The VC Dimension

Xuyuan Liu

May 29, 2024

DARTMOUTH

Table of Contents

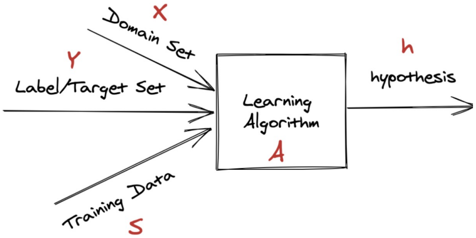
Statistical Learning Framework

Learner's Input:

- **Domain Set**: the set of objects that we wish to label, shown by $\mathcal{X}$.
- **Label Set**: Set of all possible labels, shown by $\mathcal{Y}$.
- **Training Data**: $S = ((x_1, y_1) \dots (x_m, y_m))$ is a finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$: that is, a sequence of labeled domain points.

Output:

- **Hypothesis(Predictor)** The output is a mapping function $h : \mathcal{X} \to \mathcal{Y}$ that predicts labels for new domain points. The hypothesis learned from training data S is denoted by $h_S : x \mapsto y$.



To measure the success of a model, define the *error of classifier* as the probability that it incorrectly predicts the label on a random data point from the underlying distribution.

Assuming a "correct" labeling function $f : \mathcal{X} \to \mathcal{Y}$ where $y_i = f(x_i)$, and given a probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the likelihood of an error by classifier h is assessed when points are drawn randomly from $\mathcal{D}$. The **true error (risk)** of h is thus defined as $\mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$, which can also be expressed as:

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}(\{x : h(x) \neq f(x)\}).$$

Empirical Risk Minimization (ERM)

This learning paradigm - coming up with a predictor h that minimizes *the train error(empirical risk)* $L_{\mathcal{S}}(h)-$ is called Empirical Risk Minimization(ERM).

$$L_{\mathcal{S}}(h) \stackrel{\text{def}}{=} \frac{\left|\left\{i \in [m] : h\left(x_i\right) \neq y_i\right\}\right|}{m}$$

where $[m]=\{1,\ldots,m\}$

And let h_S denote a result of applying ERM_H to S,

$$h_{\mathcal{S}} \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_{\mathcal{S}}(h).$$

PAC Learning

DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

- ϵ quantifies the algorithm's accuracy, while δ indicates the likelihood of achieving this accuracy.

PAC Learning

1 Review

DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \rightarrow \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \rightarrow \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

- ϵ quantifies the algorithm's accuracy, while δ indicates the likelihood of achieving this accuracy.
- $m_{\mathcal{H}}(\epsilon, \delta)$(sample complexity) is the minimal integer that satisfies the requirement:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

there is not always a fixed labeling function f

DEFINITION 3.3 (Agnostic PAC Learnability) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ 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 $\mathcal{D}$, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon.$$

- Define relative to some benchmark hypothesis class

Uniform Convergence

1 Review

A hypothesis class $\mathcal{H}$ has the uniform convergence property iff there is a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that, for every probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, every gap $\epsilon \in (0,1)$ and every confidence level $\delta \in (0,1)$, if $|S| \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta)$, then with probability at least $1-\delta$ over the choice of S, the inequality $|\ell_S(h) - \ell_D(h)| \leq \epsilon$ holds for all $h \in \mathcal{H}$

- As the training set size increases, a learning algorithm's performance converges to its expected performance across all training sets.

Uniform Convergence

A hypothesis class $\mathcal{H}$ has the uniform convergence property iff there is a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that, for every probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, every gap $\epsilon \in (0,1)$ and every confidence level $\delta \in (0,1)$, if $|S| \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta)$, then with probability at least $1 - \delta$ over the choice of S, the inequality $|\ell_S(h) - \ell_D(h)| \leq \epsilon$ holds for all $h \in \mathcal{H}$

- As the training set size increases, a learning algorithm's performance converges to its expected performance across all training sets.
- $m_{\mathcal{H}}^{\mathrm{UC}}$ measures the (minimal) sample complexity to obtain the uniform convergence property.

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

Fundamental Theorem

Summary

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:

- $\mathcal{H}$ is agnostic PAC learnable.
- $\mathcal{H}$ is PAC learnable.
- $\mathcal{H}$ has the uniform convergence property.
- Any ERM rule is a successful PAC learner for $\mathcal{H}$.

For finite-size class $\mathcal{H}$, we conclude sample complexity with:

$$m_{\mathcal{H}}(\epsilon, \delta) \le m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \le \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

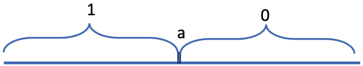
A finite-size class is PAC learnable, what about an infinite-size class?

Table of Contents

Infinite-Size Classes Can Be Learnable

Let $\mathcal{H}$ be the set of threshold functions over the real line, namely, $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a : \mathbb{R} \to \{0, 1\}$ is a function such that $h_a(x) = 1_{[x<a]}$, and $h_a(x) = 0_{[x \geq a]}$

Clearly, $\mathcal{H}$ is of infinite size.



However, $\mathcal{H}$ **is PAC learnable, using the ERM rule, with sample complexity of** $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$.

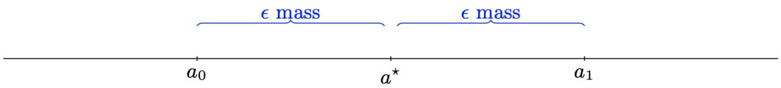
Infinite-Size Classes Can Be Learnable

$\mathcal{H}$ is PAC learnable with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$

Proof:

Given distribution $\mathcal{D}_x$ and a threshold $a^\star$ where the hypothesis $h^\star(x) = 1_{[x<a^\star]}$ achieves $L_{\mathcal{D}}(h^\star) = 0$, consider $a_0 < a^\star < a_1$:

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

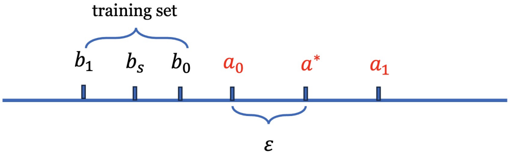


Infinite-Size Classes Can Be Learnable

$\mathcal{H}$ is PAC learnable with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$

Proof:

Let $b_0 = \max\{x \in \mathcal{S}\}$ and $b_1 = \min\{x \in \mathcal{S}\}$($\mathcal{S}$ is training set). So the threshold corresponding to an ERM hypothesis $b_{\mathcal{S}} \in (b_0, b_1)$. Therefore, a sufficient condition for $L_{\mathcal{D}}(h_{\mathcal{S}}) \geq \epsilon$ is that either $b_0 \leq a_0$ or $b_1 \geq a_1$.



$$\begin{aligned}\underset{\mathcal{S}\sim\mathcal{D}^m}{\mathbb{P}}[L_{\mathcal{D}}(h_{\mathcal{S}}) > \epsilon] &\leq \underset{\mathcal{S}\sim\mathcal{D}^m}{\mathbb{P}}[b_0 < a_0 \vee b_1 > a_1] \\ &\leq \underset{\mathcal{S}\sim\mathcal{D}^m}{\mathbb{P}}[b_0 < a_0] + \underset{\mathcal{S}\sim\mathcal{D}^m}{\mathbb{P}}[b_1 > a_1]\end{aligned}$$

Infinite-Size Classes Can Be Learnable

$\mathcal{H}$ is PAC learnable with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$

Proof: The event $b_0 < a_0$ happens only if all samples in S are outside the interval (a_0, a^*) with mass ϵ, leading to:

$$\mathbb{P}_{\mathcal{S}\sim D^m}[b_0 < a_0] = (1-\epsilon)^m \leq e^{-\epsilon m}$$

Given $m > \frac{\log(2/\delta)}{\epsilon}$, $\mathbb{P}_{\mathcal{S}\sim D^m}[b_0 < a_0] \leq \frac{\delta}{2}$. Hence:

$$\mathbb{P}_{\mathcal{S}\sim D^m}[L_{\mathcal{D}}(h_{\mathcal{S}}) > \epsilon] \leq \delta$$

Thus, $\mathcal{H}$ is PAC learnable.

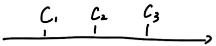
The example illustrates that while a finite $\mathcal{H}$ suffices for learnability, it is not necessary. Instead of the number of $\mathcal{H}$, maybe some type of learn-ability be more important.

Restriction

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

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

For example, we consider the threshold functions on C, we get:

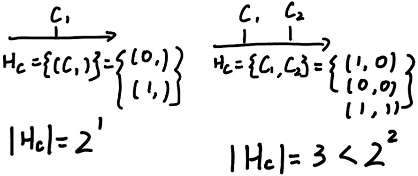


$$H_C = \{[(c_1, c_2, c_3)]\} = \left\{ \begin{array}{ll} (1,1,1) & (0,0,0) \\ (1,1,0) & \\ (1,0,0) & \end{array} \right\}$$

Thus, $|\mathcal{H}_C| = 4$

Shattering

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, $|\mathcal{H}_C| = 2^{|C|}$.



Shattering Not Shattering

Definition: (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.

Notes

To show that $\mathrm{VCdim}(\mathcal{H}) = d$ 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 $d + 1$ is not shattered by $\mathcal{H}$.

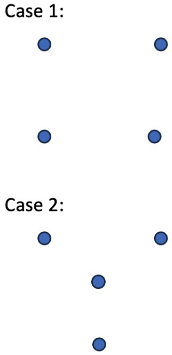
What is VC dimension?

Imagine a following game between two players α and β:

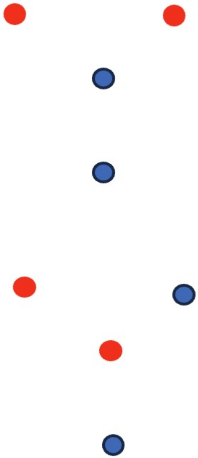
What is VC dimension?

Imagine a following game between two players α and β:

- First, player α plots $d = 4$ points on a piece of paper. She may place the points however she likes.



What is VC dimension?



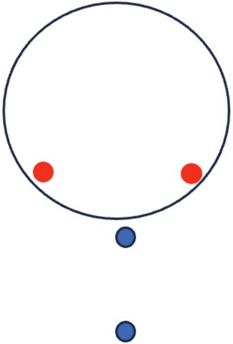
Imagine a following game between two players α and β:

- First, player α plots $d = 4$ points on a piece of paper. She may place the points however she likes.
- Next, player β marks several of the drawn points.

What is VC dimension?

Imagine a following game between two players α and β:

- First, player α plots $d = 4$ points on a piece of paper. She may place the points however she likes.
- Next, player β marks several of the drawn points.
- Finally, player α should draw a circle such that all the marked points are inside a circle, and all the unmarked points are outside.

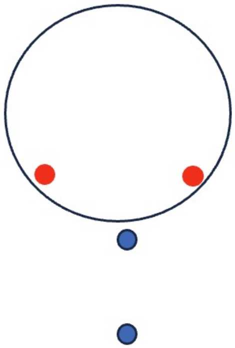


What is VC dimension?

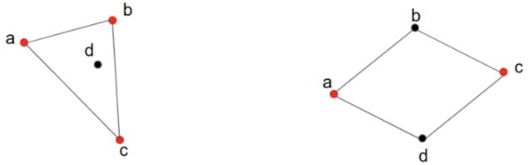
Imagine a following game between two players α and β:

- First, player α plots $d = 4$ points on a piece of paper. She may place the points however she likes.
- Next, player β marks several of the drawn points.
- Finally, player α should draw a circle such that all the marked points are inside a circle, and all the unmarked points are outside.

The player α wins if she can draw such a circle at step #3. The player β wins if making such circle is impossible.



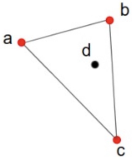
It turns out that β has a winning strategy:

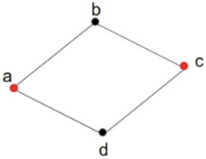


However, for $\mathrm{d} = 3$ points, it turns out that α has a winning strategy.

What is VC dimension

It turns out that β has a winning strategy:



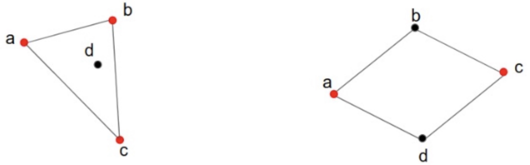


However, for $\mathrm{d} = 3$ points, it turns out that α has a winning strategy.

- The largest number d at which the game is winnable by player α is called the VC dimension of our classification set.

What is VC dimension

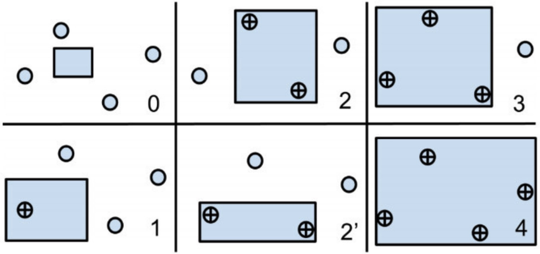
It turns out that β has a winning strategy:



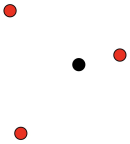
However, for $\mathrm{d} = 3$ points, it turns out that α has a winning strategy.

- The largest number d at which the game is winnable by player α is called the VC dimension of our classification set.
- Here, “circles” are our classification set.

Let $\mathcal{H}$ be the class of axis aligned rectangles,consider $|C| = 4$:

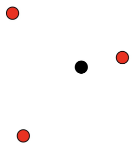


The VC dimension of the rectangle function class is at least 4.
In fact, one can prove that when $|C| = 5$, C can not be shattered by $\mathcal{H}$.



Let $\mathcal{H}$ still be the class of axis aligned rectangles, what if the 4 points are arranged like this?

What is VC dimension



Let $\mathcal{H}$ still be the class of axis aligned rectangles, what if the 4 points are arranged like this?

Notes

To show that $\mathrm{VCdim}(\mathcal{H}) = d$ we need to show that

1. There **exists** a set C of size d that is shattered by $\mathcal{H}$.

Infinite VC Dimension

Let $\mathcal{H}$ be the class of sine functions: $\{t \mapsto \sin(\omega t) : \omega \in \mathbb{R}\}$.

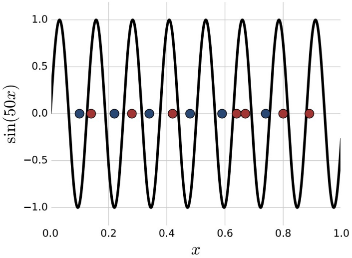


Figure 3.5
An example of a sine function (with $\omega = 50$) used for classification.

we can prove that $VCdim(\mathcal{H}) = +\infty$

Infinite VC Dimension

Let $\mathcal{H}$ be a class of infinite VC-dimension. Then, $\mathcal{H}$ is not PAC learnable.

- **Infinite Shattering**: Any finite dataset can be shattered by $\mathcal{H}$ in infinitely many ways.
- **No Constraints**: The hypothesis class can perfectly fit all data, including noise and outliers.
- **Generalization Failure**: Even with large samples, some hypotheses perfectly fit training data but fail on new data.

Finite Class

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

Remark

- A finite class implies a finite VC-dimension, but a finite VC-dimension does not guarantee a finite class.
- An infinite class does not guarantee an infinite VC-dimension, and vice versa.

For finite-size class $\mathcal{H}$, we conclude sample complexity with:

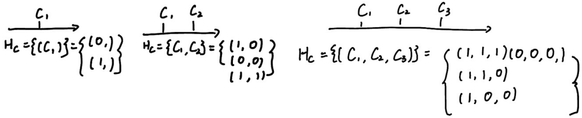
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

Can we incorporate VC-Dimension into this formula for an infinite class?

(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

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

we still consider the threshold functions on C, we get:



$\tau_H(1) = 2$ $\quad$ $\tau_H(2) = 3$ $\quad$ $\tau_H(3) = 4$

If $\mathrm{VCdim}(\mathcal{H}) = d$, when $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$. In such cases, $\mathcal{H}$ induces all possible functions from C to $\{0, 1\}$.

But What if m increase, will this function still increases exponentially?

Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$. Then, for all m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. In particular, if $m > d + 1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d \ll 2^m$.

Given the definitions of the growth function and the uniform convergence property, we can demonstrate:

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

$$|L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{\log\left(\tau_{\mathcal{H}}(2m)\right)}}{\delta\sqrt{2m}}$$

From Sauer's lemma, we know that for $m > d$, it holds that $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$, thus, for

$$|L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{\log\left(\tau_{\mathcal{H}}(2m)\right)}}{\delta\sqrt{2m}}$$

we can get :

$$|L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}} = O\left(\sqrt{\frac{\log(m/d)}{(m/d)}}\right)$$

$$|L_{\mathcal{S}}(h) - L_{\mathcal{D}}(h)| \leq \frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}} = O\left(\sqrt{\frac{\log(m/d)}{(m/d)}}\right)$$

Notes

1. If $\mathcal{H}$ has small effective size(finite VC-Dimension) then it enjoys the uniform convergence property.
2. How the ratio of m/d effect for generalization

Fundamental Theorem
Summary

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:

- $\mathcal{H}$ is agnostic PAC learnable.
- $\mathcal{H}$ is PAC learnable.
- $\mathcal{H}$ has the uniform convergence property.
- Any ERM rule is a successful PAC learner for $\mathcal{H}$.
- **$\mathcal{H}$ has a finite VC-dimension.**

Fundamental Theorem(Quantitative Version)

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. Assume that $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. Then, there are absolute constants C_1, C_2 such that:

- $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

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

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

Table of Contents

3 Discussion