Lecture 7: The VC Dimension

Hsuan-Tien Lin (林軒田)
hlin@csie.ntu.edu.tw

Department of Computer Science & Information Engineering
National Taiwan University (國立台灣大學資訊工程系)
Roadmap

1. **When Can Machines Learn?**

2. **Why Can Machines Learn?**

   **Lecture 6: Theory of Generalization**
   
   \[ E_{out} \approx E_{in} \text{ possible if } m_{\mathcal{H}}(N) \text{ breaks somewhere and } N \text{ large enough} \]

3. **How Can Machines Learn?**

4. **How Can Machines Learn Better?**

   **Lecture 7: The VC Dimension**
   
   - Definition of VC Dimension
   - VC Dimension of Perceptrons
   - Physical Intuition of VC Dimension
   - Interpreting VC Dimension
Recap: More on Growth Function

\[ m_H(N) \text{ of break point } k \leq B(N, k) = \sum_{i=0}^{k-1} \binom{N}{i} \leq N^{k-1} \]

The highest term is \( N^{k-1} \).
Recap: More on Growth Function

\( m_H(N) \) of break point \( k \) \( \leq B(N, k) = \sum_{i=0}^{k-1} \binom{N}{i} \)

highest term \( N^{k-1} \)

<table>
<thead>
<tr>
<th>( N )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>5</td>
<td>11</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>6</td>
<td>16</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>7</td>
<td>22</td>
<td>42</td>
<td>57</td>
</tr>
</tbody>
</table>
**Recap: More on Growth Function**

The VC Dimension

Definition of VC Dimension

$m_{\mathcal{H}}(N)$ of break point $k$ \leq B(N, k) = \sum_{i=0}^{k-1} \binom{N}{i}$

highest term $N^{k-1}$

<table>
<thead>
<tr>
<th>$B(N, k)$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>5</td>
<td>11</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>6</td>
<td>16</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>7</td>
<td>22</td>
<td>42</td>
<td>57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N^{k-1}$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>27</td>
<td>81</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>16</td>
<td>64</td>
<td>256</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>25</td>
<td>125</td>
<td>625</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
<td>36</td>
<td>216</td>
<td>1296</td>
</tr>
</tbody>
</table>
The VC Dimension

Definition of VC Dimension

Recap: More on Growth Function

\[ m_{\mathcal{H}}(N) \text{ of break point } k \leq B(N, k) = \sum_{i=0}^{k-1} \binom{N}{i} \]

highest term \( N^{k-1} \)

---

### Table 1: Growth Function Values

<table>
<thead>
<tr>
<th>( N )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>5</td>
<td>11</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>6</td>
<td>16</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>7</td>
<td>22</td>
<td>42</td>
<td>57</td>
</tr>
</tbody>
</table>

### Table 2: Power of Growth Function

<table>
<thead>
<tr>
<th>( N^{k-1} )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>27</td>
<td>81</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>16</td>
<td>64</td>
<td>256</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>25</td>
<td>125</td>
<td>625</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
<td>36</td>
<td>216</td>
<td>1296</td>
</tr>
</tbody>
</table>
Recap: More on Growth Function

The VC Dimension
Definition of VC Dimension

\[ m_\mathcal{H}(N) \text{ of break point } k \leq B(N, k) = \sum_{i=0}^{k-1} \binom{N}{i} \]

\[
\begin{array}{c|ccccc}
B(N, k) & 1 & 2 & 3 & 4 & 5 \\
\hline
N & 1 & 2 & 2 & 2 & 2 \\
2 & 1 & 3 & 4 & 4 & 4 \\
3 & 1 & 4 & 7 & 8 & 8 \\
4 & 1 & 5 & 11 & 15 & 16 \\
5 & 1 & 6 & 16 & 26 & 31 \\
6 & 1 & 7 & 22 & 42 & 57 \\
\end{array}
\]

\[
\begin{array}{c|ccccc}
N^{k-1} & 1 & 2 & 3 & 4 & 5 \\
\hline
k & 1 & 1 & 1 & 1 & 1 \\
2 & 1 & 2 & 4 & 8 & 16 \\
3 & 1 & 3 & 9 & 27 & 81 \\
4 & 1 & 4 & 16 & 64 & 256 \\
5 & 1 & 5 & 25 & 125 & 625 \\
6 & 1 & 6 & 36 & 216 & 1296 \\
\end{array}
\]

**provably & loosely, for } N \geq 2, k \geq 3,$

\[ m_\mathcal{H}(N) \leq B(N, k) = \sum_{i=0}^{k-1} \binom{N}{i} \leq N^{k-1} \]
Recap: More on Vapnik-Chervonenkis (VC) Bound

For \( \mathcal{H} \) and ‘statistical’ large \( \mathcal{D} \),

\[
P_{\mathcal{D}} \left[ \exists h \in \mathcal{H} \text{ s.t. } |E_{in}(h) - E_{out}(h)| > \epsilon \right] 
\leq 4m_{\mathcal{H}}(2N) \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]
For any $g = A(D) \in \mathcal{H}$ and ‘statistical’ large $D$, 

\[
P_D \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] 
\leq P_D \left[ \exists h \in \mathcal{H} \text{ s.t. } |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon \right] 
\leq 4m_\mathcal{H}(2N) \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]
Recap: More on Vapnik-Chervonenkis (VC) Bound

For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( D \), for \( N \geq 2, k \geq 3 \)

\[
\mathbb{P}_D \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \\
\leq \mathbb{P}_D \left[ \exists h \in \mathcal{H} \text{ s.t. } |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon \right] \\
\leq 4m_\mathcal{H}(2N) \exp \left(-\frac{1}{8} \epsilon^2 N\right) \\
\text{if } k \text{ exists} \\
\leq 4(2N)^{k-1} \exp \left(-\frac{1}{8} \epsilon^2 N\right)
\]
Recap: More on Vapnik-Chervonenkis (VC) Bound

For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( D \), for \( N \geq 2, k \geq 3 \)

\[
\mathbb{P}_D \left[ \left| E_{\text{in}}(g) - E_{\text{out}}(g) \right| > \epsilon \right] \\
\leq \mathbb{P}_D \left[ \exists h \in \mathcal{H} \text{ s.t. } \left| E_{\text{in}}(h) - E_{\text{out}}(h) \right| > \epsilon \right] \\
\leq 4m_\mathcal{H}(2N) \exp \left( -\frac{1}{8} \epsilon^2 N \right) \\
\leq 4(2N)^{k-1} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

if \( k \) exists

if ① \( m_\mathcal{H}(N) \) breaks at \( k \) (good \( \mathcal{H} \))

② \( N \) large enough (good \( D \))

\[ \implies \text{generalized '} E_{\text{out}} \approx E_{\text{in}} \text{'} \]
For any \( g = A(D) \in H \) and ‘statistical’ large \( D \), for \( N \geq 2, k \geq 3 \)

\[
\Pr_D \left[ |E_{in}(g) - E_{out}(g)| > \epsilon \right] \\
\leq \Pr_D \left[ \exists h \in H \text{ s.t. } |E_{in}(h) - E_{out}(h)| > \epsilon \right] \\
\leq 4m_H(2N) \exp \left( -\frac{1}{8} \epsilon^2 N \right) \\
\text{if } k \text{ exists} \\
\leq 4(2N)^{k-1} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

if ① \( m_H(N) \) breaks at \( k \) (good \( H \))

② \( N \) large enough (good \( D \))

\[ \implies \text{generalized } 'E_{out} \approx E_{in}' \], and

if ③ \( A \) picks a \( g \) with small \( E_{in} \) (good \( A \))
Recap: More on Vapnik-Chervonenkis (VC) Bound

For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( D \), for \( N \geq 2, k \geq 3 \)

\[
\mathbb{P}_D \left[ |E_{in}(g) - E_{out}(g)| > \epsilon \right] \\
\leq \mathbb{P}_D \left[ \exists h \in \mathcal{H} \text{ s.t. } |E_{in}(h) - E_{out}(h)| > \epsilon \right] \\
\leq 4m_\mathcal{H}(2N) \exp \left( -\frac{1}{8} \epsilon^2 N \right) \\
\text{if } k \text{ exists} \\
\leq 4(2N)^{k-1} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

if ① \( m_\mathcal{H}(N) \) breaks at \( k \)  \hspace{1cm} (good \( \mathcal{H} \))

② \( N \) large enough  \hspace{1cm} (good \( D \))

\[ \implies \text{generalized } 'E_{out} \approx E_{in}', \text{ and} \]

if ③ \( A \) picks a \( g \) with small \( E_{in} \)  \hspace{1cm} (good \( A \))

\[ \implies \text{learned!} \]
Recap: More on Vapnik-Chervonenkis (VC) Bound

For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( D \), for \( N \geq 2, k \geq 3 \)

\[
\mathbb{P}_D \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \\
\leq \mathbb{P}_D \left[ \exists h \in \mathcal{H} \text{ s.t. } |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon \right] \\
\leq 4m_{\mathcal{H}}(2N) \exp \left( -\frac{1}{8} \epsilon^2 N \right) \\
\text{if } k \text{ exists} \\
\leq 4(2N)^{k-1} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

if ① \( m_{\mathcal{H}}(N) \) breaks at \( k \) (good \( \mathcal{H} \))
if ② \( N \) large enough (good \( D \))

\( \implies \) probably generalized ‘\( E_{\text{out}} \approx E_{\text{in}} \)’, and

if ③ \( A \) picks a \( g \) with small \( E_{\text{in}} \) (good \( A \))

\( \implies \) probably learned! (:-) good luck)

Hsuan-Tien Lin  (NTU CSIE)  Machine Learning Foundations  3/26
VC Dimension

the formal name of break point
VC Dimension

the formal name of maximum non-break point
The VC Dimension

Definition of VC Dimension

VC Dimension

the formal name of maximum non-break point

Definition

VC dimension of $\mathcal{H}$, denoted $d_{\text{vc}}(\mathcal{H})$ is

largest $N$ for which $m_{\mathcal{H}}(N) = 2^N$

- the most inputs $\mathcal{H}$ that can shatter
The VC Dimension

Definition of VC Dimension

VC Dimension

the formal name of **maximum non**-break point

**Definition**

VC dimension of $\mathcal{H}$, denoted $d_{\text{VC}}(\mathcal{H})$ is

**largest** $N$ for which $m_{\mathcal{H}}(N) = 2^N$

- the **most** inputs $\mathcal{H}$ that can shatter
- $d_{\text{VC}} = \text{‘minimum } k’ - 1$
The VC Dimension

**Definition**

VC dimension of $\mathcal{H}$, denoted $d_{\text{VC}}(\mathcal{H})$ is

**largest** $N$ for which $m_{\mathcal{H}}(N) = 2^N$

- the **most** inputs $\mathcal{H}$ that can shatter
- $d_{\text{VC}} = \text{‘minimum } k\text{’} - 1$

$$N \leq d_{\text{VC}} \implies \mathcal{H} \text{ can shatter some } N \text{ inputs}$$
VC Dimension

The formal name of maximum non-break point

Definition

VC dimension of $\mathcal{H}$, denoted $d_{vc}(\mathcal{H})$ is

\[
\text{largest } N \text{ for which } m_{\mathcal{H}}(N) = 2^N
\]

- the most inputs $\mathcal{H}$ that can shatter
- $d_{vc} = \text{‘minimum } k’ - 1$

$N \leq d_{vc} \implies \mathcal{H} \text{ can shatter some } N \text{ inputs}$

$N > d_{vc} \implies N \text{ is a break point for } \mathcal{H}$
VC Dimension

the formal name of **maximum non-break point**

**Definition**

VC dimension of $\mathcal{H}$, denoted $d_{vc}(\mathcal{H})$ is

\[ \text{largest } N \text{ for which } m_{\mathcal{H}}(N) = 2^N \]

- the **most** inputs $\mathcal{H}$ that can shatter
- $d_{vc} = \text{‘minimum } k\text{’ } - 1$

\[ N \leq d_{vc} \implies \mathcal{H} \text{ can shatter some } N \text{ inputs} \]

\[ k > d_{vc} \implies k \text{ is a break point for } \mathcal{H} \]
VC Dimension

the formal name of maximum non-break point

Definition

VC dimension of $\mathcal{H}$, denoted $d_{\text{VC}}(\mathcal{H})$ is

- **largest** $N$ for which $m_{\mathcal{H}}(N) = 2^N$
- the **most** inputs $\mathcal{H}$ that can shatter
- $d_{\text{VC}} = \text{‘minimum } k' - 1$

$N \leq d_{\text{VC}} \implies \mathcal{H}$ can shatter some $N$ inputs

$k > d_{\text{VC}} \implies k$ is a break point for $\mathcal{H}$

if $N \geq 2, d_{\text{VC}} \geq 2$, $m_{\mathcal{H}}(N) \leq N^{d_{\text{VC}}}$
The Four VC Dimensions

- positive rays:
  \[ d_{VC} = 1 \]
  \[ m_{\mathcal{H}}(N) = N + 1 \]
The Four VC Dimensions

- **positive rays:**
  \[ d_{VC} = 1 \]
  \[ m_{\mathcal{H}}(N) = N + 1 \]

- **positive intervals:**
  \[ d_{VC} = 2 \]
  \[ m_{\mathcal{H}}(N) = \frac{1}{2} N^2 + \frac{1}{2} N + 1 \]
The Four VC Dimensions

- **positive rays:**
  \[ m_H(N) = N + 1 \]
  \[ d_{VC} = 1 \]

- **positive intervals:**
  \[ m_H(N) = \frac{1}{2} N^2 + \frac{1}{2} N + 1 \]
  \[ d_{VC} = 2 \]

- **convex sets:**
  \[ m_H(N) = 2^N \]
  \[ d_{VC} = \infty \]
The Four VC Dimensions

- **positive rays:**
  - $d_{VC} = 1$
  - $m_{\mathcal{H}}(N) = N + 1$

- **positive intervals:**
  - $d_{VC} = 2$
  - $m_{\mathcal{H}}(N) = \frac{1}{2} N^2 + \frac{1}{2} N + 1$

- **convex sets:**
  - $d_{VC} = \infty$
  - $m_{\mathcal{H}}(N) = 2^N$

- **2D perceptrons:**
  - $d_{VC} = 3$
  - $m_{\mathcal{H}}(N) \leq N^3$ for $N \geq 2$
The Four VC Dimensions

- positive rays:
  \[ d_{VC} = 1 \]
  \[ m_{\mathcal{H}}(N) = N + 1 \]

- positive intervals:
  \[ d_{VC} = 2 \]
  \[ m_{\mathcal{H}}(N) = \frac{1}{2} N^2 + \frac{1}{2} N + 1 \]

- convex sets:
  \[ d_{VC} = \infty \]
  \[ m_{\mathcal{H}}(N) = 2^N \]

- 2D perceptrons:
  \[ d_{VC} = 3 \]
  \[ m_{\mathcal{H}}(N) \leq N^3 \text{ for } N \geq 2 \]

**good:** finite \( d_{VC} \)
VC Dimension and Learning

finite $d_{VC} \implies g \text{'will' generalize } (E_{out}(g) \approx E_{in}(g))$
VC Dimension and Learning

**finite** $d_{\text{VC}} \implies g \text{ ‘will’ generalize } (E_{\text{out}}(g) \approx E_{\text{in}}(g))$

- regardless of learning algorithm $A$
VC Dimension and Learning

finite $d_{VC} \implies g$ ‘will’ generalize $(E_{out}(g) \approx E_{in}(g))$

- regardless of learning algorithm $A$
- regardless of input distribution $P$
VC Dimension and Learning

**finite** \( d_{VC} \implies g \text{ ‘will’ generalize } (E_{out}(g) \approx E_{in}(g)) \)

- regardless of learning algorithm \( \mathcal{A} \)
- regardless of input distribution \( P \)
- regardless of target function \( f \)
**VC Dimension and Learning**

**finite** $d_{vc} \implies g$ ‘will’ generalize ($E_{out}(g) \approx E_{in}(g)$)

- regardless of learning algorithm $\mathcal{A}$
- regardless of input distribution $P$
- regardless of target function $f$

```
unknown target function $f: \mathcal{X} \rightarrow \mathcal{Y}$

unknown $P$ on $\mathcal{X}$

training examples $\mathcal{D}: (x_1, y_1), \cdots, (x_N, y_N)$

learning algorithm $\mathcal{A}$

final hypothesis $g \approx f$

hypothesis set $\mathcal{H}$
```
VC Dimension and Learning

**finite** $d_{VC} \implies g \text{ ‘will’ generalize } (E_{out}(g) \approx E_{in}(g))$

- regardless of learning algorithm $A$
- regardless of input distribution $P$
- regardless of target function $f$

unknown target function $f : \mathcal{X} \rightarrow \mathcal{Y}$

unknown $P$ on $\mathcal{X}$

$x_1, x_2, \cdots, x_N$

$x$

training examples $\mathcal{D} : (x_1, y_1), \cdots, (x_N, y_N)$

learning algorithm $A$

final hypothesis $g \approx f$

‘worst case’ guarantee on generalization

hypothesis set $\mathcal{H}$
If there is a set of $N$ inputs that cannot be shattered by $\mathcal{H}$. Based only on this information, what can we conclude about $d_{\text{VC}}(\mathcal{H})$?

1. $d_{\text{VC}}(\mathcal{H}) > N$
2. $d_{\text{VC}}(\mathcal{H}) = N$
3. $d_{\text{VC}}(\mathcal{H}) < N$
4. no conclusion can be made
Fun Time

If there is a set of $N$ inputs that cannot be shattered by $\mathcal{H}$. Based only on this information, what can we conclude about $d_{VC}(\mathcal{H})$?

1. $d_{VC}(\mathcal{H}) > N$
2. $d_{VC}(\mathcal{H}) = N$
3. $d_{VC}(\mathcal{H}) < N$
4. no conclusion can be made

Reference Answer: 4

It is possible that there is another set of $N$ inputs that can be shattered, which means $d_{VC} \geq N$. It is also possible that no set of $N$ input can be shattered, which means $d_{VC} < N$. Neither cases can be ruled out by one non-shattering set.
2D PLA Revisited

linearly separable $\mathcal{D}$
2D PLA Revisited

linearly separable $\mathcal{D}$

$\downarrow$

PLA can converge
2D PLA Revisited

linearly separable $\mathcal{D}$

PLA can converge

$T$ large

$E_{\text{in}}(g) = 0$
linearly separable $\mathcal{D}$

\[ E_{in}(g) = 0 \]

PLA can converge

$T$ large

with $x_n \sim P$ and $y_n = f(x_n)$

$E_{out}(g) \approx E_{in}(g)$

$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq ...$

$\text{VC Dimension} = 3$ for linearly separable $\mathcal{D}$ with $x_n \sim P$ and $y_n = f(x_n)$ and $T$ large.
2D PLA Revisited

\[ E_{in}(g) = 0 \]

linearly separable \( D \) with \( x_n \sim P \) and \( y_n = f(x_n) \)

\[ P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq \ldots \text{ by } d_{VC} = 3 \]
The VC Dimension

VC Dimension of Perceptrons

2D PLA Revisited

linearly separable $\mathcal{D}$

$\implies$ PLA can converge

$T$ large

$E_{in}(g) = 0$

with $x_n \sim P$ and $y_n = f(x_n)$

$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq ...$ by $d_{VC} = 3$

$N$ large

$E_{out}(g) \approx E_{in}(g)$

Hsuan-Tien Lin  (NTU CSIE)
The VC Dimension

VC Dimension of Perceptrons

2D PLA Revisited

linearly separable \( \mathcal{D} \)

PLA can converge

with \( x_n \sim P \) and \( y_n = f(x_n) \)

\[ \mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq ... \]

by \( d_{VC} = 3 \)

\( T \) large

\( E_{in}(g) = 0 \)

\( N \) large

\( E_{out}(g) \approx E_{in}(g) \)

\( E_{out}(g) \approx 0 :-( \)

Hsuan-Tien Lin  (NTU CSIE)
2D PLA Revisited

linearly separable $D$

$\downarrow$

PLA can converge

$\downarrow$

$E_{in}(g) = 0$

$\downarrow$

$E_{out}(g) \approx 0 :-)$

with $x_n \sim P$ and $y_n = f(x_n)$

$\downarrow$

$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq ...$ by $d_{VC} = 3$

$\downarrow$

$N$ large

$\downarrow$

$E_{out}(g) \approx E_{in}(g)$

$T$ large

$\downarrow$

general PLA for $x$ with more than 2 features?
VC Dimension of Perceptrons

- 1D perceptron (pos/neg rays): $d_{vc} = 2$
VC Dimension of Perceptrons

- 1D perceptron (pos/neg rays): $d_{VC} = 2$
- 2D perceptrons: $d_{VC} = 3$
  - $d_{VC} \geq 3$: • •
  - $d_{VC} \leq 3$: × ○ ×
VC Dimension of Perceptrons

- 1D perceptron (pos/neg rays): $d_{VC} = 2$
- 2D perceptrons: $d_{VC} = 3$
  - $d_{VC} \geq 3$: • •
  - $d_{VC} \leq 3$: × ○ ×
- $d$-D perceptrons: $d_{VC} \geq d + 1$
VC Dimension of Perceptrons

- 1D perceptron (pos/neg rays): $d_{VC} = 2$
- 2D perceptrons: $d_{VC} = 3$
  - $d_{VC} \geq 3$: • •
  - $d_{VC} \leq 3$: × ○ ×
- $d$-D perceptrons: $d_{VC} \equiv d + 1$

Two steps:
- $d_{VC} \geq d + 1$
- $d_{VC} \leq d + 1$
What statement below shows that \( d_{VC} \geq d + 1 \)?

1. There are some \( d + 1 \) inputs we can shatter.
2. We can shatter any set of \( d + 1 \) inputs.
3. There are some \( d + 2 \) inputs we cannot shatter.
4. We cannot shatter any set of \( d + 2 \) inputs.
What statement below shows that $d_{vc} \geq d + 1$?

1. There are some $d + 1$ inputs we can shatter.
2. We can shatter any set of $d + 1$ inputs.
3. There are some $d + 2$ inputs we cannot shatter.
4. We cannot shatter any set of $d + 2$ inputs.

Reference Answer: 1

$d_{vc}$ is the maximum that $m_H(N) = 2^N$, and $m_H(N)$ is the most number of dichotomies of $N$ inputs. So if we can find $2^{d+1}$ dichotomies on some $d + 1$ inputs, $m_H(d + 1) = 2^{d+1}$ and hence $d_{vc} \geq d + 1$. 
The VC Dimension

VC Dimension of Perceptrons

$d_{VC} \geq d + 1$

There are some $d + 1$ inputs we can shatter.

• some ‘trivial’ inputs:

$$X = \begin{bmatrix}
- x_1^T \\
- x_2^T \\
- x_3^T \\
\vdots \\
- x_{d+1}^T
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 \\
1 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \ddots & 0 \\
1 & 0 & \ldots & 0 & 1
\end{bmatrix}$$

• visually in 2D:

note: $X$ invertible!
Can We Shatter $X$?

$X = \begin{bmatrix} x_1^T & \cdots & x_{d+1}^T \end{bmatrix}$

$= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & \cdots & 0 & 1 \end{bmatrix}$ invertible

to shatter ... 

for any $y = \begin{bmatrix} y_1 \\ \vdots \\ y_{d+1} \end{bmatrix}$, find $w$ such that

$\text{sign}(Xw) = y$
Can We Shatter $X$?

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_{d+1}^T \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 1 & 0 & \ldots & 0 & 1 \end{bmatrix}$$ invertible

to shatter...

for any $y = \begin{bmatrix} y_1 \\ \vdots \\ y_{d+1} \end{bmatrix}$, find $w$ such that

$$\text{sign}(Xw) = y \iff (Xw) = y$$
Can We Shatter $X$?

$$X = \begin{bmatrix}
  x_1^T \\
  x_2^T \\
  \vdots \\
  x_{d+1}^T \\
\end{bmatrix} = \begin{bmatrix}
  1 & 0 & 0 & \ldots & 0 \\
  1 & 1 & 0 & \ldots & 0 \\
  \vdots & \vdots & \vdots & \ddots & 0 \\
  1 & 0 & \ldots & 0 & 1 \\
\end{bmatrix}$$

invertible

to shatter ...

for any $y = \begin{bmatrix}
  y_1 \\
  \vdots \\
  y_{d+1} \\
\end{bmatrix}$, find $w$ such that

$$\text{sign} (Xw) = y \iff (Xw) = y \quad \iff \quad w = X^{-1}y$$
Can We Shatter $X$?

$$X = \begin{bmatrix}
-x_1^T & - & - \\
-x_2^T & - & - \\
\vdots & - & - \\
-x_{d+1}^T & - & - \\
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & 0 \\
1 & 0 & \ldots & 0 & 1 \\
\end{bmatrix} \text{ invertible}
$$

to shatter ...

for any $y = \begin{bmatrix}
y_1 \\
\vdots \\
y_{d+1}
\end{bmatrix}$, find $w$ such that

$$\text{sign}(Xw) = y \iff (Xw) = y \iff X \text{ invertible!} \iff w = X^{-1}y$$

‘special’ $X$ can be shattered $\implies d_{VC} \geq d + 1$
What statement below shows that $d_{VC} \leq d + 1$?

1. There are some $d + 1$ inputs we can shatter.
2. We can shatter any set of $d + 1$ inputs.
3. There are some $d + 2$ inputs we cannot shatter.
4. We cannot shatter any set of $d + 2$ inputs.
What statement below shows that $d_{\text{vc}} \leq d + 1$?

1. There are some $d + 1$ inputs we can shatter.
2. We can shatter any set of $d + 1$ inputs.
3. There are some $d + 2$ inputs we cannot shatter.
4. We cannot shatter any set of $d + 2$ inputs.

Reference Answer: 4

$d_{\text{vc}}$ is the maximum that $m_{\mathcal{H}}(N) = 2^N$, and $m_{\mathcal{H}}(N)$ is the most number of dichotomies of $N$ inputs. So if we cannot find $2^{d+2}$ dichotomies on any $d + 2$ inputs (i.e. break point), $m_{\mathcal{H}}(d + 2) < 2^{d+2}$ and hence $d_{\text{vc}} < d + 2$. That is, $d_{\text{vc}} \leq d + 1$. 
The VC Dimension

VC Dimension of Perceptrons

\[ d_{VC} \leq d + 1 \left( \frac{1}{2} \right) \]

A 2D Special Case

\[ X = \begin{bmatrix}
- x_1^T \\
- x_2^T \\
- x_3^T \\
- x_4^T
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1
\end{bmatrix} \]
The VC Dimension

VC Dimension of Perceptrons

\[ d_{\text{VC}} \leq d + 1 \ (1/2) \]

A 2D Special Case

\[
\begin{bmatrix}
\bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet \\
\end{bmatrix}
\]

\[
X = \begin{bmatrix}
x_1^T & \bullet \\
x_2^T & \bullet \\
x_3^T & \bullet \\
x_4^T & \bullet \\
\end{bmatrix}
= \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1 \\
\end{bmatrix}
\]

\[ ? \text{ cannot be } \times \]
The VC Dimension

**VC Dimension of Perceptrons**

\[ d_{VC} \leq d + 1 \ (1/2) \]

**A 2D Special Case**

\[
X = \begin{bmatrix}
\mathbf{x}_1^T \\
\mathbf{x}_2^T \\
\mathbf{x}_3^T \\
\mathbf{x}_4^T
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1
\end{bmatrix}
\]

\[ \mathbf{x}_4 = \mathbf{x}_2 + \mathbf{x}_3 - \mathbf{x}_1 \]

\[ \text{？ cannot be } \times \]
The VC Dimension

$\text{VC Dimension of Perceptrons}$

$d_{\text{VC}} \leq d + 1 \ (1/2)$

**A 2D Special Case**

\[ X = \begin{bmatrix}
- x_1^T \\
- x_2^T \\
- x_3^T \\
- x_4^T
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1
\end{bmatrix} \]

\[ \begin{array}{c}
\bigcirc \\
\times \\
\bigcirc
\end{array} \]

? cannot be $\times$

\[ w^T x_4 = w^T x_2 + w^T x_3 - w^T x_1 \]
The VC Dimension

VC Dimension of Perceptrons

\[ d_{vc} \leq d + 1 \ (1/2) \]

A 2D Special Case

\[
\begin{bmatrix}
\vdots & \vdots \\
\vdots & \vdots \\
\vdots & \vdots \\
\end{bmatrix}
\]

\[
X = 
\begin{bmatrix}
\mathbf{x}_1^T \\
\mathbf{x}_2^T \\
\mathbf{x}_3^T \\
\mathbf{x}_4^T \\
\end{bmatrix}
= 
\begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1 \\
\end{bmatrix}
\]

\[ \mathbf{w}^T \mathbf{x}_4 = \mathbf{w}^T \mathbf{x}_2 + \mathbf{w}^T \mathbf{x}_3 - \mathbf{w}^T \mathbf{x}_1 \]

? cannot be \( \times \)
The VC Dimension

VC Dimension of Perceptrons

\[ d_{\text{VC}} \leq d + 1 \ (1/2) \]

A 2D Special Case

\[
\begin{bmatrix}
- & x_1^T & - \\
- & x_2^T & - \\
- & x_3^T & - \\
- & x_4^T & - \\
\end{bmatrix}
= \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1 \\
\end{bmatrix}
\]

? cannot be \( \times \)

\[
w^T x_4 = w^T x_2 + w^T x_3 - w^T x_1 > 0
\]
The VC Dimension

VC Dimension of Perceptrons

\[ d_{VC} \leq d + 1 \quad (1/2) \]

A 2D Special Case

\[
X = \begin{bmatrix}
- \mathbf{x}_1^T \\
- \mathbf{x}_2^T \\
- \mathbf{x}_3^T \\
- \mathbf{x}_4^T
\end{bmatrix}
= \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1
\end{bmatrix}
\]

\[
\begin{array}\text{○ ○ ?}
\end{array} \\
\begin{array}\times \text{○}
\end{array}
\]

\[
? \text{ cannot be } \times
\]

\[
w^T \mathbf{x}_4 = w^T \mathbf{x}_2 + w^T \mathbf{x}_3 - w^T \mathbf{x}_1 > 0
\]

linear dependence restricts dichotomy
\[ d_{VC} \leq d + 1 \ (2/2) \]

**d-D General Case**

\[ X = \begin{bmatrix}
- x_1^T \\
- x_2^T \\
\vdots \\
- x_{d+1}^T \\
- x_{d+2}^T
\end{bmatrix} \]

more rows than columns:

linear dependence (some \( a_i \) non-zero)

\[ x_{d+2} = a_1 x_1 + a_2 x_2 + \ldots + a_{d+1} x_{d+1} \]
The VC Dimension

VC Dimension of Perceptrons

d_{VC} \leq d + 1 (2/2)

d-D General Case

\[ X = \begin{bmatrix}
- x_1^T - \\
- x_2^T - \\
\vdots \\
- x_{d+1}^T - \\
- x_{d+2}^T - 
\end{bmatrix} \]

more rows than columns:

linear dependence (some \( a_i \) non-zero)

\[ x_{d+2} = a_1 x_1 + a_2 x_2 + \ldots + a_{d+1} x_{d+1} \]

• can you generate \( (\text{sign}(a_1), \text{sign}(a_2), \ldots, \text{sign}(a_{d+1}), \times) \)? if so, what \( w \)?

\[
w^T x_{d+2} = a_1 w^T x_1 + a_2 w^T x_2 + \ldots + a_{d+1} w^T x_{d+1} \]

\( \circ \times \times \times \)

\( > 0 \) (contradiction!)
The VC Dimension

VC Dimension of Perceptrons

\[ d_{\text{VC}} \leq d + 1 \ (2/2) \]

**d-D General Case**

\[
X = \begin{bmatrix}
- \quad x_1^T & - \\
- \quad x_2^T & - \\
\vdots & \\
- \quad x_{d+1}^T & - \\
- \quad x_{d+2}^T & - \\
\end{bmatrix}
\]

more rows than columns:

linear dependence (some \( a_i \) non-zero)

\[ x_{d+2} = a_1 x_1 + a_2 x_2 + \ldots + a_{d+1} x_{d+1} \]

- can you generate \((\text{sign}(a_1), \text{sign}(a_2), \ldots, \text{sign}(a_{d+1}), \times)\)? if so, what \( w \)?

\[
w^T x_{d+2} = a_1 w^T x_1 + a_2 w^T x_2 + \ldots + a_{d+1} w^T x_{d+1}
\]

\[ > 0 \text{(contradiction!)} \]

‘general’ \( X \) no-shatter \( \implies d_{\text{VC}} \leq d + 1 \)
Based on the proof above, what is $d_{vc}$ of 1126-D perceptrons?

1. 1024
2. 1126
3. 1127
4. 6211

Reference Answer: 3

Well, too much fun for this section! :-)
Based on the proof above, what is $d_{vc}$ of 1126-D perceptrons?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1024</td>
</tr>
<tr>
<td>2</td>
<td>1126</td>
</tr>
<tr>
<td>3</td>
<td>1127</td>
</tr>
<tr>
<td>4</td>
<td>6211</td>
</tr>
</tbody>
</table>

Reference Answer: 3

Well, too much fun for this section! :-}
• hypothesis parameters $\mathbf{w} = (w_0, w_1, \cdots, w_d)$: creates degrees of freedom
• hypothesis parameters \( w = (w_0, w_1, \cdots, w_d) \):
  creates degrees of freedom

• hypothesis quantity \( M = |\mathcal{H}| \):
  ‘analog’ degrees of freedom
**The VC Dimension**

**Physical Intuition of VC Dimension**

---

**Degrees of Freedom**

- **hypothesis parameters** $w = (w_0, w_1, \ldots, w_d)$: creates degrees of freedom
- **hypothesis quantity** $M = |\mathcal{H}|$: ‘analog’ degrees of freedom
- **hypothesis ‘power’** $d_{vc} = d + 1$: effective ‘binary’ degrees of freedom

(modified from the work of Hugues Vermeiren on [http://www.texample.net](http://www.texample.net))
The VC Dimension

Physical Intuition of VC Dimension

Degrees of Freedom

- hypothesis parameters $w = (w_0, w_1, \cdots, w_d)$: creates degrees of freedom
- hypothesis quantity $M = |\mathcal{H}|$: ‘analog’ degrees of freedom
- hypothesis ‘power’ $d_{\text{VC}} = d + 1$: effective ‘binary’ degrees of freedom

$d_{\text{VC}}(\mathcal{H})$: powerfullness of $\mathcal{H}$

(modified from the work of Hugues Vermeiren on http://www.texample.net)
Two Old Friends

Positive Rays ($d_{VC} = 1$)

\[ h(x) = -1 \]
\[ h(x) = +1 \]

free parameters: $a$
Two Old Friends

Positive Rays ($d_{\text{VC}} = 1$)

$\begin{align*}
h(x) &= -1 \\
\ell \leq x \leq r \quad & a \quad \quad h(x) &= +1
\end{align*}$

free parameters: $a$

Positive Intervals ($d_{\text{VC}} = 2$)

$\begin{align*}
h(x) &= -1 \\
\ell \leq x \leq r \\
\quad h(x) &= +1 \\
\quad h(x) &= -1
\end{align*}$

free parameters: $\ell, r$
The VC Dimension

Physical Intuition of VC Dimension

Two Old Friends

Positive Rays ($d_{\text{VC}} = 1$)

\[ h(x) = -1 \]

\[ h(x) = +1 \]

\[ a \]

free parameters: \( a \)

Positive Intervals ($d_{\text{VC}} = 2$)

\[ h(x) = -1 \]

\[ h(x) = +1 \]

\[ h(x) = -1 \]

\[ \ell, r \]

free parameters: \( \ell, r \)

practical rule of thumb:

\[ d_{\text{VC}} \approx \#\text{free parameters} \text{ (but not always)} \]
**M and \(d_{VC}\)**

copied from Lecture 5 :-)

1. can we make sure that \(E_{out}(g)\) is close enough to \(E_{in}(g)\)?
2. can we make \(E_{in}(g)\) small enough?

**small \(M\)**

1. Yes!, \(\mathbb{P}[BAD] \leq 2 \cdot M \cdot \exp(...)
2. No!, too few choices

**large \(M\)**

1. No!, \(\mathbb{P}[BAD] \leq 2 \cdot M \cdot \exp(...)
2. Yes!, many choices
The VC Dimension

Physical Intuition of VC Dimension

$M$ and $d_{VC}$

copied from Lecture 5 :-)

1. can we make sure that $E_{out}(g)$ is close enough to $E_{in}(g)$?
2. can we make $E_{in}(g)$ small enough?

**small $M$**

1. Yes!, $\mathbb{P}[BAD] \leq 2 \cdot M \cdot \exp(\ldots)$
2. No!, too few choices

**large $M$**

1. No!, $\mathbb{P}[BAD] \leq 2 \cdot M \cdot \exp(\ldots)$
2. Yes!, many choices

**small $d_{VC}$**

1. Yes!, $\mathbb{P}[BAD] \leq 4 \cdot (2N)^{d_{VC}} \cdot \exp(\ldots)$

**large $d_{VC}$**

1. No!, too limited power
2. Yes!, lots of power

using the right $d_{VC}$ (or $H$) is important
M and $d_{VC}$

Copied from Lecture 5 :-) 

1. Can we make sure that $E_{out}(g)$ is close enough to $E_{in}(g)$?
2. Can we make $E_{in}(g)$ small enough?

**Small $M$**

1. Yes!, $P[BAD] \leq 2 \cdot M \cdot \exp(...)$
2. No!, too few choices

**Large $M$**

1. No!, $P[BAD] \leq 2 \cdot M \cdot \exp(...)$
2. Yes!, many choices

**Small $d_{VC}$**

1. Yes!, $P[BAD] \leq 4 \cdot (2N)^{d_{VC}} \cdot \exp(...)$
2. No!, too limited power

**Large $d_{VC}$**

1. No!, $P[BAD] \leq 4 \cdot (2N)^{d_{VC}} \cdot \exp(...)$
2. Yes!, lots of power
**The VC Dimension**

*Physical Intuition of VC Dimension*

**$M$ and $d_{VC}$**

Copied from Lecture 5 :-)

1. Can we make sure that $E_{out}(g)$ is close enough to $E_{in}(g)$?
2. Can we make $E_{in}(g)$ small enough?

### Small $M$

- **Yes!**, $P[\text{BAD}] \leq 2 \cdot M \cdot \exp(...)$
- **No!**, too few choices

### Large $M$

- **No!**, $P[\text{BAD}] \leq 2 \cdot M \cdot \exp(...)$
- **Yes!**, many choices

### Small $d_{VC}$

- **Yes!**, $P[\text{BAD}] \leq 4 \cdot (2N)^{d_{VC}} \cdot \exp(...)$
- **No!**, too limited power

### Large $d_{VC}$

- **No!**, $P[\text{BAD}] \leq 4 \cdot (2N)^{d_{VC}} \cdot \exp(...)$
- **Yes!**, lots of power
### The VC Dimension

**Physical Intuition of VC Dimension**

---

**$M$ and $d_{VC}$**

Copied from Lecture 5 :-)

1. Can we make sure that $E_{out}(g)$ is close enough to $E_{in}(g)$?
2. Can we make $E_{in}(g)$ small enough?

<table>
<thead>
<tr>
<th>Small $M$</th>
<th>Large $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yes!, $P[BAD] \leq 2 \cdot M \cdot \exp(...)$</td>
<td>1. No!, $P[BAD] \leq 2 \cdot M \cdot \exp(...)$</td>
</tr>
<tr>
<td>2. No!, too few choices</td>
<td>2. Yes!, many choices</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Small $d_{VC}$</th>
<th>Large $d_{VC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yes!, $P[BAD] \leq 4 \cdot (2N)^{d_{VC}} \cdot \exp(...)$</td>
<td>1. No!, $P[BAD] \leq 4 \cdot (2N)^{d_{VC}} \cdot \exp(...)$</td>
</tr>
<tr>
<td>2. No!, too limited power</td>
<td>2. Yes!, lots of power</td>
</tr>
</tbody>
</table>

---

Hsuan-Tien Lin (NTU CSIE)  
Machine Learning Foundations
### The VC Dimension

#### Physical Intuition of VC Dimension

**$M$ and $d_{\text{VC}}$**

**copied from Lecture 5 :-)**

1. can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. can we make $E_{\text{in}}(g)$ small enough?

<table>
<thead>
<tr>
<th>small $M$</th>
<th>large $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yes!, $P[\text{BAD}] \leq 2 \cdot M \cdot \exp(\ldots)$</td>
<td>1. No!, $P[\text{BAD}] \leq 2 \cdot M \cdot \exp(\ldots)$</td>
</tr>
<tr>
<td>2. No!, too few choices</td>
<td>2. Yes!, many choices</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>small $d_{\text{VC}}$</th>
<th>large $d_{\text{VC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yes!, $P[\text{BAD}] \leq 4 \cdot (2N)^{d_{\text{VC}}} \cdot \exp(\ldots)$</td>
<td>1. No!, $P[\text{BAD}] \leq 4 \cdot (2N)^{d_{\text{VC}}} \cdot \exp(\ldots)$</td>
</tr>
<tr>
<td>2. No!, too limited power</td>
<td>2. Yes!, lots of power</td>
</tr>
</tbody>
</table>

**using the right $d_{\text{VC}}$ (or $\mathcal{H}$) is important**
Origin-crossing Hyperplanes are essentially perceptrons with $w_0$ fixed at 0. Make a guess about the $d_{vc}$ of origin-crossing hyperplanes in $\mathbb{R}^d$.

1. 1
2. $d$
3. $d + 1$
4. $\infty$
Origin-crossing Hyperplanes are essentially perceptrons with $w_0$ fixed at 0. Make a guess about the $d_{vc}$ of origin-crossing hyperplanes in $\mathbb{R}^d$.

1. 1
2. $d$
3. $d + 1$
4. $\infty$

Reference Answer: 2

The proof is almost the same as proving the $d_{vc}$ for usual perceptrons, but it is the intuition ($d_{vc} \approx \#\text{free parameters}$) that you shall use to answer this quiz.
For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( D \), for \( N \geq 2 \), \( d_{VC} \geq 2 \)

\[
\mathbb{P}_D \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

BAD
For any $g = A(D) \in \mathcal{H}$ and ‘statistical’ large $D$, for $N \geq 2$, $d_{VC} \geq 2$

$$\mathbb{P}_D \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$$

BAD

Rephrase

... with probability $\geq 1 - \delta$, GOOD: $\left| E_{in}(g) - E_{out}(g) \right| \leq \epsilon$

set $\delta = 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$
VC Bound Rephrase: Penalty for Model Complexity

For any $g = A(D) \in \mathcal{H}$ and ‘statistical’ large $D$, for $N \geq 2$, $d_{VC} \geq 2$

$$\mathbb{P}_D \left[ |E_{in}(g) - E_{out}(g)| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$$

BAD

Rephrase

... with probability $\geq 1 - \delta$, GOOD: $|E_{in}(g) - E_{out}(g)| \leq \epsilon$

set $\delta = 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$

$$\frac{\delta}{4(2N)^{d_{VC}}} = \exp \left( -\frac{1}{8} \epsilon^2 N \right)$$
For any $g = A(D) \in \mathcal{H}$ and ‘statistical’ large $D$, for $N \geq 2$, $d_{vc} \geq 2$

$$\mathbb{P}_D \left[ |E_{in}(g) - E_{out}(g)| > \epsilon \right] \leq 4(2N)^{d_{vc}} \exp \left(-\frac{1}{8} \epsilon^2 N \right)$$

BAD

Rephrase

... with probability $\geq 1 - \delta$, GOOD: $|E_{in}(g) - E_{out}(g)| \leq \epsilon$

set $\delta = 4(2N)^{d_{vc}} \exp \left(-\frac{1}{8} \epsilon^2 N \right)$

$$\frac{\delta}{4(2N)^{d_{vc}}} = \exp \left(-\frac{1}{8} \epsilon^2 N \right)$$

$$\ln \left( \frac{4(2N)^{d_{vc}}}{\delta} \right) = \frac{1}{8} \epsilon^2 N$$
VC Bound Rephrase: Penalty for Model Complexity

For any $g = A(D) \in \mathcal{H}$ and ‘statistical’ large $\mathcal{D}$, for $N \geq 2, d_{vc} \geq 2$

$$\mathbb{P}_D \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq \frac{4(2N)^{d_{vc}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)}{\delta}$$

**BAD**

Rephrase

... with probability $\geq 1 - \delta$, **GOOD**: $\left| E_{in}(g) - E_{out}(g) \right| \leq \epsilon$

Set

$$\delta = \frac{4(2N)^{d_{vc}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)}{\delta}$$

$$\ln \left( \frac{4(2N)^{d_{vc}}}{\delta} \right) = \frac{1}{8} \epsilon^2 N$$

$$\sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{vc}}}{\delta} \right)} = \epsilon$$
VC Bound Rephrase: Penalty for Model Complexity

For any \( g = \mathcal{A}(\mathcal{D}) \in \mathcal{H} \) and ‘statistical’ large \( \mathcal{D} \), for \( N \geq 2 \), \( d_{\text{VC}} \geq 2 \):

\[
\mathbb{P}_D \left[ \left| E_{\text{in}}(g) - E_{\text{out}}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

\( \text{BAD} \)

Rephrase

... with probability \( \geq 1 - \delta \), \text{GOOD}!

Gen. error \( \left| E_{\text{in}}(g) - E_{\text{out}}(g) \right| \leq \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{\text{VC}}}}{\delta} \right)} \)
The VC Dimension
Interpreting VC Dimension

VC Bound Rephrase: Penalty for Model Complexity

For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( D \), for \( N \geq 2, d_{VC} \geq 2 \)

\[
\mathbb{P}_D \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

BAD

Rephrase

\[ \ldots, \text{with probability } \geq 1 - \delta, \text{GOOD!} \]

\[
\text{gen. error } \left| E_{in}(g) - E_{out}(g) \right| \leq \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{VC}}}{\delta} \right)}
\]

\[
E_{in}(g) - \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{VC}}}{\delta} \right)} \leq E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{VC}}}{\delta} \right)}
\]
For any \( g = A(D) \in H \) and ‘statistical’ large \( D \), for \( N \geq 2 \), \( d_{VC} \geq 2 \)

\[
\mathbb{P}_D \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

BAD

Rephrase

\[
\text{gen. error } \left| E_{in}(g) - E_{out}(g) \right| \leq \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{VC}}}{\delta} \right)}
\]

\[
E_{in}(g) - \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{VC}}}{\delta} \right)} \leq E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{VC}}}{\delta} \right)}
\]

\( \sqrt{\cdots} \) : penalty for model complexity

\( \Omega(N, \mathcal{H}, \delta) \)
THE VC Message

with a high probability,

\[ E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{\text{VC}}}}{\delta} \right)} \]

\[ \Omega(N, \mathcal{H}, \delta) \]
The VC Dimension

Interpreting VC Dimension

THE VC Message

with a high probability,

\[ E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{\text{VC}}}}{\delta} \right)} \]

\[ \Omega(N, H, \delta) \]

- \( d_{\text{VC}} \uparrow: E_{\text{in}} \downarrow \) but \( \Omega \uparrow \)

\( d^*_{\text{VC}} \)

Error vs. VC dimension, \( d_{\text{VC}} \)

out-of-sample error

model complexity

in-sample error
with a high probability,

\[ E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{VC}}}{\delta} \right)} + \Omega(N, \mathcal{H}, \delta) \]

- \( d_{VC} \uparrow: E_{in} \downarrow \text{ but } \Omega \uparrow \)
- \( d_{VC} \downarrow: \Omega \downarrow \text{ but } E_{in} \uparrow \)
with a high probability,

\[ E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{\text{VC}}}}{\delta} \right)} \]

\[ \Omega(N,H,\delta) \]

- \( d_{\text{VC}} \uparrow: E_{\text{in}} \downarrow \text{ but } \Omega \uparrow \)
- \( d_{\text{VC}} \downarrow: \Omega \downarrow \text{ but } E_{\text{in}} \uparrow \)
- best \( d^*_{\text{VC}} \) in the middle
The VC Dimension

Interpreting VC Dimension

THE VC Message

with a high probability,

\[ E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{\text{VC}}}}{\delta} \right)} \]

\[ \Omega(N, H, \delta) \]

- \( d_{\text{VC}} \uparrow: E_{\text{in}} \downarrow \) but \( \Omega \uparrow \)
- \( d_{\text{VC}} \downarrow: \Omega \downarrow \) but \( E_{\text{in}} \uparrow \)
- best \( d_{\text{VC}}^* \) in the middle

powerful \( H \) not always good!
VC Bound Rephrase: Sample Complexity

For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( D \), for \( N \geq 2 \), \( d_{VC} \geq 2 \)

\[
P_D \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

BAD
VC Bound Rephrase: Sample Complexity

For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( D \), for \( N \geq 2, d_{\text{VC}} \geq 2 \)

\[
P_D \left[ \left| E_{\text{in}}(g) - E_{\text{out}}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8}\epsilon^2 N \right)
\]

BAD

Given specs \( \epsilon = 0.1, \delta = 0.1, d_{\text{VC}} = 3 \), want

\[
4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8}\epsilon^2 N \right) \leq \delta
\]

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.82 \times 10^7</td>
</tr>
</tbody>
</table>
VC Bound Rephrase: Sample Complexity

For any $g = A(D) \in \mathcal{H}$ and ‘statistical’ large $D$, for $N \geq 2$, $d_{VC} \geq 2$

$$\mathbb{P}_D \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$$

BAD

Given specs $\epsilon = 0.1$, $\delta = 0.1$, $d_{VC} = 3$, want $4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \leq \delta$

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\text{bound}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$2.82 \times 10^7$</td>
</tr>
<tr>
<td>1,000</td>
<td>$9.17 \times 10^9$</td>
</tr>
</tbody>
</table>
The VC Dimension

Interpreting VC Dimension

VC Bound Rephrase: Sample Complexity

For any \( g = A(\mathcal{D}) \in \mathcal{H} \) and ‘statistical’ large \( \mathcal{D} \), for \( N \geq 2 \), \( d_{\text{VC}} \geq 2 \)

\[
\mathbb{P}_\mathcal{D} \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \leq 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

BAD

given specs \( \epsilon = 0.1, \delta = 0.1, d_{\text{VC}} = 3 \), want \( 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \leq \delta \)

<table>
<thead>
<tr>
<th>( N )</th>
<th>bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>( 2.82 \times 10^7 )</td>
</tr>
<tr>
<td>1,000</td>
<td>( 9.17 \times 10^9 )</td>
</tr>
<tr>
<td>10,000</td>
<td>( 1.19 \times 10^8 )</td>
</tr>
</tbody>
</table>
VC Bound Rephrase: Sample Complexity

For any $g = A(D) \in \mathcal{H}$ and ‘statistical’ large $D$, for $N \geq 2$, $d_{VC} \geq 2$

$$\mathbb{P}_{D} \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$$

BAD

Given specs $\epsilon = 0.1$, $\delta = 0.1$, $d_{VC} = 3$, want $4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \leq \delta$

<table>
<thead>
<tr>
<th>$N$</th>
<th>bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$2.82 \times 10^7$</td>
</tr>
<tr>
<td>1,000</td>
<td>$9.17 \times 10^9$</td>
</tr>
<tr>
<td>10,000</td>
<td>$1.19 \times 10^8$</td>
</tr>
<tr>
<td>100,000</td>
<td>$1.65 \times 10^{-38}$</td>
</tr>
</tbody>
</table>
VC Bound Rephrase: Sample Complexity

For any \( g = A(D) \in \mathcal{H} \) and ‘statistical’ large \( \mathcal{D} \), for \( N \geq 2 \), \( d_{\text{VC}} \geq 2 \)

\[
\mathbb{P}_{\mathcal{D}} \left[ \left| E_{\text{in}}(g) - E_{\text{out}}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

BAD

given specs \( \epsilon = 0.1, \delta = 0.1, d_{\text{VC}} = 3 \), want \( 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \leq \delta \)

\begin{center}
\begin{tabular}{|c|c|}
\hline
\( N \) & bound \( \times 10^{\pm} \) \\
\hline
100 & 2.82 \times 10^7 \\
1,000 & 9.17 \times 10^9 \\
10,000 & 1.19 \times 10^8 \\
100,000 & 1.65 \times 10^{-38} \\
29,300 & 9.99 \times 10^{-2} \\
\hline
\end{tabular}
\end{center}
VC Bound Rephrase: Sample Complexity

For any \( g = A(\mathcal{D}) \in \mathcal{H} \) and ‘statistical’ large \( \mathcal{D} \), for \( N \geq 2 \), \( d_{VC} \geq 2 \)

\[
P_{\mathcal{D}} \left[ |E_{in}(g) - E_{out}(g)| > \varepsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \varepsilon^2 N \right)
\]

BAD

Given specs \( \varepsilon = 0.1, \delta = 0.1, d_{VC} = 3 \), want \( 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \varepsilon^2 N \right) \leq \delta \)

<table>
<thead>
<tr>
<th>( N )</th>
<th>bound</th>
<th>sample complexity: need ( N \approx 10,000d_{VC} ) in theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.82 \times 10^7</td>
<td>( 9.17 \times 10^9 )</td>
</tr>
<tr>
<td>1,000</td>
<td>9.17 \times 10^9</td>
<td>( 1.19 \times 10^8 )</td>
</tr>
<tr>
<td>10,000</td>
<td>1.19 \times 10^8</td>
<td>( 1.65 \times 10^{-38} )</td>
</tr>
<tr>
<td>100,000</td>
<td>1.65 \times 10^{-38}</td>
<td>( 9.99 \times 10^{-2} )</td>
</tr>
<tr>
<td>29,300</td>
<td>9.99 \times 10^{-2}</td>
<td></td>
</tr>
</tbody>
</table>

Practical rule of thumb:

\( N \approx 10d_{VC} \) often enough!
Looseness of VC Bound

\[ P_D \left[ |E_{in}(g) - E_{out}(g)| > \epsilon \right] \leq 4(2N)^{d_{vc}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \]

theory: \( N \approx 10,000d_{vc} \); practice: \( N \approx 10d_{vc} \)
Looseness of VC Bound

\[
P_D \left[ \left| E_{\text{in}}(g) - E_{\text{out}}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

theory: \( N \approx 10,000d_{\text{VC}} \); practice: \( N \approx 10d_{\text{VC}} \)

Why?

- Hoeffding for unknown \( E_{\text{out}} \)

any distribution, any target
The VC Dimension

Interpreting VC Dimension

Looseness of VC Bound

\[ P_D \left[ |E_{in}(g) - E_{out}(g)| > \epsilon \right] \leq 4(2N)^{d_{VC}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \]

theory: \( N \approx 10,000d_{VC} \); practice: \( N \approx 10d_{VC} \)

Why?

- Hoeffding for unknown \( E_{out} \) any distribution, any target
- \( m_{\mathcal{H}}(N) \) instead of \( |\mathcal{H}(x_1, \ldots, x_N)| \) ‘any’ data
Looseness of VC Bound

\[ \mathbb{P}_D \left[ \left| E_{\text{in}}(g) - E_{\text{out}}(g) \right| > \epsilon \right] \leq 4(2N)^{d_{\text{vc}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \]

theory: \( N \approx 10,000d_{\text{vc}} \); practice: \( N \approx 10d_{\text{vc}} \)

Why?

- Hoeffding for unknown \( E_{\text{out}} \)  
  any distribution, any target
- \( m_{\mathcal{H}}(N) \) instead of \( |\mathcal{H}(x_1, \ldots, x_N)| \)  
  ‘any’ data
- \( N^{d_{\text{vc}}} \) instead of \( m_{\mathcal{H}}(N) \)  
  ‘any’ \( \mathcal{H} \) of same \( d_{\text{vc}} \)
Looseness of VC Bound

\[
\mathbb{P}_D \left[ |E_{in}(g) - E_{out}(g)| > \epsilon \right] \leq 4(2N)^{d_{vc}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)
\]

theory: \( N \approx 10,000d_{vc} \); practice: \( N \approx 10d_{vc} \)

Why?

- Hoeffding for unknown \( E_{out} \) any distribution, any target
- \( m_{\mathcal{H}}(N) \) instead of \( |\mathcal{H}(x_1, \ldots, x_N)| \) ‘any’ data
- \( N^{d_{vc}} \) instead of \( m_{\mathcal{H}}(N) \) ‘any’ \( \mathcal{H} \) of same \( d_{vc} \)
- union bound on worst cases any choice made by \( A \)
Looseness of VC Bound

\[ \mathbb{P}_D \left( \left| E_{\text{in}}(g) - E_{\text{out}}(g) \right| > \epsilon \right) \leq 4(2N)^{d_{\text{vc}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \]

theory: \( N \approx 10,000d_{\text{vc}} \); practice: \( N \approx 10d_{\text{vc}} \)

Why?

- Hoeffding for unknown \( E_{\text{out}} \)
- \( m_{\mathcal{H}}(N) \) instead of \( |\mathcal{H}(x_1, \ldots, x_N)| \)
- \( N^{d_{\text{vc}}} \) instead of \( m_{\mathcal{H}}(N) \)
- union bound on worst cases

—but hardly better, and ‘similarly loose for all models’

any distribution, any target
‘any’ data
‘any’ \( \mathcal{H} \) of same \( d_{\text{vc}} \)
any choice made by \( A \)
Looseness of VC Bound

\[ \mathbb{P}_D \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \leq 4(2N)^{d_{\text{VC}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \]

theory: \( N \approx 10,000d_{\text{VC}} \); practice: \( N \approx 10d_{\text{VC}} \)

Why?
- Hoeffding for unknown \( E_{\text{out}} \) any distribution, any target
- \( m_{\mathcal{H}}(N) \) instead of \( |\mathcal{H}(x_1, \ldots, x_N)| \) ‘any’ data
- \( N^{d_{\text{VC}}} \) instead of \( m_{\mathcal{H}}(N) \) ‘any’ \( \mathcal{H} \) of same \( d_{\text{VC}} \)
- union bound on worst cases any choice made by \( \mathcal{A} \)

—but hardly better, and ‘similarly loose for all models’

philosophical message of VC bound
important for improving ML
Consider the VC Bound below. How can we decrease the probability of getting BAD data?

$$\mathbb{P}_D \left[ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \right] \leq 4(2N)^{d_{\text{vc}}} \exp \left( -\frac{1}{8} \epsilon^2 N \right)$$

1. decrease model complexity $d_{\text{vc}}$
2. increase data size $N$ a lot
3. increase generalization error tolerance $\epsilon$
4. all of the above
Consider the VC Bound below. How can we decrease the probability of getting BAD data?

\[ P_D \left[ |E_{in}(g) - E_{out}(g)| > \epsilon \right] \leq 4(2N)^{d_{vc}} \exp \left( -\frac{1}{8} \epsilon^2 N \right) \]

1. Decrease model complexity \( d_{vc} \)
2. Increase data size \( N \) a lot
3. Increase generalization error tolerance \( \epsilon \)
4. All of the above

Reference Answer: 4

Congratulations on being Master of VC bound! :-)

Hsuan-Tien Lin (NTU CSIE)
The VC Dimension

Interpreting VC Dimension

Summary

1. When Can Machines Learn?
2. **Why** Can Machines Learn?

Lecture 6: Theory of Generalization

Lecture 7: The VC Dimension

- Definition of VC Dimension
  - maximum non-break point
- VC Dimension of Perceptrons
  - $d_{VC}(\mathcal{H}) = d + 1$
- Physical Intuition of VC Dimension
  - $d_{VC} \approx \# \text{free parameters}$
- Interpreting VC Dimension
  - loosely: model complexity & sample complexity

- next: more than noiseless binary classification?

3. How Can Machines Learn?
4. How Can Machines Learn Better?