Statistical Learning Theory

Omar Montasser

Lecture 3

PAC Learning and VC Theory

Probably Approximately Correct (PAC)

Definition. A hypothesis class H is realizably-PAC-learnable if there exists a learning rule A such that $\forall(\epsilon,\delta)\in(0,1)^2, \exists m(\epsilon,\delta)\in\mathbb{N}, \forall D$ s.t. $\inf_{h\in H} L_D(h)=0$,

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

Definition. A hypothesis class H is agnostically-PAC-learnable if there exists a learning rule A such that $\forall(\epsilon,\delta)\in(0,1)^2, \exists m(\epsilon,\delta)\in\mathbb{N}, \forall D$,

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

RESEARCH CONTRIBUTIONS

Artificial Intelligence and Language Processing

David Waltz Editor

A Theory of the Learnable

L. G. VALIANT

Leslie Valiant

The Growth Function

- For $C = (x_1, x_2, \ldots, x_m) \in X^m$, define the restriction (or projection) of H onto C:
  - $H|_C = \left\{ \left(h(x_1), h(x_2), \ldots, h(x_m) \right) \mid h \in H \right\}$.
- $\Gamma_H(m) = \max_{C \in X^m} |H|_C|$.
- Examples:
  - $X = \{1,\ldots,100\}, H = \{\pm 1\}^X$: $\Gamma_H(m) = \min(2^m, 2^{100})$.
  - $X = \{1,\ldots,2^{100}\}, H = \{\mathbf{1}[x \leq \theta] \mid \theta \in \{1,\ldots,2^{100}\}\}$: $\Gamma_H(m) = \min(m+1, 2^{100})$.

Vapnik-Chervonenkis Dimension

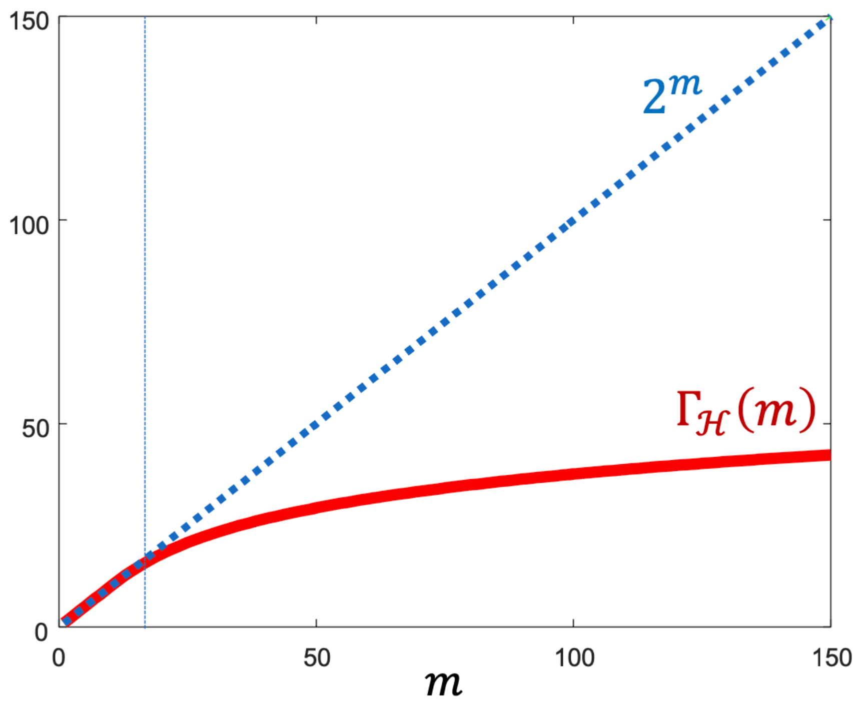
- $C = \{x_1, \ldots, x_m\}$ is **shattered** by H if $|H|_C| = 2^m$, i.e., the projection contains all 2^m labelings:
  - $\forall y_1, \ldots, y_m \in \{\pm 1\}, \exists h \in H$ s.t. $\forall_{1 \le i \le m} h(x_i) = y_i$.
- The VC-dimension of H, denoted $\mathrm{vc}(H)$, is the largest number of points that can be shattered by H:
  - $\mathrm{vc}(H) = \max\{m \in \mathbb{N} : \Gamma_H(m) = 2^m\}$.
- If H is *infinite* and $\forall m, \Gamma_H(m) = 2^m$ then we say $\mathrm{vc}(H)$ is infinite.

Sauer-Shelah-Perles Lemma

Lemma. If $\mathrm{vc}(H) = d$, then for all m:

$$\Gamma_H(m) \le \sum_{i=0}^{d} \binom{m}{i}.$$

In particular, when $m > d$, $\Gamma_H(m) \le (em/d)^d = O(m^d)$.



Sauer-Shelah-Perles Lemma

Lemma. If $\mathrm{vc}(H) = d$, then for all m:

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

In particular, when $m > d$, $\Gamma_H(m) \leq (em/d)^d = O(m^d)$.

Refinement (Pajor 1985). If $\mathrm{vc}(H) = d$, then for all $C \in X^m$:

$$|H|_C| \leq |\{B \subseteq C : H \text{ shatters } B\}|.$$

Sauer-Shelah-Perles Lemma

Refinement (Pajor 1985). If $\mathrm{vc}(H) = d$, then for all $C \in X^m$:

$$|H|_C| \leq |\{B \subseteq C : H \text{ shatters } B\}|.$$

Proof Sketch.

- Base case when $m = 1$ holds.
- Induction: suppose statement holds for $k < m$.
- Let $C = \{x_1, \ldots, x_m\} \in X^m$ and $C' = \{x_2, \ldots, x_m\}$. Consider:

$$A = \{(y_2, \ldots, y_m) : (+1, y_2, \ldots, y_m) \in H|_C \vee (-1, y_2, \ldots, y_m) \in H|_C\}$$

and

$$B = \{(y_2, \ldots, y_m) : (+1, y_2, \ldots, y_m) \in H|_C \wedge (-1, y_2, \ldots, y_m) \in H|_C\}.$$

- Verify that $|H|_C| = |A| + |B|$.
- By induction, it holds that

$$|A| = |H|_{C'}| \leq |\{B \subseteq C' : H \text{ shatters } B\}| = |\{B \subseteq C : x_1 \notin B \wedge H \text{ shatters } B\}|.$$

- Let $H' \subseteq H$ be defined as

$$H' = \{h \in H : \exists h' \in H \text{ s.t. } (1 - h'(x_1), h'(x_2), \ldots, h'(x_m)) = (h(x_1), h(x_2), \ldots, h(x_m))\}.$$

- Observe that $B = H'|_{C'}$. Thus, by induction,

$$\begin{aligned} |B| &= |H'_{C'}| \leq |\{B \subseteq C' : H' \text{ shatters } B\}| = |\{B \subseteq C' : H' \text{ shatters } B \cup \{x_1\}\}| \\ &= |\{B \subseteq C : x_1 \in B \wedge H' \text{ shatters } B\}| \leq |\{B \subseteq C : x_1 \in B \wedge H \text{ shatters } B\}|. \end{aligned}$$

- Combining the above, we have

$$\begin{aligned} |H|_C| &= |A| + |B| \\ &\leq |\{B \subseteq C : x_1 \notin B \wedge H \text{ shatters } B\}| + |\{B \subseteq C : x_1 \in B \wedge H \text{ shatters } B\}| \\ &= |\{B \subseteq C : H \text{ shatters } B\}|. \end{aligned}$$

Summary so far …

Recall

Theorem. For any hypothesis class H, any (realizable) distribution D, any $(\epsilon, \delta) \in (0,1)^2$, with sample complexity

$$m(\epsilon,\delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon}\right) \text{ or } m(\epsilon,\delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon^2}\right)$$

with probability $\geq 1-\delta$ over $S \sim D^{m(\epsilon,\delta)}$:

$$\forall h \in H : L_S(h) = 0 \Longrightarrow L_D(h) \leq \epsilon \text{ or } \forall h \in H : |L_D(h) - L_S(h)| \leq \epsilon\,.$$

Sauer-Shelah-Perles Lemma. If $\mathrm{vc}(H) = d$, then for all m:

$$\Gamma_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq (em/d)^d = O(m^d).$$

When $m > d$

Corollaries. Any hypothesis class H with finite VC dimension is

- realizably-PAC-learnable using ERM with sample complexity
$$m(\epsilon,\delta) = O\left(\frac{\mathrm{vc}(H)\ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}\right).$$
- agnostically-PAC-learnable using ERM with sample complexity
$$m(\epsilon,\delta) = O\left(\frac{\mathrm{vc}(H) + \ln(1/\delta)}{\epsilon^2}\right).$$

Questions …

- Are there classes H with *infinite* VC dimension that are PAC-learnable?
- Can we learn with *fewer* samples than VC dimension?
- Are there any hypothesis classes that are *not* PAC-learnable?

Plan

Theorem. For any hypothesis class H, any (realizable) distribution D, any $(\epsilon, \delta) \in (0,1)^2$, with sample complexity

$$m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon}\right) \text{ or } m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon^2}\right)$$

with probability $\geq 1 - \delta$ over $S \sim D^{m(\epsilon,\delta)}$:

$$\forall h \in H : L_S(h) = 0 \implies L_D(h) \leq \epsilon \text{ or } \forall h \in H : |L_D(h) - L_S(h)| \leq \epsilon\,.$$

- Questions:
  - Are there classes H with *infinite* VC dimension that are PAC-learnable?
  - Can we learn with *fewer* samples than VC dimension?
  - Are there any hypothesis classes that are *not* PAC-learnable?
  - What about computation?

Theorem. For any hypothesis class H, any (realizable) distribution D, any $(\epsilon, \delta) \in (0,1)^2$, with sample complexity

$$m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon}\right) \text{or } m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon^2}\right)$$

with probability $\geq 1 - \delta$ over $S \sim D^{m(\epsilon,\delta)}$:

$$\forall h \in H : L_S(h) = 0 \Longrightarrow L_D(h) \leq \epsilon \text{ or } \forall h \in H : |L_D(h) - L_S(h)| \leq \epsilon\,.$$

Proof.

- Given a set $S = \{(x_i, y_i)\}_{i=1}^m$ of m examples, define the event $A_S = \{\, \exists h \in H : L_D(h) > \epsilon \ \wedge \ L_S(h) = 0\}$.
    - Our goal is to show that $\mathbb{P}_{S\sim D^m}[A_S] \leq \delta$.
- Now, consider drawing two sets S, S' of m examples each. Define the event $B_{S,S'} = \{\, \exists h \in H : L_{S'}(h) > \epsilon/2 \ \wedge \ L_S(h) = 0\}$.
- Claim: $\mathbb{P}_{S,S'\sim D^m}[B_{S,S'}] \geq \frac{1}{2}\mathbb{P}_{S\sim D^m}[A_S]$. Why?
    - $\mathbb{P}_{S,S'\sim D^m}[B_{S,S'}] = \mathbb{P}_{S\sim D^m}[A_S]\mathbb{P}_{S,S'\sim D^m}[B_{S,S'} \mid A_S]$, and $\mathbb{P}_{S,S'\sim D^m}[B_{S,S'} \mid A_S] \geq \frac{1}{2}$ by a Chernoff bound as long as $m > 8/\epsilon$.
- Thus, it suffices to show that $\mathbb{P}_{S,S'\sim D^m}[B_{S,S'}] \leq \delta/2$.

Theorem. For any hypothesis class H, any (realizable) distribution D, any $(\epsilon, \delta) \in (0,1)^2$, with sample complexity

$$m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon}\right) \text{or } m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon^2}\right)$$

with probability $\geq 1 - \delta$ over $S \sim D^{m(\epsilon,\delta)}$:

$$\forall h \in H : L_S(h) = 0 \implies L_D(h) \leq \epsilon \text{ or } \forall h \in H : |L_D(h) - L_S(h)| \leq \epsilon\,.$$

Proof.

- Now consider a 3rd experiment. Draw a set S'' of $2m$ examples, then randomly partition S'' into two sets S, S' each of size m.
- Define event $C_{S'',S,S'} = \{\, \exists h \in H : L_{S'}(h) > \epsilon/2 \ \wedge \ L_S(h) = 0\}$.
- Claim: $\mathbb{P}_{S''\sim D^m,S,S'}[C_{S'',S,S'}] = \mathbb{P}_{S,S'\sim D^m}[B_{S,S'}]$.
- Thus, it suffices to show that $\mathbb{P}_{S''\sim D^m,S,S'}[C_{S'',S,S'}] \leq \delta/2$.
- We will actually prove that for any (fixed) S'', $\mathbb{P}_{S,S'}[C_{S'',S,S'}] \leq \delta/2$.

Theorem. For any hypothesis class H, any (realizable) distribution D, any $(\epsilon, \delta) \in (0,1)^2$, with sample complexity

$$m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon}\right) \text{or } m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon^2}\right)$$

with probability $\geq 1 - \delta$ over $S \sim D^{m(\epsilon,\delta)}$:

$$\forall h \in H : L_S(h) = 0 \implies L_D(h) \leq \epsilon \text{ or } \forall h \in H : |L_D(h) - L_S(h)| \leq \epsilon\ .$$

Proof.

- To show that for any (fixed) S'' of $2m$ examples, $\mathbb{P}_{S,S'}[C_{S'',S,S'}] \leq \delta/2$.
    - **Key idea**: Once S'' is fixed, we only need to consider the projection/restriction of H onto the x's that appear in S''. In other words, there are at most $\Gamma_H(2m)$ labelings that we need to consider.
    - For each such labeling, we will show that the chance of being perfect on S but error $\geq \epsilon/2$ on S' is low. Then, we apply a union bound.
- Fix a labeling $h \in H|_{S''}$. We can assume that h makes at least $\epsilon m/2$ mistakes on S'', otherwise the probability of the bad event is zero.
- When we randomly split S'' into S and S', what's the chance that all these mistakes fall into S'?

Theorem. For any hypothesis class H, any (realizable) distribution D, any $(\epsilon, \delta) \in (0,1)^2$, with sample complexity

$$m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon}\right) \text{or } m(\epsilon, \delta) = O\left(\frac{\ln\left(\Gamma_H(2m)\right) + \ln(1/\delta)}{\epsilon^2}\right)$$

with probability $\geq 1 - \delta$ over $S \sim D^{m(\epsilon,\delta)}$:

$$\forall h \in H : L_S(h) = 0 \implies L_D(h) \leq \epsilon \text{ or } \forall h \in H : |L_D(h) - L_S(h)| \leq \epsilon \, .$$

Proof.

- To show that for any (fixed) S'' of $2m$ examples, $\mathbb{P}_{S,S'}[C_{S'',S,S'}] \leq \delta/2$.
- h makes at least $\epsilon m/2$ mistakes on S''. When we randomly split S'' into S and S', what's the chance that all these mistakes fall into S'?
    - Consider partitioning S'' by randomly pairing the points together $(a_1, b_1), \ldots, (a_m, b_m)$. Then, for each pair (a_i, b_i), flip a coin: if heads then a_i goes to S and b_i goes to S', if tails then a_i goes to S' and b_i goes to S.
    - Observe that if there is any pair (a_i, b_i) where h makes a mistake on both then the chance is zero. Otherwise, the probability that all mistakes fall in S' is at most $\left(\frac{1}{2}\right)^{\epsilon m/2}$.
- By a union bound over all labelings in $H|_{S''}$, $\mathbb{P}_{S,S'}[C_{S'',S,S'}] \leq \Gamma_H(2m)2^{-\epsilon m/2}$.
- To conclude the proof, choose m large enough so that $\Gamma_H(2m)2^{-\epsilon m/2} \leq \delta/2$.

Statistical No Free Lunch

Theorem. For any hypothesis class H, any learning rule A, and any $\epsilon < 1/4$, there exists a (realizable) distribution D, such that if

$$m < \frac{\mathrm{vc}(H) - 1}{8\epsilon},$$

then

$$\mathbb{E}_{S\sim D^m}\left[L_D(A(S))\right] \geq \epsilon\,.$$

Proof Sketch.

- Pick $d = \mathrm{vc}(H)$ shattered points.
- Define a marginal distribution P with probability mass $1 - 4\epsilon$ on one point, and mass $4\epsilon/(d-1)$ on the remaining points.
- Pick a random labeling from the 2^d possible target functions. Then,

$$\begin{aligned}
\mathbb{E}_{S\sim D^m}\left[L_D(A(S))\right] &= \mathbb{P}\{\text{mistake on test point}\} \\
&\geq \frac{1}{2}\mathbb{P}\{\text{test point not in } S\} \\
&\geq \frac{1}{2}4\epsilon\left(1 - \frac{4\epsilon}{d-1}\right)^m \geq 2\epsilon\left(1 - \frac{m4\epsilon}{d-1}\right) \geq 2\epsilon\left(1 - \frac{1}{2}\right) = \epsilon\,.
\end{aligned}$$

Fundamental Theorem of Statistical Learning

Theorem. For any hypothesis class H with finite VC dimension:

- H is (realizably)-PAC-learnable with sample complexity
$$m(\epsilon, \delta) = \Theta\left(\frac{\mathrm{vc}(H) + \ln(1/\delta)}{\epsilon}\right).$$
ERM incurs a multiplicative $\log(1/\epsilon)$.
- H is (agnostically)-PAC-learnable with sample complexity
$$m(\epsilon, \delta) = \Theta\left(\frac{\mathrm{vc}(H) + \ln(1/\delta)}{\epsilon^2}\right).$$
Achieved by ERM.
- H satisfies the uniform convergence property with sample complexity
$$m(\epsilon, \delta) = \Theta\left(\frac{\mathrm{vc}(H) + \ln(1/\delta)}{\epsilon^2}\right).$$

Key Takeaways / Implications

- What is learnable? VC classes.
- How to learn? ERM.
- Tight quantitative understanding of sample complexity.

Leslie Valiant

Question (Harvard, 1984): What is PAC-Larnable?

Leslie Valiant

Question (Harvard, 1984): What is PAC-Larnable?

Answer (Moscow, 1971): H is learnable iff it has finite VC dimension.

Alexey Chervonenkis Vladimir Vapnik

Leslie Valiant

Question (Harvard, 1984): What is PAC-Larnable?

1986

Answer (Moscow, 1971): *H* is learnable iff it has finite VC dimension.

Alexey Chervonenkis   Vladimir Vapnik

Leslie Valiant

Question (Harvard, 1984): What is PAC-Larnable?

1986

Answer (Moscow, 1971): *H* is learnable iff it has finite VC dimension.

Alexey Chervonenkis Vladimir Vapnik

Valiant's actual question: What is efficiently PAC-Larnable?