PAC Learning

Yinan Huang

May 30, 2021

Contents

1 PAC Learning

1.1 Basic Concepts of Machine Learning

The basic components of a learning machine are:

- *Input of learning machine*:
  - Domain set $\mathcal{X}$: we want to make prediction based on samples $x \in \mathcal{X}$. $x \in \mathcal{X}$ is also called covariate in stats.
  - Label set $\mathcal{Y}$: where the prediction outcome lies in. It is also called response in stats. For example, $\mathcal{Y} = \{0, 1\}$ for binary classification and $\mathcal{Y} = \mathbb{R}$ for regression.
  - Training set $S = \{(x_i, y_i)|i = 1, 2, ..., m\}$: samples from $\mathcal{X} \times \mathcal{Y}$. Though it is called 'set', but in fact S can contains identical samples so it is better to think S as sequence.
- *Output of learning machine*: a prediction rule $f_S : \mathcal{X} \to \mathcal{Y}$ relying on training set S. Basically learning machine will search through a hypothesis space H consisting of all kinds of f and returns $f_S \in \mathcal{H}$ based on training set S with certain algorithm.
- *Data generative model D*: samples (x_i, y_i) i.i.d. drawn from joint distribution D. Usually we write $(X, Y) \sim D$ and $S \sim D^m$. The generative model is unknown to the learning machine. Sometimes we consider the deterministic label, that is $X \sim D_X$ and $Y = f(X)$.
- *Metric of evaluation*: A loss function $L : \mathcal{H} \times \mathcal{Y} \to \mathbb{R}_+$ should be specified. The evaluation of a prediction rule f can thereby be evaluated by risk (expected loss):

$$R_D(f) = \mathbb{E}_{(X,Y)\sim D} L(f, Y). \tag{1.1}$$

- *Learning algorithm*: receives training samples S and return a predictor f_S. Since the goal of learning machine in general is to find (learn) the risk minimizer: $f^* \triangleq \arg\min_f R_D(f)$ (though is unknown since we do not know D), usually the learning algorithm is **empirical risk minimizer**:

$$f_S \triangleq \arg\min_{f\in\mathcal{H}} R_S(f) = \arg\min_{f\in\mathcal{H}} \frac{1}{m} \sum_i^m L(f(x), y_i), \tag{1.2}$$

Statistical learning theory is to study the relation between f_S (what our machine learned) and f^* (what we expect to learn). For simplicity, in the following discussion of PAC learning, we focus on binary classification problem with 0-1 loss:

$$L(f(x), y) \triangleq \mathbb{1}_{f(x)\neq y} \tag{1.3}$$

whose expectation is exactly accuracy.

1.2 PAC Learnablity

Under what conditions can f_S be close to f^*? And how to evaluate the error? The PAC learnablity formally defines the measure of the closeness.

Definition 1.1 (Realizability of $\mathcal{H}$)**.** *Suppose label is deterministic. Hypothesis space $\mathcal{H}$ is said to have **realizability**, if there exists $f^* \in \mathcal{H}$ such that*

$$R_D(f^*) = 0. \tag{1.4}$$

Definition 1.2 (PAC Learnablity)**.** *A hypothesis space $\mathcal{H}$, given loss function L, is said to be **Probably Approximately Correct (PAC) learnable**, if there exists a sample complexity function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm, for all $\delta, \epsilon \in (0,1)$ and any generative model $X \sim D$, $Y = f(X)$, assuming realizability holds, if we feed $m > m_{\mathcal{H}}(\epsilon, \delta)$ training samples i.i.d. generated from (D, f), the learning algorithm returns a prediction rule f_S that satisfies the following:*

$$P_{S \sim D^m}(R_D(f_S) > \epsilon) < \delta. \tag{1.5}$$

So PAC learnablity says if $\mathcal{H}$ is PAC learnable, then we can always find a learning algorithm (whatever it is) that returns a suboptimal prediction rule (in the sense that it is slightly more risky than optimal one) with bounded probability given number of samples.

Claim 1.3. *It turns out if we restrict to a binary classfication with deterministic label (that is, $Y = f^*(X) \in \{0,1\}$) and use misclassification rate as risk function, then any finite hypothesis with realizability ($f^* \in \mathcal{H}$) is PAC learnable.*

Proof 1.4. *To prove this, first notice that $\min_f R_D(f) = R_D(f^*) = 0$, and thus $R_S(f^*) = 0$ (because it is a deterministic label). Since $0 \leq R_S(f_S) = \min_{f \in \mathcal{H}} R_S(f) \leq R_S(f^*) = 0$, so $R_S(f_S) = 0$. To bound the probability of $R_D(f_S) > \epsilon$, notice that*

$$R_S(f_S) = 0, R_D(f_S) > \epsilon \quad \Rightarrow \quad \exists f \in \mathcal{H}, R_S(f) = 0, R_D(f) > \epsilon. \tag{1.6}$$

So

$$\begin{aligned} P_{S \sim D^m}(R_D(f_S) > \epsilon) &\leq P(\exists f \in \mathcal{H}, R_S(f) = 0, R_D(f) > \epsilon) = P(\cup_{f \in \mathcal{H}_{\mathcal{B}}} R_S(f) = 0) \\ &\leq \sum_{f \in \mathcal{H}_{\mathcal{B}}} P(R_S(f) = 0) = |\mathcal{H}_{\mathcal{B}}| \, (1 - P(f(X) \neq Y))^m \\ &= |\mathcal{H}_{\mathcal{B}}| \, (1 - R_D(f))^m \leq |\mathcal{H}_{\mathcal{B}}| \, (1 - \epsilon)^m \leq |\mathcal{H}| \, e^{-m\epsilon}, \end{aligned} \tag{1.7}$$

where $\mathcal{H}_{\mathcal{B}} \triangleq \{f \in \mathcal{H} | R_D(f) > \epsilon\}$. Therefore if we let

$$m_{\mathcal{H}}(\epsilon, \delta) = \frac{1}{\epsilon} \log \frac{|\mathcal{H}|}{\delta}, \tag{1.8}$$

then it satisfies the PAC conditions. So any finite $\mathcal{H}$ is PAC learnable in this deterministic binary classification case.

In the next section we will show that finite $\mathcal{H}$ has uniform convergence property and thus is PAC learnable.

PAC learning can be generalized to stochastic label $X \times Y \sim D$ and we have the so-called agnostic PAC learning.

Definition 1.5 (Agnostic PAC Learnability)**.** *to be **Agnostic Probably Approximately Correct (PAC) learnable**, if there exists a sample complexity function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm, for all $\delta, \epsilon \in (0,1)$ and any generative model $(X, Y) \sim D$, if we feed $m > m_{\mathcal{H}}(\epsilon, \delta)$ training samples i.i.d. generated from D, the learning algorithm returns a prediction rule f_S that satisfies the following:*

$$P_{S \sim D^m}(R_D(f_S) > \min_{f \in \mathcal{H}} R_D(f) + \epsilon) < \delta. \tag{1.9}$$

Later in the fundamental theorems of statistical learning we will show that PAC learnability is actually equivalent to agnostic PAC Learnability for binary classification with 0-1 loss (and also true for regression with MSE, etc.).

1.3 Summary

We introduce the formal definition of a learnable hypothesis space by PAC learnability. Basically PAC learnability says we can control the accuracy and the confidence of our leaning result by increasing sample size.

2 Uniform Convergence Learning

PAC learnablity says $R_D(f_S)$ should be close to $\min_{f\in\mathcal{H}} R_D(f)$. Notice that $f_S = \arg\min_{f\in\mathcal{H}} R_S(f)$ and so we can imagine that this is true if R_S is close to R_D. Under the protection of law of large numbers, R_S converges to R_D almost surely as sample size $m \to \infty$. We carefully study the relation between R_S and R_D in this section.

2.1 Uniform Convergence

Notice that R_S depends on training samples S, so the closeness of R_S to R_D really relies on S. The goodness of S is defined by the following.

Definition 2.1 (ϵ-representative)**.** *Given data generative model $(X, Y) \sim D$, hypothesis space $\mathcal{H}$ and loss function L, traning samples S is said to be ϵ-representative, if for all $f \in \mathcal{H}$,*

$$|R_S(f) - R_D(f)| \leq \epsilon. \tag{2.1}$$

Lemma 2.2. *If S is $\frac{\epsilon}{2}$-representative (w.r.t. $(\mathcal{H}, D, L)$), then we the risk of ERM bounded by*

$$R_D(f_S) \leq \min_{f\in\mathcal{H}} R_D(f) + \epsilon. \tag{2.2}$$

Proof 2.3. *The proof is direct:*

$$R_D(f_S) \leq R_S(f_S) + \frac{\epsilon}{2} \leq R_S(f^*) + \frac{\epsilon}{2} \leq R_D(f^*) + \epsilon \tag{2.3}$$

where $f^* = \arg\min_{f\in\mathcal{H}} R_D(f)$ and thus finishes the proof.

The lemma above is exactly says that if R_S is close to R_D, then $R_D(f_S)$ is close to $\min_{f\in\mathcal{H}} R_D(f)$, which is we want in PAC learning. Concretely, we define the uniform convergence property of $\mathcal{H}$ by the following:

Definition 2.4. *We say $\mathcal{H}$ has uniform convergence property, if there exists a sample complexity $m_{\mathcal{H}}^{UC} : (0, 1)^2 \to \mathbb{N}$, for any $\epsilon, \delta \in (0, 1)$ and any distribution D over (X, Y), if we have sample size $m > m_{\mathcal{H}}^{UC}(\epsilon, \delta)$, then*

$$P(S \text{ is } \epsilon\text{-representative}) > 1 - \delta. \tag{2.4}$$

Intuitively, we can control the goodness of R_S (in terms of how close it is to R_D) by increasing sample size. Since if R_S is close to R_D, then $R_D(f_S)$ is close to $\min_{f\in\mathcal{H}} R_D(f)$, so it is not surprising that uniform convergence property leads to PAC learnability, which is stated by the following theorem.

Theorem 2.5 (Uniform convergence implies PAC learnability)**.** *If $\mathcal{H}$ has uniform convergence property, then $\mathcal{H}$ is agnostic PAC learnable with ERM.*

Proof 2.6. *Recall that uniform convergence means for any distribution $(X, Y) \sim D$ and any $\epsilon, \delta \in (0, 1)$, when sample size $m > m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$, we have*

$$P(S \text{ is } \frac{\epsilon}{2}\text{-representative}) > 1 - \delta. \tag{2.5}$$

By the definition of $\frac{\epsilon}{2}$-representative, we then have

$$P(\forall f \in \mathcal{H}, |R_S(f) - R_D(f)| \le \frac{\epsilon}{2}) > 1 - \delta. \tag{2.6}$$

We have shown in lemma 2.2 that this implies

$$P(\forall f \in \mathcal{H}, R_D(f_S \le \min_{f \in \mathcal{H}} R_D(f) + \epsilon)) > 1 - \delta, \tag{2.7}$$

which is exactly agnostic PAC learnability. So there exists a sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \le m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$ so that we can achive PAC learnability with ERM.

As an example, we are going to show that any finite $\mathcal{H}$ has uniform convergence property.

Claim 2.7. *Finite $\mathcal{H}$ ($|\mathcal{H}| < \infty$) has uniform convergence property and thus is PAC learnable.*

Proof 2.8. *To show uniform convergence, we need to prove the following:*

$$P(\forall f \in \mathcal{H}, |R_S(f) - R_D(f)| < \epsilon) > 1 - \delta, \tag{2.8}$$

which is equivalent to (by taking the contropositive)

$$P(\exists f \in \mathcal{H}, |R_S(f) - R_D(f) > \epsilon|) < \delta. \tag{2.9}$$

Notice that $\{S | \exists f \in \mathcal{H}, |R_S(f) - R_D(f)| > \epsilon\} = \cup_{f \in \mathcal{H}} \{S || R_S(f) - R_D(f)| > \epsilon\}$, so we obtain a upper bound

$$P(\exists f \in \mathcal{H}, |R_S(f) - R_D(f) > \epsilon|) \le \sum_{f \in \mathcal{H}} P(|R_S(f) - R_D(f)| > \epsilon). \tag{2.10}$$

Suppose risk function is 0-1 loss, then by Hoeffding's inequality, we have

$$P(|R_S(f) - R_D(f)| > \epsilon) \le 2e^{-2m\epsilon^2}. \tag{2.11}$$

Therefore

$$P(\exists f \in \mathcal{H}, |R_S(f) - R_D(f) > \epsilon|) \le 2|\mathcal{H}|e^{-2m\epsilon^2}. \tag{2.12}$$

So if we let $m > \frac{\log(2|\mathcal{H}/\delta)|}{2\epsilon^2}$, then we achive uniform convergence. The simple complexity of uniform convergence is $m_{\mathcal{H}}^{UC}(\epsilon, \delta) \le \frac{\log(2|\mathcal{H}/\delta)|}{2\epsilon^2}$. Since finite $\mathcal{H}$ has uniform convergence property, hence it is agnostic PAC learnable.

2.2 Summary

We introduce the uniform convergence of $\mathcal{H}$. Intuitively, uniform convergence means R_S is close to R_D and thus it implies PAC learnability. Actually the fundamental theorems of statistical learning show that uniform convergence is equivalent to PAC learnability, not just a sufficient condition.

3 VC-dimension

In the last section we introduct uniform convergence and show that uniform convergence leads to PAC learnability. But the uniform convergence property of a infinite $\mathcal{H}$ is difficult to compute, and thus we need a new characteristic to explain why a $\mathcal{H}$ is PAC learnable.

3.1 No-Free-Lunch theorem

In order to understand why a $\mathcal{H}$ is PAC learnable, let us first focus on when it fails to be PAC learnable. Consider a binary classification problem, the largest hypothesis space is the one contains all possible functions $\mathcal{H}_{all} = \{f|f : \mathcal{X} \to \{0, 1\}\}$. $\mathcal{X}$ is a infinite set (if not so, then $\mathcal{H}_{all}$ is finite and therefore is PAC learnable due to uniform convergence) and intuitively $\mathcal{H}_{all}$ should not be PAC learnable due to its great complexity. The No-Free-Lunch theorem formally explain the reason behind of why $\mathcal{H}_{all}$ is not PAC learnable.

Theorem 3.1 (No-Free-Lunch theorem)**.** *Let $\mathcal{X}$ be a infinite set and $\mathcal{H}_{all} = \{f|f : \mathcal{X} \to \{0, 1\}\}$. Then $\mathcal{H}_{all}$ is not PAC learnable with resepect to 0-1 loss.*

Proof 3.2. *The details are omitted. The key idea to find an adversary distribution D such that the error $R_D(f_S)$ cannot be controlled. Intuitively, for any sample size m, we can always find a subset $C \subset \mathcal{X}$ with $|C| = 2m$ and create an adversary distribution D on C. Notice that $\mathcal{H}_{all}$* ***contains all functions from** C **to*** $\{0, 1\}$ *and learning algorithm only receives at most half of the samples from C, therefore it cannot have a good generalization over distribution D on C.*

As we can see in the proof, the main idea of why $\mathcal{H}_{all}$ fails to be PAC learnable is because $\mathcal{H}_{all}$ contains all possible functions from a subset $C \subset \mathcal{X}$ to $\{0, 1\}$, and thereby an adversary distribution can be generated to fail $\mathcal{H}_{all}$. This analysis is also true for any $\mathcal{H}$. To further play with the idea, we define shattering and VC-dimension.

3.2 Shattering and VC-dimension

Definition 3.3. *A finite subset $C = \{c_1, c_2, ..., c_{|C|}\} \in \mathcal{X}$ is said to be* ***shattered*** *by $\mathcal{H}$, if $|\mathcal{H}_C| = 2^{|C|}$ where $\mathcal{H}_C = \{(f(c_1), f(c_2), ..., f(c_{|C|}))|f \in \mathcal{H}\}$. That is, the restriction of $\mathcal{H}$ on C contains all functions from C to $\{0, 1\}$.*

Definition 3.4 (VC-dimension)**.** *The* ***VC-dimension*** *of $\mathcal{H}$ is defined by the cardinality of the largest subset $C \subset \mathcal{X}$ that is shattered by $\mathcal{H}$. That is,*

$$\textit{VC-dim}(\mathcal{H}) = \max\{|C||C \subset \mathcal{X} \textit{ and } C \textit{ is shattered by } \mathcal{H}\}. \tag{3.1}$$

VC-dimension characterizes the capacity of $\mathcal{H}$. From the similar idea of No-Free-Lunch theorem we can show that if VC-dim$(\mathcal{H}) = \infty$, then $\mathcal{H}$ is not PAC learnable.

Theorem 3.5. *If VC-dim$(\mathcal{H}) = \infty$, then $\mathcal{H}$ is not PAC learnable.*

Proof 3.6. *The proof is similar to No-Free-Lunch theorem. Assume $\mathcal{H}$ is PAC learnable, then for any given $\epsilon, \delta \in (0, 1)$, we can find a sample size m to let $R_D(f_S)$ be controlled under ϵ, δ. However, since the VC-dimension of $\mathcal{H}$ is infinity, we can always find a subset $C \subset \mathcal{X}$ with $|C| = 2m$ that is shattered by $\mathcal{H}$. Since C is shattered by $\mathcal{H}$, an adversary distribution can be constructed to fail f_S. Then we obtain a contradiction and therefore $\mathcal{H}$ should not be PAC learnable.*

Here are some examples of VC-dimension.

Example 3.7. *Consider $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{f_a(x) = \mathbb{1}_{x<a} | a \in \mathbb{R}\}$ is all threshold function. The VC-dimension of $\mathcal{H}$ is 1 since for any $C = c_1$ we can always let $a = c_1 + 1$ and $a = c_1 - 1$ to shatter C while $C = c_1, c_2$ cannot be shattered by $\mathcal{H}$.*

Example 3.8. *Consider $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{f_{a,b}(x) = \mathbb{1}_{x<b}\mathbb{1}_{x>a} | a, b \in \mathbb{R}\}$. The VC-dimension of $\mathcal{H}$ is 2, since any $C = c_1, c_2$ is shattered by $\mathcal{H}$ and any $C = c_1, c_2, c_3$ is not shattered by $\mathcal{H}$.*

3.3 Summary

VC-dimension is defined to portray the capacity of a hypothesis space. We also show that if VC-dimension is infinity, then the corresponding $\mathcal{H}$ is not PAC learnable. The idea of uniform convergence, VC-dimension and PAC learnablity will be perfectly connected together by the fundamental theorems of statistical learning in the next section.

4 Fundamental Theorems of Statistical Learning

The fundamental theorems of statistical learning build a exact relation between PAC learnability, uniform convergence and VC-dimension.

Theorem 4.1 (Fundamental theorems of statistical learning)**.** *Let H be hypothesis space consisting of functions from $\mathcal{X}$ to $\{0, 1\}$ (binary classification) and the loss function be 0-1 loss. The following statements are equivalent:*

(1) $\mathcal{H}$ has uniform convergence property.

(2) $\mathcal{H}$ is agnostic PAC learnable with ERM.

(3) $\mathcal{H}$ is agnostic PAC learnable.

(4) $\mathcal{H}$ is PAC learnable.

(5) $\mathcal{H}$ is PAC learnable with ERM.

(6) VC-dim$(\mathcal{H}) < \infty$.

Proof 4.2. *Most the proof can be easily done by using what we learned in the previous sections. (1) imples (2) by our previous proof. (2) implies (3) by definition. (3) implies (4), (2) implies (5) since agnostic PAC learning is a generalized version of PAC learning. (4),(5) implies (6) by taking the contropositive of statement "infinite VC-dimension implies $\mathcal{H}$ is not PAC learnable". The left one is to prove (6) implies (1), whose proof is convoluted and is omitted here. By all of this, we show the equivalent of these statements.*

The theorem says uniform convergence and finite VC-dimension are actually equivalent condition of (agnostic) PAC learnability. Note that it is the theorem for binary classification of 0-1 loss, and there may exist task where ERM fails or uniform convergence cannot characterize PAC learnability.

The power of VC-dimension is more than this. We can even quantify the sample complexity from VC-dimension.

Theorem 4.3. *Let H be hypothesis space consisting of functions from $\mathcal{X}$ to $\{0, 1\}$ (binary classification) and the loss function be 0-1 loss. Then the sample complexity for agnostic PAC learnability is bounded by*

$$C_1 \frac{\textit{VC-dim}(\mathcal{H}) + \log 1/\delta}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{\textit{VC-dim}(\mathcal{H}) + \log 1/\delta}{\epsilon^2} \tag{4.1}$$

where C_1, C_2 are some constants.