The VC-Dimension

Yujia Liu

School of Mathematical Sciences & Department of Computer Science
Peking University

December 24, 2020

Review

Notations

- $\mathcal{X}$: a set of instances
- $\mathcal{Y}$: a set of labels
- $\mathcal{H}$: a set of hypotheses
- S: a training set ($S \subset \mathcal{X}$)
- D: a distribution over $\mathcal{X} \times \mathcal{Y}$

Review

- Loss function $l : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \longrightarrow \mathbb{R}_+$
- Empirical risk
  $L_S(h) := \frac{1}{N} \sum_{i=1}^{N} l(h, x_i, y_i), S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$
- True risk $L_D(h) := \mathbb{E}_{(x,y)\sim D}[l(h, x, y)]$
- ERM learning $h_S := \arg\min_{h\in\mathcal{H}} L_S(h)$

Agnostic PAC learnable

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution D over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by D, the algorithm returns a hypothesis h such that, with probability of at least $1-\delta$ over the choice of the m training examples,

$$L_D(h) \leq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon$$

Uniform Convergence

We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t.a domain $\mathcal{X} \times \mathcal{Y}$ and a loss function l) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \longrightarrow \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution D over $\mathcal{X} \times \mathcal{Y}$, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d. according to D, then, with probability of at least $1 - \delta$, S is ϵ-representative.

$$\forall h \in \mathcal{H}, |L_S(h) - L_D(h)| \leq \epsilon$$

Finite classes are agnostic learnable

Let $\mathcal{H}$ be a finite hypothesis class. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \frac{log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

Furthermore, the class is agonostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \frac{2log(2|\mathcal{H}|/\delta)}{\epsilon^2}$$

'$\mathcal{H}$ is finite' is not necessary !

Finite classes are agnostic learnable

Let $\mathcal{H}$ be a finite hypothesis class. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \frac{log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

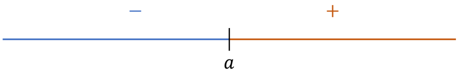
Furthermore, the class is agonostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \frac{2log(2|\mathcal{H}|/\delta)}{\epsilon^2}$$

'$\mathcal{H}$ is finite' is not necessary !

An example: Infinite-size classes can be learnable

$$\mathcal{H} = \{The\ thresholds\ of\ the\ real\ line\}$$

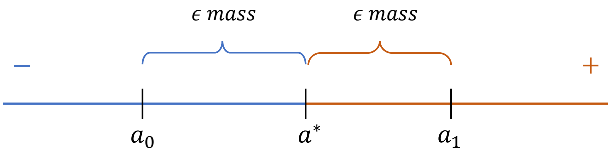


For any ϵ, δ, we need to find a function $m_{\mathcal{H}}(\epsilon, \delta)$, such that when $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, we have $L_D(h_s) \leq \epsilon$ with the probability of at least $1 - \delta$, where

$h_s = \text{argmin}_{h \epsilon \mathcal{H}} L_S(h)$ (ERM Learning)

An example: Infinite-size classes can be learnable

- Proof: Let a^* be a threshold such that h^* achieves $L_D(h^*) = 0$.



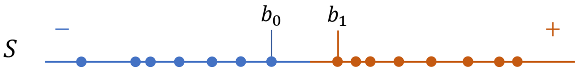
$$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^*)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^*, a_1)] = \epsilon$$

PS.

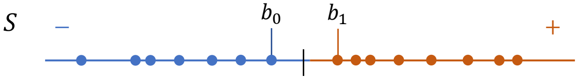
If $\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (-\infty, a^*)] \leq \epsilon$, we set $a_0 = -\infty$.

If $\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^*, +\infty)] \leq \epsilon$, we set $a_1 = +\infty$.

An example: Infinite-size classes can be learnable



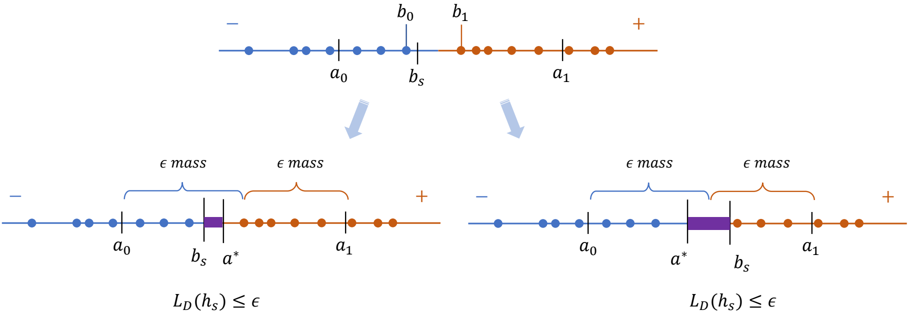
h_s=$\text{argmin}_{h \epsilon \mathcal{H}} L_S(h)$ (ERM Learning)



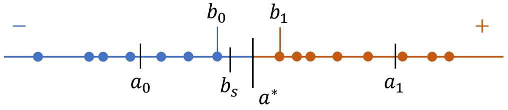
$b_0 < b_s < b_0$

An example: Infinite-size classes can be learnable

If $b_0 \geq a_0$ and $b_1 \leq a_1$



An example: Infinite-size classes can be learnable



$$b_0 \geq a_0 \text{ and } b_1 \leq a_1 \quad \Longrightarrow \quad L_D(h_s) \leq \epsilon$$

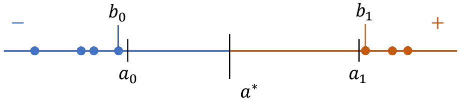
the converse-negative proposition

$$L_D(h_s) > \epsilon \quad \Longrightarrow \quad b_0 < a_0 \text{ or } b_1 > a_1$$

$$\mathbb{P}_{S\sim\mathcal{D}}[L_D(h_s) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}}[b_0 < a_0 \text{ or } b_1 > a_1]$$

$$\leq \mathbb{P}_{S\sim\mathcal{D}}[b_0 < a_0] + \mathbb{P}_{S\sim\mathcal{D}}[b_1 > a_1]$$

An example: Infinite-size classes can be learnable



$$\mathbb{P}_{S\sim\mathcal{D}}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}}[\forall\,(x,y)\in S, x \notin (a_0, a^*)] = (1-\epsilon)^m \le e^{-\epsilon m}$$

$$\mathbb{P}_{S\sim\mathcal{D}}[b_1 > a_1] = \mathbb{P}_{S\sim\mathcal{D}}[\forall (x,y)\in S, x \notin (a^*, a_1)] = (1-\epsilon)^m \le e^{-\epsilon m}$$

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}}[L_D(h_s) > \epsilon] &\le \mathbb{P}_{S\sim\mathcal{D}}[b_0 < a_0 \text{ or } b_1 > a_1] \\ &\le \mathbb{P}_{S\sim\mathcal{D}}[b_0 < a_0] + \mathbb{P}_{S\sim\mathcal{D}}[b_1 > a_1] \\ &\le 2e^{-\epsilon m}\end{aligned}$$

When $m > \log(2/\delta)/\epsilon$

$$\mathbb{P}_{S\sim\mathcal{D}}[L_D(h_s) > \epsilon] \le \delta$$

A Question

What is the sufficient condition for learnability?

The VC-Dimension of $\mathcal{H}$ is finite!

A Question

What is the sufficient condition for learnability?

The VC-Dimension of $\mathcal{H}$ is finite!

What is the VC-Dimension?

Before introducing the VC-dimension, we need to learn some definitions:

- *(Restriction $\mathcal{H}$ to C.)* Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, c_2, \cdots, c_m\} \subset \mathcal{X}$. 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

$$\mathcal{H}_C = \{(h(c_1), h(c_2), \cdots, h(c_m)) : h \in \mathcal{H}\}$$

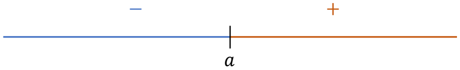
where we represent each function from C to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

What is the VC-Dimension?

Before introducing the VC-dimension, we need to learn some definitions:

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

$\mathcal{H} = \{The\ thresholds\ of\ the\ real\ line\}$



1. $C = \{c_1\}$ √
2. $C = \{c_1, c_2\}\ (c_1 < c_2)$ ×

What is the VC-Dimension?

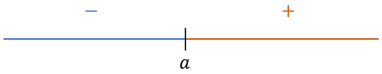
(VC-Dimension.) The VC-Dimension of a hypothesis class $\mathcal{H}$, denoted $\mathrm{VCdim}(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite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

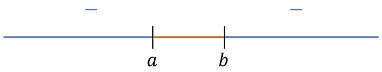
The VC-Dimension is d:

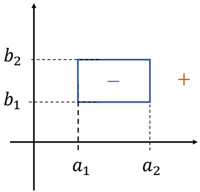
- There exists a set of d points that can be shattered
- There is no set of $d+1$ points that can be shattered

Some examples for the VC-Dimension

- Threshold Functions
- Intervals
- Axis Aligned Rectangles
- The number of parameters
- Finite Classes







$$\mathcal{H} = \{x \mapsto sgn(sin(\theta x)) : \theta \in \mathbb{R}\}$$

- The number of parameters in $\mathcal{H}$ is 1, but VCdim($\mathcal{H}$)$= \infty$.
- (Lemma) If $0.x_1 x_2 x_3 \cdots$, is the binary expansion of $x \in (0,1)$, then for any natural number m, $sgn(sin(2^m \pi x)) = (1 - x_m)$.
- For $\forall d \in \mathbb{N}$, we construct a set of d points: (1) the label of c_i is y_i; (2) $c_i = 2^{-i} (i = 1, 2, \cdots, d)$. Set $\theta = \pi(\sum_{i=1}^{d} (1 - y_i) 2^i)$

The VC-Dimension & The number of parameters

$$\mathcal{H} = \{x \mapsto sgn(sin(\theta x)) : \theta \in \mathbb{R}\}$$

- The number of parameters in $\mathcal{H}$ is 1, but VCdim($\mathcal{H}$)$= \infty$.
- (Lemma) If $0.x_1x_2x_3\cdots$, is the binary expansion of $x \in (0, 1)$, then for any natural number m, $sgn(sin(2^m \pi x)) = (1 - x_m)$.
- For $\forall d \in \mathbb{N}$, we construct a set of d points: (1) the label of c_i is y_i; (2) $c_i = 2^{-i}(i = 1, 2, \cdots, d)$. Set $\theta = \pi(\sum_{i=1}^{d}(1 - y_i)2^i)$

The VC-Dimension & The number of parameters

$$\mathcal{H} = \{x \mapsto sgn(sin(\theta x)) : \theta \in \mathbb{R}\}$$

- The number of parameters in $\mathcal{H}$ is 1, but VCdim($\mathcal{H}$)$= \infty$.
- (Lemma) If $0.x_1 x_2 x_3 \cdots$, is the binary expansion of $x \in (0,1)$, then for any natural number m, $sgn(sin(2^m \pi x)) = (1 - x_m)$.
- For $\forall d \in \mathbb{N}$, we construct a set of d points: (1) the label of c_i is y_i; (2) $c_i = 2^{-i} (i = 1, 2, \cdots, d)$. Set $\theta = \pi(\sum_{i=1}^{d}(1 - y_i)2^i)$

The VC-Dimension & The number of parameters

$$\mathcal{H} = \{x \mapsto sgn(sin(\theta x)) : \theta \in \mathbb{R}\}$$

- The number of parameters in $\mathcal{H}$ is 1, but VCdim($\mathcal{H}$)$= \infty$.
- (Lemma) If $0.x_1x_2x_3\cdots$, is the binary expansion of $x \in (0,1)$, then for any natural number m, $sgn(sin(2^m\pi x)) = (1 - x_m)$.
- For $\forall d \in \mathbb{N}$, we construct a set of d points: (1) the label of c_i is y_i; (2) $c_i = 2^{-i}(i = 1, 2, \cdots, d)$. Set $\theta = \pi(\sum_{i=1}^{d}(1 - y_i)2^i)$

Questions to be answered

- Is 'finite VC-Dimension' is a necessary and sufficient condition for the PAC learnability?
- Why we come to the VC-Dimension?

Review

(No Free Lunch.)

Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$.

Assume there is a set $C \subset \mathcal{X}$ of size $2m$ that can be shattered by $\mathcal{H}$. Let $S \subset C$ of size m be a training set.

Then, for any learning algorithm, A, there exist a distribution D over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in \mathcal{H}$.

Such that $L_D(h) = 0$ but with probability of at least $1/7$ over the choice of $S \sim D$, we have that $L_D(A(S)) \geq 1/8$.

PAC learnable $\Rightarrow$ *VCdim*$(\mathcal{H})$ is finite

Proof. *VCdim*$(\mathcal{H})$ is infinite $\Rightarrow$ $\mathcal{H}$ is not PAC learnable

Since $\mathcal{H}$ has an infinite VC-Dimension, for any training set size m, there existed a shattered set of $2m$, and the claim follows by *No Free Lunch Theorem* ($\mathcal{H}$ is not learnable).

$VCdim(\mathcal{H})$ is finite $\Rightarrow$ PAC learnable

- Step 1: $VCdim(\mathcal{H})$ is finite $\Rightarrow \tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$
- Step 2: $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \Rightarrow \mathcal{H}$ has the uniform convergence property
- Step 3: $\mathcal{H}$ has the uniform convergence property $\Rightarrow \mathcal{H}$ is PAC learnable

Growth Function

The growth function measures the maximal "effective" size of $\mathcal{H}$ on a set of m examples.

- (Growth Function.) Let $\mathcal{H}$ be a hypothesis class. Then the growth function of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as:

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCdim(\mathcal{H}) = d < +\infty$. Then, for all $m \geq d$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. Furthermore, $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.

- $|\mathcal{H}_C| = O(|C|^d)$.
- The size of $\mathcal{H}_C$ grows polynomially rather than exponentially with $|C|$.

Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCdim(\mathcal{H}) = d < +\infty$. Then, for all $m \geq d$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. Furthermore, $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.

- $|\mathcal{H}_C| = O(|C|^d)$.
- The size of $\mathcal{H}_C$ grows polynomially rather than exponentially with $|C|$.

Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCdim(\mathcal{H}) = d < +\infty$. Then, for all $m \geq d$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. Furthermore, $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.

- $|\mathcal{H}_C| = O(|C|^d)$.
- The size of $\mathcal{H}_C$ grows polynomially rather than exponentially with $|C|$.

Proof of Sauer's Lemma

A claim: For any $C = \{c_1, c_2, \cdots, c_m\}$, we have

$$\forall \mathcal{H}, |\mathcal{H}_C| \leq |B \subset C : \mathcal{H} \textit{ shatters } B|$$

Proof: (the mathematical induction)

(1) When $m = 1, \checkmark$
(2) Suppose the inequality holds for sets of size $k < m$
(3) Prove the inequality holds for sets of size m.

Proof of Sauer's Lemma – Claim (3)

Fix $\mathcal{H}$ and $C = \{c_1, c_2, \cdots, c_m\}$. Denote $C^{'} = \{c_2, \cdots, c_m\}$. In addition, define the following two sets:

$$Y_0 = \{(y_2, y_3, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathcal{H}_C \vee (1, y_2, \cdots, y_m) \in \mathcal{H}_C\}$$

$$Y_1 = \{(y_2, y_3, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \cdots, y_m) \in \mathcal{H}_C\}$$

We have:
(1) $|\mathcal{H}_C| = |Y_0| + |Y_1|$
(2) $Y_0 = \mathcal{H}_{C'}$. Moreover,

$$\begin{aligned} |Y_0| = |\mathcal{H}_{C'}| &\le |\{B \subset C^{'} : \mathcal{H} \text{ shatters } B\}| \\ &= |\{B \subset C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}| \end{aligned}$$

Proof of Sauer's Lemma – Claim (3)

$C = \{c_1, c_2, \cdots, c_m\}$ and $C' = \{c_2, \cdots, c_m\}$

$$Y_0 = \{(y_2, y_3, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathcal{H}_C \vee (1, y_2, \cdots, y_m) \in \mathcal{H}_C\}$$

$$Y_1 = \{(y_2, y_3, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \cdots, y_m) \in \mathcal{H}_C\}$$

Next, we define $\mathcal{H}' \subset \mathcal{H}$ to be

$$\begin{aligned}\mathcal{H}' = &\{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t.} \\ &\quad (1 - h'(c_1), h'(c_2), \cdots, h'(c_m)) = (h(c_1), h(c_2), \cdots, h(c_m))\}\end{aligned}$$

Then, it is obvious that $Y_1 = \mathcal{H}'_{C'}$. Moreover,

$$\begin{aligned}|Y_1| = |\mathcal{H}'_{C'}| &\le |\{B' \subset C' : \mathcal{H}' \text{ shatters } B'\}| = |\{B' \subset C' : \mathcal{H}' \text{ shatters } B' \cup c_1\}| \\ &= |\{B \subset C : c_1 \in B \wedge \mathcal{H}' \text{ shatters } B\}| \\ &\le |\{B \subset C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}|\end{aligned}$$

Proof of Sauer's Lemma – Claim (3)

$C = \{c_1, c_2, \cdots, c_m\}$ and $C' = \{c_2, \cdots, c_m\}$

$$Y_0 = \{(y_2, y_3, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathcal{H}_C \vee (1, y_2, \cdots, y_m) \in \mathcal{H}_C\}$$

$$Y_1 = \{(y_2, y_3, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \cdots, y_m) \in \mathcal{H}_C\}$$

Next, we define $\mathcal{H}' \subset \mathcal{H}$ to be

$$\begin{aligned}\mathcal{H}' =& \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t.} \\ & (1 - h'(c_1), h'(c_2), \cdots, h'(c_m)) = (h(c_1), h(c_2), \cdots, h(c_m))\}\end{aligned}$$

Then, it is obvious that $Y_1 = \mathcal{H}'_{C'}$. Moreover,

$$\begin{aligned}|Y_1| = |\mathcal{H}'_{C'}| \le |\{B' \subset C' : \mathcal{H}' \text{ shatters } B'\}| &= |\{B' \subset C' : \mathcal{H}' \text{ shatters } B' \cup c_1\}| \\ &= |\{B \subset C : c_1 \in B \wedge \mathcal{H}' \text{ shatters } B\}| \\ &\le |\{B \subset C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}|\end{aligned}$$

Proof of Sauer's Lemma

Overall, we have shown that

$$\begin{aligned}
|\mathcal{H}_C| &= |Y_0| + |Y_1| \\
&\leq |\{B \subset C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}| + |\{B \subset C : c_1 \in B \wedge \mathcal{H}^{'} \text{ shatters } B\}| \\
&= |\{B \subset C : \mathcal{H} \text{ shatters } B\}|
\end{aligned}$$

If $VCdim(\mathcal{H}) \leq d$, then no set whose size is larger than d can be shattered by $\mathcal{H}$. Therefore

$$|\{B \subset C : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{m}{i}$$

$VCdim(\mathcal{H})$ is finite $\Rightarrow$ The uniform convergence

Theorem (6.11)

Let $\mathcal{H}$ be a class of hypothesis and let $\tau_{\mathcal{H}}$ be its growth function. The loss is 0-1 loss. Then, for every D and every $\delta \in (0,1)$, with probability of at least $1-\delta$ over the choice of $S \sim D$ we have

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

Proof.

(1) $\mathbb{E}_{S\sim D}[\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)|] \leq \frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}$

(2) Markov's inequality (Section B.1) □

$VCdim(\mathcal{H})$ is finite $\Rightarrow$ The uniform convergence

- Sauer's lemma: when $m > d$, we have $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$
- Theorem 6.11: with probability of at least $1 - \delta$,

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

- Combining these two conclusions, we have,

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{d\log(2em/d)}}{\delta\sqrt{2m}}$$

$VCdim(\mathcal{H})$ is finite $\Rightarrow$ The uniform convergence

- Sauer's lemma: when $m > d$, we have $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$
- Theorem 6.11: with probability of at least $1 - \delta$,

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

- Combining these two conclusions, we have,

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{d\log(2em/d)}}{\delta\sqrt{2m}}$$

$VCdim(\mathcal{H})$ is finite $\Rightarrow$ The uniform convergence

- Assume $\sqrt{d\log(2em/d)} \geq 4$,
$$|L_D(h) - L_S(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}}$$
- To ensure the proceeding is at most ϵ, we need,
$$m \geq \frac{2d\log(m)}{(\delta\epsilon)^2} + \frac{2d\log(2e/d)}{(\delta\epsilon)^2}$$
- (Lemma A.2 in Appendix A) There exists a function $f(\epsilon, \delta)$ which is a sufficient condition for the proceeding to hold
- Finally, we can let $m_{\mathcal{H}}^{UC}(\epsilon, \delta) = f(\epsilon, \delta)$. Then the uniform convergence property of $\mathcal{H}$ is proved

$VCdim(\mathcal{H})$ is finite $\Rightarrow$ The uniform convergence

- Assume $\sqrt{d\log(2em/d)} \geq 4$,

$$|L_D(h) - L_S(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}}$$

- To ensure the proceeding is at most ϵ, we need,

$$m \geq \frac{2d\log(m)}{(\delta\epsilon)^2} + \frac{2d\log(2e/d)}{(\delta\epsilon)^2}$$

- (Lemma A.2 in Appendix A) There exists a function $f(\epsilon, \delta)$ which is a sufficient condition for the proceeding to hold
- Finally, we can let $m_{\mathcal{H}}^{UC}(\epsilon, \delta) = f(\epsilon, \delta)$. Then the uniform convergence property of $\mathcal{H}$ is proved

$VCdim(\mathcal{H})$ is finite $\Rightarrow$ The uniform convergence

- Assume $\sqrt{d \log(2em/d)} \geq 4$,

$$|L_D(h) - L_S(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}}$$

- To ensure the proceeding is at most ϵ, we need,

$$m \geq \frac{2d\log(m)}{(\delta\epsilon)^2} + \frac{2d\log(2e/d)}{(\delta\epsilon)^2}$$

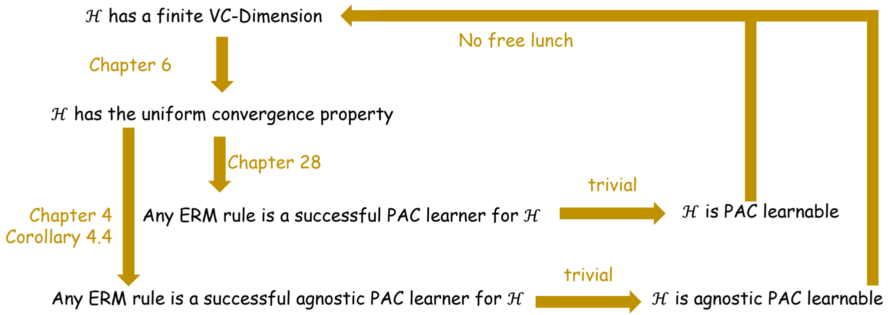
- (Lemma A.2 in Appendix A) There exists a function $f(\epsilon, \delta)$ which is a sufficient condition for the proceeding to hold
- Finally, we can let $m_{\mathcal{H}}^{UC}(\epsilon, \delta) = f(\epsilon, \delta)$. Then the uniform convergence property of $\mathcal{H}$ is proved

$VCdim(\mathcal{H})$ is finite $\Rightarrow$ The uniform convergence

- Assume $\sqrt{d\log(2em/d)} \geq 4$,
$$|L_D(h) - L_S(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}}$$
- To ensure the proceeding is at most ϵ, we need,
$$m \geq \frac{2d\log(m)}{(\delta\epsilon)^2} + \frac{2d\log(2e/d)}{(\delta\epsilon)^2}$$
- (Lemma A.2 in Appendix A) There exists a function $f(\epsilon, \delta)$ which is a sufficient condition for the proceeding to hold
- Finally, we can let $m_{\mathcal{H}}^{UC}(\epsilon, \delta) = f(\epsilon, \delta)$. Then the uniform convergence property of $\mathcal{H}$ is proved

Conclude

(The Fundamental Theorem of Statistical Learning) Let the loss function be the $0 - 1$ loss.



Conclude

(The Fundamental Theorem of Statistical Learning - Quantitative Version)
Let the loss function be the $0-1$ loss. Assume $VCdim(\mathcal{H}) = d < +\infty$.
Then,

	Uniform Convergence	Agnostic PAC Learnable	PAC Learnable
$m_{\mathcal{H}}(\epsilon, \delta)$	$\Theta(\frac{d + log(1/\delta)}{\epsilon^2})$	$\Theta(\frac{d + log(1/\delta)}{\epsilon^2})$	$O(\frac{d\ log(1/\epsilon) + log(1/\delta)}{\epsilon^2})$