Fundamental Theorem of PAC Learning

PAC Learnability

Definition 1 *(PAC learnability) Let $\mathcal{C}$ be a class of boolean functions $f : \{0,1\}^n \rightarrow 0,1$. We say that $\mathcal{C}$ is PAC-learnable if there exists an algorithm $\mathcal{L}$ such that*
for every $f \in \mathcal{C}$
for any probability distribution $\mathcal{D}$
for any ϵ $(where 0 \leq \epsilon < \frac{1}{2})$
for any δ $(where 0 \leq \delta < 1$
algorithm $\mathcal{L}$ on input ϵ and δ and a set of random examples picked from any probability distribution $\mathcal{D}$ outputs at least with a probability $1 - \delta$, concept h such that $error(h, f) \leq \epsilon$.

Results on PAC Learnability

- Finite hypothesis classes are PAC learnable under realizability assumption
- Finite hypothesis classes are agnostic PAC learnable
- Hypothesis classes with finite VC dimension are PAC learnable

Sample Complexity of ERM Learner

- Finite hypothesis classes (realizable)

$$m(\varepsilon, \delta) = 1/\ \varepsilon\ log(|H|/\ \delta)$$

- Finite hypothesis classes (agnostic)

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

- Infinite hypothesis classes with finite VC dimension *d*

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

A Corollary on Shattering

COROLLARY 6.4 *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 $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Corollary

Corollary 6.4 tells us that if $\mathcal{H}$ shatters some set C of size $2m$ then we cannot learn $\mathcal{H}$ using m examples. Intuitively, if a set C is shattered by $\mathcal{H}$, and we receive a sample containing half the instances of C, the labels of these instances give us no information about the labels of the rest of the instances in C – every possible labeling of the rest of the instances can be explained by some hypothesis in $\mathcal{H}$. Philosophically,

If someone can explain every phenomenon, his explanations are worthless.

THEOREM 6.6 *Let $\mathcal{H}$ be a class of infinite VC-dimension. Then, $\mathcal{H}$ is not PAC learnable.*

No Free Lunch Theorem

For every learner, there is a task (target function) on which it fails, even though the task can be successfully learned by another learner.

No earner is universally superior to others.

THEOREM 5.1 (No-Free-Lunch) *Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
2. *With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

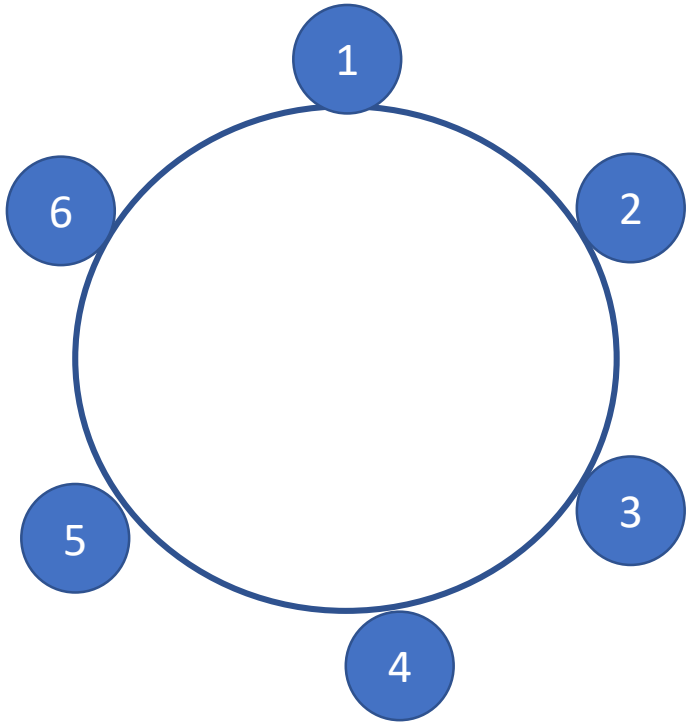
Fundamental Theorem of PAC Learning

THEOREM 6.7 (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 learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

Fundamental Theorem Justification

- 1 → 2 (uniform convergence)
- 2 → 3 (trivial)
- 3 → 4 (trivial)
- 2 → 5 (trivial)
- 4 → 6 (no free lunch)
- 5 → 6 (no free lunch)
- 6 → 1 (Sauer lemma)



Quantitative Bounds

THEOREM 6.8 (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*

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