Statistical Learning Theory Assignment 1

Zimeng Yu

September 27, 2025

Problem 1: VC Dimension of Half Spaces

(a) show $vc(\mathcal{H}) \geq d$

This means we need to find a set of d points that can be shattered by $\mathcal{H}$.

$$\mathcal{H} = \{x \mapsto \text{sign}(\langle w, x \rangle) | w \in \mathbb{R}^d\}.$$

Build a sample set $S = \{x_1, x_2, \ldots, x_d\}$, where x_i is the vector with the i-th element equal to 1, and others 0; i.e., $x_i = e_i$ for $i = 1, \ldots, d$.

The classifier is $h(x_i) = \text{sign}(\langle w, x_i \rangle)$.

$$\langle w, x_i \rangle = \sum_{j=1}^{d} w_j \cdot (e_i)_j = w_i.$$

So, $h(x_i) = \text{sign}(w_i)$.

Goal: For any labeling $y = (y_1, y_2, \ldots, y_d) \in \{\pm 1\}^d$, we need to find a vector $w = (w_1, \ldots, w_d)$ such that for all i, $\text{sign}(w_i) = y_i$.

Let $w_i = y_i$, i.e., $w = y$. Since $y_i \in \{\pm 1\}$, we can set $w_i = y_i$. Then $\text{sign}(w_i) = \text{sign}(y_i) = y_i$. Such a w exists for any labeling y.

Therefore, the set S is shattered by $\mathcal{H}$. By definition of VC dimension, $vc(\mathcal{H}) \geq d$. Proof completed.

(b) show $vc(\mathcal{H}) \leq d$

For any set of $d + 1$ points in a d-dimensional space, $\{x_1, x_2, \ldots, x_{d+1}\}$, the set is linearly dependent. This means there exist scalars $a_1, a_2, \ldots, a_{d+1}$, not all zero, such that:

$$\sum_{i=1}^{d+1} a_i x_i = 0.$$

Consider the specific labeling $y_i = \text{sign}(a_i)$ for all i where $a_i \neq 0$.

Assume for contradiction that this set of $d + 1$ points can be shattered by $\mathcal{H}$. This implies there exists a w that can realize the labeling $y_i = \text{sign}(a_i)$. So, for every i with $a_i \neq 0$, we have:

$$\text{sign}(\langle w, x_i \rangle) = y_i = \text{sign}(a_i).$$

This means that the product $a_i \langle w, x_i \rangle$ must be positive for all i where $a_i \neq 0$. Since not all a_i are zero, at least one such term exists and is positive. All other non-zero terms are also positive.

Therefore, the sum must be strictly positive:

$$\sum_{i=1}^{d+1} a_i \langle w, x_i \rangle > 0.$$

However, from the linear dependence of the points, we have:

$$\sum_{i=1}^{d+1} a_i x_i = 0.$$

Taking the dot product with w on both sides:

$$\left\langle w, \sum_{i=1}^{d+1} a_i x_i \right\rangle = \langle w, 0 \rangle$$

$$\sum_{i=1}^{d+1} a_i \langle w, x_i \rangle = 0.$$

But from our assumption, we derived that $\sum_{i=1}^{d+1} a_i \langle w, x_i \rangle > 0$. This is a contradiction ($0 > 0$ is false).

So, the assumption that the set could be shattered is false. Specifically, the labeling $y_i = \text{sign}(a_i)$ cannot be realized. This proves that $vc(\mathcal{H}) \leq d$.

In summary, from (a) and (b), we have $vc(\mathcal{H}) = d$.

Problem 2 (VC Dimension of Composition)

- $\mathcal{X}$: an arbitrary instance space
- $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$: an arbitrary binary hypothesis class where $vc(\mathcal{H}) = d$,
- $k \in \mathbb{N}$, boolean function $f : \{\pm 1\}^k \to \{\pm 1\}$
- for binary class function f:

$$\mathcal{H}_f^k = \{x \mapsto f(h_1(x), \ldots, h_k(x)) \mid h_1, \ldots, h_k \in \mathcal{H}\}$$

Show $vc(\mathcal{H}_f^k) \leq O(dk \log(k))$ use Sauer-Shelah-Perles Lemma.

Step 1

Set $S = \{x_1, \ldots, x_m\}$, $m \in \mathbb{N}$.
The number of ways a hypothesis $h \in \mathcal{H}$ can label the set S is given by the growth function $\Pi_{\mathcal{H}}(m)$.
By Sauer-Shelah-Perles Lemma, for class $\mathcal{H}$ with $vc(\mathcal{H}) = d$,

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d \quad \text{(by Hint)}.$$

The number of ways the composite class can label S is bounded by:

$$\Pi_{\mathcal{H}_f^k}(m) \leq (\Pi_{\mathcal{H}}(m))^k \leq \left(\left(\frac{em}{d}\right)^d\right)^k = \left(\frac{em}{d}\right)^{dk}.$$

Step 2

Show that for a sufficiently large m, $\Pi_{\mathcal{H}_f^k}(m) < 2^m$. This implies the class cannot shatter any set of size m, and thus $vc(\mathcal{H}_f^k) < m$.

$$\left(\frac{em}{d}\right)^{dk} < 2^m$$

Take $\log_2$ on both sides:

$$\log_2\left(\left(\frac{em}{d}\right)^{dk}\right) < \log_2(2^m)$$
$$dk \cdot \log_2\left(\frac{em}{d}\right) < m$$
$$\text{expand:} \quad dk \cdot (\log_2(e) + \log_2(m) - \log_2(d)) < m$$

As suggested by the target bound $O(dk\log(k))$, let's check if a choice of m proportional to $dk\log(k)$ works. Assume m is large enough, let's find a constant C for

$$m = C \cdot dk\log_2(k).$$

Substitute this into the inequality:

$$dk \cdot (\log_2 e + \log_2 m - \log_2 d) < C \cdot dk\log_2(k)$$
$$\log_2 e + \log_2(C \cdot dk \cdot \log_2 k) - \log_2 d < C\log_2(k)$$
$$\log_2 e + \log_2 C + \log_2 d + \log_2 k + \log_2(\log_2 k) - \log_2 d < C\log_2 k$$
$$\log_2 e + \log_2 C + \log_2 k + \log_2(\log_2 k) < C\log_2 k$$

Divide by $\log_2 k$:

$$\frac{\log_2 e + \log_2 C}{\log_2 k} + 1 + \frac{\log_2(\log_2 k)}{\log_2 k} < C$$

As $k \to \infty$, the left-hand side approaches 1. For the inequality to hold for large k, we need $1 < C$. Thus, any constant $C > 1$ will work for sufficiently large k and d values. This shows that for $m = O(dk\log k)$, the set cannot be shattered, so $vc(\mathcal{H}_f^k) < m$, which implies $vc(\mathcal{H}_f^k) = O(dk\log k)$.

Thus, if we choose $m = C \cdot dk\log_2(k)$, the number of ways to label a set of m points is strictly less than 2^m, which proves that no set of size m can be scattered by the hypothesis class $\mathcal{H}_f^k$.

By definition of VC dimension, $vc(\mathcal{H}_f^k) < m$. Since m is on the order of $dk\log(k)$,

$$vc(\mathcal{H}_f^k) \leq O(dk\log k).$$

The proof is complete.

Problem 3

Given $\mathcal{H} = \{x \mapsto \text{sign}(\sin(\theta x)) \mid \theta \in \mathbb{R}\}$, show $vc(\mathcal{H})$ is infinite.

- $\text{sign}(\sin(\theta x)) = +1 \iff \sin(\theta x) > 0 \iff \theta x \in (2k\pi, (2k+1)\pi)$
- $\text{sign}(\sin(\theta x)) = -1 \iff \sin(\theta x) < 0 \iff \theta x \in ((2k+1)\pi, (2k+2)\pi)$

Hint: $\sin(\pi(n+a)) = (-1)^n \sin(\pi a)$. The sign of $\sin(\theta x)$ is determined by the parity of n.

Rewrite the hypothesis $h(x)$ as: $\text{sign}(\sin(\theta x)) = (-1)^{\lfloor \theta x/\pi \rfloor}$.

For a better structured proof, choose the sample set:

$$S = \{x_j = 2^{j-1} \mid j = 1, 2, 3, \ldots, d\} = \{1, 2, 4, \ldots, 2^{d-1}\}$$

For an arbitrary labeling $(y_1, y_2, \ldots, y_d)$, define the target parity bit b_j:

$$b_j = \frac{1 - y_j}{2} \in \{0, 1\}, \text{ for } y_j \in \{\pm 1\}.$$

If $y_j = 1$, $b_j = 0$; if $y_j = -1$, $b_j = 1$.

Our goal is to find a θ such that $\lfloor \theta x_j/\pi \rfloor \pmod 2 = b_j$ for all j.

Define $z = \theta/\pi$. Rewrite the condition as: $\lfloor z \cdot 2^{j-1} \rfloor \pmod 2 = b_j$.

- For $j = 1$: $\lfloor z \cdot 2^0 \rfloor \pmod 2 = \lfloor z \rfloor \pmod 2 = b_1$
- For $j = 2$: $\lfloor z \cdot 2^1 \rfloor \pmod 2 = b_2$
- For $j = 3$: $\lfloor z \cdot 2^2 \rfloor \pmod 2 = b_3$
- $\ldots$
- For j: $\lfloor z \cdot 2^{j-1} \rfloor \pmod 2 = b_j$

This suggests that the bit of z at position $2^{-(j-1)}$ should be equal to b_j. Let's construct z based on this:

$$z = \frac{\theta}{\pi} = \sum_{j=1}^{d} b_j 2^{-(j-1)}.$$

Check the construction. Let $I_j = \lfloor z \cdot 2^{j-1} \rfloor$:

$$\begin{aligned} I_j &= \left\lfloor 2^{j-1} \sum_{k=1}^{d} b_k 2^{-(k-1)} \right\rfloor = \left\lfloor \sum_{k=1}^{d} b_k 2^{j-k} \right\rfloor \\ &= \lfloor (b_1 2^{j-1} + b_2 2^{j-2} + \ldots + b_j 2^0) + (b_{j+1} 2^{-1} + \ldots + b_d 2^{j-d}) \rfloor \end{aligned}$$

The integer part of the expression inside the floor is $\sum_{k=1}^{j} b_k 2^{j-k}$. The parity of this integer is determined by the last term, $b_j 2^0 = b_j$. Therefore, $I_j \pmod 2 = b_j$.

Then $\sin(\theta x_j) = \sin(\pi \cdot z \cdot 2^{j-1}) = 0$.

Let $z = \sum_{j=1}^{d} b_j 2^{-(j-1)} + 2^{-(d+1)}$, and let $\theta = \pi z$.

$$\begin{aligned} z \cdot 2^{j-1} &= \left(\sum_{k=1}^{d} b_k 2^{-(k-1)} + 2^{-(d+1)} \right) 2^{j-1} \\ &= \sum_{k=1}^{d} b_k 2^{j-k} + 2^{j-d-2} \\ &= \underbrace{(b_1 2^{j-1} + \ldots + b_j 2^0)}_{\text{Integer Part } I_j} + \underbrace{(b_{j+1} 2^{-1} + \ldots + b_d 2^{j-d} + 2^{j-d-2})}_{\text{Fractional Part } F_j} \end{aligned}$$

For the Fractional Part F_j, it is strictly larger than 0 and less than 1. The maximum possible value is less than $\sum_{k=1}^{\infty} 2^{-k} = 1$. So, $0 < F_j < 1$.

So, $\lfloor z \cdot 2^{j-1} \rfloor = I_j$, and the parity is correct (b_j). Also, $z \cdot 2^{j-1}$ is never an integer, so $\sin(\theta x_j)$ is never zero.

In summary, the set $S = \{1, 2, 4, \ldots, 2^{d-1}\}$ can be shattered. Since we can do this for an arbitrarily large d, the VC dimension of $\mathcal{H}$ is infinite.

Problem 4 (Optimistic Bounds, Bernstein Inequality)

Let $\mathcal{H} \subseteq Y^{\mathcal{X}}$ be a finite hypothesis class and D be an arbitrary distribution over $\mathcal{X} \times Y$. Let $\delta \in (0,1)$, $m \in \mathbb{N}$.

(a) Bernstein's Inequality to prove Equation (3)

The goal is to prove the uniform convergence bound:

$$\forall h \in \mathcal{H} : |L_S(h) - L_D(h)| \leq c_1 \sqrt{\frac{L_D(h)\log(|\mathcal{H}|/\delta)}{m}} + c_2 \frac{\log(|\mathcal{H}|/\delta)}{m}$$

This starts with a bound on $|\frac{1}{m}\sum X_i|$. From Bernstein's Inequality, we first derive a bound for a single variable's average:

$$\left|\frac{1}{m}\sum_{i=1}^{m} X_i\right| \leq \sqrt{\frac{2\mathbb{E}[X_1^2]\log(2/\delta)}{m}} + \frac{2M\log(2/\delta)}{3m}$$

To get this, we bound $\mathbb{P}[\sum X_i \geq t]$. We apply Bernstein's inequality with $t = m\epsilon$. We know $\sum_{i=1}^{m} \mathbb{E}[X_i^2] = m\mathbb{E}[X_1^2]$ (since the variables are i.i.d.) and for each variable, $|X_i| \leq M$.

Plug in and then get:

$$\mathbb{P}\left[\sum_{i=1}^{m} X_i \geq m\epsilon\right] \leq \exp\left(-\frac{(m\epsilon)^2/2}{m\mathbb{E}[X_1^2] + M(m\epsilon)/3}\right) = \exp\left(-\frac{m\epsilon^2/2}{\mathbb{E}[X_1^2] + M\epsilon/3}\right)$$

Using a union bound for the absolute value, the probability of deviation ϵ is:

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i\right| \geq \epsilon\right] \leq 2 \cdot \exp\left(-\frac{m\epsilon^2/2}{\mathbb{E}[X_1^2] + M\epsilon/3}\right)$$

Set this probability to δ:

$$\delta = 2 \cdot \exp\left(-\frac{m\epsilon^2/2}{\mathbb{E}[X_1^2] + M\epsilon/3}\right)$$

Rearrange to solve for ϵ:

$$\left(\frac{m}{2}\right)\epsilon^2 - \left(\frac{M\ln(2/\delta)}{3}\right)\epsilon - (\mathbb{E}[X_1^2]\ln(2/\delta)) = 0$$

Solving this quadratic for ϵ gives the result:

$$\epsilon \leq \sqrt{\frac{2\mathbb{E}[X_1^2]\log(2/\delta)}{m}} + \frac{2M\log(2/\delta)}{3m}$$

$$\epsilon \leq \sqrt{\frac{2\mathbb{E}[X_1^2]\log(2/\delta)}{m}} + \frac{2M\log(2/\delta)}{3m}$$

Proof completed.

(b) Use Equation 4 and the union bound over Hto show Equation 3

To bound the deviation $|L_S(h) - L_D(h)|$, define a random variable X_i. Let ℓ denote the 0-1 loss:

$$X_i^{(h)} = \ell(h(z_i), y_i) - L_D(h)$$

$$\frac{1}{m}\sum_{i=1}^{m} X_i^{(h)} = \frac{1}{m}\sum_{i=1}^{m} \ell(h(z_i), y_i) - L_D(h) = L_S(h) - L_D(h)$$

Check conditions:

1. Mean: $\mathbb{E}[X_i^{(h)}] = \mathbb{E}[\ell(h(z_i), y_i)] - L_D(h) = L_D(h) - L_D(h) = 0$.
2. Boundedness: For 0-1 loss, $\ell \in \{0, 1\}$. Since $0 \leq L_D(h) \leq 1$, $X_i^{(h)}$ is bounded. A simple bound is $M = 1$.
3. Variance: The variance is $L_D(h)(1-L_D(h))$. Since $(1-L_D(h)) \leq 1$, we can use the upper bound $\mathbb{E}[(X_i^{(h)})^2] \leq L_D(h)$.

Use Equation (4) for a fixed h with failure probability δ':

$$|L_S(h) - L_D(h)| \leq \sqrt{\frac{2L_D(h)\log(2/\delta')}{m}} + \frac{2\log(2/\delta')}{3m}$$

Use the union bound: set $\delta' = \delta/|\mathcal{H}|$. Then,

$$\log(2/\delta') = \log(2|\mathcal{H}|/\delta)$$

Thus, for the final uniform bound, which holds with probability at least $1 - \delta$ for all $h \in \mathcal{H}$:

$$|L_S(h) - L_D(h)| \leq \sqrt{\frac{2L_D(h)\log(2|\mathcal{H}|/\delta)}{m}} + \frac{2\log(2|\mathcal{H}|/\delta)}{3m}$$

This matches Equation (3), with $c_1 = \sqrt{2}$, $c_2 = 2/3$. Proof completed.

(c)

For an upper bound on $L_D(\hat{h})$, let $B(h)$ be the bound term (the right-hand side) from Equation (3).

$$L_D(\hat{h}) \leq L_S(\hat{h}) + B(\hat{h})$$

The true best hypothesis is h^*. By definition, $\hat{h}$ is the empirical risk minimizer (minimizes sample loss):

$$L_S(\hat{h}) \leq L_S(h^*)$$

Substituting this into the first inequality:

$$L_D(\hat{h}) \leq L_S(h^*) + B(\hat{h})$$

Now consider the sample loss of h^*, $L_S(h^*)$, and use Equation (3) on it:

$$L_S(h^*) \leq L_D(h^*) + B(h^*)$$

Substituting this back into our chain of inequalities:

$$L_D(\hat{h}) \leq (L_D(h^*) + B(h^*)) + B(\hat{h})$$

$$L_D(\hat{h}) \leq L_D(h^*) + B(h^*) + B(\hat{h})$$

Now,

$$L_D(\hat{h}) \leq L_D(h^*) + \left(c_1 \sqrt{\frac{L_D(h^*) \log'}{m}} + c_2 \frac{\log'}{m} \right) + \left(c_1 \sqrt{\frac{L_D(\hat{h}) \log'}{m}} + c_2 \frac{\log'}{m} \right)$$

where $\log' = \log(|\mathcal{H}|/\delta)$.

We can argue that $B(\hat{h})$ can be upper-bounded by a term involving $L_D(h^*)$. Since $L_D(\hat{h})$ should be close to $L_D(h^*)$, we can approximate the bound on $\hat{h}$ with the bound on h^*:

$$L_D(\hat{h}) \approx L_D(h^*) + 2 \times \mathrm{B}(h^*)$$

Then:

$$L_D(\hat{h}) \leq L_D(h^*) + 2c_1 \sqrt{\frac{L_D(h^*) \log(|\mathcal{H}|/\delta)}{m}} + 2c_2 \frac{\log(|\mathcal{H}|/\delta)}{m}$$

This matches Equation (5), with new constants $c_3 = 2c_1$ and $c_4 = 2c_2$.

Proof complete.

Appendix

I asked Gemini for following questions as a supplementary, check this link: Gemini Prompt.