PAC Learnability and Complexity

20170745 Jaehui Hwang

Table of Contents

Table of Contents

Quick Review : ϵ-δ Argument (Calculus I / Analysis I)

Definition 1

A function f defined on a set X of real numbers has the **limit** L at x_0, i.e.,

$$\lim_{x \to x_0} f(x) = L,$$

if, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\left| f(x) - L \right| < \epsilon, \text{ whenever } x \in X \text{ and } 0 < |x - x_0| < \delta.$$

Quick Review : ϵ-δ Argument (Calculus I / Analysis I)

Why ϵ-δ argument is important?

Quick Review : ϵ-δ Argument (Calculus I / Analysis I)

Why ϵ-δ argument is important?

Because it provides δ for every $\epsilon > 0$, regardless of the magnitude of ϵ.

Why ϵ-δ argument is important?

Because it provides δ for every $\epsilon > 0$, regardless of the magnitude of ϵ.

In other words, we can think that there exists a function $f : \epsilon \to \delta$ such that it satisfies the condition.

Definition 2

- The set S of all possible outcomes of an experiment a way that in each trial of the experiment one and only one of the outcomes (events) in the set will occur, we call the set S a **sample space** for the experiment. Each element S is called a **simple outcome**, or **simple event**.
- An **event E** is defined to be any subset of S (including the empty set and the sample space S). Event E is a **simple event** if it contains only one element and a **compound event** if it contains more than one element.
- We say that **an event E occurs** if any of the simple events in E occurs.

Definition 3

Given a probability assignment for the simple events in a sample space S, we define the **probability of an arbitrary event** E, denoted by $\mathbb{P}(E)$, as follows:

- If E is the empty set, then $\mathbb{P}(E) = 0$.
- If E is a simple event, i.e. $E = \{e_i\}$, then $\mathbb{P}(E) = \mathbb{P}(e_i)$ as defined previously.
- If E is a compound event, then $\mathbb{P}(E)$ is the sum of the probabilities of all the simple events in E.
- If E is the sample space S, then $\mathbb{P}(E) = \mathbb{P}(S) = 1$.

Example : In a family with 3 children, excluding multiple births, what is the probability of having exactly 2 girls? Assume that a boy is as likely as a girl at each birth.

Example : In a family with 3 children, excluding multiple births, what is the probability of having exactly 2 girls? Assume that a boy is as likely as a girl at each birth.

- First we determine the sample space S:

$$S = \{GGG, GGB, GBG, BGG, GBB, BGB, BBG, BBB\}$$

Example : In a family with 3 children, excluding multiple births, what is the probability of having exactly 2 girls? Assume that a boy is as likely as a girl at each birth.

- First we determine the sample space S:

$$S = \{GGG, GGB, GBG, BGG, GBB, BGB, BBG, BBB\}$$

- Since a boy is as likely as a girl at each birth, each of the 8 outcomes in S is equally likely; so each outcome has probability $\frac{1}{8}$.

Example : In a family with 3 children, excluding multiple births, what is the probability of having exactly 2 girls? Assume that a boy is as likely as a girl at each birth.

- First we determine the sample space S:

$$S = \{GGG, GGB, GBG, BGG, GBB, BGB, BBG, BBB\}$$

- Since a boy is as likely as a girl at each birth, each of the 8 outcomes in S is equally likely; so each outcome has probability $\frac{1}{8}$.
- There exists only 3 cases, GGB, GBG, BGG. Thus the probability of having exactly 2 girls are $\frac{1}{8} \times 3 = \frac{3}{8}$.

Definition 4

The expected value, also called the **expectation** or **mean**, of a random variable is its average value weighted by its probability distribution.

The expected value or mean of a random variable X is written as $\mathbb{E}(X)$.

Example : What is the expectation value of rolling a 6-sided die?

Example : What is the expectation value of rolling a 6-sided die?

Answer : The mean of a discrete random variable is defined as

$$\mathbb{E}(X) = \sum_{x \in X} x p(x),$$

where $X = \{1, 2, 3, 4, 5, 6\}$. Therefore,

$$\mathbb{E}(X) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = 3.5.$$

Table of Contents

Basic Definitions

Definition 5

We denote by X the set of all possible **examples** or **instances**. X is also sometimes referred to as the **input space**.

Definition 6

The set of all possible **labels** or **target values** is denoted by Y.

Basic Definitions

To make the problems easier, we will limit ourselves to the case where Y is reduced to two labels,

$$Y = \{0, 1\}.$$

which corresponds to the so-called **binary classification**.

Basic Definitions

Definition 7

A **concept** $c : X \to Y$ is a mapping from X to Y.

Definition 8

A **concept class** is a set of concepts we may wish to learn and is denoted by C.

Basic Definitions

We assume that examples are independently and identically distributed (i.i.d.) according to some fixed but unknown distribution D.

Definition 9

We call a fixed set of possible concepts as a **hypothesis set**, H.

Question : What is the difference between hypothesis set and concept class?

Definition : Learning Problem

Definition 10(Learning Problem)

A learner considers a hypothesis set H, which might not necessarily coincide with C. It receives a sample $S = (x_1, ..., x_m)$ drawn i.i.d. according to D as well as the labels $(c(x_1), ..., c(x_m))$, which are based on a specific target concept $c \in C$ to learn. **Learning problem** is a task to use the labeled sample S to select a hypothesis $h_S \in H$ that has a small error with respect to the concept c.

Definition : Learning Problem

Intuitively, we can assume that for $c \in C$ is a goal (model) to learn, and $h \in H$ is a 'incomplete' model.

Definition : Learning Problem

Intuitively, we can assume that for $c \in C$ is a goal (model) to learn, and $h \in H$ is a 'incomplete' model.

Then how can we measure the error terms between h and c?

Definition : Error

Definition 11(Generalized Error)

Given a hypothesis $h \in H$, a target concept $c \in C$, and an underlying distribution D, the **generalization error** or **risk** of h is defined by

$$R(h) = \underset{x \sim D}{\mathbb{P}}[h(x) \neq c(x)] = \underset{x \sim D}{\mathbb{E}}[1_{h(x) \neq c(x)}],$$

where 1_ω is the indicator function of the event ω.

Definition 12(Empirical Error)

Given a hypothesis $h \in H$, a target concept $c \in C$, and a sample $S = (x_1, ..., x_m)$, the **empirical error** or **empirical risk** of h is defined by

$$\hat{R}_s(h) = \frac{1}{n} \sum_{i=1}^{m} 1_{h(x) \neq c(x)}.$$

Definition : PAC-Learning

The following introduces the Probably Approximately Correct (PAC) learning framework.

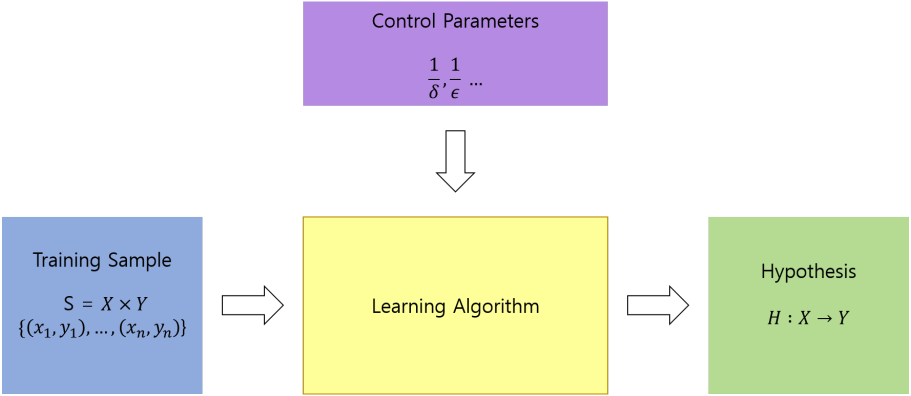
Definition 13(PAC-Learning)

A concept class C is said to be **PAC-learnable** if there exists an algorithm A and a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions D on X and for any target concept $c \in C$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$\underset{S \sim D^m}{\mathbb{P}}[R(h_S) \leq \epsilon] \geq 1 - \delta.$$

If A further runs in $poly(1/\epsilon, 1/\delta, n, size(c))$, then C is said to be **efficiently PAC-learnable**. When such an algorithm A exists, it is called a **PAC-learning algorithm** for C.

Definition : PAC-Learning



Example : Consider the case where the set of instances are points in the plane, $X = \mathbb{R}^2$, and the concept class C is the set of all axis-aligned rectangles lying in $\mathbb{R}^2$.

Example : Consider the case where the set of instances are points in the plane, $X = \mathbb{R}^2$, and the concept class C is the set of all axis-aligned rectangles lying in $\mathbb{R}^2$.

The learning problem consists of determining with small error a target axis-aligned rectangle using the labeled training sample. We will show that the concept class of axis-aligned rectangles is PAC-learnable.

Definition : PAC-Learning

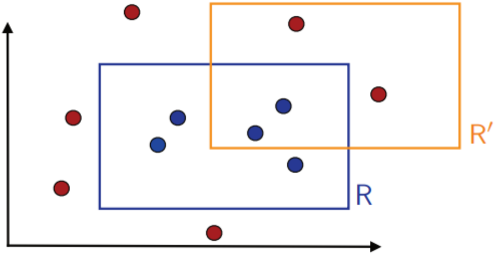


Figure: Target concept R and possible hypothesis R'. Circles represent training instances. A blue circle is a point labeled with 1, since it falls within the rectangle R. Others are red and labeled with 0.

Definition : PAC-Learning

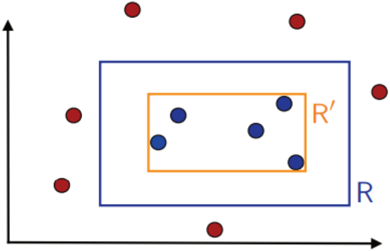
Theorem 14

The concept class of axis-aligned rectangles is PAC-learnable.

Theorem 14

The concept class of axis-aligned rectangles is PAC-learnable.

Proof : To show that the concept class is PAC-learnable, we describe a simple PAC-learning algorithm A. Given a labeled sample S, the algorithm consists of returning the tightest axis-aligned rectangle $R' = R_S$ containing the points labeled with 1.



Proof(continued) :

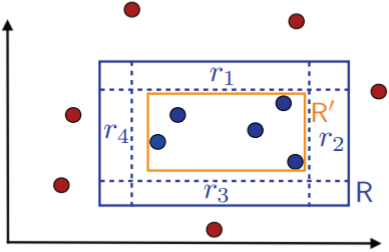
Let $R \in C$ be a target concept. Fix $\epsilon > 0$. Let $\mathbb{P}[R]$ denote the probability mass of the region defined by R, that is the probability that a point randomly drawn according to D falls within R.

Since errors made by our algorithm can be due only to points falling inside R, we can assume that $\mathbb{P}[R] > \epsilon$.

Definition : PAC-Learning

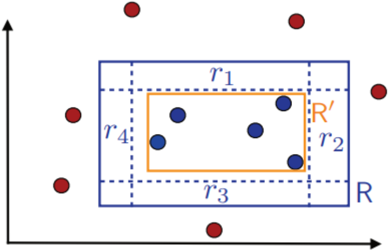
Proof(continued) :

Now we can define four rectangular regions r_1, r_2, r_3, and r_4 along the sides of R, each with probability at least $\epsilon/4$. These regions can be constructed by starting with the full rectangle R and then decreasing the size by moving one side as much as possible while keeping a distribution mass of at least $\epsilon/4$.



Proof(continued) :

Let l, r, b, and t be the four real values defining R : $R = [l, r] \times [b, t]$. Then, for example, the left rectangle r_4 is defined by $r_4 = [l, s_4] \times [b, t]$, with $s_4 = \inf\{s : P[[l, s] \times [b, t]] \geq \epsilon/4\}$.
The probability of the region $\overline{r_4} = [l, s_4] \times [b, t]$ obtained from r_4 by excluding the rightmost side is at most $\epsilon/4$. r_1, r_2, r_3 and $\overline{r_1}, \overline{r_2}, \overline{r_3}$ are defined in a similar way.



Definition : PAC-Learning

Proof(continued) :

As a result, we can write

$$\begin{aligned}
\mathop{\mathbb{P}}_{S\sim D^m}[R(h_S) > \epsilon] &\leq \mathop{\mathbb{P}}_{S\sim D^m}[\cup_{i=1}^{4}\{R_S \cap r_i = \emptyset\}] \\
&\leq \sum_{i=1}^{4} \mathop{\mathbb{P}}_{S\sim D^m}[\{R_S \cap r_i = \emptyset\}] \\
&\leq 4(1-\epsilon/4)^m \\
&\leq 4\exp(-m\epsilon/4),
\end{aligned}$$

from $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$.

Proof(continued) :

Now, For any $\delta > 0$, to ensure that $\underset{S\sim D^m}{\mathbb{P}}[R(h_S) > \epsilon] \leq \delta$, we can impose

$$4\exp(-m\epsilon/4) \leq \delta \Leftrightarrow m \geq \frac{4}{\epsilon}\log\frac{4}{\delta}.$$

Thus, for any $\epsilon > 0$ and $\delta > 0$, if the sample size m is greater than $\frac{4}{\epsilon}\log\frac{4}{\delta}$, then $\underset{S\sim D^m}{\mathbb{P}}[R(h_S) > \epsilon] \leq \delta$, which proves that the concept class of axis-aligned rectangles is PAC-learnable. $\square$

Table of Contents

Definition of Agnostic PAC-Learning

Now we generalize the definition of PAC-Learning.

Definition 15(Agnostic PAC-Learning)

Let H be a hypothesis set. A is an **agnostic PAC-learning algorithm** if there exists a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions D over $S = X \times Y$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$\mathbb{P}_{S\sim D^m}[R(h_S) - \min_{h\in H} R(h) \leq \epsilon] \geq 1 - \delta.$$

If A further runs in $poly(1/\epsilon, 1/\delta, n)$, then it is said to be an **efficient agnostic PAC-learning algorithm**.

Definition of Agnostic PAC-Learning

Now we generalize the definition of PAC-Learning.

Definition 15(Agnostic PAC-Learning)

Let H be a hypothesis set. A is an **agnostic PAC-learning algorithm** if there exists a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions D over $S = X \times Y$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$\underset{S \sim D^m}{\mathbb{P}}[R(h_S) - \min_{h \in H} R(h) \leq \epsilon] \geq 1 - \delta.$$

If A further runs in $poly(1/\epsilon, 1/\delta, n)$, then it is said to be an **efficient agnostic PAC-learning algorithm**.

Question : What is the difference between 'agnostic' PAC-learning and PAC-learning?

Quick Review : VC Dimension

Definition 16(Growth Function)

The **growth function** $\Pi_H : \mathbb{N} \to \mathbb{N}$ for a hypothesis set H is defined by:

$$\forall m \in \mathbb{N}, \Pi_H(m) = \max_{\{x_1,\ldots x_m\} \subseteq X} \left| \left\{ \left(h(x_1), \ldots, h(x_m) \right) : h \in H \right\} \right| .$$

Definition 17(VC Dimension)

The **VC-dimension** of a hypothesis set H is the size of the largest set that can be shattered by H:

$$\mathrm{VCdim}(H) = \max \left\{ m : \Pi_H(m) = 2^m \right\}.$$

Do you remember the definition of 'shattered'?

Quick Review : Sauer's Lemma

Theorem 18(Sauer's Lemma)

Let H be a hypothesis set with $\text{VCdim}(H) = d$. Then, for all $m \in \mathbb{N}$, the following inequality holds:

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

Question : Sauer's lemma suggests an 'upper bound' of the generalized error. But how?

The significance of Sauer's lemma can be seen by the following theorem, which remarkably shows that growth function only exhibits two types of behavior: either $\mathrm{VCdim}(H) = d < +\infty$, in which case $\Pi_H(m) = O(m^d)$, or $\mathrm{VCdim}(H) = +\infty$, in which case $\Pi_H(m) = 2^m$.

Theorem 19

Let H be a hypothesis set with $\mathrm{VCdim}(H) = d$. Then for all $m \geq d$,

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d = O(m^d).$$

Proof : The proof begins by using Sauer's lemma.

$$
\begin{aligned}
\Pi_H(m) &\leq \sum_{i=0}^{d} \binom{m}{i} \\
&\leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \\
&\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \\
&= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^{i} \\
&= \left(\frac{m}{d}\right)^{d} \left(1 + \frac{d}{m}\right)^{m} \leq \left(\frac{m}{d}\right)^{d} e^{d}.
\end{aligned}
$$

Upper Bound for the Generalization Error

Theorem 20

Let H be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H$:

$$R(h) \leq \hat{R}_s(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

In other words, the form of this generalization bound is

$$R(h) \leq \hat{R}_s(h) + O\left(\sqrt{\frac{\log (m/d)}{(m/d)}}\right)$$

Lower Bound for the Generalization Error

Until now, I presented an upper bound on the generalization error.

Lower Bound for the Generalization Error

Until now, I presented an upper bound on the generalization error.

Then how about the lower bound? What if there does not exists 'enough lower bound' for any learning algorithms?

Lower Bound for the Generalization Error

Until now, I presented an upper bound on the generalization error.

Then how about the lower bound? What if there does not exists 'enough lower bound' for any learning algorithms?

Big Picture : In some situation this really happens. In other words, I will introduce the condition which is not agnostic PAC-learnable.

Lower Bound for the Generalization Error

Theorem 21

Let H be a hypothesis set with VC-dimension $d > 1$. Then, for any $m \geq 1$ and any learning algorithm A, there exists a distribution D over $X \times \{0, 1\}$ such that:

$$\mathbb{P}_{S \sim D^m}\left[R(h_S) - \inf_{h \in H} R(h) > \sqrt{\frac{d}{320m}}\right] \geq 1/64.$$

Equivalently, for any learning algorithm, the sample complexity verifies

$$m \geq \frac{d}{320\epsilon^2}.$$

Lower Bound for the Generalization Error

Corollary 22

With an infinite(unlimited) VC-dimension, agnostic PAC-learning is not possible.

Corollary 22

With an infinite(unlimited) VC-dimension, agnostic PAC-learning is not possible.

Proof : The previous theorem shows that for any algorithm A(in the non-realizable case), there exists a 'bad' distribution over $S = X \times \{0, 1\}$ such that the error of the hypothesis returned by A is a constant times $\sqrt{\frac{d}{m}}$ with some constant probability. The VC-dimension appears as a critical quantity in learning in this general setting as well. In particular, with an infinite VC-dimension, agnostic PAC-learning is not possible. $\square$

Table of Contents

Definition of Learning Problem

Let X be the **input space**. Let Y be the set of **target values**.

The learning problem is to find a hypothesis $h \in H$ with small generalization error

$$R(h) = \mathop{\mathbb{P}}_{(x,y)\sim D}[h(x) \neq y] = \mathop{\mathbb{E}}_{(x,y)\sim D}[1_{h(x)\neq y}].$$

Definition of PAC Learning

Definition(PAC-Learning)

A concept class C is said to be **PAC-learnable** if there exists an algorithm A and a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions D on X and for any target concept $c \in C$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$\mathbb{P}_{S \sim D^m}[R(h_S) \leq \epsilon] \geq 1 - \delta.$$

If A further runs in $poly(1/\epsilon, 1/\delta, n, size(c))$, then C is said to be **efficiently PAC-learnable**. When such an algorithm A exists, it is called a **PAC-learning algorithm** for C.

Definition of PAC Learning

Theorem

The concept class of axis-aligned rectangles is PAC-learnable.

Main idea of proof : We can construct a function which grows slower than polynomial such that it satisfies the below condition :

$$\underset{S\sim D^m}{\mathbb{P}}[R(h_S) > \epsilon] \leq \delta \Leftrightarrow \underset{S\sim D^m}{\mathbb{P}}[R(h_S) \leq \epsilon] \geq 1 - \delta.$$

Definition of Agnostic PAC-Learning

Definition(Agnostic PAC-Learning)

Let H be a hypothesis set. A is an **agnostic PAC-learning algorithm** if there exists a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions D over $S = X \times Y$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$\mathop{\mathbb{P}}_{S \sim D^m}[R(h_S) - \min_{h \in H} R(h) \leq \epsilon] \geq 1 - \delta.$$

If A further runs in $poly(1/\epsilon, 1/\delta, n)$, then it is said to be an efficient **agnostic PAC-learning algorithm**.

Condition for Agnostic PAC-Learning is Not Possible

Definition(VC dimension)

The **VC-dimension** of a hypothesis set H is the size of the largest set that can be shattered by H:

$$\mathrm{VCdim}(H) = \max\{m : \Pi_H(m) = 2^m\}.$$

Theorem

With an infinite(unlimited) VC-dimension, agnostic PAC-learning is not possible.

References

- Foundations of Machine Learning, 2nd edition. (Chap 2 $\sim$ 3)
  https://cs.nyu.edu/~mohri/mlbook/
- Wikipedia, PAC Learning
  https://en.wikipedia.org/wiki/Probably_approximately_correct_learning
- CS492(F) Computational Learning Theory(2021F) in KAIST, by professor Hongseok Yang
  https://github.com/hongseok-yang/CLT21

Thank you.