9 VC-dimension

Monday, March 9, 2020 9:04 PM

Thm 5.9 (Radon): Any set $S \subseteq \mathbb{R}^d$ with $|S| \geq d+2$ can be partitioned into two disjoint subsets A and B such that $\text{convex}(A) \cap \text{convex}(B) \neq \emptyset$.

proof: WLOG, assume $|S| = d+2$, with $S = \{\vec{a}_1, \ldots, \vec{a}_{d+2}\}$

Let $A = [\vec{a}_1\ \vec{a}_2\ \vec{a}_3\ \ldots\ \vec{a}_{d+2}] \in \mathbb{R}^{d \times (d+2)}$.

Let $B = \begin{bmatrix} A \\ \vec{1}^T \end{bmatrix}$. $\text{rank}(B) \leq d+1$, so columns are lin. dep. Let $B = [\vec{b}_1 \ \ldots\ \vec{b}_{d+2}]$.

Let $\vec{x} = (x_1, x_2, \ldots, x_{d+2})^T$ s.t. $B\vec{x} = 0$.

WLOG, say $x_1, \ldots, x_s \geq 0$ and $x_{s+1}, \ldots, x_{d+2} < 0$.

Normalize $\vec{x}$ s.t. $\sum_{i=1}^{s} |x_i| = 1$.

Then $$\sum_{i=1}^{s} |x_i| \vec{b}_i = \sum_{i=s+1}^{d+2} |x_i| \vec{b}_i$$

$$\Rightarrow \sum_{i=1}^{s} |x_i| \vec{a}_i = \sum_{i=s+1}^{d+2} |x_i| \vec{a}_i \quad \text{and} \quad \sum_{i=1}^{s} |x_i| = \sum_{i=s+1}^{d+2} |x_i| = 1$$

Thus, both sides are convex combinations of disjoint columns of A.

The convex hulls of the two sets of corresponding pts intersect.

But then it is impossible to have a linear separator of these two sets, so half-planes cannot shatter a $(d+2)$-pt set.

$$\Rightarrow VC(\mathcal{H}) < d+2$$

$$\Rightarrow VC(\mathcal{H}) = d+1.$$

Ex. $X = \mathbb{R}^d$, $\mathcal{H} = \{\vec{x} \mid |\vec{x} - \vec{x}_0| \leq r\}$ spheres.

$VC(\mathcal{H}) < d+2$ because if we can put spheres around two disjoint sets,

$VC(\mathcal{H}) < d+2$ because if we can put spheres around two disjoint sets, then those sets are also divided by a hyperplane, and $VC(\{\text{half-spaces}\}) = d+1$.

$VC(\mathcal{H}) \geq d+1$ by the same construction as half-spaces.

Let $S = \{\vec{0}, \vec{e}_1, \ldots, \vec{e}_d\}$.

Given a subset $A \subseteq S$, choose the ball center $\vec{a}_0 = \sum_{a \in A} \vec{a}$.

Then $|\vec{a}_0 - a| = \sqrt{|A|-1}$ $\forall\, a \in A$ and $a \neq 0$

$|\vec{a}_0 - a| = \sqrt{|A|+1}$ $\forall\, a \notin A$ and $a \neq 0$.

$|\vec{a}_0| = \sqrt{|A|}$.

So we can choose a radius so that this ball contains exactly A.

Define: For any set system $(X, \mathcal{H})$, the shatter function

$$\pi_{\mathcal{H}}(n) = \max_{|A|=n} \{A \cap h\}_{h \in \mathcal{H}},$$

the maximum number of subsets of any set A, $|A|=n$, that can be expressed as $A \cap h$ for $h \in \mathcal{H}$.

Note: $\pi_{\mathcal{H}}(n) = 2^n$ for $n \leq VC(\mathcal{H})$.

Notation: $\binom{n}{\leq d} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{d} \leq n^d + 1$. ($\leq n^d$ if $d > 1$)

(because consider choosing n items d times with duplicates.)

Lemma 5.10 (Sauer): For any set system $(X, \mathcal{H})$ of VC-dim. $\leq d$,

$$\pi_{\mathcal{H}}(n) \leq \binom{n}{\leq d} \quad \forall n.$$

proof. Note that for $n \geq 1$ and $d \geq 1$,

$$\binom{n}{\leq d} = \binom{n-1}{\leq d-1} + \binom{n-1}{\leq d}$$

If we choose the first element, then

If we don't choose the first element,

If we choose the first element, then we can choose at most $d-1$ of the remaining $n-1$.

If we don't choose the first element, then we can choose at most d of the remaining n.

We will prove by induction on n and d.

Base case: $d = 0$.

Suppose $h_1 \neq h_2 \in \mathcal{H}$. Then $\exists\, a \in X$ s.t. $a \in h_1$, $a \notin h_2$.

But then $\mathcal{H}$ shatters $\{a\} \Rightarrow d \geq 1$

$\Rightarrow \mathcal{H} = \{h\}$.

Any set $A \subseteq X$ has $|2^A| = 2^{|A|}$, so only set of size 0 can be shattered if $|\mathcal{H}| = 1$, i.e. only $\emptyset$ can be shattered.

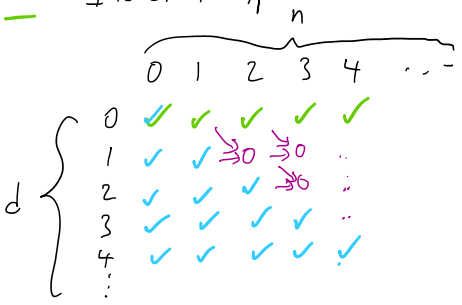
$\Rightarrow \pi_{\mathcal{H}}(n) = 1 = \binom{n}{\leq 0}$.

Base case: $n \leq d$.

By definition any set $|A| = n$ can be shattered, so $\pi_{\mathcal{H}}(n) = 2^n$.

$$\binom{n}{\leq d} = \underbrace{\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n}}_{= 2^n} + \underbrace{\binom{n}{n+1} + \cdots + \binom{n}{d}}_{= 0} = 2^n.$$

General n, d: Induction hypothesis: assume true for $(n-1, d-1)$ and $(n-1, d)$.



Select $A \subseteq X$ with $|A| = n$ s.t. $\pi_{\mathcal{H}}(n)$ subsets of A can be expressed as $A \cap h$ for $h \in \mathcal{H}$

a "maximally shatterable" A.

WLOG, assume $X = A$ and replace $h \in \mathcal{H}$ by $h \cap A$, removing duplicate sets. Then $|\mathcal{H}| = \pi_{\mathcal{H}}(n)$ and each $h \in \mathcal{H}$ is a subset of A. (i.e. $\mathcal{S} = (A, \{h \cap A\}_{h \in \mathcal{H}})$)

Need only control $|\mathcal{H}| \leq \binom{n}{\leq d}$.

Need only control $|\mathcal{H}| \le \binom{n}{\le d}$.

Remove u from set A and from each set in $\mathcal{H}$.

i.e. Let $\mathcal{S}_1 = (A - \{u\}, \mathcal{H}_1)$, where $\mathcal{H}_1 = \{h - \{u\} \mid h \in \mathcal{H}\}$

Note: $|\mathcal{H}_1| = \pi_{\mathcal{H}_1}(n)$. Clearly, $|\mathcal{H}_1| \ge \pi_{\mathcal{H}_1}(n)$. Suppose $\pi_{\mathcal{H}_1}(n) < |\mathcal{H}_1|$.

Then $\exists\, h_1' \ne h_2' \in \mathcal{H}_1$ s.t. $h_1' \cap (A - \{u\}) = h_2' \cap (A - \{u\})$

$\Rightarrow (h_1 - \{u\}) \cap (A - \{u\}) = (h_2 - \{u\}) \cap (A - \{u\})$, $h_1, h_2 \in \mathcal{H}$.

$\Rightarrow (h_1 \cap A) - \{u\} = (h_2 \cap A) - \{u\}$

$\Rightarrow h_1 - \{u\} = h_2 - \{u\}$

$\Rightarrow h_1' = h_2'$. $\rightarrow\!\leftarrow$

For $h \subseteq A - \{u\}$, if exactly one of h and $h \cup \{u\}$ is in $\mathcal{H}$, then h contributes one set to both $\mathcal{H}$ and $\mathcal{H}_1$.

If both h and $h \cup \{u\}$ are in $\mathcal{H}_1$, then h contributes two sets to $\mathcal{H}$ but only one set to $\mathcal{H}_1$.

Thus, $|\mathcal{H}| - |\mathcal{H}_1| = \left| \{(h_1, h_2) \mid h_1 - h_2 = \{u\}\}_{h_1, h_2 \in \mathcal{H}} \right|$. i.e. pairs of sets that differ only by u.

Let $\mathcal{S}_2 = (A - \{u\}, \mathcal{H}_2)$ where $\mathcal{H}_2 = \{h \mid u \notin h,\ h \in \mathcal{H},\ h \cup \{u\} \in \mathcal{H}\}$

Note: $|\mathcal{H}_2| = \pi_{\mathcal{H}_2}(n-1)$. Clearly, $|\mathcal{H}_2| \ge \pi_{\mathcal{H}_2}(n-1)$. Suppose $\pi_{\mathcal{H}_2}(n-1) < |\mathcal{H}_2|$.

Then $\exists\, h_1' \ne h_2' \in \mathcal{H}_2$ s.t. $h_1' \cap (A - \{u\}) = h_2' \cap (A - \{u\})$

$\Rightarrow h_1' = h_2'$. $\rightarrow\!\leftarrow$.

Then $|\mathcal{H}| = |\mathcal{H}_1| + |\mathcal{H}_2|$

$\Rightarrow \pi_{\mathcal{H}}(n) = \pi_{\mathcal{H}_1}(n-1) + \pi_{\mathcal{H}_2}(n-1)$

Note: $VC(\mathcal{H}_1) \le d$ because otherwise, $\mathcal{H}_1$ would shatter a set of cardinality $d+1$, and $\mathcal{H}$ would also shatter that set.

Note: $VC(\mathcal{H}_2) \le d-1$ because if $\mathcal{H}_2$ shattered $B \subseteq A - \{u\}$ with $|B| \ge d$, then $B \cup \{u\}$ would be shattered by $\mathcal{H}$.

By the induction hypothesis, $\pi_{\mathcal{H}_1}(n-1) \le \binom{n-1}{\le d}$

$$\pi_{\mathcal{H}_2}(n-1) \le \binom{n-1}{\le d-1}.$$

Thus, $\pi_{\mathcal{H}}(n) \le \binom{n-1}{\le d} + \binom{n-1}{\le d-1} = \binom{n}{\le d}$

So far, we have only considered VC-dimension on single concepts. Often, in ML, we want to combine together multiple concepts.

Lemma 5.11 Suppose $(X, \mathcal{H}_1)$ and $(X, \mathcal{H}_2)$ are set systems on the same X.

Then $\pi_{\mathcal{H}_1 \cap \mathcal{H}_2}(n) \leq \pi_{\mathcal{H}_1}(n) \cdot \pi_{\mathcal{H}_2}(n)$,

where $\mathcal{H}_1 \cap \mathcal{H}_2 = \{ h_1 \cap h_2 \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2 \}$.

proof. Let $A \subseteq X$, $|A| = n$. Let $S = \{ A \cap h \mid h \in \mathcal{H}_1 \cap \mathcal{H}_2 \}$.

By definition, $S = \{ A \cap (h_1 \cap h_2) \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2 \}$

$\Rightarrow S = \{ (A \cap h_1) \cap (A \cap h_2) \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2 \}$.

$\Rightarrow |S| \leq |\{ A \cap h_1 \mid h_1 \in \mathcal{H}_1 \}| \cdot |\{ A \cap h_2 \mid h_2 \in \mathcal{H}_2 \}|$

Choose A s.t. $|S| = \pi_{\mathcal{H}_1 \cap \mathcal{H}_2}(n)$.

Then $\pi_{\mathcal{H}_1 \cap \mathcal{H}_2}(n) \leq \pi_{\mathcal{H}_1}(n) \pi_{\mathcal{H}_2}(n)$.

This allows us to take the Boolean AND of concepts.

i.e. if $X = \mathbb{R}^d$, and $\mathcal{H}_1 = \{\text{half-spaces}\}$ and $\mathcal{H}_2 = \{\text{half-spaces}\}$,

$\mathcal{H}_1 \cap \mathcal{H}_2 = \{\text{intersection of two half-spaces}\}$

$= \{\text{half-space 1 AND half-space 2}\}$.

Can extend to Boolean ANDs of many concepts.

Define Given a concept class $\mathcal{H}$, a Boolean function f, and an integer k, define the set

$$\text{comb}_f(h_1, \dots, h_k) = \{ x \in X \mid f(h_1(x), \dots, h_k(x)) = 1 \},$$

where $h_i(x) = 1$ iff $x \in h_i$.

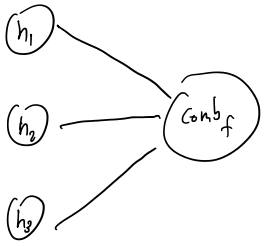
Ex. f is the AND function $\Rightarrow \text{comb}_f(h_1, \dots, h_k) = \prod h_i$

Ex. f is the majority-vote function $\Rightarrow \text{comb}_f(h_1, \dots, h_k) = \left\lfloor \sum h_i / k \right\rfloor$.

Interpretation: 

"depth-2"

Interpretation:



"depth-2" neural network

Define: $COMB_{f,k}(\mathcal{H}) = \{comb_f(h_1,\dots,h_k) \mid h_i \in \mathcal{H}\}$ a new concept class.

Lemma 5.12 For any Boolean function f, hypothesis class $\mathcal{H}$, integer k,

$$\pi_{COMB_{f,k}(\mathcal{H})}(n) \le \pi_{\mathcal{H}}(n)^k.$$

proof. Same reasoning as 5.11.

Theorem 5.13 If $VC(\mathcal{H}) = V$, then for any Boolean function f and integer k,

$$VC\left(COMB_{f,k}(\mathcal{H})\right) = O(kV\log(kV)).$$

proof. Let $n = VC(COMB_{f,k}(\mathcal{H}))$. We use n because it'll be the size of a shattered set.

By def., $\exists$ set S of n points shattered by $COMB_{f,k}(\mathcal{H})$.

By Sauer's Lemma, $\pi_{\mathcal{H}}(n) \le \binom{n}{\le V} \le n^V$, so there are at most n^V ways of partitioning S using sets in $\mathcal{H}$.

But each set in $COMB_{f,k}(\mathcal{H})$ is determined by k sets in $\mathcal{H}$, so there are at most $(n^V)^k = n^{kV}$ ways of partitioning the points using $COMB_{f,k}(\mathcal{H})$.

Since S is shattered, we must have $2^n \le n^{kV} \Rightarrow n \le kV\log_2(n)$.

If $n \ge 16$, $\log_2(n) \le \sqrt{n} \Rightarrow kV\log_2(n) \le kV\sqrt{n}$

$\Rightarrow n \le kV\sqrt{n}$

$n \le (kV)^2$.

Plugging back in, $n \le kV\log_2(kV)^2 = 2kV\log_2(kV)$. ▨

Theorem 5.14 (Key Thm): Let $(X, \mathcal{H})$ be a set system, D a probability distribution over X, and let n be an integer satisfying

Theorem 3.11. Let $(X, \mathcal{H})$ be a [illegible] ν a probability distribution over X, and let n be an integer satisfying

$$n \geq \frac{2}{\varepsilon}\left[\log_2 2\pi_{\mathcal{H}}(2n) + \log_2 \frac{1}{\delta}\right].$$

Let S_1 consist of n points drawn from D, possibly with repetition.

With prob. $\geq 1-\delta$, every set in $\mathcal{H}$ of probability mass $\geq \varepsilon$ intersects S_1.

proof. Let A be the event: $\exists\, h \in \mathcal{H}$ with $\mu(h) \geq \varepsilon$ s.t. $h \cap S_1 = \emptyset$.

Draw a second set S_2 of n points from D.

Let B be the event: $\exists\, h \in \mathcal{H}$ with $h \cap S_1 = \emptyset$ but $|h \cap S_2| \geq \frac{\varepsilon}{2} \cdot n$

If $h \cap S_1 = \emptyset$, and $\mu(h) \geq \varepsilon$, then $\mathbb{E}|h \cap S_2| = \sum_{x \in S_2} \mathbb{1}_h \geq n\varepsilon$.

By Markov, $\mathbb{P}\left(|h \cap S_2| \geq \frac{\varepsilon n}{2}\right) \geq \frac{1}{2}$.

$$\Rightarrow \quad \text{Prob}(B \mid A) \geq \frac{1}{2}. \quad \Rightarrow \quad \text{Prob}(B) \geq \frac{1}{2}\text{Prob}(A)$$

Thus, to prove $\text{Prob}(A) \leq \delta$, it would suffice to show $\text{Prob}(B) \leq \frac{\delta}{2}$.

Consider drawing a list S_3 of $2n$ points and then randomly partitioning into lists S_1 and S_2. Clearly, this yields the same probability distribution.

Let's hold off on partitioning S_3.

Note that $|\{S_3 \cap h \mid h \in \mathcal{H}\}| \leq \pi_{\mathcal{H}}(2n)$ (even if $|\mathcal{H}| = \infty$)

So $\text{Prob}(B) \leq \sum_{h' \in \{S_3 \cap h \mid h \in \mathcal{H}\}} \text{Prob}\left(|S_1 \cap h'| = 0 \text{ AND } |S_2 \cap h'| \geq \frac{\varepsilon}{2} n\right)$

$$\leq \pi_{\mathcal{H}}(2n) \cdot \text{Prob}\left(|S_1 \cap h'| = 0 \text{ AND } |S_2 \cap h'| \geq \frac{\varepsilon}{2} n\right) \quad \forall\, h'.$$

So to prove $\text{Prob}(B) \leq \frac{\delta}{2}$, it suffices to show

$$\text{Prob}\left(|S_1 \cap h'| = 0 \text{ AND } |S_2 \cap h'| \geq \frac{\varepsilon}{2} n\right) \leq \frac{\delta}{2\,\pi_{\mathcal{H}}(2n)}.$$

Consider h'. $|h'| \geq \frac{\varepsilon}{2} n$. Then $|S_1 \cap h'| < \frac{\varepsilon}{2} n$.

Prob($|S_1 \cap h'| = 0$ AND ...) $\leq 2\pi_{\mathcal{H}}(2n)$

Case: $|h'| < \frac{\varepsilon}{2} n$. Then $|S_2 \cap h'| < \frac{\varepsilon}{2} n$.

Case: $|h'| \geq \frac{\varepsilon}{2} n$. Then $\underset{\substack{\text{random} \\ \text{partition} \\ \text{of } S_3}}{\text{Prob}}\left(|S_1 \cap h'| = 0\right) \leq \left(\frac{1}{2}\right)^{\frac{\varepsilon n}{2}}$ because each item in h' has a $\frac{1}{2}$ chance of falling in S_1 or S_2.

$$\leq 2^{-\log_2 2\pi_{\mathcal{H}}(2n) + \log_2 \delta} = \frac{\delta}{2\pi_{\mathcal{H}}(2n)}.$$

Thus, $\text{Prob}(B) \leq \frac{\delta}{2} \Rightarrow \text{Prob}(A) \leq \delta$.

Proof technique where we picked S_1 and S_2 2 different ways is known as "double sampling" or the "ghost sample" method. We postpone random choices until later, like we did in the percolation theory proofs.

Consider a target concept c^* such as spam emails and a set of hypotheses $\mathcal{H}$ which are sets of emails we claim are spam.

Let $\mathcal{H}' = \{\underbrace{h \Delta c^*}_{\text{symmetric difference}} \mid h \in \mathcal{H}\}$ be the collection of error regions of hypotheses in $\mathcal{H}$.

Note that $\mathcal{H}$ and $\mathcal{H}'$ have the same VC-dimension and shatter function.

Thm 5.15 (sample bound): For any class $\mathcal{H}$ and distribution D, if a training sample S is drawn from D of size

$$n \geq \frac{2}{\varepsilon}\left[\log\left(2\pi_{\mathcal{H}}(2n)\right) + \log\frac{1}{\delta}\right],$$

then w.p. $\geq 1-\delta$, every $h \in \mathcal{H}$ with true error $\text{err}_D(h) \geq \varepsilon$ has $\text{err}_S(h) > 0$. Equivalently, every $h \in \mathcal{H}$ with training error $\text{err}_S(h) = 0$ has $\text{err}_D(h) < \varepsilon$.

proof. Apply Thm 3.14 to $\mathcal{H}' = \{h \Delta c^* \mid h \in \mathcal{H}\}$.

Thm 5.16 (growth function uniform convergence)

For any class $\mathcal{H}$ and distribution D, if a training sample S is drawn from D of size

$$n \geq \frac{8}{\varepsilon^2}\left[\ln\left(2\pi_{\mathcal{H}}(2n)\right) + \ln\frac{1}{\delta}\right],$$

then w.p. $1-\delta$, every $h \in \mathcal{H}$ will have $\left|err_S(h) - err_D(h)\right| \leq \varepsilon$.

proof. Similar to last couple theorems. See book.

Corollary 5.17 For any class $\mathcal{H}$ and distribution D, a training sample S of size

$$O\left(\frac{1}{\varepsilon}\left[VC(\mathcal{H})\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]\right)$$

is sufficient to ensure w.p. $1-\delta$ that every $h \in \mathcal{H}$ with true error $err_D(h) \geq \varepsilon$ has training error $err_S(h) > 0$.

Equivalently, every $h \in \mathcal{H}$ with $err_S(h) = 0$ has $err_D(h) < \varepsilon$.

VC-dim is one measure of the complexity of a set system, which allows proving generalization guarantees. There are others, such as Shannon entropy and Rademacher complexity (how well a concept class $\mathcal{H}$ can fit random noise).

These types of guarantees give us hope that we can train a ML algorithm on a small sample of data and then make useful predictions elsewhere.