STATISTICAL LEARNING THEORY

Notes for Review

Zimeng Yu
219 Prospect St
Department of Statistics and Data Science
Sep, 2025

Contents

1 PAC Learning and VC Theory

Theorem 1.1. This is a theorem.

Proposition 1.2. This is a proposition.

Principle 1.3. This is a principle.

1.1 Lecture 1 PAC Learning and VC Theory

- Empirical error: training error, the error in training set. Measure the performance of a model on a specific, finite dataset, typically the training data
- Expected error: the true, theoretical performance of a model on all possible data points drawn from the underlying data distribution.
- Mapping from X to Y: describes a rule or function that connects each element in a set X (the domain) to one and only one element in a set Y (the codomain or range).

Statistical Learning Framework

- Domain X.
  - Each $x \in X$ is called an "instance". For example: $X = \mathbb{R}^d$.
- Label Space Y.
  - Example: $Y = \{\pm 1\}$, $Y = \mathbb{R}$.
- Unknown source distribution D over $X \times Y$.
  - Our assumption on the data generating process ("nature" or "reality").
- Goal: find a predictor $h : X \to Y$ achieving small *expected error* $L_D(h) = \mathbb{P}_{(x,y)\sim D}\{h(x) \neq y\}$.
- Based on an i.i.d. training sample $S = \{(x_i, y_i)\}_{i=1}^m$ drawn from D (can also write $S \sim D^m$).

Figure 1: Definition

1.2 Lecture 2 PAC Learning and VC Theory

Power Rule: $\frac{d}{dx}x^n = n \cdot x^{n-1}$
Exponent Rule: $\frac{d}{dx}a^u = a^u \cdot \ln a \cdot \frac{du}{dx}$

2 Non-Uniform Learning

3 Non-Uniform Learning

4 Non-Uniform Learning

5 Non-Uniform Learning

Last course: efficient property learning can be available or not? three-term DNFs and CNFs.

RP: pol

NP: efficient

$RP \neq NP$

We assume that RP ≠ NP. Then, for any family $\{H_n\}_{n\in\mathbb{N}}$:
- If $\text{vc}(H_n) \leq \text{poly}(n)$ and there is a poly-time algorithm implementing Consistent Learning, then family H_n is efficiently-properly-PAC learnable.
- If solving $\text{CONS}_{H_n}(S)$ is NP-hard, then family H_n is *not* efficiently-properly-PAC learnable.

Figure 2: Proper Learning

Improper learning? Not efficient, canot rely on $RP \neq NP$. "If cryptography(Cryptography, or cryptology is the practice and study of techniques for secure communication in the presence of adversarial behavior.) is possible, then efficient learning is impossible. This is based on cryptographic assumptions."

$F : \{0,1\}^n \to \{0,1\}$

Benefit: add more assumptions

Relizable vs. Agnostic Learning

Definition. A family of classes $\{H_n\}_{n\in\mathbb{N}}$ is efficiently-properly-PAC-learnable in the realizable setting if there exists a learning algorithm A such that
$\forall n \forall (\epsilon,\delta) \in (0,1)^2, \exists m(n,\epsilon,\delta) \in \mathbb{N}, \forall D$ s.t. $\inf_{h\in H} L_D(h) = 0$,

$$\mathbb{P}_{S\sim D^{m(\epsilon,\delta)}}\left\{L_D(A(S)) \leq \epsilon\right\} \geq 1-\delta,$$

and A runs in time polynomial in $n, 1/\epsilon, \log(1/\delta)$, and A always outputs a predictor in H_n.

Definition. A family of classes $\{H_n\}_{n\in\mathbb{N}}$ is efficiently-properly-PAC-learnable in the agnostic setting if there exists a learning algorithm A such that
$\forall n \forall (\epsilon,\delta) \in (0,1)^2, \exists m(n,\epsilon,\delta) \in \mathbb{N}, \forall D$

$$\mathbb{P}_{S\sim D^{m(\epsilon,\delta)}}\left\{L_D(A(S)) \leq \inf_{h\in H_n} L_D(h) + \epsilon\right\} \geq 1-\delta,$$

and A runs in time polynomial in $n, 1/\epsilon, \log(1/\delta)$, and A always outputs a predictor in H_n.

Figure 3: Relizable vs. Agnostic Learning

$$\mathrm{ERM}_{H_n}(S) = \arg\min_{h \in H_n} \frac{1}{|S|} \sum_{(x,y)\in S} \mathbf{1}\{h(x_i) \neq y_i\}$$

Claim.
- If $\mathrm{vc}(H_n) \leq \mathrm{poly}(n)$, and
- there is a poly-time algorithm implementing ERM for $\{H_n\}_{n\in\mathbb{N}}$,

then $\{H_n\}_{n\in\mathbb{N}}$ is efficiently-agnostically-properly-PAC-Learnable.

For a family $\{H_n\}_{n\in\mathbb{N}}$ consider the decision problem:

$$\mathrm{AGREEMENT}_{H_n}(S,k) = 1 \text{ iff } \exists h \in H_n, L_S(h) \leq 1 - \frac{k}{|S|}.$$

Claim. If H_n is efficiently-agnostically-properly-PAC-learnable then $\mathrm{AGREEMENT}_{H_n} \in \mathrm{RP}$.

Corollary. If $\mathrm{RP} \neq \mathrm{NP}$ and $\mathrm{AGREEMENT}_{H_n}$ is NP-hard, then H_n is not efficiently-agnostically-properly-PAC-learnable.

Figure 4: Conditions for Efficient Agnostic Learning

- Poly-time functions? No! (Not even in the realizable case)
- Poly-size depth-2 neural networks? No! (Not even in the realizable case)
- Halfspaces (linear predictors)?
  - $X_n = \{0,1\}^n, H_n = \{x \mapsto \mathbf{1}[\langle w, x\rangle > 0] : w \in \mathbb{R}^n\}$.
  - Claim: $\mathrm{AGREEMENT}_{H_n}$ is NP-hard.
  - No!
- Conjunctions? No!
- Unions of segments on the line?
  - $X_n = [0,1], H_n = \left\{x \mapsto \vee_{i=1}^n \mathbf{1}[a_i \leq x \leq b_i] \mid a_i, b_i \in [0,1]\right\}$.
  - Yes! Efficiently Properly Agnostically PAC Learnable.

Figure 5: What is Efficiently Properly Agnostically PAC Learnable?

- Agnostic Learning for Halfspaces: Not. Implication: Not efficiently backward learnable because it is NP-Hard.
- $AGREEMENT_{H_n}(S, K) = 1$ the agreement can be 0 or 1; k is the param for sample. Given the dataset with size k, if there exist any $L_s(h)$ could be over 1- $\frac{k}{|S|}$
- Unions of segments on the line? Yes!

Surrogate Loss: some loss can be better and more idea, like be converge. e.g.: Hing Loss in margin; cross-entrepy loss.

One major challenge in machine learning theory is to reconcile the success of deep learning methods (based on local search procedures, e.g. Stochastic Gradient Descent) with worst-case hardness results. (e.g.: sometimes in worst case it is not backward learnable but can still be minimized for its loss, in that case, it will be unexplainable.) "It must be that learning problems where deep learning succeeds are not worst-case in nature, but understanding what makes these problems efficiently learnable is a major open research direction."

5.1 Non-Uniform Bias

- So far: a uniform prior over H, each $h \in H$ is equally likely. Sometimes there exist other assumed prior.
- Instead: prior $p(h)$ that encodes "preference" or "bias".

5.1.1 Shorter Description

$$\mathrm{MDL}_p(S) = \arg\max_{L_S(h)=0} p(h) = \arg\min_{L_S(h)=0} |d(h)|.$$

Theorem. For prior p over a countable H s.t. $\sum_h p(h) \le 1$ (e.g., $p(h) = 2^{-|d(h)|}$ for a prefix-free d), any $\delta \in (0,1), m \in \mathbb{N}$, and any distribution D s.t. $\exists h^\star \in H, L_D(h^\star) = 0$, with probability $\ge 1-\delta$ over $S \sim D^m$,

$$L_D(\mathrm{MDL}_p(S)) \le \frac{\ln\left(\frac{1}{p(h^\star)}\right) + \ln(1/\delta)}{m} = \frac{|d(h^\star)|\ln(2) + \ln(1/\delta)}{m}.$$

Proof. For any $h \in H$ such that $L_D(h) > \epsilon_h := \frac{\ln(1/p(h)) + \ln(1/\delta)}{m}$. Then,
$\mathbb{P}_{S\sim D^m}\{L_S(h) = 0\} \le (1-\epsilon_h)^m \le e^{-\epsilon_h m} = p(h)\cdot\delta$. By a union bound,
$\mathbb{P}_{S\sim D^m}\{\exists h, L_D(h) > \epsilon_h : L_S(h) = 0\} \le \sum_{h\in H} \mathbb{P}_{S\sim D^m}\{L_S(h) = 0\} \le \sum_{h\in H} p(h)\delta \le \delta$. Finally,
since $\exists h^\star \in H, L_D(h^\star) = 0$, it follows that $p\left(\mathrm{MDL}_p(S)\right) \ge p(h^\star)$.

Figure 6: Bias to Shorter Description

- MDL: Minimum Description Length.
- For two models $H_1 and H_2$, Shorter description will prefer a higher weight.
- Shorter, the learning loss will decrease faster, and vice versa. And the upper bound for MDL_p wants to make sure.

Why no contradiction to Fundamental Theorem? According to the Fundamental Theorem of PAC Learning, a class is only learnable if its VC dimension is finite. So, how can MDL learn a class with an infinite VC dimension? VC dimension → not back-learnerable.

If the true hypothesis $h*$ is simple (has a short description), MDL can learn it efficiently with a relatively small sample size.

If the true hypothesis $h*$ is complex (has a long description), MDL will require a much larger sample size to learn it effectively.

5.1.2 Structural Risk Minimization

Given a prior p over H, with probability $\geq 1-\delta$ over $S \sim D^m$:

$$L_D(h) \leq L_S(h) + \sqrt{\frac{\ln(1/p(h)) + \ln(2/\delta)}{2m}}.$$

Minimized by ERM — Minimized by MDL

$$\mathrm{SRM}_p(S) = \arg\min_h L_S(h) + \sqrt{\frac{\ln(1/p(h))}{2m}}$$

Fit data — Match the prior / simple / short description

Theorem. For prior p over a countable H s.t. $\sum_h p(h) \leq 1$, any distribution D, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1-\delta$ over $S \sim D^m$:

$$L_D(\mathrm{SRM}_p(S)) \leq \inf_h \left(L_D(h) + 2\sqrt{\frac{\ln(1/p(h)) + \ln(2/\delta)}{2m}} \right).$$

Figure 7: Structural Risk Minimization

- uncountable classes? $vc(h) = \infty$, Previously, non-uniform learning was applied to countable classes, where you could assign a prior probability p(h) to each individual hypothesis h. But with an uncountable class, this approach fails
- Answer 1: use a prior over hypothesis classes. The solution presented is to shift the level of abstraction. Instead of defining a prior over individual hypotheses (h), you define a prior over countable subclasses.

Theorem. For $H = \cup_{r \in \mathbb{N}} H_r$ and prior $p(H_r)$ s.t. $\sum_r p(H_r) \leq 1$, any distribution D, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1 - \delta$ over $S \sim D^m$:

$$L_D(\text{SRM}_p(S)) \leq \inf_{r, h \in H_r} \left(L_D(h) + c \sqrt{\frac{\text{vc}(H_r) + \ln(1/p(H_r)) + \ln(1/\delta)}{m}} \right).$$

Figure 8: Prior Over Hypothesis Classes

5.1.3 Prior Over Hypothesis Classes

- The main theorem provides a bound on the generalization error for a learning algorithm that uses Structural Risk Minimization (SRM) with a prior.
- **Reduces to Standard SRM**: When the hypothesis classes H_r are countable, their VC dimension is 0. The formula simplifies to the standard SRM guarantee for countable classes, which is the basis for MDL.
- **Reduces to Empirical Risk Minimization (ERM)**: If all the prior probability is concentrated on a single class H_r, the prior term in the formula becomes zero. The theorem then reduces to the standard ERM principle, where the learner simply minimizes the training error within that one chosen class.
- The complexity of the subclass it belongs to, measured by its VC dimension $vc(H_r)$.
- $U_{r \in N}$: countable union of simpler subclasses.
- H_r: degree-r polynomials.
- The final part of the slide applies the theorem to the example of learning sign-of-polynomial functions. By defining the H_r classes as polynomials of degree r and using a prior like $p(H_r) = 2^{-r}$, the theorem provides a concrete sample complexity bound. This bound is non-uniform; it shows that the amount of data needed to learn a hypothesis h depends on its degree deg(h), confirming that simpler hypotheses require less data to be learned effectively. This elegantly resolves the paradox of "unlearnable" classes with infinite VC dimension.

5.1.4 PAC-Bayes

- w.p.: will probably
- K-L divergence: $KL(Q\|P) = \sum Q(i) log \frac{Q(i)}{P(i)}$

- So far we have used a discrete prior / distribution over hypotheses, or discrete prior over hypothesis classes (in MDL and SRM).
- What about arbitrary distributions / priors over uncountable H?
- Consider randomized (average) predictor h_Q defined as:
  - $h_Q(x) = y$ w.p. $\mathbb{P}_{h\sim Q}(h(x) = y)$.
  - $L_D(h_Q) = \mathbb{E}_{(x,y)\sim D}\mathbb{E}_{h\sim Q}\mathbf{1}[h(x) \neq y] = \mathbb{E}_{h\sim Q}L_D(h)$.

Theorem. For any class H and any prior distribution P over H, any distribution D over $X \times Y$, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1-\delta$ over $S \sim D^m$:

$$\forall \text{ posterior dist'ns } Q \text{ over } H : \mid L_D(h_Q) - L_S(h_Q) \mid \leq \sqrt{\frac{\mathrm{KL}(Q \mid\mid P) + \log(2m/\delta)}{2(m-1)}}.$$

Figure 9: PAC-Bayes

References