Theory behind Application: PAC, VC-Dimension and Related Topics

Richard Liu

June 5, 2020

Source

- Shai Shalev-Shwartz, Shai Ben-David, *Understanding Machine Learning: From Theory to Algorithms*

Section 1

Introduction

Introduction

- We have introduced ways to do machine learning in practice, however, why does it work? Why not work?
  - No versatile model: No-Free-Lunch Principle.
  - Why training error only is insufficient?
  - Bias-Variance Decomposition: How to understand?
- This chapter will focus mainly on some interesting theorems for explaining these interesting and confusing phenomenons.

Section 2

Overfitting

Empirical Risk Minimization (ERM): Terminologies

- Training set S
- Distribution $\mathcal{D}$ (Unknown)
- Target Function f
- Predictor h (Difference?)

Empirical Risk Minimization (ERM)

Definition 1: Training Error

Suppose that there exists a training set S sampled from an unknown distribution $\mathcal{D}$ and labeled by some target function f and the predictor based on the sample set S ($h_S : \mathcal{X} \to \mathcal{Y}$), then we define

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

be the training error.

Empirical Risk Minimization (ERM)

Definition 2: Generalization Error

Define the error of a prediction rule $h : \mathcal{X} \to \mathcal{Y}$ be

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

Proposition 1

Let $\mathcal{H}$ be a class of binary classifiers over a domain $\mathcal{X}$, $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let f be the target hypothesis in $\mathcal{H}$, Then fix some $h \in \mathcal{H}$, we have

$$\mathbb{E}_{S|_x\sim\mathcal{D}^m}[L_S(h)] = L_{(\mathcal{D},f)}(h)$$

- What does it mean?

Empirical Risk Minimization (ERM): Overfitting

Definition 3: ERM

$$\mathrm{ERM}_{\mathcal{H}}(S) \in \arg\min_{h \in \mathcal{H}} L_S(h)$$

- Problem: Overfitting, which means high $L_D(h_S)$ but $L_S(h_S) = 0$.
- An example?
- Why discuss training error is enough? Where is the test set?

Empirical Risk Minimization (ERM): Overfitting

- Goal: Restrict $\mathcal{H}$, this is the source of model complexity.

Subsection 1

Theorem 1: Finite Hypothesis Classes

Prevent Overfitting: Finite Case

Assumption 1

Assume there exists $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$.

- This means we can train a model h to achieve a generalization error 0 in the finite hypotheses class.

Theorem 1

In finite hypotheses classes case, if we have taken sufficient large number of i.i.d samples, then with a high probability we could bound the generalization error to a sufficient small number.

- Proof?

Preparation for Proof

- Suppose we have sampled the training data $S|_x = (x_1, \ldots, x_m)$.
- Suppose the confidence parameter is denoted by δ.
- Suppose we want to bound the generalization error to ϵ.
- Suppose $\mathcal{D}^m$ be the distribution of sampled m-tuples.
- Goal: Find the upper bound of

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\})$$

As a reminder, h_S is the predictor found by ERM principle.

Proof

- Let "bad" hypotheses be

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)(h)} > \epsilon\}$$

- If h_S is a bad hypothesis, then it must belong to $\mathcal{H}_B$, also, it must satisfy $L_S(h) = 0$ (Why?). So we have

$$\{S|_x : L_{\mathcal{D},f}(h_S) > \epsilon\} \subset M, M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$$

Proof

- Up to now, we have got

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)(h_S)} > \epsilon\}) \leq \mathcal{D}^m(\cup_{h\in\mathcal{H}_B}\{S|_x : L_S(h) = 0\})$$

$$\leq \sum_{h\in\mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\})$$

- Note that

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) = \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\})$$

$$\leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

So the upper bound is $|\mathcal{H}|e^{-\epsilon m}$.

Summary

- If we take $m = \frac{\ln(|\mathcal{H}/\delta|)}{\epsilon}$, then the probability of making mistakes (ERM principle chooses a bad hypothesis) is only δ. This is acceptable for solving overfitting problem.
- The assumption MUST hold in finite hypotheses classes.

Section 3

More Formal Definition: PAC Learnability

Simple Definition of PAC

Definition 4: Probably Approximately Correct (PAC) Learnability

If there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and for every $\epsilon, \delta \in (0,1)$, every distribution $\mathcal{D}$ and every $f : \mathcal{X} \to \{0,1\}$, assumption holds w.r.t. $\mathcal{H}, \mathcal{D}, f$, and if $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, the algorithm will return a predictor h such that with probability of at least $1 - \delta$, the generalization error is less than ϵ. Then the hypotheses class $\mathcal{H}$ is PAC learnable.

- $m_{\mathcal{H}}(\epsilon, \delta)$ is the minimal sample complexity . In the previous case, it is $\frac{\ln(|\mathcal{H}/\delta|)}{\epsilon}$.

Subsection 1

Theorem 2: Bayes Optimal Predictor

Bayes Optimal Predictor

- What if we remove assumption 1?
  - Bayes Optimal Predictor!

Definition 5: Bayes Optimal Predictor

BOP is defined as

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Bayes Optimal Predictor

Theorem 2

For every classifier g, we have $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Proof

$L_{\mathcal{D}}(f_{\mathcal{D}}) = \mathbb{P}(f_{\mathcal{D}}(x) \neq y) = \mathbb{E}(I\{f_{\mathcal{D}}(x) \neq y\})$
$= \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{P}(f_{\mathcal{D}}(x) = 0)\mathbb{P}(y = 1|x) + \mathbb{P}(f_{\mathcal{D}}(x) = 1)\mathbb{P}(y = 0|x)]$
For each specific x_0, if $P(y = 1|x_0) \geq \frac{1}{2}$, then we should let $f_{\mathcal{D}}(x_0) = 1$, vise versa. $\square$

An Extension of PAC Learnability

Definition 6: Agnostic PAC Learnability

$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$ and hold other assumptions unchanged compared with definition 4.

Section 4

More about B-V Decomposition

Subsection 1

Theorem 3: No-Free-Lunch

No-Free-Lunch

- Question: Whether there exists a learning algorithm A and a training set size m, such that for every distribution $\mathcal{D}$, if A receives m i.i.d. examples from $\mathcal{D}$, there is a high chance that it outputs a predictor h that has a low risk?
  - Unfortunately, no.
  - No-Free-Lunch theorem.

No-Free-Lunch

Theorem 3

Let A be any learning algorithm for the task of binary classification w.r.t. the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that
1. There exists a function f such that $L_{\mathcal{D}}(f) = 0$.
2. With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$

- Proof?

Proof: Part 1

- Let C be a subset of $\mathcal{X}$ of size $2m$, then there are $T = 2^{2m}$ possible functions from C to $\{0, 1\}$, let $\mathcal{D}_i$ be the distribution defined by

$$\mathcal{D}_i(\{(x, y)\}) = \begin{cases} 1/|C| & y = f_i(x) \\ 0 & \text{otherwise} \end{cases}$$

  Then obviously $L_{\mathcal{D}_i}(f_i) = 0$.

- If we could prove that under such distribution, the second conclusion holds, then theorem 3 is shown to be true.

Proof: Part 2

- We will show that

$$\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m} [L_{\mathcal{D}_i}(A(S))] \geq \frac{1}{4}$$

  for every algorithm A. This means that there exists one learning task f such that $\mathbb{E}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(A'(S))] \geq \frac{1}{4}$. The remaining part could be solved by Markov Inequality (shown later).

- Note that there are $k = (2m)^m$ possible permutations of m examples from C, denoted by $S_1, \ldots, S_k$ (k different datasets). Denote $S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m)))$ be the j-th dataset labeled by f_i.

Proof: Part 3

- For uniform sampling we have

$$\mathbb{E}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] = \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i))$$

- Taking maximum yields

$$LHS \geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i))$$

(Why?) We only need to consider the behavior of i.

Proof: Part 4

- Let $v_1, \ldots, v_p$ be the samples in C that do not appear in S_j, then we have $p \geq m$ and we have

$$L_{\mathcal{D}_i}(h) = \frac{1}{2m} \sum_{x \in C} I\{h(x) \neq f_i(x)\} \geq \frac{1}{2p} \sum_{r=1}^{p} I\{h(v_r) \neq f_i(v_r)\}$$

So for the same reasons, we have

$$\frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{2} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} I\{A(S_j^i)(v_r) \neq f_i(v_r)\}$$

$$= \frac{1}{4} \quad (\textit{Why?})$$

Markov Inequality

Lemma 1: Markov Inequality

Let X be a non-negative random variable, then for any $\alpha > 0$, we have

$$\mathbb{P}[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}$$

Proof

Define $f(X) = \begin{cases} 1 & X \geq \alpha \\ 0 & \text{otherwise} \end{cases}$, then we have $f(X) \leq X/\alpha$, so $\mathbb{E}[f(X)] \leq \frac{\mathbb{E}[X]}{\alpha}$, $\mathbb{E}[f(X)] = \mathbb{P}[X \geq \alpha]$ yields the result. $\square$

Proof: Part 5

- By Markov Inequality and $\mathbb{E}(\theta) \geq \frac{1}{4}$, we could prove the result by showing that $\mathbb{P}(\theta \geq \frac{1}{8}) \geq \frac{1}{7}$ (How?)
- No-Free-Lunch Theorem guarantees that no information will lead to no positive result.

Section 5

VC-Dimension

Introduction

- Question: Infinite hypotheses space = Not PAC learnable?
  - Fortunately, no.
  - An example?

Background

Let $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a(x) = I\{x < a\}$

Example

Proposition 2

Let $\mathcal{H}$ be the space defined before, then it is PAC learnable using the ERM rule with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \ln(\frac{\frac{2}{\delta}}{\epsilon}) \rceil$

Proof (Part 1)

Let a^* be a threshold such that $L_{\mathcal{D}}(h^*) = 0$. Also, let $\mathbb{P}_{x \sim \mathcal{D}_x}[x \in (a_0, a^*)] = \mathbb{P}_{x \sim \mathcal{D}_x}[x \in (a^*, a_1)] = \epsilon$. Let $b_0 = \max\{x : (x, 1) \in S\}$ and $b_1 = \min\{x : (x, 0) \in S\}$, then $b_S \in (b_0, b_1)$, where b_S corresponds to the ERM hypothesis.

Example

Proof (Part 2)

Note that if we want $L_{\mathcal{D}}(h_S) \leq \epsilon$, we must let $b_0 \geq a_0, b_1 \leq a_1$, so we have

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0, b_1 > a_1]$$

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

For we have $\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = (1-\epsilon)^m \leq e^{-\epsilon m}$ and similar to the dual probability, we could conclude the proof. $\square$

VC Dimension

- Idea: Observe what $\mathcal{H}$ behaves like on a subset C.
  - Recall: we have proved the No-Free-Lunch theorem by such method, we want to prevent such things happening again.

Definition 7: Restriction of $\mathcal{H}$ on 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 $\mathcal{H}$, which is

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

VC Dimension

Definition 8: 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, $|\mathcal{H}_C| = 2^{|C|}$.

- What does it mean?

VC Dimension

Example 1

Consider the hypotheses discussed before, then if we take $C = \{c_1\} \subset \mathbb{R}$, then C is shattered by $\mathcal{H}$. However, if we take $C = \{c_1, c_2\} \subset \mathbb{R}$, this is not the case.

Proposition 3

Assume that there exists a set $C \subset \mathcal{X}$ of size $2m$ that is shattered by $\mathcal{H}$, then the No-Free-Lunch Theorem holds.

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

VC Dimension

Definition 9: VC-Dimension

The VC-Dimension is defined as the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$, denoted by $\mathrm{VCdim}(\mathcal{H})$

VC Dimension: Examples

Example 1

See page 41, we have $\text{VCdim}(\mathcal{H}) = 1$.

Example 2

Consider the class of intervals, which means $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ and $h_{a,b}(x) = I\{x \in (a, b)\}$, then we have $\text{VCdim}(\mathcal{H}) = 2$.

VC Dimension: Examples

Example 3

For a finite class, we have $\mathrm{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.

Theorem 4

If $\mathcal{H}$ has a finite VC-Dimension, then it is PAC-learnable.

- Any other real examples?

Thank you!