Lecture 14: VC Dimension, Validation, Neural Networks, Kernels (workshop)

Nisha Chandramoorthy

October 5, 2023

VC dimension

- A hypothesis class restricted to some C *shatters* C if any binary function on C is in the class

VC dimension

- A hypothesis class restricted to some C *shatters* C if any binary function on C is in the class
- VC dimension, $\mathrm{VCdim}(\mathcal{H})$: maximal size of $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.

VC dimension

- A hypothesis class restricted to some C *shatters* C if any binary function on C is in the class
- VC dimension, $\mathrm{VCdim}(\mathcal{H})$: maximal size of $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.
- Eg 1. VCdim of threshold functions on $\mathbb{R}$ is 1.

VC dimension

- A hypothesis class restricted to some C *shatters* C if any binary function on C is in the class
- VC dimension, $\mathrm{VCdim}(\mathcal{H})$: maximal size of $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.
- Eg 1. VCdim of threshold functions on $\mathbb{R}$ is 1.
- Eg 2: VCdim of indicator functions on intervals of $\mathbb{R}$ is 2.

VC dimension

- A hypothesis class restricted to some C *shatters* C if any binary function on C is in the class
- VC dimension, $\mathrm{VCdim}(\mathcal{H})$: maximal size of $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.
- Eg 1. VCdim of threshold functions on $\mathbb{R}$ is 1.
- Eg 2: VCdim of indicator functions on intervals of $\mathbb{R}$ is 2.
- Eg 3: VCdim of a finite class $\mathcal{H} \leqslant \log_2 |\mathcal{H}|$

Generalization bounds based on VC dimension

- $\mathcal{H} = \{h_\theta(x) = \sin(\theta x) : \theta \in \mathbb{R}\}$.

Generalization bounds based on VC dimension

- $\mathcal{H} = \{h_\theta(x) = \sin(\theta x) : \theta \in \mathbb{R}\}$.
- VCdim is ∞

Generalization bounds based on VC dimension

- $\mathcal{H} = \{h_\theta(x) = \sin(\theta x) : \theta \in \mathbb{R}\}$.
- VCdim is ∞
- Binary classification generalization for 0-1 loss over class $\mathcal{H}$ with VCdim = d: there exist constants $C_1, C_2 > 0$ such that

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leqslant m_{\mathcal{H}}(\epsilon, \delta) \leqslant C_2 \frac{d + \log(1/\delta)}{\epsilon}$$

- Consistent with bound from lecture 2.

- Consistent with bound from lecture 2.
- $m_{\mathcal{H}} = \dfrac{\log(|\mathcal{H}|/\delta)}{\epsilon}$
- $\mathcal{D}^m(S : R(h_S) \geqslant \epsilon) \leqslant |\mathcal{H}_b| e^{-\epsilon m} \leqslant |\mathcal{H}| e^{-\epsilon m}$.

- Consistent with bound from lecture 2.
- $m_{\mathcal{H}} = \dfrac{\log(|\mathcal{H}|/\delta)}{\epsilon}$
- $\mathcal{D}^m(S : R(h_S) \geqslant \epsilon) \leqslant |\mathcal{H}_b| e^{-\epsilon m} \leqslant |\mathcal{H}| e^{-\epsilon m}$.

VC dimension contd

Cross Validation

- Recall Hoeffding's inequality: $X_1, \cdots, X_m$ iid sequence with $P(a \leqslant X \leqslant b) = 1$. Then, with probability $\geqslant 1 - \delta$,

$$|\frac{1}{m}\sum_{i=1}^{m} X_i - EX| < (b-a)\sqrt{\frac{\log(2/\delta)}{2m}}$$

Cross Validation

- Recall Hoeffding's inequality: $X_1, \cdots, X_m$ iid sequence with $P(a \leqslant X \leqslant b) = 1$. Then, with probability $\geqslant 1 - \delta$,

$$|\frac{1}{m}\sum_{i=1}^{m} X_i - EX| < (b-a)\sqrt{\frac{\log(2/\delta)}{2m}}$$

- For any set V of size m, when loss $\in (0,1)$, by Hoeffding's inequality,

$$|R_V(h) - R(h)| \leqslant \sqrt{\frac{\log(2/\delta)}{2m}}$$

- k-fold cross validation to choose models (e.g., regularization parameters):
  - Divide given set S into k subsets (folds)
  - For each parameter, each fold: run learning algorithm on union of all folds except one; calculate test/validation loss on fold
  - Run algorithm on S using parameter with minimum total test/validation loss.

Models/representations, algorithms, statistical principles

- how to make and test conjectures about how large language models (LLMs) learn?

Models/representations, algorithms, statistical principles

- how to make and test conjectures about how large language models (LLMs) learn?
- "how to train them better (more efficiently)" – number of practical questions perhaps benefit from theory

Models/representations, algorithms, statistical principles

- how to make and test conjectures about how large language models (LLMs) learn?
- “how to train them better (more efficiently)” – number of practical questions perhaps benefit from theory
- AI safety, fair and ethical ethical use, combining with other domain knowledge (e.g., physics, chemistry etc).... and many more!

Models/representations, algorithms, statistical principles

- how to make and test conjectures about how large language models (LLMs) learn?
- "how to train them better (more efficiently)" – number of practical questions perhaps benefit from theory
- AI safety, fair and ethical ethical use, combining with other domain knowledge (e.g., physics, chemistry etc).... and many more!
- Perhaps biggest contribution advance to LLMs: transformers and their training.

(partial) History - trace back from transformers (source:Wikipedia)

- Transformer architecture: 2017, Google Brain [Vaswani et al]
- Deep learning, unsupervised learning 2010s (e.g., GANs 2014)...
- ImageNet: 2009, Fei Fei Li
- Long-short term memory (LSTM) architecture: 1997, [Hochreiter and Schmidhuber]
- Convolutional NNs: (inspired from) 1979 work by [Fukushima]; Recurrent neural networks: 1982 [Hopfield]
- ...
- Automatic Differentiation: 1970 [Linnainmaa]
- ...
- First neural networks: 1950s [Minsky and others]

Fully connected Neural Networks

- Neuron: input $\sum_j w_j h_j$; output $\sigma(\sum_j w_j h_j)$
- Organized into layers of depth l and width n
- Graph: V, E, σ, w; weight function.

Fully connected Neural Networks

- Neuron: input $\sum_j w_j h_j$; output $\sigma(\sum_j w_j h_j)$
- Organized into layers of depth l and width n
- Graph: V, E, σ, w; weight function.
- VC dim of $\mathcal{H}_{V,E,\mathrm{sign}} \leqslant C|E| \log |E|$

Fully connected Neural Networks

- Neuron: input $\sum_j w_j h_j$; output $\sigma(\sum_j w_j h_j)$
- Organized into layers of depth l and width n
- Graph: V, E, σ, w; weight function.
- VC dim of $\mathcal{H}_{V,E,\mathrm{sign}} \leqslant C|E|\log|E|$
  - Proof: Growth function $\tau_{\mathcal{H}}(m) = \max_{C,|C|=m} |\mathcal{H}|_C|$

Fully connected Neural Networks

- Neuron: input $\sum_j w_j h_j$; output $\sigma(\sum_j w_j h_j)$
- Organized into layers of depth l and width n
- Graph: V, E, σ, w; weight function.
- VC dim of $\mathcal{H}_{V,E,\mathrm{sign}} \leqslant C|E|\log|E|$
  - Proof: Growth function $\tau_{\mathcal{H}}(m) = \max_{C,|C|=m} |\mathcal{H}|_C|$ Fact 1: Let $\mathcal{H} = \mathcal{H}_l \circ \cdots \circ \mathcal{H}_1$. Then, $\tau_{\mathcal{H}}(m) \leqslant \prod_{t=1}^{l} \tau_{\mathcal{H}_t}(m)$.

Fully connected Neural Networks

- Neuron: input $\sum_j w_j h_j$; output $\sigma(\sum_j w_j h_j)$
- Organized into layers of depth l and width n
- Graph: V, E, σ, w; weight function.
- VC dim of $\mathcal{H}_{V,E,\mathrm{sign}} \leqslant C|E|\log|E|$
  - Proof: Growth function $\tau_{\mathcal{H}}(m) = \max_{C,|C|=m} |\mathcal{H}|_C|$ Fact 1: Let $\mathcal{H} = \mathcal{H}_l \circ \cdots \circ \mathcal{H}_1$. Then, $\tau_{\mathcal{H}}(m) \leqslant \prod_{t=1}^{l} \tau_{\mathcal{H}_t}(m)$.
  - Fact 2: Let $\mathcal{H} = \mathcal{H}^{(1)} \cdots \circ \mathcal{H}^{(n)}$. Then, $\tau_{\mathcal{H}}(m) \leqslant \prod_{t=1}^{l} \tau_{\mathcal{H}^{(t)}}(m)$.

Fully connected Neural Networks

- Neuron: input $\sum_j w_j h_j$; output $\sigma(\sum_j w_j h_j)$
- Organized into layers of depth l and width n
- Graph: V, E, σ, w; weight function.
- VC dim of $\mathcal{H}_{V,E,\mathrm{sign}} \leqslant C|E| \log |E|$
  - Proof: Growth function $\tau_{\mathcal{H}}(m) = \max_{C,|C|=m} |\mathcal{H}|_C|$ Fact 1: Let $\mathcal{H} = \mathcal{H}_l \circ \cdots \circ \mathcal{H}_1$. Then, $\tau_{\mathcal{H}}(m) \leqslant \prod_{t=1}^{l} \tau_{\mathcal{H}_t}(m)$.
  - Fact 2: Let $\mathcal{H} = \mathcal{H}^{(1)} \cdots \circ \mathcal{H}^{(n)}$. Then, $\tau_{\mathcal{H}}(m) \leqslant \prod_{t=1}^{l} \tau_{\mathcal{H}^{(t)}}(m)$.
  - Fact 3: Sauer's Lemma: $\tau_{\mathcal{H}}(m) = (em/d)^d$, where $d \geqslant$ VCdim $(\mathcal{H})$

<u>Restriction</u>:

$h \in \mathcal{H}$

C

$h|_C(x_i) = h(x_i) \qquad x_i \in C$

$\mathcal{H}|_C = \{ h|_C : h \in \mathcal{H} \}$

Restriction of $\mathcal{H}$ to a set C.

→ $C = \{x_1, x_2, x_3\}$

$h|_C$ should take values

$\mathcal{H}$

$$\left\{ \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \dots, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

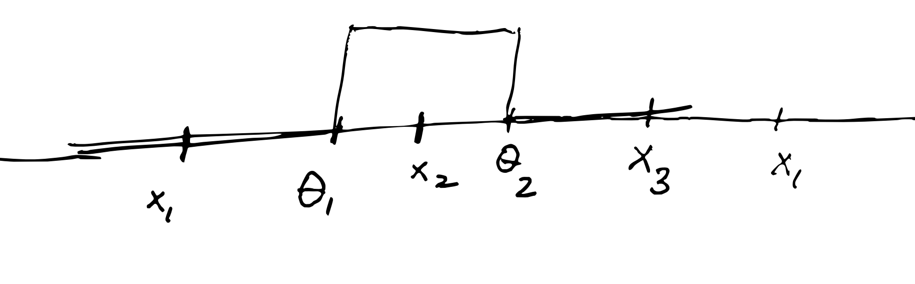
$|\mathcal{H}|_C|$ should have at least 8 elements if $\mathcal{H}$ can shatter C.

→ If $VC\dim(\mathcal{H}) = m$

- $\exists$ some set $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$ and $|C| = m$
- $\nexists$ a set $C \subseteq \mathcal{X}$ of size $|C| = m+1$ that is shattered by $\mathcal{H}$.

→ $VC\dim(\mathcal{H}) = 1$

$\mathcal{H} = \{ x \to h_\theta(x) : \theta \in \mathbb{R}$, $h_\theta(x) = 1$ if $x < \theta$ and $h_\theta(x) = 0$ o.w.$\}$



- $VC\dim(\mathcal{H}) \geq 1$
- $VC\dim(\mathcal{H}) < 2$

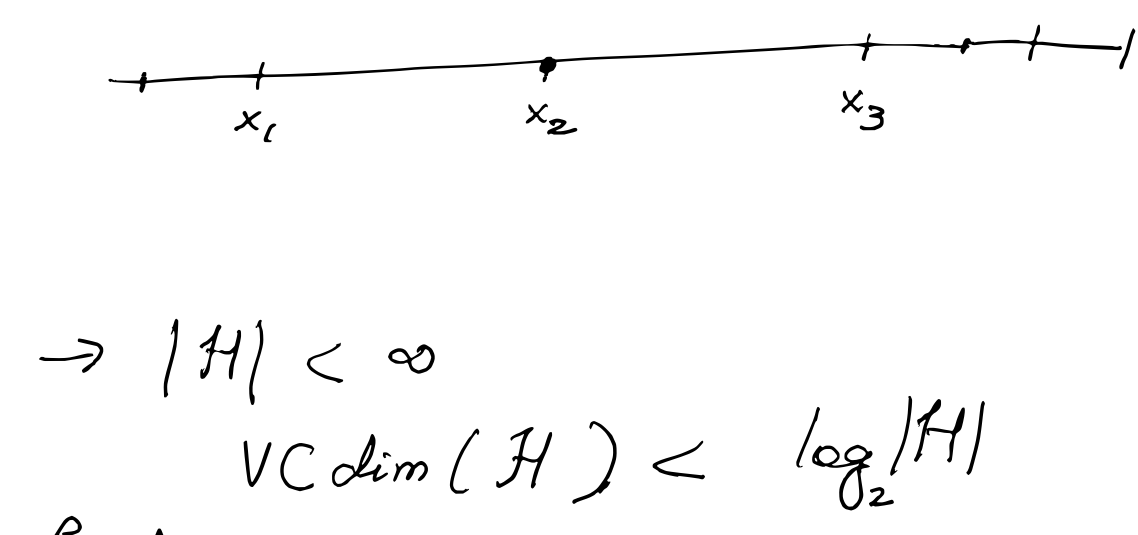
$C = \{x_1, x_2\} \qquad x_2 > x_1$ WLOG

$h|_C(x_1) = 0$ and $h|_C(x_2) = 1$

Not possible if $x_2 > x_1$.

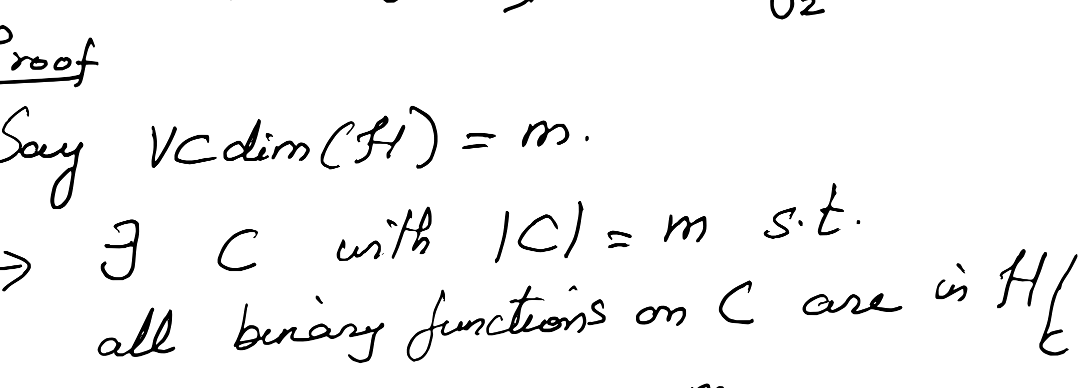
→ $VC\dim(\mathcal{H}) = 2$

$$H = \{ h_{\theta_1,\theta_2}(x) = \mathbb{1}_{(\theta_1,\theta_2)}(x) : \theta_1 < \theta_2,\ \theta_1, \theta_2 \in \mathbb{R} \}$$



0, 1, 0

1, 0, 1 → Not possible



→ $|H| < \infty$

$$VC\dim(\mathcal{H}) < \log_2 |\mathcal{H}|$$

<u>Proof</u>

Say $VC\dim(\mathcal{H}) = m$.

→ $\exists\ C$ with $|C| = m$ s.t. all binary functions on C are in $\mathcal{H}|_C$.

$$|\mathcal{H}| \geq 2^m$$

$$m \leq \log_2 |\mathcal{H}|$$

→ $VC\dim(\mathcal{H}) = d$ (parameter dim)

$\mathcal{H}$ ← halfspaces

- $\mathcal{H} = \{ x \to \langle w, x \rangle : w \in \mathbb{R}^d, x \in \mathbb{R}^d \}$
- Suppose $VC\dim(\mathcal{H}) = d$

$\exists\ x_1, x_2, \dots, x_d$ is shattered by $\mathcal{H}$.

and

$\nexists\ x_1, x_2, \dots, x_{d+1}$ that are shattered by $\mathcal{H}$.

- Assume that $\exists\ x_1, x_2, \dots, x_{d+1}$ that is shattered by $\mathcal{H}$. → show this is not possible.

- $$\begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \\ x_{d1} & \cdots & x_{dd} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \underbrace{\begin{bmatrix} \cdot \\ \vdots \\ \cdot \end{bmatrix}}_{h|_C} \in \mathbb{R}^d$$

if $x_1, \dots, x_d$ are linearly independent, yes! $\Rightarrow VC\dim(\mathcal{H}) \geq d$.

- $$\begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \\ x_{d+1\,1} & & x_{(d+1)d} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$x_{d+1} = \sum_{i=1}^{d} a_i x_i \qquad (\because x_i \in \mathbb{R}^d)$$

$$x_{d+1} = a_d x_d$$

$$\langle x_{d+1}, w \rangle = a_d \langle x_d, w \rangle$$

$$\text{sgn}(a_d) < 0$$

- $x_{d+1} = \sum_{i=1}^{d} a_i x_i$

$\exists$ some $a_i \in \mathbb{R}^{d+1}$

$$\sum_{i=1}^{d+1} a_i x_i = 0 \quad \to \text{(A)}$$

$P = \{ i : a_i > 0 \} \qquad N = \{ i : a_i < 0 \}$

$$\sum_{i \in P} a_i x_i = \sum_{i \in N} |a_i| x_i \quad \text{(A)}$$

if $x_1, x_2, \dots, x_{d+1}$ is shattered by $\mathcal{H}$, $\exists\ w$ s.t. for any $i \in P$, $\langle x_i, w \rangle > 0$ and $i \in N$, $\langle x_i, w \rangle < 0$ ✓

$$0 < \langle \sum_{i \in P} a_i x_i, w \rangle = \langle \sum_{i \in N} |a_i| x_i, w \rangle$$

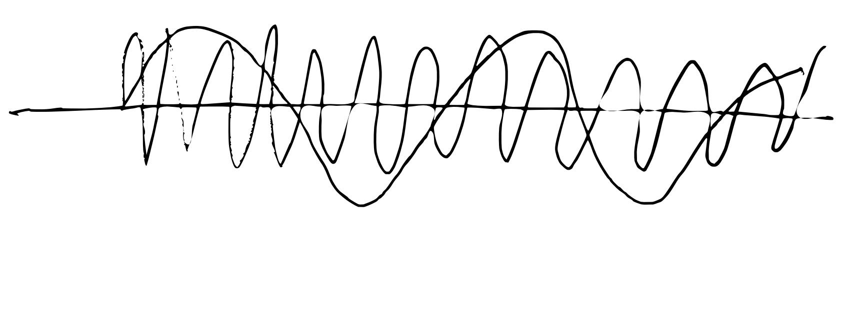
$$\langle \sum_{i \in P} a_i x_i, w \rangle = \sum_{i \in P} a_i \langle x_i, w \rangle > 0$$

$$= \sum_{i \in N} |a_i| \langle x_i, w \rangle$$

$$< 0$$

Contradiction!

→ $\mathcal{H} = \{ x \to \sin\theta x : \theta \in \mathbb{R} \}$



→

with $\mathbb{P}_r$ (over $S \sim D^m$) $> 1-\delta$,

$$R(h) < R_S(h) + \frac{\mathbb{C}(\mathcal{H})}{m} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

Sample Complexity: $m_{\mathcal{H}}(\varepsilon, \delta)$
as the minimal number of samples s.t. $|R(h) - R_S(h)| < \varepsilon$
with $P_r = 1-\delta$.

$$m_{\mathcal{H}}(\varepsilon, \delta) = O\left(\frac{\log|\mathcal{H}| + \log(1/\delta)}{\varepsilon}\right)$$

Recap of Proof from Lec 2

$\rightarrow$ $\mathcal{H}_b = \{ h \in \mathcal{H} : R(h) > \varepsilon \}$

$\rightarrow$ Realizability assumption

$$\min_{h \in \mathcal{H}} R_S(h) = 0$$

h_S is ERM for $\mathcal{H}, S$.

$$\mathbb{P}r\left(R_S(h_S) = 0 \text{ and } R(h_S) > \varepsilon\right)$$

$$\leq (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

$\rightarrow$ Next time:

$$\text{VC dim}(\text{FCNN with } |E|) \leq C|E|\log|E|$$