Review of Lecture 6

- \( m_H(N) \) is polynomial
- if \( H \) has a break point \( k \)

<table>
<thead>
<tr>
<th>( N )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>...</th>
</tr>
</thead>
</table>
| 1 | 1 | 2 | 2 | 2 | 2 | 2 | ...
| 2 | 1 | | | | | | |
| 3 | 1 | | | | | | |
| 4 | 1 | | | | | | |
| 5 | 1 | | | | | | |
| 6 | 1 | | | | | | |

\[
m_H(N) \leq \sum_{i=0}^{k-1} \binom{N}{i} \] maximum power is \( N^{k-1} \)

**The VC Inequality**

**Hoeffding Inequality**

\[
P \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 2 M e^{-2 \epsilon^2 N}
\]

**Union Bound**

\[
P \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 4 m_H(2N) e^{-\frac{1}{8} \epsilon^2 N}
\]

**VC Bound**

\[
P \left[ \left| E_{in}(g) - E_{out}(g) \right| > \epsilon \right] \leq 2 m_H(N) e^{-\frac{1}{M} \epsilon^2 N}
\]
Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 7: The VC Dimension
Outline

- The definition
- VC dimension of perceptrons
- Interpreting the VC dimension
- Generalization bounds
Definition of VC dimension

The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\text{VC}}(\mathcal{H})$, is

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

"the most points $\mathcal{H}$ can shatter"

$N \leq d_{\text{VC}}(\mathcal{H}) \implies \mathcal{H}$ can shatter $N$ points

$k > d_{\text{VC}}(\mathcal{H}) \implies k$ is a break point for $\mathcal{H}$
The growth function

In terms of a break point $k$:

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

In terms of the VC dimension $d_{VC}$:

$$m_\mathcal{H}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i}$$

maximum power is $N^{d_{VC}}$
Examples

- $\mathcal{H}$ is positive rays:
  \[ d_{VC} = 1 \]

- $\mathcal{H}$ is 2D perceptrons:
  \[ d_{VC} = 3 \]

- $\mathcal{H}$ is convex sets:
  \[ d_{VC} = \infty \]
VC dimension and learning

\[ d_{\text{VC}}(\mathcal{H}) \text{ is finite } \implies g \in \mathcal{H} \text{ will generalize} \]

- Independent of the learning algorithm
- Independent of the input distribution
- Independent of the target function
VC dimension of perceptrons

For \( d = 2 \), \( d_{\text{VC}} = 3 \)

In general, \( d_{\text{VC}} = d + 1 \)

We will prove two directions:

\[ d_{\text{VC}} \leq d + 1 \]

\[ d_{\text{VC}} \geq d + 1 \]
Here is one direction

A set of $N = d + 1$ points in $\mathbb{R}^d$ shattered by the perceptron:

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

$X$ is invertible
Can we shatter this data set?

For any \( \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} \), can we find a vector \( \mathbf{w} \) satisfying

\[
\text{sign}(X\mathbf{w}) = \mathbf{y}
\]

**Easy!** Just make \( X\mathbf{w} = \mathbf{y} \)

which means \( \mathbf{w} = X^{-1}\mathbf{y} \)
We can shatter these $d + 1$ points

This implies what?

[a] $d_{VC} = d + 1$

[b] $d_{VC} \geq d + 1$ ✓

[c] $d_{VC} \leq d + 1$

[d] No conclusion
Now, to show that $d_{vc} \leq d + 1$

We need to show that:

[a] There are $d + 1$ points we cannot shatter

[b] There are $d + 2$ points we cannot shatter

[c] We cannot shatter any set of $d + 1$ points

[d] We cannot shatter any set of $d + 2$ points \(\checkmark\)
Take any $d + 2$ points

For any $d + 2$ points,

$$x_1, \ldots, x_{d+1}, x_{d+2}$$

More points than dimensions $\implies$ we must have

$$x_j = \sum_{i \neq j} a_i x_i$$

where not all the $a_i$'s are zeros
So?

\[ x_j = \sum_{i \neq j} a_i x_i \]

Consider the following dichotomy:

- \( x_i \)'s with non-zero \( a_i \) get \( y_i = \text{sign}(a_i) \)
- \( x_j \) gets \( y_j = -1 \)

No perceptron can implement such dichotomy!
Why?

\[ x_j = \sum_{i \neq j} a_i x_i \implies w^T x_j = \sum_{i \neq j} a_i w^T x_i \]

If \( y_i = \text{sign}(w^T x_i) = \text{sign}(a_i) \), then \( a_i w^T x_i > 0 \)

This forces \( w^T x_j = \sum_{i \neq j} a_i w^T x_i > 0 \)

Therefore, \( y_j = \text{sign}(w^T x_j) = +1 \)
Putting it together

We proved \( d_{VC} \leq d + 1 \) and \( d_{VC} \geq d + 1 \)

\[ d_{VC} = d + 1 \]

What is \( d + 1 \) in the perceptron?

It is the number of parameters \( w_0, w_1, \ldots, w_d \)
Outline

- The definition
- VC dimension of perceptrons
- Interpreting the VC dimension
- Generalization bounds
1. Degrees of freedom

Parameters create degrees of freedom

# of parameters: *analog* degrees of freedom

\(d_{VC}\): equivalent ‘*binary*’ degrees of freedom
The usual suspects

Positive rays \( d_{VC} = 1 \):

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

Positive intervals \( d_{VC} = 2 \):

\[
h(x) = -1 \quad \quad h(x) = +1 \quad \quad h(x) = -1
\]
Not just parameters

Parameters may not contribute degrees of freedom:

\[ x \xrightarrow{} \_ \xrightarrow{} \_ \xrightarrow{} \_ \xrightarrow{} y \]

\( d_{VC} \) measures the **effective** number of parameters
2. Number of data points needed

Two small quantities in the VC inequality:

\[ P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2N} \]

If we want certain \( \epsilon \) and \( \delta \), how does \( N \) depend on \( d_{\text{VC}} \)?

Let us look at \( N^d e^{-N} \)
Fix $N^d e^{-N} = \text{small value}$

How does $N$ change with $d$?

**Rule of thumb:**

$$N \geq 10 \, d_{\text{VC}}$$
Outline

- The definition
- VC dimension of perceptrons
- Interpreting the VC dimension
- Generalization bounds
Rearranging things

Start from the VC inequality:

$$\mathbb{P}[|E_{out} - E_{in}| > \epsilon] \leq \frac{4m_\mathcal{H}(2N)e^{-\frac{1}{8}\epsilon^2 N}}{\delta}$$

Get $\epsilon$ in terms of $\delta$:

$$\delta = 4m_\mathcal{H}(2N)e^{-\frac{1}{8}\epsilon^2 N} \implies \epsilon = \sqrt{\frac{8}{N} \ln \frac{4m_\mathcal{H}(2N)}{\delta}}$$

With probability $\geq 1 - \delta$,

$$|E_{out} - E_{in}| \leq \Omega(N, \mathcal{H}, \delta)$$
Generalization bound

With probability $\geq 1 - \delta$,

$$E_{out} - E_{in} \leq \Omega$$

$$\Rightarrow$$

With probability $\geq 1 - \delta$,

$$E_{out} \leq E_{in} + \Omega$$