

# Chapter 2 VC-dimension

Siheng Zhang  
zhangsiheng@cvte.com

October 7, 2020

This part corresponds to **Chapter 2-5 in UML**, and mainly answers the following questions:

- The necessary and sufficient condition of PAC learnability.
- 

## Contents

# 1 The VC-dimension

## 1.1 Shattering

Consider the set of threshold functions over the real line  $\mathcal{H} = \{h_a(x) = \mathbb{1}_{[x \leq a]}, a \in \mathbb{R}\}$ . Let  $a^*$  be the threshold such that  $L_{\mathcal{D}}(h^*) = 0$ . Let  $a_0 < a^* < a_1$  such that:

$$\mathbb{P}_{x \sim \mathcal{D}_x} [x \in (a_0, a^*)] = \mathbb{P}_{x \sim \mathcal{D}_x} [x \in (a^*, a_1)] = \epsilon$$

If  $\mathcal{D}_x(-\infty, a^*) \leq \epsilon$ , we set  $a_0 = -\infty$ , and similarly for  $a_1$ .

Given a training set  $S$ , let  $b_0 = \max\{x : (x, 1) \in S\}$  (if no example is positive then  $b_0 = -\infty$ ), and  $b_1 = \min\{x : (x, 0) \in S\}$  (if no example is negative then  $b_1 = \infty$ ). Let  $b_S$  be the threshold of an ERM hypothesis  $h_S$ , which implies  $b_S \in (b_0, b_1)$ , then we have

$$\mathbb{P}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(h_S) < \epsilon] \leq \mathbb{P}_{S \sim \mathcal{D}^m} [b_0 < a_0] + \mathbb{P}_{S \sim \mathcal{D}^m} [b_1 > a_1]$$

Each term on the right-side is bounded by  $(1 - \epsilon)^m \leq e^{-\epsilon m}$ . Let  $m > \log(2/\delta)/\epsilon$ , then the left-side is bounded by  $\delta$ . As a result, the hypothesis class is PAC-learnable.

The example above shows that: **finiteness is not a necessary condition for learnability**, and hence we turn to the definition of **shattering**, which describes the ability of a hypothesis set to cover the training set.

The definition of VC-dimension is motivated from the No-Free-Lunch theorem: without restricting the hypothesis class, for any learning algorithm, an **adversary** can construct a distribution for which the learning algorithm will perform poorly, while there is another learning algorithm that will succeed on the same distribution. To make any algorithm fail, the **adversary** used the power of choosing a target function from the set of all possible labelling functions.

When considering PAC learnability of a hypothesis class  $\mathcal{H}$ , the **adversary** is restricted to constructing distributions for which some hypothesis  $h \in \mathcal{H}$  achieves a zero risk. Since we are considering distributions that are concentrated on elements of  $C$ , we should study how  $h \in \mathcal{H}$  behaves on  $C$ .

**Definition** (Restriction of  $\mathcal{H}$  to  $C$ ): The restriction of  $\mathcal{H}$  to  $C$  is the set of functions from  $C$  to  $\{0, 1\}$  that can be derived from  $\mathcal{H}$ . That is,

$$\mathcal{H}_C = \{(h(c_1), \dots, h(c_m)) : h \in \mathcal{H}\} \quad (1)$$

where we represent each function from  $C$  to  $\{0, 1\}$  as a vector in  $\{0, 1\}^{|C|}$ .

**Definition** (Shattering): A hypothesis class  $\mathcal{H}$  shatters a finite set  $C \in \mathcal{X}$  if the restriction of  $\mathcal{H}$  to  $C$  is the set of all functions from  $C$  to  $\{0, 1\}$ . That is,  $|\mathcal{H}_C| = 2^{|C|}$ .

**Corollary 1** Let  $\mathcal{H}$  be a hypothesis class of functions from  $\mathcal{X}$  to  $\{0, 1\}$ . Let  $m$  be a training set size. Assume that there exists a set  $C \subset \mathcal{X}$  of size  $2m$  that is shattered by  $\mathcal{H}$ . Then, for any learning algorithm,  $A$ , there exist a distribution  $\mathcal{D}$  over  $\mathcal{X} \times \{0, 1\}$  and a predictor  $h \in \mathcal{H}$  such that  $L_{\mathcal{D}}(h) = 0$  but with probability of at least  $1/7$  over the choice of  $S \sim \mathcal{D}^m$  we have that  $L_{\mathcal{D}}(A(S)) \geq 1/8$ .

The corollary shows that **whenever if  $\mathcal{H}$  shatters some set  $C$  of size  $2m$ , then we cannot learn  $\mathcal{H}$  by using  $m$  examples**. This leads us directly to the definition of the VC dimension.

## 1.2 The VC-dimension

**Definition** (VC-dimension): The VC-dimension of a hypothesis class  $\mathcal{H}$ , denoted  $\text{VCdim}(\mathcal{H})$ , is the maximal size of a set  $C \subset \mathcal{X}$  that can be shattered by  $\mathcal{H}$ . If  $\mathcal{H}$  can shatter sets of arbitrarily large size we say that  $\mathcal{H}$  has infinite VC-dimension.

**Theorem 1** If  $\mathcal{H}$  is a class of infinite VC-dimension, then  $\mathcal{H}$  is not PAC learnable.

### 1.2.1 Examples

To calculate the VC-dimension for a hypothesis set, we should show that:

- There **exists** a subset of size  $d$  that can be shattered;
- **Every** subset of size  $d + 1$  can not be shattered.

### 1 Threshold functions

$$\mathcal{H} = \{\mathbb{1}_{x \leq a} : a \in \mathbb{R}\}$$

For an arbitrary set  $C = \{c\}$ ,  $\mathcal{H}$  shatters  $C$ , therefore  $\text{VCdim}(\mathcal{H}) \geq 1$ ; for an arbitrary set  $C = \{c_1, c_2\}$ , where  $c_1 \leq c_2$ , any threshold that assigns 0 to  $c_1$  must assign 0 to  $c_2$ . In other words, not all functions from  $C$  to  $\{0, 1\}$  are included by  $\mathcal{H}_C$ . So,  $\mathcal{H}$  does not shatter  $C$ .

### 2 Intervals

$$\mathcal{H} = \{\mathbb{1}_{x \in (a, b)} : a < b, a, b \in \mathbb{R}\}$$

Denote the set  $C = \{c_1, c_2\}$ . If we take  $a > c_2$  or  $b < c_2$ , then we have  $h_{a,b}(c_1) = 0, h_{a,b}(c_2) = 0$ ; if we take  $c_1 < a < c_2 < b$ , then we have  $h_{a,b}(c_1) = 0, h_{a,b}(c_2) = 1$ ; if we take  $a < c_1 < b < c_2$ , then we have  $h_{a,b}(c_1) = 1, h_{a,b}(c_2) = 0$ ; if we take  $a < c_1 < c_2 < b$ , then we have  $h_{a,b}(c_1) = 1, h_{a,b}(c_2) = 1$ . Therefore,  $\mathcal{H}_C$  is the set of all functions from  $C$  to  $\{0, 1\}^2$ .

Take the set  $C = \{c_1, c_2, c_3\}$ , without loss of generalization, let the labels be  $(1, 0, 1)$ , therefore  $\mathcal{H}$  does not shatter  $C$ .

Hence,  $\text{VCdim}(\mathcal{H}) = 2$ .

### 3 Axis Aligned Rectangles

$$\mathcal{H} = \{\mathbb{1}_{(a_1 \leq x_1 \leq a_2, b_1 \leq x_2 \leq b_2)} : a_1 < a_2, b_1 < b_2\}$$

Any set with 4 points can be shattered. Take the set with 5 points. Suppose that there is 1 point (labelled as 0) surrounded by 4 points (labelled as 1), it cannot be shattered. Hence,  $\text{VCdim}(\mathcal{H}) = 4$ .

### 4 Finite class

Let  $\mathcal{H}$  be a finite class. Then, clearly, for any set  $C$  we have  $|\mathcal{H}_C| \leq |\mathcal{H}|$  and thus it cannot be shattered if  $|\mathcal{H}| < 2^{|C|}$ . This implies that  $\text{VCdim}(\mathcal{H}) < \log_2 |\mathcal{H}|$ .

*remark1:* In the previous examples, the VC-dimension happened to equal the number of parameters defining. This is not always true. See exercise ? for detail.

## 2 Fundamental theorem of PAC learning

**Theorem 2** Let  $\mathcal{H}$  be a hypothesis class of functions from a domain  $\mathcal{X}$  to  $\{0, 1\}$  and let the loss function be the 0-1 loss. Then, the following are equivalent:

1. The hypothesis class has uniform convergence property.

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) = O\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right) \quad (2)$$

2. Any ERM rule is a successful agnostic PAC learner for the hypothesis class.
3. The hypothesis class is agnostic PAC learnable.

$$m_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right) \quad (3)$$

4. The hypothesis class is PAC learnable.

$$m_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}\right) \quad (4)$$

5. Any ERM rule is a successful PAC learner for the hypothesis class.
6. The hypothesis class has a finite VC-dimension.

*remark1:* 1-2-3-4-5-6 are all learned. The leaving part is 6-1, which is solved below.

## 3 Effective size of a hypothesis class

## 4 Non-uniform learnability

“non-uniform learnability” allows the sample size to be non-uniform with respect to the different hypotheses with which the learner is competing.

A hypothesis is  $(\epsilon, \delta)$ -competitive with another if

## 5 Summary

## 6 Exercises and solutions

*To be continue...*

*Chapter 3. Bayesian-PAC*

*Chapter 4. Generalization in Deep Learning*