Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d - 1$ unknowns.

Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d - 1$ unknowns.
- In a 64-bit floating point representation, size of hypothesis class: $|\mathcal{H}| = 2^{64d}$ 2^{64}

Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d-1$ unknowns.
- In a 64-bit floating point representation, size of hypothesis class: $|\mathcal{H}| = 2^{64d}$
- How many samples do we need to guarantee $\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\gamma$ to hold with probability at least $1-\delta$?

$k = O(2^d)$

$$m \geq O\left(\frac{1}{\gamma^2}\log\frac{2^{64d}}{\delta}\right) = O\left(\frac{d}{\gamma^2}\log\frac{1}{\delta}\right) = O_{\gamma,\delta}(d)$$

Infinite hypothesis class: Challenges

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d - 1$ unknowns.
- In a 64-bit floating point representation, size of hypothesis class: $|\mathcal{H}| = 2^{64d}$
- How many samples do we need to guarantee $\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\gamma$ to hold with probability at least $1 - \delta$?

$$m \geq O\left(\frac{1}{\gamma^2} \log \frac{2^{64d}}{\delta}\right) = O\left(\frac{d}{\gamma^2} \log \frac{1}{\delta}\right) = O_{\gamma,\delta}(d)$$

Infinite hypothesis class: Challenges

$f(x),\ g(x)$

$f(x) = O(g(x))$

$\exists\, c \in \mathbb{R},\ |f(x)| \le c \cdot g(x)$

$f(x) = x^2 + 4x - 5$

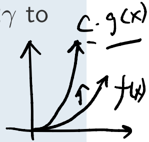
$\hookrightarrow O(x^2)$.

Can we apply the same theorem to infinite $\mathcal{H}$?

Example

- Suppose $\mathcal{H}$ is parameterized by d real numbers. e.g. $\theta = [\theta_1, \theta_2, \ldots, \theta_d] \in \mathbb{R}^d$ in linear regression with $d - 1$ unknowns.
- In a 64-bit floating point representation, size of hypothesis class: $|\mathcal{H}| = 2^{64d}$
- How many samples do we need to guarantee $\epsilon(\hat{h}) \le \epsilon(h^*) + 2\gamma$ to hold with probability at least $1 - \delta$?

$$m \ge O\left(\frac{1}{\gamma^2} \log \frac{2^{64d}}{\delta}\right) = O\left(\frac{d}{\gamma^2} \log \frac{1}{\delta}\right) = \boxed{O_{\gamma,\delta}(d)}$$



To learn **well**, *the number of samples has to be linear in d*

Infinite hypothesis class: Challenges

Size of $\mathcal{H}$ depends on the choice of parameterization

Example

$\underline{2n+2}$ parameters:

$$h_{u,v} = \mathbf{1}\{(\underline{u_0^2 - v_0^2}) + (\underline{u_1^2 - v_1^2})x_1 + \ldots + (\underline{u_n^2 - v_n^2})x_n \geq 0\}$$

is equivalent the hypothesis with $n+1$ parameters:

$$h_\theta(x) = \mathbf{1}\{\underbrace{\theta_0 + \boxed{\theta_1} x_1 + \ldots + \theta_n x_n}_{} \geq 0\}$$

n+1 parameters.

$\underbrace{u_i \quad v_i}$

Infinite hypothesis class: Challenges

Size of $\mathcal{H}$ depends on the choice of parameterization

Example

$2n+2$ parameters:

$$h_{u,v} = \mathbf{1}\{(u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + \ldots + (u_n^2 - v_n^2)x_n \geq 0\}$$

is equivalent the hypothesis with $n+1$ parameters:

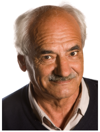
$$h_\theta(x) = \mathbf{1}\{\theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n \geq 0\}$$

We need a complexity measure of a hypothesis class invariant to parameterization choice

Infinite hypothesis class: Vapnik-Chervonenkis theory

VC - dimension

A computational learning theory developed during 1960-1990 explaining the learning process from a statistical point of view.



Alexey Chervonenkis (1938-2014), Russian mathematician

Vladimir Vapnik (Facebook AI Research, Vencore Labs)
Most known for his contribution in statistical learning theory

Shattering a point set

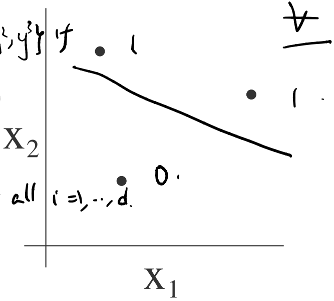
- Given d points $x^{(i)} \in \mathcal{X}$, $i = 1, \ldots, d$, $\mathcal{H}$ **shatters** S if $\mathcal{H}$ can realize any labeling on S.

$\forall$ labeling ℓ on S.
$\exists\ h \in H$, h realizes ℓ

Example: $S = \{x^{(1)}, x^{(2)}, x^{(3)}\}$ where $x^{(i)} \in \mathbb{R}^2$.

$y^i \in \{0,1\}$

H realize a labeling $\{y^1, y^2, y^3\}$ if
$\exists\ h \in H$, h perfectly predicts y^1, y^2, y^3
$h(x^i) = y^i$ for all $i = 1, \ldots, d$.



2^3

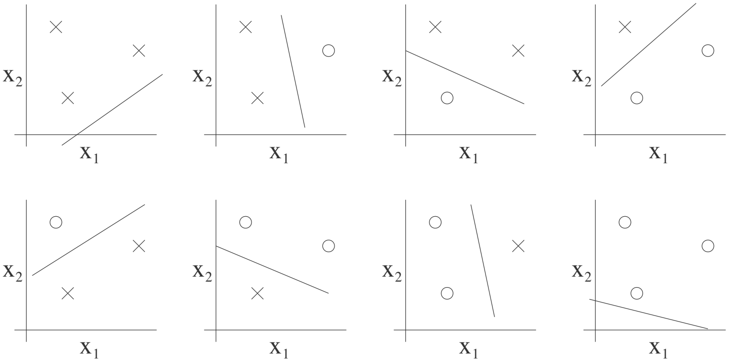
Suppose $y^{(i)} \in \{0, 1\}$, how many possible labelings does S have?

Shattering a point set

- Example: Let $\mathcal{H}_{LTF,2}$ be the linear threshold function in $\mathbb{R}^2$ (e.g. in the perceptron algorithm)

$\mathcal{H}_{LTF,2} \leftarrow$ linear threshold function

$$h(x) = \begin{cases} 1 & w_1x_1 + w_2x_2 \geq b \\ 0 & \text{otherwise} \end{cases}$$



$$\mathcal{H}_{LTF,2} \boxed{\text{shatters}} S = \{x^{(1)}, x^{(2)}, x^{(3)}\}$$

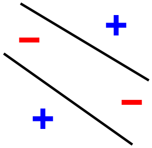
VC Dimension

The **Vapnik-Chervonenkis** dimension of $\mathcal{H}$, or $VC(\mathcal{H})$, is the cardinality of the largest set shattered by $\mathcal{H}$.

- Example: $VC(H_{LTF,2}) = 3$

$VC(H_{LTF,2}) \geq 3.$

$VC(H_{LTF,2}) < 4.$

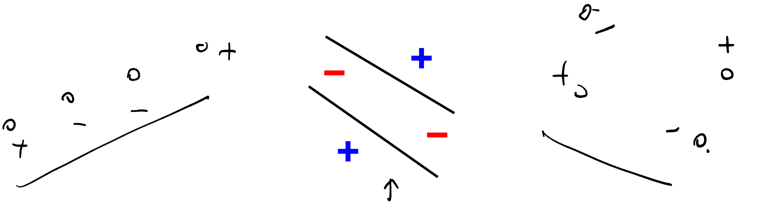


$\mathcal{H}_{LTF}$ can not shatter 4 points: for any 4 points, label points on the diagonal as '+'. (See Radon's theorem)

VC Dimension

The **Vapnik-Chervonenkis** dimension of $\mathcal{H}$, or $VC(\mathcal{H})$, is the cardinality of the largest set shattered by $\mathcal{H}$.

- Example: $VC(H_{LTF,2}) = 3$



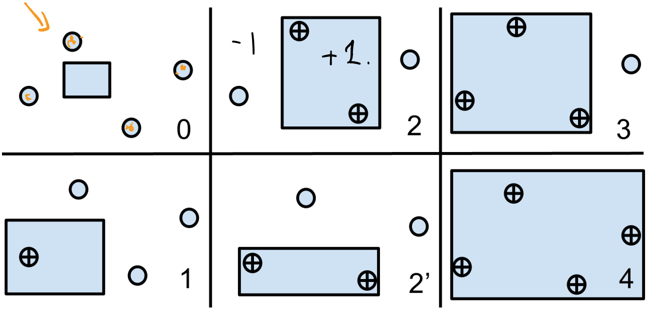
$\mathcal{H}_{LTF}$ can not shatter 4 points: for any 4 points, label points on the diagonal as '+'. (See Radon's theorem)

S

- To show $VC(\mathcal{H}) \geq d$, it's sufficient to find **one** set of d points shattered by $\mathcal{H}$
- To show $VC(\mathcal{H}) < d$, need to prove $\mathcal{H}$ doesn't shatter any set of d points

VC Dimension

- Example: $VC(AxisAlignedRectangles) = 4$



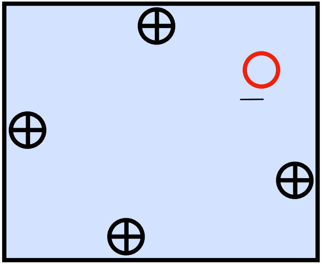
Axis-aligned rectangles can shatter 4 points. $VC(AxisAlignedRectangles) \boxed{\geq 4}$ ✓

∃ set S. st. AAR shatters S.

VC(AAR) < 5.

VC Dimension

- Example: $VC(AxisAlignedRectangles) = 4$



For any 5 points, label topmost, bottommost, leftmost and rightmost points as "+".
$VC(AxisAlignedRectangles) < 5$

VC(AAR) = 4.

Discussion on VC Dimension

More VC results of common $\mathcal{H}$:

- $VC(ConstantFunctions) = 0$.

$\boxed{h(x) = 0.}$

H

$\boxed{h(x) = c}$

$VC(H) < 0$.

for all set S, H doesn't shatter S.

$\exists$ some labeling that H can't realize.

Discussion on VC Dimension

VC(H) < 0. for any S, H can't shatter it.

More VC results of common $\mathcal{H}$: h(x)=1.

- $VC(ConstantFunctions) = 0$ H={h(x)=0}.
- $VC(PositiveHalf\text{-}Lines) = 1, \mathcal{X} = \mathbb{R}$

$$h(x) = \begin{cases} 1 & x \geq c \\ 0 & x < c \end{cases}$$

VC(H) ≥ 1.

VC(H) < 2.

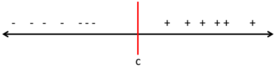
- $VC(Intervals) = 2, \mathcal{X} = \mathbb{R}$
- $VC(LTF \text{ in } \mathbb{R}^n) = n + 1, \mathcal{X} = \mathbb{R}^n$ ← *prove this at home!*

Discussion on VC Dimension

$\mathcal{H}=$

More VC results of common $\mathcal{H}$:

- $VC(ConstantFunctions) = 0$
- $VC(PositiveHalf\text{-}Lines) = 1, \mathcal{X} = \mathbb{R}$



- $VC(Intervals) = 2, \mathcal{X} = \mathbb{R}$
- $VC(LTF \text{ in } \mathbb{R}^n) = n+1, \mathcal{X} = \mathbb{R}^n$ ← *prove this at home!*

Proposition 2

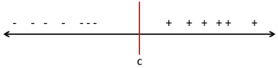
If $\mathcal{H}$ is finite, VC dimension is related to the cardinality of $\mathcal{H}$:

$$VC(\mathcal{H}) \leq \log|\mathcal{H}|$$

Discussion on VC Dimension

More VC results of common $\mathcal{H}$:

- $VC(ConstantFunctions) = 0$
- $VC(PositiveHalf\text{-}Lines) = 1, \mathcal{X} = \mathbb{R}$



- $VC(Intervals) = 2, \mathcal{X} = \mathbb{R}$
- $VC(LTF\ in\ \mathbb{R}^n) = n + 1, \mathcal{X} = \mathbb{R}^n$ ← *prove this at home!*

Proposition 2

If $\mathcal{H}$ is finite, VC dimension is related to the cardinality of $\mathcal{H}$:

$$VC(\mathcal{H}) \leq log|\mathcal{H}|$$

Proof. Let $d = VC|\mathcal{H}|$. There must exists a shattered set of size d on which H realizes all possible labelings. Every labeling must have a corresponding hypothesis, then $|\mathcal{H}| \geq 2^d$ □

$\log|H| \geq d = VC|H|.$

Learning bound for infinite $\mathcal{H}$

Theorem 6

Given $\mathcal{H}$, let $d = VC(\mathcal{H})$.

- *With probability at least $1-\delta$, we have that for all h*

$$|\epsilon(h) - \hat{\epsilon}(h)| \leq O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

Learning bound for infinite $\mathcal{H}$

Theorem 6

Given $\mathcal{H}$, let $d = VC(\mathcal{H})$.

- *With probability at least $1-\delta$, we have that for all h*

$$|\epsilon(h) - \hat{\epsilon}(h)| \leq O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

- *Thus, with probability at least $1-\delta$, we also have*

ERM

$$\underline{\epsilon(\hat{h})} \leq \underline{\epsilon(h^*)} + \underbrace{O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)}$$

Learning bound for infinite $\mathcal{H}$

Corollary 7

For $|\epsilon(h) - \hat{\epsilon}(h)| \leq \gamma$ *to hold for all* $h \in \mathcal{H}$ *with probability at least* $1 - \delta$*, it suffices that* $m = O_{\gamma,\delta}(d)$.

Learning bound for infinite $\mathcal{H}$

Corollary 7

For $|\epsilon(h) - \hat{\epsilon}(h)| \leq \gamma$ to hold for all $h \in \mathcal{H}$ with probability at least $1 - \delta$, it suffices that $m = O_{y,\delta}(d)$.

Remarks

- Sample complexity using $\mathcal{H}$ is linear in $VC(\mathcal{H})$
- For "most"[a] hypothesis classes, the VC dimension is linear in terms of parameters
- For algorithms minimizing training error, # training examples needed is roughly linear in number of parameters in $\mathcal{H}$.

[a] Not always true for deep neural networks

VC Dimension of Deep Neural Networks

Theorem 8 (Cover, 1968; Baum and Haussler, 1989)

Let $\mathcal{N}$ be an arbitrary feedforward neural net with w weights that consists of linear threshold activations, then $VC(\mathcal{N}) = O(w \log w)$.

ReLu.

VC Dimension of Deep Neural Networks

Theorem 8 (Cover, 1968; Baum and Haussler, 1989)

Let $\mathcal{N}$ be an arbitrary feedforward neural net with w weights that consists of linear threshold activations, then $VC(\mathcal{N}) = O(w \log w)$.

Recent progress

- For feed-forward neural networks with piecewise-linear activation functions (e.g. ReLU), let w be the number of parameters and l be the number of layers, $VC(\mathcal{N}) = O(wl \log(w))$ [Bartlett et. al., 2017]

For a N with w parameters, the larger the l,
the higher VC(N)

Bartlett and W. Maass (2003) Vapnik-Chervonenkis Dimension of Neural Nets
Bartlett et. al., (2017) Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks.

VC Dimension of Deep Neural Networks

Theorem 8 (Cover, 1968; Baum and Haussler, 1989)

Let $\mathcal{N}$ be an arbitrary feedforward neural net with w weights that consists of linear threshold activations, then $VC(\mathcal{N}) = O(w \log w)$.

Recent progress

- For feed-forward neural networks with piecewise-linear activation functions (e.g. ReLU), let w be the number of parameters and l be the number of layers, $VC(\mathcal{N}) = O(wl \log(w))$ [Bartlett et. al., 2017]
- *Among all networks with the same size (number of weights), more layers have larger VC dimension* , thus more training samples are needed to learn a deeper network

Bartlett and W. Maass (2003) Vapnik-Chervonenkis Dimension of Neural Nets
Bartlett et. al., (2017) Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks.