VC Dimension

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 20

OOPS

Order of Presentations:

Tu	We	Th
#2	#3	#1
#4	#5	

Vapnic-Chernonenkis (VC) dimension

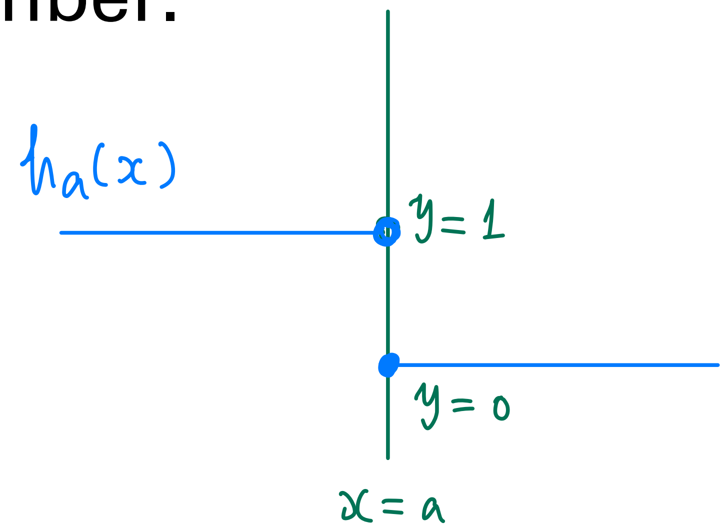
infinite-size classes may be learnable

- According to No-Free-Lunch theorem, if there is no restriction on the hypothesis class $\mathcal{H}$ ($\mathcal{H}$ contains all functions from $\mathcal{X}$ to $\{0,1\}$), then for any learning algorithm, there exists a distribution on which it performs poorly.

- Is it because that $|\mathcal{H}| = \infty$? Let's look at the following example.

infinite-size classes may be learnable

- Let $\mathcal{H}$ be the set of threshold functions over the real line:
  - $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ where $h_a(x) = \mathbf{1}_{\{x<a\}}(x)$.
  - Then $|\mathcal{H}| = \infty$ since a can be any real number.



- Claim: this $\mathcal{H}$ is PAC-learnable

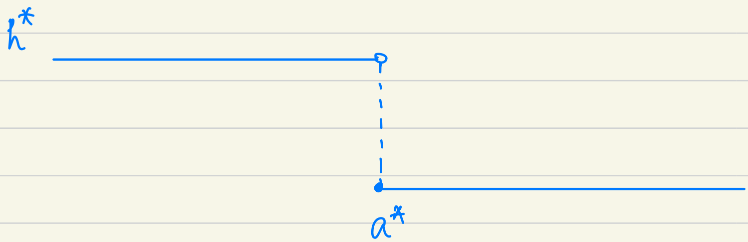
Claim: $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ where $h_a(x) = \mathbb{1}_{\{x<a\}}(x)$ is PAC-learnable.

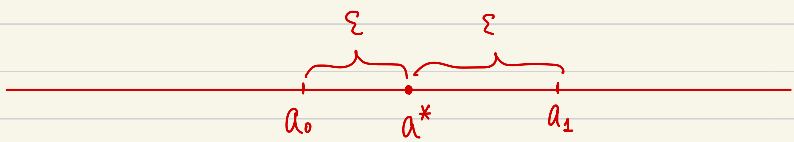
Pf: Let a^* be a threshold s.t. the hypothesis

$$h^*(x) := h_{a^*}(x) = \mathbb{1}_{\{x<a^*\}}(x) \quad \text{achieves}$$

$$L_{\mathcal{D}}(h^*) = 0.$$

(This exists by the realizability assumption.)





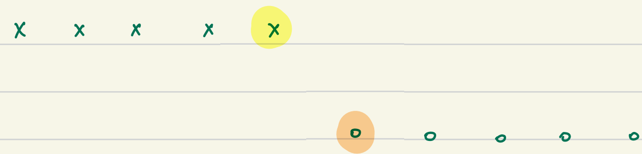
Let $a_0 < a^* < a_1$ be such that

$$\mathbb{P}_{\mathcal{D}}(x \in (a_0, a^*)) = \mathbb{P}_{\mathcal{D}}(x \in (a^*, a_1)) = \varepsilon$$

(we can also say $\mathcal{D}(x \in (a_0, a^*)) = \mathcal{D}(x \in (a^*, a_1)) = \varepsilon$.)

If a_0 does not exist, then take $a_0 = -\infty$;
if a_1 does not exist, then take $a_1 = +\infty$.

Given a training set S.



Let $b_0 = \max_{x \in \mathbb{R}} \{ (x, 1) \in S \}$

"The maximal x in $S|_x$ whose label is 1."

$b_1 = \min_{x \in \mathbb{R}} \{ (x, 0) \in S \}$

"The minimal x in $S|_x$ whose label is 0."

If no b_0 exists, take $b_0 = -\infty$;

if no b_1 exists, take $b_1 = +\infty$.

Let b_S minimize the training loss. That is,

$$L_S(h_{b_S}) = 0$$

For brevity, denote $h_S := h_{b_S}$.

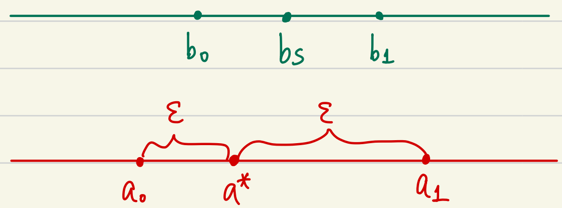
b_S is a correct threshold for all data in the traing set S.

Also, $b_0 < b_S < b_1$.

Therefore, a sufficient condition for

$$L_{\mathcal{D}}(h_S) \leq \varepsilon \quad \text{(approximately correct)}$$

is that both $b_0 \geq a_0$ and $b_1 \leq a_1$.



In other words,

$$\mathbb{P}_{\mathcal{D}^m}\left(L_{\mathcal{D}}(h_S) > \varepsilon\right) \leq \mathbb{P}_{\mathcal{D}^m}\left(b_0 < a_0 \text{ or } b_1 > a_1\right)$$

$$\leq \mathbb{P}_{\mathcal{D}^m}(b_0 < a_0) + \mathbb{P}_{\mathcal{D}^m}(b_1 > a_1).$$

Note that $b_0 < a_0$ if and only if all the data in S are not in (a_0, a^*). Namely,

$$\mathbb{P}_{\mathcal{D}^m}(b_0 < a_0) = \mathbb{P}_{\mathcal{D}^m}\left(\text{for any } x \in S|_x,\ x \notin (a_0, a^*)\right)$$

$$\leq (1-\varepsilon)^m \quad \text{where} \quad m = |S| .$$

Similarly, $\mathbb{P}_{\mathcal{D}^m}(b_1 > a_1) \leq (1-\varepsilon)^m$.

Together, $\mathbb{P}_{\mathcal{D}^m}(L_{\mathcal{D}}(h_S) > \varepsilon) \leq 2(1-\varepsilon)^m < 2e^{-\varepsilon m}$

since $1-\varepsilon < e^{-\varepsilon}$ for $\varepsilon > 0$.

Setting $2e^{-\varepsilon m} \leq \delta$ yields $m \geq \frac{1}{\varepsilon} \log\left(\frac{2}{\delta}\right)$.

If $m \geq \frac{1}{\varepsilon} \log\left(\frac{2}{\delta}\right)$, then $\mathbb{P}_{\mathcal{D}^m}(L_{\mathcal{D}}(h_S) > \varepsilon) < \delta$.

That is, if $m \geq \frac{1}{\varepsilon} \log\left(\frac{2}{\delta}\right)$, then

$$\mathbb{P}_{\mathcal{D}^m}(L_{\mathcal{D}}(h_S) \leq \varepsilon) \geq 1-\delta .$$

Hence, $\mathcal{H}$ is PAC-learnable with a sample complexity

$$m_{\mathcal{H}} \leq \left\lceil \frac{1}{\varepsilon} \log\left(\frac{2}{\delta}\right) \right\rceil$$

☺

restriction of hypothesis class

- In order to characterize learnability, we need the following definitions.
- In the proof of NFL, we used a set $C \subset \mathcal{X}$

Definition (Restriction of $\mathcal{H}$ to C)

Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, \cdots, c_m\} \subset \mathcal{X}$ .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,

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

shattering

Definition (Restriction of $\mathcal{H}$ to C)

Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, \cdots, c_m\} \subset \mathcal{X}$ .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,

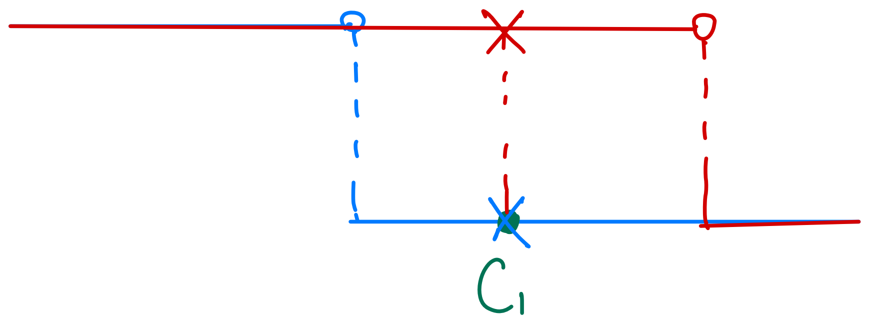
$$\mathcal{H}_C = \{(h(c_1), \cdots, h(c_m)) : h \in \mathcal{H}\}$$

Definition (Shattering)

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

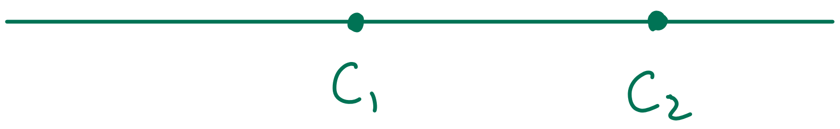
example of shattering

- Let $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ where $h_a(x) = \mathbf{1}_{\{x<a\}}(x)$
- Let $C = \{c_1\}$. Does $\mathcal{H}$ shatter C?



example of shattering

- Let $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ where $h_a(x) = \mathbf{1}_{\{x<a\}}(x)$
- Let $C = \{c_1\}$. Does $\mathcal{H}$ shatter C?
- What about $C = \{c_1, c_2\}$? $(c_1 < c_2)$



	c_1	c_2
✓	0	0
✓	1	1
✓	1	0
✗	0	1

$\mathcal{H}$ does not shatter C.

shattering

- If $\mathcal{H}$ shatters some set C of size $2m$, then we cannot learn $\mathcal{H}$ using m examples.

- A corollary of NFL: 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 exists 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 $L_{\mathcal{D}}(A(S)) \geq 1/8$.

VC (Vapnik-Chervonenkis) dimension

- The **VC-dimension** of a hypothesis class $\mathcal{H}$, denoted by $\mathrm{VCdim}(\mathcal{H})$, 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 $\mathcal{H}$ has infinite VC dimension.

VC (Vapnik-Chervonenkis) dimension

- The **VC-dimension** of a hypothesis class $\mathcal{H}$, denoted by $\mathrm{VCdim}(\mathcal{H})$, 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 $\mathcal{H}$ has infinite VC dimension.
- If $\mathrm{VCdim}(\mathcal{H}) = \infty$, then $\mathcal{H}$ is not PAC learnable.

VC (Vapnik-Chervonenkis) dimension

- The **VC-dimension** of a hypothesis class $\mathcal{H}$, denoted by $\mathrm{VCdim}(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.

- To show $\mathrm{VCdim}(\mathcal{H}) = d$ we need to show that
  1. There exists a set C of size d that is shattered by $\mathcal{H}$
  2. Every set C of size $d + 1$ cannot be shattered by $\mathcal{H}$

VC dim: example 1

- Let $\mathcal{H}$ be the set of threshold functions
  - $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ where $h_a(x) = \mathbf{1}_{\{x<a\}}(x)$
- Take $C = \{c_1\}$ for some c_1
  - Take $a = c_1 + 1$,then $h_a(c_1) = 1$
  - Take $a = c_1 - 1$,then $h_a(C_1) = 0$

$\mathcal{H}$ shatters C

- Consider $C' = \{c_1, c_2\}$ for any $c_1 < c_2$
  - No h_a can maps c_1 to 0 and c_2 to 1

$\mathcal{H}$ does not shatter C'

$\boxed{\mathrm{VCdim}(\mathcal{H}) = 1}$

Questions?

Reference

- *VC dimension*
    - *[S-S] Ch 6.1-6.3*