Machine Learning Theory (CS 6783)

Lecture 6 : VC dimension recap and continuted

1 Recap

1. For any set $V \subset \mathbb{R}^n$:

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mathbf{v}\in V}\sum_{t=1}^{n}\epsilon_t\mathbf{v}[t]\right] \le \frac{1}{n}\sqrt{2\left(\sup_{\mathbf{v}\in V}\sum_{t=1}^{n}\mathbf{v}^2[t]\right)\log|V|}$$

2. We defined growth function as $\Pi(\mathcal{F}, n) = \max_{x_1,\ldots,x_n}\left|\mathcal{F}_{|x_1,\ldots,x_n}\right|$

3. VC dimension : size of largest set of input instances we can shatter

$$\mathrm{VC}(\mathcal{F}) = \max\{d : \Pi(\mathcal{F}, d) = 2^d\}$$

4. VC/Sauer/Shelah Lemma : $\Pi(\mathcal{F}, n) \le \sum_{i=0}^{\mathrm{VC}(\mathcal{F})}\binom{n}{i}$

5. $\mathcal{F}$ is learnable if and only if $\mathrm{VC}(\mathcal{F}) < \infty$

   (a) $\mathcal{V}_n^{\mathrm{stat}}(\mathcal{F}) \le \sqrt{\frac{\mathrm{VC}(\mathcal{F})\log n}{n}}$

   (b) $\mathrm{VC}(\mathcal{F}) = \infty \implies \mathcal{V}_n^{\mathrm{stat}}(\mathcal{F}) \ge 1/4$

2 VC dimension examples and utility lemmas

1. To show $\mathrm{VC}(\mathcal{F}) \ge d$ show that you can at least pick d points $x_1, \ldots, x_d$ that can be shattered.

2. To show that $\mathrm{VC}(\mathcal{F}) \le d$ show that no configuration of $d + 1$ points can be shattered.

Examples : Intervals, axis aligned rectangle, lines, convex polygons

Claim 1. *VC dimension of half-spaces in* $\mathbb{R}^d$ *is* $d + 1$

Proof. We consider half-spaces that map vector in $\mathbb{R}^d$ to $\{\pm 1\}$. That is

$$\mathcal{F} = \{\mathbf{x} \mapsto \mathrm{sign}\left(\mathbf{f}^\top\mathbf{x} + f_0\right) : \mathbf{f} \in \mathbb{R}^d, f_0 \in \mathbb{R}\}$$

We prove the statement as follows.

1. $\text{VC}(\mathcal{F}) \geq d + 1$:
   We can shatter the points $\mathbf{e}_1, \ldots, \mathbf{e}_d, \mathbf{0}$. To see this, note that given any $y_1, \ldots, y_{d+1} \in \{\pm 1\}^{d+1}$, if we consider $f \in \mathcal{F}$ given by $f_0 = y_{d+1}$ and for all $i \in [d]$, $\mathbf{f}[i] = y_i - y_{d+1}$. Hence note that, $f(\mathbf{0}) = \text{sign}\left(\mathbf{f}^\top \mathbf{0} + f_0\right) = \text{sign}(y_{d+1}) = y_{d+1}$. Also, for any $i \in [d]$, $f(\mathbf{e}_i) = \text{sign}\left(\mathbf{f}^\top \mathbf{e}_i + f_0\right) = \text{sign}\left(y_i - y_{d+1} + y_{d+1}\right) = y_i$.

2. $\text{VC}(\mathcal{F}) < d + 2$:
   By Radon theorem, any set of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two disjoint subsets whose convex hulls have a non-empty intersection. Label one of these partitions $+1$ and other -1. No half-space can successfully label points in the intersection.

□

Claim 2. *For any binary hypothesis class $\mathcal{F}$,*

$$\text{VC}(\mathcal{F}) \leq \log_2 |\mathcal{F}|$$

Proof. Note that for any d, $\Pi(\mathcal{F}, d) \leq |\mathcal{F}|$. From the definition of VC dimension, we have, $\text{VC}(\mathcal{F}) = \max\{d : \Pi(\mathcal{F}, d) = 2^d\}$. Hence $2^{\text{VC}(\mathcal{F})} \leq |F|$ □

Claim 3. *Consider any fixed function $g : \{\pm 1\}^k \mapsto \{\pm 1\}$. For every $i \in [k]$, let $\mathcal{F}_i$ be a some class of binary hypotheses mapping from input space $\mathcal{X}$. Let $\mathcal{G} = \{x \mapsto g(f_1(x), \ldots, f_k(x)) : f_i \in \mathcal{F}_i\}$. Then*

$$\text{VC}(\mathcal{G}) \leq 2k \ \text{VC}(\mathcal{F}) \log(5k)$$

Proof.

$$\Pi(\mathcal{G}, n) \leq \prod_{i=1}^{k} \Pi(\mathcal{F}_i, n) \leq (en)^{\sum_{i=1}^k \text{VC}(\mathcal{F}_i)}$$

By definition of VC dimension, we have that $\text{VC}(\mathcal{G}) = \max\{d : \Pi(\mathcal{G}, d) = 2^d\}$. Hence,

$$2^{\text{VC}(\mathcal{G})} \leq (e \ \text{VC}(\mathcal{G}))^{\sum_{i=1}^k \text{VC}(\mathcal{F}_i)}$$

Hence

$$\text{VC}(\mathcal{G}) \leq \log_2 (e \ \text{VC}(\mathcal{G})) \sum_{i=1}^{k} \text{VC}(\mathcal{F}_i)$$

Hence, we can conclude that $\text{VC}(\mathcal{G}) \leq \sum_{i=1}^k \text{VC}(\mathcal{F}_i) \ \log\left(\sum_{i=1}^k \text{VC}(\mathcal{F}_i) \cdot \log\left(\sum_{i=1}^k \text{VC}(\mathcal{F}_i)\right)\right)$ □