N}$, we have*

$$\mathbb{E}_S\left[\sup_{h \in \mathcal{H}}(\hat{R}_S(h) - R(h))\right] \leq 2\mathrm{Rad}_n(\mathcal{H})$$

Lemma 1 can be used to derive the PAC learning bound for ERM, which is showed in the next section. The proof of Lemma 1 is given below.

Proof

$$\begin{aligned}
LHS &= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(\hat{R}_S(h) - \mathbb{E}_{S'}[\hat{R}_{S'}(h)])\right] \because S' \text{ drawn from } \mathbb{P}^n_{x,y} \text{ and independent with } S \\
&= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{S'}[\hat{R}_S(h) - \hat{R}_{S'}(h)]\right] \\
&\leq \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}[\hat{R}_S(h) - \hat{R}_{S'}(h)]\right] \because \sup \mathbb{E} \leq \mathbb{E} \sup \text{ (subadditivity)} \\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}[\ell(h(x_i), y_i) - \ell(h(x_i'), y_i')]\right] \\
&= \mathbb{E}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}[\sigma_i\ell(h(x_i), y_i) - \sigma_i\ell(h(x_i'), y_i')]\right] \because \text{the symmetry of the two datasets } S, S' \\
&= \mathbb{E}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}[\sigma_i\ell(h(x_i), y_i)] + \frac{1}{n}\sum_{i=1}^{n}[-\sigma_i\ell(h(x_i'), y_i')]\right] \\
&\leq \mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}[\sigma_i\ell(h(x_i), y_i)]\right] + \mathbb{E}_{S',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}[-\sigma_i\ell(h(x_i'), y_i')]\right] \because \sup(a+b) \leq \sup a + \sup b \\
&= \mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}[\sigma_i\ell(h(x_i), y_i)]\right] + \mathbb{E}_{S',\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}[\sigma_i\ell(h(x_i'), y_i')]\right] \because \sigma \text{ is distributed symmetrically} \\
&= 2\mathrm{Rad}_n(\mathcal{H})
\end{aligned}$$

□

2 PAC Learning Bound for ERM

Theorem 1. *Let $\mathcal{H}$ be a hypothesis calss with finite $\mathrm{Rad}_n(\mathcal{H})$. Let $\hat{h}$ be obtained via ERM using an i.i.d dataset of n samples. Let $\epsilon > 0$. Then, there exist universal constants c_1, c_2 such that with probability at least $1 - 2e^{-2n\epsilon^2}$*

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

We will prove this theorem in the next homework. The following ideas may be helpful in this proof.

- For the case that $\exists h^* \in \mathcal{H}$ such that $R(h^*) = \inf_{h \in \mathcal{H}} R(h)$. We can do the following decomposition:

$$R(\hat{h}) - [illegible] + \hat{R}(\hat{h}) - R(h^*) \leq \underbrace{R(\hat{h}) - \hat{R}(\hat{h})}_{T_1} + \underbrace{\hat{R}(h^*) - R(h^*)}_{T_2}$$

  By McDiarmid's i[illegible]d both T_1 and T_2.

- We also need to c[illegible]ase that $\nexists h^* \in \mathcal{H}$ such that $R(h^*) = \inf_{h \in \mathcal{H}} R(h)$, which will not be showed he[illegible]

3 Growth Function

While the above bound in useful, computing $\mathrm{Rad_n}(\mathcal{H})$ can be difficult for general hypothesis classes. Hence, we will relate the Radamacher complexity to the VC dimension which is easier to bound. For this, we will first define the *growth function.*

Definition 2. *Restriction of $\mathcal{H}$ to S*

Given a sample $\mathrm{S} = \{(\mathrm{x_1}, \mathrm{y_1}), ..., (\mathrm{x_n}, \mathrm{y_n})\}$, *and a hypothesis space $\mathcal{H}$, define*

$$\mathcal{L}(S, \mathcal{H}) = \{[\ell(h(x_1), y_1), ..., \ell(h(x_n), y_n)] \mid h \in \mathcal{H}\}$$

to be the set of all possible loss vectors of S given $\mathcal{H}$, i.e. all possible loss vectors we can generate from S by iterating over all $h \in \mathcal{H}$.

For $0-1$ loss, each datapoint in the sample can take on one of two values since $\ell(h(x_i), y_i) \in \{0, 1\}$. This allows us to easily bound the cardinality of $\mathcal{L}$ for $0-1$ loss as

$$|\mathcal{L}(S, \mathcal{H})| \leq 2^n$$

Now, let us go through a few examples of $\mathcal{L}$.

Example 2. Let $S = \{(x_1 = -1, y_1 = 0), (x_2 = 1, y_2 = 1)\}$ and $\mathcal{H}_{one-sided} = \{h_a(x) = \mathbb{1}_{\{x \geq a\}} \mid \forall a \in \mathbb{R}\}$

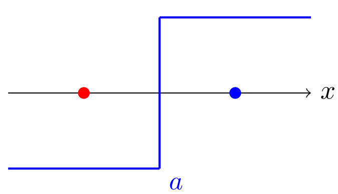


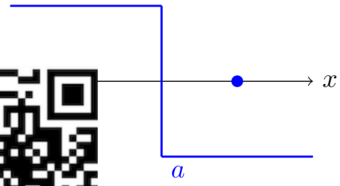
Figure 2: An example of a $h \in \mathcal{H}_{one-sided}$ that gives us a $[0, 0]$ loss vector.

be the set of all "one-sided threshold functions." Then

$$\mathcal{L}(S, \mathcal{H}_{one-sided}) = \{[0, 1], [1, 0], [0, 0]\}$$

Since we can either misclassify a single point or no points, but it is not possible to misclassify both points with this hypothesis class.

Example 3.



Figu[...] $\in \mathcal{H}_{two-sided}$ that gives us a $[0, 0]$ loss vector

Let $S = \{(x_1 = -1,$ [...]$)\}$ and $\mathcal{H}_{two-sided} = \mathcal{H}_{one-sided} \cup \{h'_a(x) = \mathbb{1}_{\{x<a\}} \mid \forall a \in \mathbb{R}\}$ be the set of all "two-sided threshold functions." Then

$$\mathcal{L}(S, \mathcal{H}_{two-sided}) = \{[0, 1], [1, 0], [0, 0], [1, 1]\}$$

Example 4.
Let $S = \{(x_1 = 0, y_1 = 0), (x_2 = 0, y_2 = 1)\}$ and $\mathcal{H} = \mathcal{H}_{one-sided}$. Then

$$\mathcal{L}(S, \mathcal{H}_{one-sided}) = \{[0, 1], [1, 0]\}$$

because we can only classify one of the two points correctly at the same time.

In all of these examples we have that $|\mathcal{L}(S, \mathcal{H})| \leq 2^n$ as expected.

Definition 3 (Growth Function). *Given $n \in N$ and a hypothesis space $\mathcal{H}$, the growth function is defined as*

$$g(n, \mathcal{H}) = \max_{S;|S|=n} |\mathcal{L}(S, \mathcal{H})| \leq 2^n$$

which corresponds to the maximum number of loss vectors that can be constructed from a sample of n data points.

Let's go through a few more examples to show how the growth function behaves under various conditions.

Example 5. Let $\mathcal{H} = \mathcal{H}_{one-sided}$. Starting with $n = 1$, we can see that

$$\mathcal{L}(S, \mathcal{H}_{one-sided}) = \{[1], [0]\}$$

and

$$g(1, \mathcal{H}_{one-sided}) = |\mathcal{L}(S, \mathcal{H}_{one-sided})| = 2 \leq 2^1$$

For $n = 2$ we can draw on our work from 3 to show that

$$\begin{aligned} g(2, \mathcal{H}_{one-sided}) &= |\mathcal{L}(S, \mathcal{H}_{one-sided})| \\ &= |\{[0, 1], [1, 0], [0, 0]\}| \\ &= 2 \\ &\leq 2^2 \end{aligned}$$

And for $n = 3$

$|\{[0,0,1],[0,1,0],[1,1,0],[0,0,0]\}|$

4

2^3

Example 6.
Running through th... $_{o-sided}$ we get

$$g(1, \mathcal{H}_{two-sided}) = 2 \leq 2^1$$
$$g(2, \mathcal{H}_{two-sided}) = |\{[0,1],[1,0],[0,0],[1,1]\}| = 4 \leq 2^2$$
$$g(3, \mathcal{H}_{two-sided}) = |\{[0,0,1],[0,1,0],[1,1,0],[1,0,1],[0,0,0],[1,1,1]\}| = 6 \leq 2^3$$

Example 7.

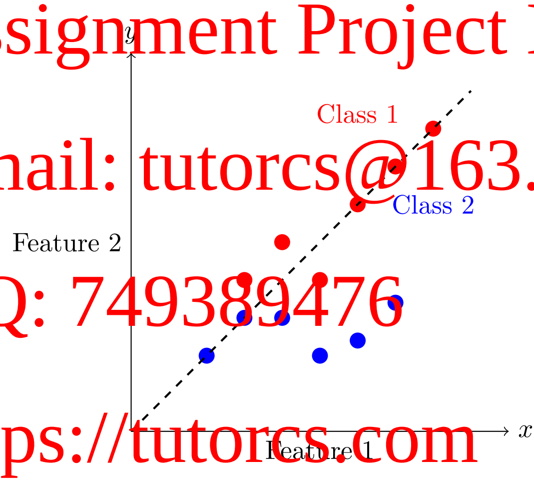


Figure 4: An example $h \in \mathcal{H}_{\text{2D linear}}$

Now, let us briefly consider the hypothesis space of 2D linear classifiers $\mathcal{H}_{\text{2D linear}} = \{\text{2D linear classifiers}\}$. For this class it can be shown that

$$g(1, \mathcal{H}_{\text{2D linear}}) = 2$$
$$g(2, \mathcal{H}_{\text{2D linear}}) = 4$$
$$g(3, \mathcal{H}_{\text{2D linear}}) = 8$$
$$g(4, \mathcal{H}_{\text{2D linear}}) = 14$$

which is notable greater than the growth function values for the other spaces.

This is because the hypothesis class $\mathcal{H}_{\text{2D linear}}$ is more flexible than the two threshold function spaces we have previously examined. In fact, the different space have 1, 2, and 3 degrees of freedom respectively.

- $\mathcal{H}_{one-sided}$: where we place the threshold a
- $\mathcal{H}_{two-sided}$: where we place the threshold a, and which side of the threshold corresponds to each class

- $\mathcal{H}_{\text{2D linear}}$: the slope, the intercept, and which class will be on either side of the boundary

Interestingly enough [illegible] dom correspond to the n at which $g(n, \mathcal{H})$ stops hitting the upper bound of 2^n.