Lecture 6: Theory of Generalization

Presentation · April 2019

1 author:

Alaa Tharwat
Fachhochschule Bielefeld

Lecture 6: Theory of Generalization

Alaa Tharwat

Lecture 6: Theory of Generalization

- Review of Lecture 5
- Proof $m_H(N)$ is polynomial
- VC boundary

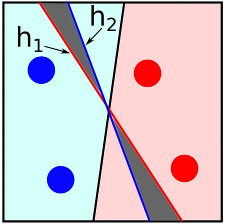
Lecture 6: Theory of Generalization

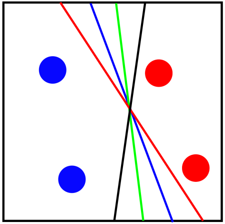
- Review of Lecture 5
- Proof $m_H(N)$ is polynomial
- VC boundary

- Bad events are very overlapping
- Dichotomies are mini-hypotheses: $h(x_1, x_2, \ldots, x_N) \rightarrow \{-1, +1\}$
- number of dichotomies $|H(x_1, x_2, \ldots, x_N)|$ is at most 2^N, this is called **growth function**, and it is denoted by $m_H(N)$
- Instead of using M, use $m_H(N)$ which is polynomial (the proof is in this lecture)

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

- No break point $\Rightarrow m_H(N) = 2^N$
- Any break point $\Rightarrow m_H(N)$ is $\boxed{\textbf{polynomial} \text{ in } N}$





Lecture 6: Theory of Generalization

- Review of Lecture 5
- Proof $m_H(N)$ is polynomial
- VC boundary

- To show that $m_H(N)$ is polynomial, means that $m_H(N) \leq \cdots \leq \ldots$ a polynomial
- The term $B(N,k)$ represents the maximum number of dichotomies on N points, with break point k, where B denotes the binomial. Or, maximum number of rows given N points and no k columns have all possible patterns
- In the table, $B(N,k) = \alpha + 2\beta$
- In the S_1 group, all samples/rows appear once from x_1 to x_N
- Each row in S_2^+ appears also in S_2^-, but the difference is in x_N. Hence, from x_1 to x_{N-1}, S_2^+ and S_2^- are identical
- What is α and β?

		# of rows	x_1	x_2	....	x_{N-1}	x_N
	S_1	α	+1	+1	....	+1	+1
			-1	+1	....	+1	-1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	-1	-1
			-1	+1	....	-1	+1
S_2	S_2^+	β	+1	-1	....	+1	+1
			-1	-1	....	+1	+1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	+1	+1
			-1	-1	....	-1	+1
	S_2^-	β	+1	-1	....	+1	-1
			-1	-1	....	+1	-1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	+1	-1
			-1	-1	....	-1	-1

- For the columns $(x_1, x_2, \dots, x_{N-1})$, there are $\alpha + \beta$ different dichotomies
- As mentioned before, no subset of size k can be shattered; hence, $\alpha + \beta \leq B(N-1, k)$

		# of rows	x_1	x_2	....	x_{N-1}	x_N
	S_1	α	+1	+1	....	+1	+1
			-1	+1	....	+1	-1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	-1	-1
			-1	+1	....	-1	+1
S_2	S_2^+	β	+1	-1	....	+1	+1
			-1	-1	....	+1	+1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	+1	+1
			-1	-1	....	-1	+1
	S_2^-	β	+1	-1	....	+1	-1
			-1	-1	....	+1	-1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	+1	-1
			-1	-1	....	-1	-1

- S_2 has $S_2^+ \cup S_2^-$ rows, i.e. $\beta + \beta$ dichotomies
- β dichotomies on $(x_1, x_2, \ldots, x_{N-1})$ with x_N paired
- $B(N,k)$ 'no shatter' any k inputs $\Rightarrow$ β 'no shatter' $k-1$ inputs
- $\beta \leq B(N-1, k-1)$

		# of rows	x_1	x_2	....	x_{N-1}	x_N
	S_1	α	+1	+1	....	+1	+1
			-1	+1	....	+1	-1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	-1	-1
			-1	+1	....	-1	+1
S_2	S_2^+	β	+1	-1	....	+1	+1
			-1	-1	....	+1	+1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	+1	+1
			-1	-1	....	-1	+1
	S_2^-	β	+1	-1	....	+1	-1
			-1	-1	....	+1	-1
			⋮	⋮	⋮	⋮	⋮
			+1	-1	....	+1	-1
			-1	-1	....	-1	-1

- $B(N,k) = \alpha + 2\beta$
- $\alpha + \beta \leq B(N-1,k)$
- $\beta \leq B(N-1,k-1)$

$$B(N,k) \leq B(N-1,k) + B(N-1,k-1)$$

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i} \text{(The proof next slides)}$$

B(4,3)=11

x_1	x_2	x_3	x_4
-1	-1	-1	-1
-1	-1	-1	+1
-1	-1	+1	-1
-1	-1	+1	+1
-1	+1	-1	-1
-1	+1	-1	+1
-1	+1	+1	-1
-1	+1	+1	+1
+1	-1	-1	-1
+1	-1	-1	+1
+1	-1	+1	-1
+1	-1	+1	+1
+1	+1	-1	-1
+1	+1	-1	+1
+1	+1	+1	-1
+1	+1	+1	+1

	# of rows	x_1	x_2	x_3	x_4
S_1	α	-1 +1 +1	+1 -1 +1	+1 +1 -1	-1 -1 -1
S_2^+	β	-1 -1 -1 +1	-1 -1 +1 -1	-1 +1 -1 -1	-1 -1 -1 -1
S_2^-	β	-1 -1 -1 +1	-1 -1 +1 -1	-1 +1 -1 -1	+1 +1 +1 +1

≤ B(3,3)=7 + B(3,2)=4

x_1	x_2	x_3
-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1

x_1	x_2	x_3
-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1

$$
\begin{aligned}
B(N,k) &\le B(N-1,k) + B(N-1,k-1) \\
&\le \sum_{i=0}^{k-1} \binom{N-1}{i} + \sum_{i=0}^{k-2} \binom{N-1}{i} \\
&\le 1 + \sum_{i=1}^{k-1} \binom{N-1}{i} + \sum_{i=1}^{k-1} \binom{N-1}{i-1} \\
&\le 1 + \sum_{i=1}^{k-1} \left[\binom{N-1}{i} + \binom{N-1}{i-1} \right] \\
&\le 1 + \sum_{i=1}^{k-1} \binom{N-1+1}{i} = \sum_{i=0}^{k-1} \binom{N}{i}
\end{aligned}
$$

Picking i object from N distinct objects is ($B(N,i) = \binom{N}{i}$)

- a specific object is included $B(N-1,i-1) = \binom{N-1}{i-1}$
- a specific object is excluded $B(N-1,i) = \binom{N-1}{i}$ (so, you still have to choose i objects from $N-1$ objects)

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

- The maximum power is $N^{k-1} \Rightarrow$ polynomial
- The solution of problem 2.5 proves that $m_H(N) \leq N^{k-1} + 1$ or $m_H(N) \leq N^{d_{VC}} + 1$

- If $N = 1$ means that we have one point and hence at maximum we have two dichotomies (initial condition, first row)
- If $k = 1$ means we are not allowed to have two different patterns for each column. Hence, we have one unique value for each column; i.e., one sample (initial condition, first column)
- $k \geq 2$, so we have two or more unique samples
- On board ($N = 3$ and $k = 2$, this is the break point for the 2D perceptron)

N-1,k-1 N-1,k

k

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	...
4	1	5	11	...	...	...	...
5	1	6	...				
6	1	7	...				
⋮	⋮	⋮	⋮				

N,k

- The figure below shows all the possible dichotomies, i.e., without adding the breakpoint restriction, when $N = 4$
- Assume the breakpoint is two, i.e. $k = 2$. This means that given any two points, we cannot generate all dichotomies. This is the reason why the highlighted rows cannot be generated and hence we have only five rows after adding that restriction (i.e. $k = 2$) $B(4,2) \leq \sum_{i=0}^{2-1} \binom{4}{i} = \binom{4}{0} + \binom{4}{1} = 5$, *and the number of rows/dichotomies is already five*

x_1	x_2	x_3	x_4
-1	-1	-1	-1
-1	-1	-1	+1
-1	-1	+1	-1
-1	-1	+1	+1
-1	+1	-1	-1
-1	+1	-1	+1
-1	+1	+1	-1
-1	+1	+1	+1
+1	-1	-1	-1
+1	-1	-1	+1
+1	-1	+1	-1
+1	-1	+1	+1
+1	+1	-1	-1
+1	+1	-1	+1
+1	+1	+1	-1
+1	+1	+1	+1

	# of rows	x_1	x_2	x_3	x_4
S_1	α	-1	-1	+1	-1
		-1	+1	-1	-1
		+1	-1	-1	-1
S_2^+	β	-1	-1	-1	+1
S_2^-	β	-1	-1	-1	-1

- H is positive rays: break point $k = 2$,

$$m_H(N) = N + 1 \leq \sum_{i=0}^{2-1} \binom{N}{i}$$

$$\text{where } \sum_{i=0}^{2-1} \binom{N}{i} = \cancelto{1}{\binom{N}{0}} + \cancelto{N}{\binom{N}{1}} = N + 1$$

- H is positive intervals: break point $k=3$,

$$m_H(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 \leq \sum_{i=0}^{3-1} \binom{N}{i}$$

where $$\sum_{i=0}^{3-1} \binom{N}{i} = \cancelto{1}{\binom{N}{0}} + \cancelto{N}{\binom{N}{1}} + \cancelto{\frac{N(N-1)}{2}}{\binom{N}{2}} = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

- H is 2D perceptron: break point $k=4$,

$$m_H(N) = ? \leq \sum_{i=0}^{4-1} \binom{N}{i}$$

where $$\sum_{i=0}^{4-1} \binom{N}{i} = \cancelto{1}{\binom{N}{0}} + \cancelto{N}{\binom{N}{1}} + \cancelto{\frac{N(N-1)}{2}}{\binom{N}{2}} + \cancelto{\frac{N(N-1)(N-1)}{6}}{\binom{N}{3}} = \frac{1}{6}N^3 + \frac{5}{6}N + 1$$

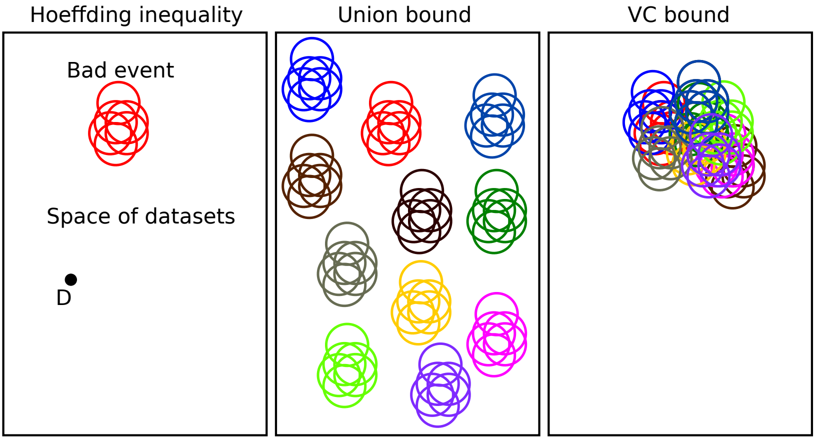
Lecture 6: Theory of Generalization

- Review of Lecture 5
- Proof $m_H(N)$ is polynomial
- VC boundary

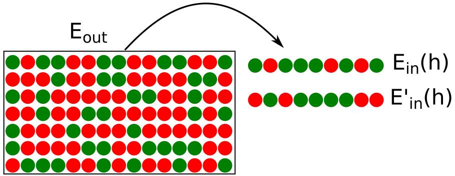
$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

We need to replace M with $m_H(N)$

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2m_H(N)e^{-2\epsilon^2 N}$$


Hoeffding inequality
Bad event
Space of datasets
D
Union bound
VC bound

- With multiple bins, the tracking between E_{in} and E_{out} becomes looser and looser
- Instead of one sample, we can take two samples D and D', and E'_{in} is the in-sample error for D'
- E_{in} and E'_{in} track E_{out} and hence E_{in} tracks E'_{in}



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

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

This is called, the $\boxed{\text{Vapnik-Chervonenkis}}$ inequality