Recap: VC dimension

This 

Recall definitions:

\[ \text{sgn}(U) := \bigg\{ \text{sgn}(u_1), \ldots, \text{sgn}(u_n) \bigg\} : U \subseteq \mathbb{F}^n, \]

\[ \text{Sh}(\mathcal{F}; \text{sgn}) := \left\| \text{sgn}(\mathcal{F}) \right\|, \]

\[ \text{Sh}(\mathcal{F}; n) := \max \sup_{S \subseteq \mathcal{F}} \text{Sh}(\mathcal{F}; \text{sgn}) \quad \text{in worst-case...} \]

\[ \text{VC}(\mathcal{F}) := \max \left\{ \sum_{i \in \mathbb{Z}_0} \text{Sh}(\mathcal{F}; i) = 2^i \right\}. \]

\[ \text{emp} \text{ (Sauer-Shelah)} \text{ set } V \subseteq \text{VC}(\mathcal{F}). \]

\[ \text{Sh}(\mathcal{F}; n) \leq \left( \frac{2^n}{|\mathcal{F}|} \right) \text{ n} \in V \text{ o.w.} \]

\[ \text{Sh}(\mathcal{F}; n) \leq n^\text{th}. \]

Theorem ("VC theorem"): \( \text{w/ pr > 1-\epsilon, any } \mathcal{F} \text{ satisfies } \)

\[ R_2(\mathcal{F}) \leq R_2(\mathcal{F}) + \frac{1}{\epsilon} \text{Rad}(\text{sgn}(\mathcal{F})) + 3 \sqrt{\frac{\ln(\frac{1}{\epsilon})}{2n}} \]

where \( R_2(\mathcal{F}) = \text{Pr}[	ext{Bartlett}(V, \mathcal{F})] \text{ and } \)

\[ \text{Rad}(\text{sgn}(\mathcal{F})) \leq \sqrt{2n \ln(\frac{1}{\epsilon} \text{Sh}(\mathcal{F}; \text{sgn}))}, \]

\[ \ln(\text{Sh}(\mathcal{F}; \text{sgn})) \leq \ln(\text{Sh}(\mathcal{F}; n)) \leq \text{VC}(\mathcal{F})(n+1). \]
**Theorem.** Set $f : \{ x \mapsto \text{sgn}(\langle w, x \rangle - b) : w \in \mathbb{R}^d, b \in \mathbb{R} \}$.

Then $\text{VC}(f) = d+1$.

**Remarks.**

- **Convention to include offset.** Homework will (?) discuss removing/adding.

- By Sauer–Shelah, $\text{Sh}(\mathbb{R}; n) \leq \frac{2^n}{n!} + 1$.

- Anthony–Bartlett chapter 3 gives an exact computation (an equality!) for $\text{Sh}(\mathbb{R}; n)$.

- Let's compare to a Rademacher complexity.

  $\text{Rad}(\text{sgn}(\cdot), \mathbb{R}) \leq \sqrt{\frac{d \ln(2n)}{2n \cdot d \cdot \ln(n+1)}}$.

  $\text{Rad}(\{ x \mapsto \langle w, x \rangle : w \in \mathbb{R}^d \}, \| \cdot \|_2) \leq \frac{\sqrt{d \ln n}}{\| [e_1, \ldots, e_n] \|_2}$.

- VC lower bound will use standard basis vectors, thus $\| [e_1, \ldots, e_n] \|_2 = \sqrt{n}$.

- One is scale-sensitive, other is not.

- One gives rise to standard $l_2$-reg, other does not!

  ($\beta$ for usual sets do we have good complexity measures & regularization?)
Proof.

**Lower Bound.** \( \text{VC}(\mathcal{F}) \geq d + 1 \).

Suffices to show \( \exists S := \{ e_1, \ldots, e_d \} \) with \( \text{Sh}(\mathcal{F}_S) = 2^{d+1} \).

Pick \( S = \{ e_1, \ldots, e_d \} \cup \{ 0, \ldots, 0 \} \).

Given any \( P \subseteq S \), consider predictor \( x \mapsto \text{sgn}(\langle w, x \rangle - b) \) where corresponding parameters \((w, b)\) satisfy:

\[
\begin{align*}
    w &:= 2 \cdot \mathbb{1}[e_i \in P] - 1, \\
    b &:= \frac{1}{2} - \mathbb{1}[0 \in P].
\end{align*}
\]

Then:

\[
\begin{align*}
    \text{sgn}(\langle w, e_i \rangle - b) &= \text{sgn}(2 \cdot \mathbb{1}[e_i \in P] - 1 - b) = 2 \cdot \mathbb{1}[e_i \in P] - 1, \\
    \text{sgn}(\langle w, 0 \rangle - b) &= \text{sgn}(1 - b) = 2 \cdot \mathbb{1}[0 \in P] - 1.
\end{align*}
\]

Hence, \((w, b)\) achieves labelling \( P \).

**Upper bound.** \( \text{VC}(\mathcal{F}) \leq d + 2 \).

Consider any \( S \subseteq \mathbb{R}^d \) with \( |S| = d + 2 \).

By Radon lemma (proved next), \( \exists P, N \).

exists partition of \( S \) into nonempty \( P, N \) with \( \text{conv}(P) \cap \text{conv}(N) \neq \emptyset \).

Therefore, any linear predictor labelling \( P \) as positive

must also label at least one point of \( N \) as positive.

\( S \) has arbitrary \( \Rightarrow \) \( \text{Sh}(\mathcal{F}_{d+2}) < 2^{d+2} \Rightarrow \text{VC}(\mathcal{F}) < d + 2 \).
Theorem ("Radon's lemma").

Given \( S \subseteq \mathbb{R}^{d+2} \) with \( |S| = d+2 \),

exists partition into nonempty subsets \( P, N \) with \( \text{conv}(P) \cap \text{conv}(N) \neq \emptyset \).

Proof:

Let \( S := \{ x_1, \ldots, x_{d+2} \} \) be given, define \( u_i := x_i - x_{d+2} \).

\((u_1, \ldots, u_{d+1}) \) must be linearly dependent.

Existence:

\( \sum_{i=1}^{d+3} a_i x_i = 0 \) \( \forall a_i \neq 0 \) (i.e., \( a_i = 1 \))

\[ \sum_{i=1}^{d+3} a_i x_i = 0 \]

\[ \Rightarrow \quad x_i - x_{d+2} = \sum_{i=d+1}^{d+3} a_i (x_i - x_{d+2}) \]

\[ \Rightarrow \quad 0 = \sum_{i=d+1}^{d+3} a_i x_i \]

Not all \( P_i = 0 \), and \( \sum b_i = 0 \).

Set \( P := \{ x_i : b_i > 0 \} \), \( N := \{ x_i : b_i < 0 \} \).

Note neither is empty.

Set \( b := \sum_{i \in P} b_i > 0 = -\sum_{i \in N} b_i \).

Since \( 0 = \sum b_i x_i = \sum_{i \in P} b_i x_i + \sum_{i \in N} b_i x_i \), then \( \frac{0}{b} = \sum_{i \in P} \frac{b_i}{b} x_i + \sum_{i \in N} \frac{b_i}{b} x_i \)

\[ \Rightarrow \quad \sum_{i \in P} \frac{b_i}{b} x_i = \sum_{i \in N} \frac{b_i}{b} x_i = 0 \]

Thus \( \exists \in \text{conv}(P) \cap \text{conv}(N) \).
Consider iterating previous construction:

A neural network where nodes have activation $\text{sgn}(\cdot)$.

To study this, consider a more elaborate object than simply the output.

**Definition.** Given a sample of size $n$ and an LTF network with $m$ nodes, define a $m \times m$ activation matrix $A := \{a_{ij}\}$, where $a_{ij} = \text{sgn}(c)$ for node $i$ on example $j$.

Let $\text{Act}(\mathcal{F})$ denote the set of all activation matrices achieved with LTF architecture/loss $\mathcal{F}$ on sample $S$.

**Remark.** Since the last column of $A$ is the labelling,

$|\text{Act}(\mathcal{F})| \geq \text{sh}(n, m)$.

- $\text{Act}(\mathcal{F})$ tells us more and seems a nice complexity measure on its own. But we don't get an estimate of it from a single training algorithm run (in contrast with say, Lipschitz constant).

- We can do this for ReLU too! Now $A_{ij} := 1$ if node $j$ is active on example $i$. 
Theorem. Consider LTF networks $\mathcal{F}$ with finite parameters (weight coordinates & biases). Then $\forall n \geq 1$

$$Sh(\mathcal{F}; n) \leq \max_{S, 1 \leq i \leq n} |Act(\mathcal{F}; S)| \leq (n+1)^p,$$

when $p > 12,$ $VC(\mathcal{F}) \leq 6 \text{ ph}(p).$

Proof. Topologically sort nodes, let $p_1, \ldots, p_m$ be the number of parameters per each.

We'll construct partitions $W_1, U_1, \ldots, U_m$ of the weight space (in detail, $U_i$ partitions weight space of nodes $S_i$)

so that activation pattern of nodes $W_1 \cup \cdots \cup U_i$ is fixed within each partition cell.

For convenience define $S_0 = \emptyset, S_1 \in \mathcal{I}.$

By induction it'll be shown that $|S_i| \leq (n+1)^{\frac{i-1}{12}}.$

Since $|Act(\mathcal{F}; S)| \leq \sum_{S_m}$, this gives first bound.

Base case. $|S_0| = 1 = (n+1)^0.$

Inductive step. Fix any cell $C$ in $S_i$; for those choices of weights in layers $S_i$, the activation pattern is fixed. Therefore node $i_l$ receives a fixed input (which are a combination of original input coordinates & intermediate node outputs). By VC-dim of LTF 4.Sauer-Shelah, we can further refine $C$ into $(n+1)^{0.12}$ pieces which activation of nodes is fixed.

Doing this for each $C \in S_i$, we have $|S_i| \leq 15 \cdot 15 \cdot \ldots \cdot 15 \leq (n+1)^{0.12}.$
Proof continued (of LTF net VC dim).

It remains to prove the VC dimension from $\mathcal{S}(\mathcal{F}, n)$.

Note

\[ VC(\mathcal{F}) < n \]

\[ \iff \forall i \leq n \quad \mathcal{S}(\mathcal{F}; i) < 2^i \]

\[ \iff \forall i \leq n \quad (i+1) p < 2^i \]

\[ \iff \forall i \leq n \quad (i+1) p \ln(i+1) < i \ln(2) \]

\[ \iff \forall i \leq n \quad p < \frac{i \ln(2)}{\ln(i+1)} \]

\[ \iff p < \frac{n \ln(2)}{\ln(n+1)} \]

If $n = 6 \rho \ln(\rho)$,

\[ \frac{n \ln(2)}{\ln(n+1)} = \frac{n \ln(2)}{\ln(2n)} = \frac{6 \rho \ln(\rho) \ln(2)}{\ln(2n) + \ln(\rho) + \ln(\rho)} \]

\[ \geq \frac{6 \rho \ln(\rho) \ln(2)}{3 \ln(\rho)} > p \]

Remark: Had to do "$\forall i \leq n$"!

VC dim is really a "sup"; a weird function class which is 0 for even-sized data and 1 for even-sized odd sizes still has infinite VC.
Remark.

* lower bound \( \Omega(p \ln (\frac{\#K}{m}) \)

(Anthony - Artelt ch6; no proof).

But these lower bounds are specific network architectures.

\[ \Rightarrow \text{really nots interpolation hard!} \]

* Other bounds

<table>
<thead>
<tr>
<th>arbitrary concave concave</th>
<th>VC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RELU</td>
<td>( \tilde{O}(\Psi \ln(\Psi/c)) )</td>
</tr>
<tr>
<td>sigmoid</td>
<td>( \tilde{O}(p^2 m^2) ) (also ( \Omega(pL) ))</td>
</tr>
</tbody>
</table>

But lower bounds, again, are deceiving fixed architecture!

Also: bit complex thus a work in!