PAC Learnability(Cont.)

Spring 2025

Outline

The VC-Dimension

- Our goal:
  - figure out which classes $\mathcal{H}$ are PAC learnable, and
  - characterize exactly the sample complexity of learning a given hypothesis calss
- Recall that finite classes are learnable.
  - Every $\mathcal{H}$ is PAC learnable with sample complexity

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

  - $\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity

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

- Let $\mathcal{H}$ be *the set of all functions* from an *infinite domain set* $\mathcal{X}$ to $\{0, 1\}$. Then, $\mathcal{H}$ is not PAC learnable.
- Can infinite-size classes be learnable?

Finiteness vs Infiniteness:

- Consider $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a(x) = \mathbb{I}(x < a)$.
- $\mathcal{H}$ is of infinite size.
- $\mathcal{H}$ is PAC learnable, using the ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$.

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, $\mathcal{H}_C = \{(h(c_1), \cdots, h(c_m)) : h \in \mathcal{H}\}$, 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 $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$. That is,

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

Consider $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ again.

- Consider $C_1 = \{c_1\}$, then $h_{c_1+1}(c_1) = 1$ and $h_{c_1-1}(c_1) = 0$. So, $\mathcal{H}$ shatters C_1;
- Consider $C_2 = \{c_1, c_2\}$, where $c_1 < c_2$, then $h_a(c_1) = 0$ implies $h_a(c_2) = 0$. So, $h' : C_2 \to \{0, 1\}$ is not icluded in $\mathcal{H}_{C_2}$, and C_2 is not shattered by $\mathcal{H}$.

Corollary

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 $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$.

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.

- 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 $\mathcal{H}$.
- Let $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, then $\mathrm{VCdim}(\mathcal{H}) = 1$.
- Let $\mathcal{H}$ be a finite class, then $\mathrm{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.

Theorem

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

The converse is also true: A finite VC-dimension guarantees learnability.

Theorem (The Fundamental Theorem of Statistical Learning)

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

Theorem (The Fundamental Theorem of Statistical Learning-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:

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*

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

2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

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

3. *$\mathcal{H}$ is PAC learnable with sample complexity*

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

Definition (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

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

- If $\mathrm{VCdim}(\mathcal{H}) = d < \infty$, then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$.
- How does the growth function increase when $m > d$?

Lemma (Sauer-Shelah-Perles)

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

Theorem (6.11)

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

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

Theorem (The Fundamental Theorem of Statistical Learning)

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

Proof of the Fundamental Theorem

It suffices to prove that if the VC-dimension is finite then the uniform convergence property holds.

- From Sauer's lemma we have that for $m > d$, $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$. Combining this with Theorem 6.11 we obtain that with probability of at least $1 - \delta$,

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

For simplicity assume that $\sqrt{d \log(2em/d)} \geq 4$; hence,

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

- To ensure that $\frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}}$ is at most ϵ we need that

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

- Notice that $x \geq 4a\log(2a) + 2b \Rightarrow x \geq a\log(x) + b$ for $a \geq 1$ and $b > 0$(LEMMA A.2), then a sufficient condition for the preceding to hold is that

$$m \geq 4\frac{2d}{(\delta\epsilon)^2}\log(\frac{4d}{(\delta\epsilon)^2}) + \frac{4d\log(2e/d)}{(\delta\epsilon)^2}.$$

- Hence, $\mathcal{H}$ has the uniform convergence property.

Nonuniform Learnability

Spring 2025

Outline

- Recall the definition of Agnostic PAC Learnability:

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a learning algorithm, A, and a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that, for every $\delta, \epsilon \in (0,1)$ and for every distribution $\mathcal{D}$, if $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, then with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ it holds that

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

- This implies that for every $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon.$$

- Competitiveness:

We say that a hypothesis h is (ϵ, δ)-competitive with another hypothesis h' if, with probability higher than $(1 - \delta)$,

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

- $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ vs $m \geq m_{\mathcal{H}}(\epsilon, \delta, h)$?

Definition (Nonuniform Learnability)

A hypothesis class $\mathcal{H}$ is nonuniformly learnable if there exists a learning algorithm, A, and a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0,1)$ and for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ then for every distribition $\mathcal{D}$, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, it holds that

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon.$$

- Agnostic PAC Learnability VS Nonuniform Learnability
  - Both two notions require that the output hypothesis will be (ϵ, δ)-competitive with every other hypothesis in the class.
  - The difference is the question of whether the sample size may depend on the hypothesis h to which the error of $A(S)$ is compared.

Characterizing Nonuniform Learnability

- Recall that uniform convergence is sufficient for agnostic PAC learnabilitty.
- Can we generalize this to nonuniform learnability?

Theorem (7.3)

Let $\mathcal{H}$ be a hypothesis class that can be written as a countable union of hypothesis classes, $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ enjoys the uniform convergence property. Then, $\mathcal{H}$ is nonuniformly learnable.

Theorem (7.2)

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable if and only if it is a countable union of agnostic PAC learnable hypothesis classes.

Proof of Theorem 7.2.

- Assume that $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ is agnostic PAC learnable. Using the fundamental theorem of statistical learning, it follows that each $\mathcal{H}_n$ has the uniform convergence property. By Theorem 7.3, $\mathcal{H}$ is nonuniformly learnable.
- For the other direction, assume that $\mathcal{H}$ is nonuniformly learnable using some algorithm A. For every $n \in \mathbb{N}$, let

$$\mathcal{H}_n = \{h \in \mathcal{H} : m_{\mathcal{H}}^{NUL}(1/8, 1/7, h) \leq n\}.$$

Clearly, $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$.

- (Cont.) In addition, for any distribution $\mathcal{D}$ that satisfies the realizability assumption with respect to $\mathcal{H}_n$, with probability of at least $1 - 1/7$ over $S \sim \mathcal{D}^n$ we have that

$$L_D(A(S)) \leq 1/8.$$

Using the fundamental theorem of statistical learning, this implies that the VC-dimension of $\mathcal{H}_n$ must be finite, and therefore $\mathcal{H}_n$ is agnostic PAC learnable.

- There are hypothesis classses that are nonuniform learnble but are not agnostic PAC learnable.
  - Consider $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$:
    - for every $n \in \mathbb{N}$, $\mathcal{H}_n$ is the class of polynomial classifiers of degree n.
    - $\mathrm{VCdim}(\mathcal{H}_n) = n + 1$.
    - $\mathcal{H}_n$ is agnostic PAC learnable.
    - $\mathrm{VCdim}(\mathcal{H}) = \infty$, where $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$.
    - $\mathcal{H}$ is nonuniform learnable.
    - $\mathcal{H}$ IS NOT agnostic PAC learnable.
- This implies that nonuniform learnablity is a strict relaxation of agnostic PAC learnability.

Outline

Theorem (7.4)

Let $w : \mathbb{N} \to [0,1]$ *be a function such that* $\sum_{n=1}^{\infty} w(n) \leq 1$. *Let* $\mathcal{H}$ *be a hypothesis class that can be written as* $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$, *where for each n,* $\mathcal{H}_n$ *satisfies the uniform convergence property with a sample complexity function* $m^{UC}_{\mathcal{H}_n}$. *Let* $\epsilon_n : \mathbb{N} \times (0,1) \to (0,1)$ *defined as*

$$\epsilon_n(m, \delta) = \min\{\epsilon \in (0,1) : m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) \leq m\}.$$

Then, for every $\delta \in (0,1)$ *and distribution* $\mathcal{D}$, *with probability of at least* $1 - \delta$ *over choice of* $S \sim \mathcal{D}^m$, *the following bound holds (simultaneously) for every* $n \in \mathbb{N}$ *and* $h \in \mathcal{H}_n$,

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta).$$

Therefore, for every $\delta \in (0,1)$ *and distribution* $\mathcal{D}$, *with probability of at least* $1 - \delta$ *it holds that*

$$\forall h \in \mathcal{H},\ L_{\mathcal{D}}(h) \leq L_S(h) + \min_{n : h \in \mathcal{H}_n} \epsilon_n(m, w(n) \cdot \delta).$$

Proof of Theorem 7.4.

- For each n define $\delta_n = w(n)\delta$. Applying the assumption of uniform convergence, we obtain that if we fix n in advance, then with probability of at least $1 - \delta_n$ over choice of $S \sim \mathcal{D}^m$,

$$\forall h \in \mathcal{H}_n, \ |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, \delta_n).$$

- Applying the union bound over $n = 1, 2, \ldots$, we obtain that with probability of at least

$$1 - \sum_n \delta_n = 1 - \delta(\sum_n w(n)) \geq 1 - \delta,$$

the preceding holds for all n, which concludes our proof.

Denote

$$n(h) = \min\{n : h \in \mathcal{H}_n\},$$

and then

$$\forall h \in \mathcal{H},\ L_{\mathcal{D}}(h) \leq L_S(h) + \min_{n:h\in\mathcal{H}_n} \epsilon_n(m, w(n) \cdot \delta).$$

implies that

$$L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta).$$

The Structural Risk Minimization paradigm searches for h that minimizes this bound.

Structural Risk Minimization (SRM)

prior knowledge:
$\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$, where for each n, $\mathcal{H}_n$ has the uniform convergence property with $m^{UC}_{\mathcal{H}_n}$;
$w : \mathbb{N} \to [0, 1]$ where $\sum_n w(n) \leq 1$.
define:

$$\epsilon_n(m, \delta) = \min\{\epsilon \in (0, 1) : m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) \leq m\};$$

$$n(h) = \min\{n : h \in \mathcal{H}_n\}.$$

input: training set $S \sim \mathcal{D}^m$, confidence δ
output: $h \in \operatorname{argmin}_{h \in \mathcal{H}}[L_S(h) + \epsilon_{n(h)}(m, w(n(h))\delta)]$

Theorem (7.5)

Let $\mathcal{H}$ be a hypothesis class such that $\mathcal{H} = \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ has the uniform convergence property with sample complexity $m^{UC}_{\mathcal{H}_n}$. Let $w : \mathbb{N} \to [0, 1]$ be a weighting function such that $w(n) = \frac{6}{n^2\pi^2}$. Then, $\mathcal{H}$ is nonuniformly learnable using the SRM rule with rate

$$m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h) \leq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, \frac{6\delta}{(\pi n(h))^2}).$$

Proof of Theorem 7.5.

- Let A be the SRM algorithm with respect to the weighting function w. For every $h \in H$, ϵ, and δ, let

$$m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon, w(n(h))\delta).$$

- Using the fact that $\sum_n w(n) = 1$,we can apply Theorem 7.4 to get that, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, we have that for every $h' \in \mathcal{H}$,

$$L_{\mathcal{D}}(h') \leq L_S(h') + \epsilon_{n(h')}(m, w(n(h'))\delta).$$

  The preceding holds in particular for the hypothesis $A(S)$.
- By SRM, we obtain that

$$\begin{aligned} L_{\mathcal{D}}(A(S)) &\leq \min_{h'}[L_S(h') + \epsilon_{n(h')}(m, w(n(h'))\delta)] \\ &\leq L_S(h) + \epsilon_{n(h)}(m, w(n(h))\delta). \end{aligned}$$

- Finally, if $m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, w(n(h))\delta)$ then clearly

$$\epsilon_{n(h)}(m, w(n(h))\delta) \leq \epsilon/2.$$

- In addition, from the uniform convergence property of each $\mathcal{H}_n$ we have that with probability of more than $1 - \delta$,

$$L_S(h) \leq L_{\mathcal{D}}(h) + \epsilon/2.$$

- Combining all the preceding we obtain that

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon,$$

which concludes the proof.

NOTE THAT the previous theorem also proves Theorem 7.3.