Machine Learning Foundations
(機器學習基石)

Lecture 6: Theory of Generalization

Hsuan-Tien Lin (林軒田)

htlin@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 5: Training versus Testing
effective price of choice in training: **(wishfully) growth function** $m_{\mathcal{H}}(N)$ with **a break point**

Lecture 6: Theory of Generalization
- Restriction of Break Point
- Bounding Function: Basic Cases
- Bounding Function: Inductive Cases
- A Pictorial Proof

3 How Can Machines Learn?

4 How Can Machines Learn Better?

The Four Break Points

growth function $m_{\mathcal{H}}(N)$: max number of dichotomies

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

  $\circ\times$ $m_{\mathcal{H}}(2) = 3 < 2^2$: break point at 2

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

  $\circ\times\circ$ $m_{\mathcal{H}}(3) = 7 < 2^3$: break point at 3

- convex sets: $\qquad\qquad\qquad\qquad\qquad\qquad\qquad m_{\mathcal{H}}(N) = 2^N$

  $\circ \quad \begin{matrix} \circ & \times \\ \times & \circ \end{matrix}$ $m_{\mathcal{H}}(N) = 2^N$ always: no break point

- 2D perceptrons: $\qquad\qquad\qquad\qquad m_{\mathcal{H}}(N) < 2^N$ **in some cases**

  $\times \begin{matrix} \circ \\ \circ \end{matrix} \times$ $m_{\mathcal{H}}(4) = 14 < 2^4$: break point at 4

break point $k \implies$ break point $k + 1, \ldots$
what else?

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = ** 3)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

1 dichotomy   , shatter any two points? **no**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = 3**)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

2 dichotomies , shatter any two points? **no**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	✕

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = 3**)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

3 dichotomies , shatter any two points? **no**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	✕
○	✕	○

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = 3**)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

4 dichotomies , shatter any two points? **yes**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	✕
○	✕	○
○	✕	✕

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = ** 3)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

4 dichotomies , shatter any two points? **no**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	✕
○	✕	○
✕	○	○

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = 3**)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

5 dichotomies , shatter any two points? **yes**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	✕
○	✕	○
✕	○	○
✕	○	✕

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = 3**)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

5 dichotomies , shatter any two points? **yes**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	✗
○	✗	○
✗	○	○
✗	✗	○

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = 3**)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

5 dichotomies , shatter any two points? **yes**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	✗
○	✗	○
✗	○	○
✗	✗	✗

Restriction of Break Point (1/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible = 3**)

maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$ and $k = 2$?

maximum possible so far: **4 dichotomies**

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	✗
○	✗	○
✗	○	○
:-(	:-(	:-(

Restriction of Break Point (2/2)

what 'must be true' when **minimum break point** $k = 2$

- $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
- $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition
  (so **maximum possible =** 3)
- $N = 3$: **maximum possible =** $4 \ll 2^3$

—break point k **restricts maximum possible** $m_{\mathcal{H}}(N)$ **a lot** for $N > k$

idea: $m_{\mathcal{H}}(N)$
  $\leq$ maximum possible $m_{\mathcal{H}}(N)$ given k
  $\leq$ $poly(N)$

Fun Time

When minimum break point $k = 1$, what is the maximum possible $m_{\mathcal{H}}(N)$ when $N = 3$?

1 1   **2** 2   **3** 4   **4** 8

Reference Answer: ①

Because $k = 1$, the hypothesis set cannot even shatter one point. Thus, every 'column' of the table cannot contain both ○ and ×. Then, after including the first dichotomy, it is not possible to include any other different dichotomy. Thus, the maximum possible $m_{\mathcal{H}}(N)$ is 1.

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	×	○
○	×	×

Bounding Function

bounding function $B(N, k)$:

maximum possible $m_{\mathcal{H}}(N)$ when break point = k

- combinatorial quantity:
  maximum number of length-N vectors with $(\circ, \times)$
  while **'no shatter' any length-k** subvectors
- irrelevant of the details of $\mathcal{H}$
  e.g. $B(N, 3)$ bounds both
  - positive intervals ($k = 3$)
  - 1D perceptrons ($k = 3$)

new goal: $B(N, k) \leq poly(N)$**?**

Table of Bounding Function (1/4)

					k			
$B(N,k)$		1	2	3	4	5	6	...
	1							
	2		**3**					
	3		**4**					
N	4							
	5							
	6							
	⋮							

Known

- $B(2,2) = 3$ (maximum < 4)
- $B(3,2) = 4$ ('pictorial' proof previously)

Table of Bounding Function (2/4)

	$B(N, k)$	1	2	3	4	5	6	...
	1	**1**						
	2	**1**	3					
	3	**1**	4					
N	4	**1**						
	5	**1**						
	6	**1**						
	⋮	**⋮**						

(column header k spans columns 1–6)

Known

- $B(N, 1) = 1$ (see previous quiz)

Table of Bounding Function (3/4)

					k			
	$B(N, k)$	1	2	3	4	5	6	...
	1	1	**2**	**2**	**2**	**2**	**2**	...
	2	1	3	**4**	**4**	**4**	**4**	...
	3	1	4		**8**	**8**	**8**	...
N	4	1				**16**	**16**	...
	5	1					**32**	...
	6	1						...
	⋮	⋮						

Known

- $B(N, k) = 2^N$ for $N < k$

  —including all dichotomies not violating 'breaking condition'

Table of Bounding Function (4/4)

	$B(N, k)$	1	2	3	4	5	6	...
				k				
	1	**1**	2	2	2	2	2	...
	2	1	**3**	4	4	4	4	...
	3	1	4	**7**	8	8	8	...
N	4	1			**15**	16	16	...
	5	1				**31**	32	...
	6	1					**63**	...
	⋮	⋮						⋱

Known

- $B(N, k) = 2^N - 1$ for $N = k$
  —**removing a single dichotomy** satisfies 'breaking condition'

more than halfway done! :-)

Fun Time

For the 2D perceptrons, which of the following claim is true?

1 minimum break point $k = 2$

2 $m_{\mathcal{H}}(4) = 15$

3 $m_{\mathcal{H}}(N) < B(N, k)$ when $N = k =$ minimum break point

4 $m_{\mathcal{H}}(N) > B(N, k)$ when $N = k =$ minimum break point

Reference Answer: ③

As discussed previously, minimum break point for 2D perceptrons is 4, with $m_{\mathcal{H}}(4) = 14$. Also, note that $B(4, 4) = 15$. So bounding function $B(N, k)$ can be 'loose' in bounding $m_{\mathcal{H}}(N)$.

Estimating $B(4, 3)$

	$B(N, k)$	1	2	3	4	5	6	...
	1	1	2	2	2	2	2	...
	2	1	3	4	4	4	4	...
	3	1	4	7	8	8	8	...
N	4	1		**?**	15	16	16	...
	5	1				31	32	...
	6	1					63	...
	⋮	⋮						⋱

The column header k spans columns 1 through 6.

Motivation

- $B(4, 3)$ shall be related to $B(3, ?)$
  —'adding' one point from $B(3, ?)$

next: reduce $B(4, 3)$ to $B(3, ?)$

'Achieving' Dichotomies of $B(4, 3)$

after checking all 2^{2^4} sets of dichotomies, **the winner is** . . .

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
01	○	○	○	○
02	✗	○	○	○
03	○	✗	○	○
04	○	○	✗	○
05	○	○	○	✗
06	✗	✗	○	✗
07	✗	○	✗	○
08	✗	○	○	✗
09	○	✗	✗	○
10	○	✗	○	✗
11	○	○	✗	✗

					k		
	$B(N, k)$	1	2	3	4	5	6
	1	1	2	2	2	2	2
	2	1	3	4	4	4	4
	3	1	4	7	8	8	8
N	4	1		**11**	15	16	16
	5	1				31	32
	6	1					63

how to reduce $B(4, 3)$ to $B(3, ?)$ cases?

Reorganized Dichotomies of $B(4, 3)$

after checking all 2^{2^4} sets of dichotomies, **the winner is** . . .

	x_1	x_2	x_3	x_4
01	○	○	○	○
02	✗	○	○	○
03	○	✗	○	○
04	○	○	✗	○
05	○	○	○	✗
06	✗	✗	○	✗
07	✗	○	✗	○
08	✗	○	○	✗
09	○	✗	✗	○
10	○	✗	○	✗
11	○	○	✗	✗

$\Longrightarrow$

	x_1	x_2	x_3	x_4
01	○	○	○	○
05	○	○	○	✗
02	✗	○	○	○
08	✗	○	○	✗
03	○	✗	○	○
10	○	✗	○	✗
04	○	○	✗	○
11	○	○	✗	✗
06	✗	✗	○	✗
07	✗	○	✗	○
09	○	✗	✗	○

orange: pair; **purple**: single

Estimating Part of $B(4, 3)$ (1/2)

$$B(4, 3) = 11 = 2\alpha + \beta$$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
α	○	○	○
	×	○	○
	○	×	○
	○	○	×
β	×	×	○
	×	○	×
	○	×	×

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
2α	○	○	○	○
	○	○	○	×
	×	○	○	○
	×	○	○	×
	○	×	○	○
	○	×	○	×
	○	○	×	○
	○	○	×	×
β	×	×	○	×
	×	○	×	○
	○	×	×	○

- $\alpha + \beta$: dichotomies on $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$
- $B(4, 3)$ 'no shatter' any 3 inputs
  $\Longrightarrow \alpha + \beta$ 'no shatter' any 3

$$\alpha + \beta \leq B(3, 3)$$

Estimating Part of $B(4, 3)$ (2/2)

$$B(4, 3) = 11 = 2\alpha + \beta$$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
	○	○	○
α	✕	○	○
	○	✕	○
	○	○	✕

- α: dichotomies on $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ with $\mathbf{x}_4$ **paired**
- $B(4, 3)$ 'no shatter' any 3 inputs $\implies \alpha$ 'no shatter' any 2

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
	○	○	○	○
	○	○	○	✕
	✕	○	○	○
2α	✕	○	○	✕
	○	✕	○	○
	○	✕	○	✕
	○	○	✕	○
	○	○	✕	✕
	✕	✕	○	✕
β	✕	○	✕	○
	○	✕	✕	○

$$\alpha \leq B(3, 2)$$

Putting It All Together

$$B(4,3) = 2\alpha + \beta$$
$$\alpha + \beta \leq B(3,3)$$
$$\alpha \leq B(3,2)$$
$$\Rightarrow B(4,3) \leq B(3,3) + B(3,2)$$

				k			
	$B(N,k)$	1	2	3	4	5	6
	1	1	2	2	2	2	2
	2	1	3	4	4	4	4
	3	1	4	7	8	8	8
N	4	1	≤ 5	11	15	16	16
	5	1	≤ 6	≤ 16	≤ 26	31	32
	6	1	≤ 7	≤ 22	≤ 42	≤ 57	63

now have **upper bound** of bounding function

Putting It All Together

$$
\begin{aligned}
B(N, k) &= 2\alpha + \beta \\
\alpha + \beta &\leq B(N-1, k) \\
\alpha &\leq B(N-1, k-1) \\
\Rightarrow B(N, k) &\leq B(N-1, k) + B(N-1, k-1)
\end{aligned}
$$

	$B(N, k)$	1	2	3	k 4	5	6
	1	1	2	2	2	2	2
	2	1	3	4	4	4	4
	3	1	4	7	8	8	8
N	4	1	≤ 5	11	15	16	16
	5	1	≤ 6	≤ 16	≤ 26	31	32
	6	1	≤ 7	≤ 22	≤ 42	≤ 57	63

now have **upper bound** of bounding function

Bounding Function: The Theorem

$$B(N, k) \leq \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{highest term } N^{k-1}}$$

- simple induction using **boundary and inductive formula**
- for fixed k, $B(N, k)$ upper bounded by *poly*(N)
  $\implies m_{\mathcal{H}}(N)$ **is** *poly*(N) **if break point exists**

'$\leq$' can be '$=$' actually,
go play and prove it if math lover! :-)

The Three Break Points

$$B(N, k) \leq \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{highest term } N^{k-1}}$$

- positive rays: $m_{\mathcal{H}}(N) = N + 1 \leq N + 1$
  - $\circ\times$ $m_{\mathcal{H}}(2) = 3 < 2^2$: break point at 2
- positive intervals: $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 \leq \frac{1}{2}N^2 + \frac{1}{2}N + 1$
  - $\circ\times\circ$ $m_{\mathcal{H}}(3) = 7 < 2^3$: break point at 3
- 2D perceptrons: $m_{\mathcal{H}}(N) = ? \leq \frac{1}{6}N^3 + \frac{5}{6}N + 1$

  $\times \,{}^{\circ}_{\circ}\, \times$ $m_{\mathcal{H}}(4) = 14 < 2^4$: break point at 4

can bound $m_{\mathcal{H}}(N)$ by only **one break point**

Fun Time

For 1D perceptrons (positive and negative rays), we know that $m_{\mathcal{H}}(N) = 2N$. Let k be the minimum break point. Which of the following is not true?

1. $k = 3$

2. for some integers $N > 0$, $m_{\mathcal{H}}(N) = \sum_{i=0}^{k-1} \binom{N}{i}$

3. for all integers $N > 0$, $m_{\mathcal{H}}(N) = \sum_{i=0}^{k-1} \binom{N}{i}$

4. for all integers $N > 2$, $m_{\mathcal{H}}(N) < \sum_{i=0}^{k-1} \binom{N}{i}$

Reference Answer: ③

The proof is generally trivial by listing the definitions. For ②, $N = 1$ or 2 gives the equality. One thing to notice is ④: the upper bound can be 'loose'.

BAD Bound for General $\mathcal{H}$

want:

$$\mathbb{P}\Big[\exists h \in \mathcal{H} \text{ s.t. } \big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \epsilon\Big] \leq 2 \quad m_{\mathcal{H}}(\ N) \cdot \exp\left(-2 \qquad \epsilon^2 N\right)$$

actually, when *N* large enough,

$$\mathbb{P}\Big[\exists h \in \mathcal{H} \text{ s.t. } \big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \epsilon\Big] \leq 2 \cdot 2 m_{\mathcal{H}}(2N) \cdot \exp\left(-2 \cdot \frac{1}{16} \epsilon^2 N\right)$$

next: **sketch** of proof

Step 1: Replace E_{out} by E'_{in}

$$\frac{1}{2}\mathbb{P}\Big[\exists h \in \mathcal{H} \text{ s.t. } \big|E_{in}(h) - E_{out}(h)\big| > \epsilon\Big]$$

$$\leq \quad \mathbb{P}\Big[\exists h \in \mathcal{H} \text{ s.t. } \big|E_{in}(h) - E'_{in}(h)\big| > \frac{\epsilon}{2}\Big]$$

- $E_{in}(h)$ finitely many, $E_{out}(h)$ infinitely many —replace the evil E_{out} first
- how? sample verification set $\mathcal{D}'$ of size N to calculate E'_{in}
- BAD h of $E_{in} - E_{out}$ $\overset{\text{probably}}{\Longrightarrow}$ BAD h of $E_{in} - E'_{in}$
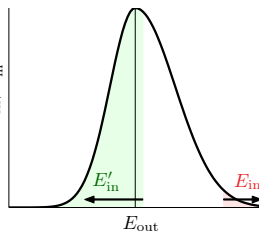
evil E_{out} removed by
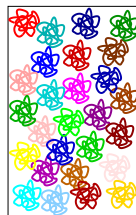verification with '**ghost data**'

Step 2: Decompose $\mathcal{H}$ by Kind

$$\begin{aligned}
\text{BAD} &\leq 2\mathbb{P}\Big[\exists h \in \mathcal{H} \text{ s.t. } \big|E_{\text{in}}(h) - E'_{\text{in}}(h)\big| > \frac{\epsilon}{2}\Big] \\
&\leq 2m_{\mathcal{H}}(2N)\mathbb{P}\Big[\text{fixed } h \text{ s.t. } \big|E_{\text{in}}(h) - E'_{\text{in}}(h)\big| > \frac{\epsilon}{2}\Big]
\end{aligned}$$

- E_{in} with $\mathcal{D}$, E'_{in} with $\mathcal{D}'$
  —now $m_{\mathcal{H}}$ comes to play
- how? infinite $\mathcal{H}$ becomes
  $|\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}'_1, \ldots, \mathbf{x}'_N)|$
  kinds
- union bound on $m_{\mathcal{H}}(2N)$ kinds
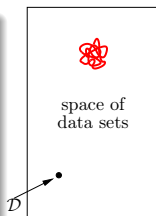
(a) Hoeffding Inequality   (b) Union Bound   (c) Now

space of
data sets

$\mathcal{D}$

use $m_{\mathcal{H}}(2N)$ to calculate BAD-overlap properly

Step 3: Use Hoeffding without Replacement

$$
\begin{aligned}
\text{BAD} &\leq 2m_{\mathcal{H}}(2N)\mathbb{P}\left[\text{fixed } h \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right] \\
&\leq 2m_{\mathcal{H}}(2N) \cdot 2\exp\left(-2\left(\frac{\epsilon}{4}\right)^2 N\right)
\end{aligned}
$$

- consider bin of $2N$ examples, choose N for E_{in}, leave others for E'_{in}
  $|E_{\text{in}} - E'_{\text{in}}| > \frac{\epsilon}{2} \Leftrightarrow \left|E_{\text{in}} - \frac{E_{\text{in}} + E'_{\text{in}}}{2}\right| > \frac{\epsilon}{4}$
- so? just 'smaller bin', 'smaller ϵ', and Hoeffding without replacement

sample for E_{in}

small bin
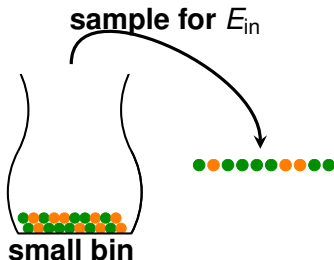
use Hoeffding after zooming to fixed h

That's All!

Vapnik-Chervonenkis (VC) bound:

$$\mathbb{P}\Big[\exists h \in \mathcal{H} \text{ s.t. } \big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \epsilon\Big]$$

$$\leq \quad 4m_{\mathcal{H}}(2N) \exp\left(-\frac{1}{8}\epsilon^2 N\right)$$

- replace E_{out} by E_{in}'
- decompose $\mathcal{H}$ by kind
- use Hoeffding without replacement

2D perceptrons:
- break point? 4
- $m_{\mathcal{H}}(N)$? $O(N^3)$

learning with 2D perceptrons feasible! :-)

Summary

1 When Can Machines Learn?

2 **Why** Can Machines Learn?

> ### Lecture 5: Training versus Testing
>
> ### Lecture 6: Theory of Generalization
>
> - Restriction of Break Point
>   **break point 'breaks' consequent points**
> - Bounding Function: Basic Cases
> $B(N, k)$ **bounds** $m_{\mathcal{H}}(N)$ **with break point** k
> - Bounding Function: Inductive Cases
> $B(N, k)$ **is** $poly(N)$
> - A Pictorial Proof
> $m_{\mathcal{H}}(N)$ **can replace** M **with a few changes**

- **next: how to 'use' the break point?**

3 How Can Machines Learn?

4 How Can Machines Learn Better?