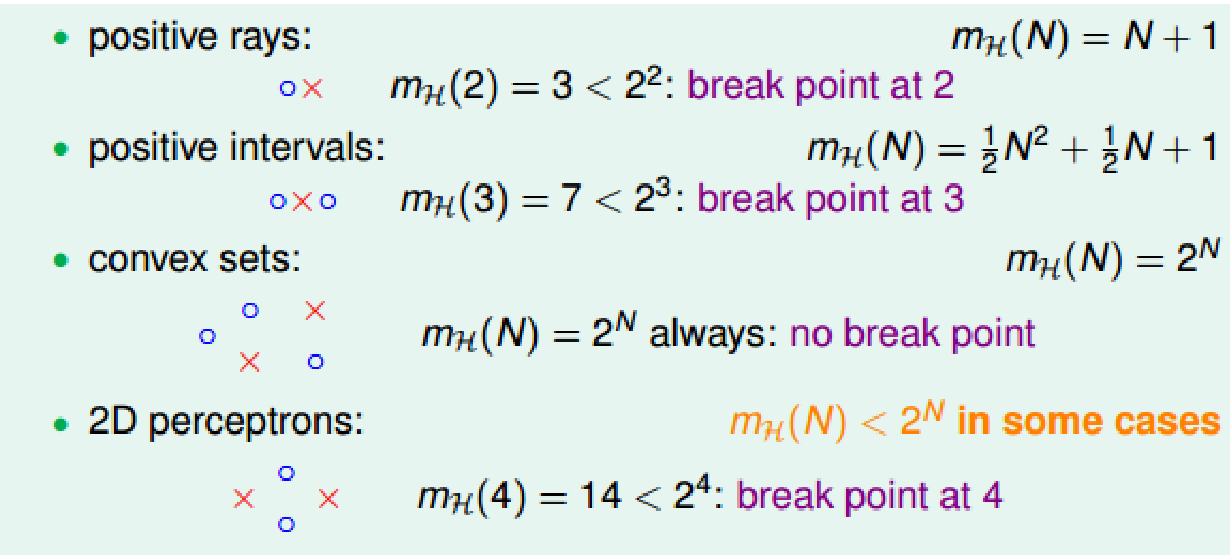
台湾大学林轩田机器学习基石课程学习笔记6 -- Theory of Generalization

> 作者：红色石头
> 微信公众号：AI有道（ID：redstonewill）

上一节课，我们主要探讨了当M的数值大小对机器学习的影响。如果M很大，那么就不能保证机器学习有很好的泛化能力，所以问题转换为验证M有限，即最好是按照多项式成长。然后通过引入了成长函数$m_H(N)$和dichotomy以及break point的概念，提出2D perceptrons的成长函数$m_H(N)$是多项式级别的猜想。这就是本节课将要深入探讨和证明的内容。

一、Restriction of Break Point

我们先回顾一下上节课的内容，四种成长函数与break point的关系：



下面引入一个例子，如果k=2，那么当N取不同值的时候，计算其成长函数$m_H(N)$是多少。很明显，当N=1时，$m_H(N)$=2,；当N=2时，由break point为2可知，任意两点都不能被shattered（shatter的意思是对N个点，能够分解为2^N种dichotomies）；$m_H(N)$最大值只能是3；当N=3时，简单绘图分析可得其$m_H(N) = 4$，即最多只有4种dichotomies。

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 $N > k$

所以，我们发现当N>k时，break point限制了$m_H(N)$值的大小，也就是说影响成长函数$m_H(N)$的因素主要有两个：

- 抽样数据集N
- break point k（这个变量确定了假设的类型）

那么，如果给定N和k，能够证明其$m_H(N)$的最大值的上界是多项式的，则根据霍夫丁不等式，就能用$m_H(N)$代替M，得到机器学习是可行的。所以，证明$m_H(N)$的上界是poly(N)，是我们的目标。

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

二、Bounding Function: Basic Cases

现在，我们引入一个新的函数：bounding function，B(N,k)。Bound Function指的是当break point为k的时候，成长函数$m_H(N)$可能的最大值。也就是说B(N,k)是$m_H(N)$的上界，对应$m_H(N)$最多有多少种dichotomy。那么，我们新的目标就是证明：

$$B(N,k) \leq poly(N)$$

这里值得一提的是，B(N,k)的引入不考虑是1D postive intrervals问题还是2D perceptrons问题，而只关心成长函数的上界是多少，从而简化了问题的复杂度。

bounding function $B(N, k)$:
maximum possible $m_{\mathcal{H}}(N)$ when break point = k

- combinatorial quantity:
  maximum number of length-N vectors with (∘, ×)
  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 ($k = 3$)

求解B(N,k)的过程十分巧妙：

- 当k=1时，B(N,1)恒为1。
- 当N < k时，根据break point的定义，很容易得到$B(N,k)=2^N$。
- 当N = k时，此时N是第一次出现不能被shatter的值，所以最多只能有2^N-1个dichotomies，则$B(N,k)=2^N-1$。

	$B(N,k)$	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	...
	⋮	⋮						⋱

到此，bounding function的表格已经填了一半了，对于最常见的N>k的情况比较复杂，推导过程下一小节再详细介绍。

三、Bounding Function: Inductive Cases

N > k的情况较为复杂，下面给出推导过程：

以B(4,3)为例，首先想着能否构建B(4,3)与B(3,x)之间的关系。

首先，把B(4,3)所有情况写下来，共有11组。也就是说再加一种dichotomy，任意三点都能被shattered，11是极限。

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

对这11种dichotomy分组，目前分成两组，分别是orange和purple，orange的特点是，x1,x2和x3是一致的，x4不同并成对，例如1和5，2和8等，purple则是单一的，x1,x2,x3都不同，如6,7,9三组。

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

$\Rightarrow$

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

orange: pair; **purple**: single

将Orange去掉x4后去重得到4个不同的vector并成为α，相应的purple为β。那么$B(4,3) = 2\alpha + \beta$，这个是直接转化。紧接着，由定义，B(4,3)是不能允许任意三点shatter的，所以由α和β构成的所有三点组合也不能shatter（alpha经过去重），即$\alpha + \beta \leq B(3,3)$。

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

- $\alpha + \beta$: 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

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

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

另一方面，由于α中x4是成对存在的，且α是不能被任意三点shatter的，则能推导出α是不能被任意两点shatter的。这是因为，如果α是不能被任意两点shatter，而x4又是成对存在的，那么x1、x2、x3、x4组成的α必然能被三个点shatter。这就违背了条件的设定。这个地方的推导非常巧妙，也解释了为什么会这样分组。此处得到的结论是 $\alpha \leq B(3,2)$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
α	o	o	o
	×	o	o
	o	×	o
	o	o	×

- 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
  $\Longrightarrow \alpha$ 'no shatter' any 2

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

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

由此得出B(4,3)与B(3,x)的关系为：

$$
\begin{aligned}
B(4,3) &= 2\alpha + \beta \\
\alpha + \beta &\leq B(3,3) \\
\alpha &\leq B(3,2) \\
\Rightarrow B(4,3) &\leq B(3,3) + B(3,2)
\end{aligned}
$$

最后，推导出一般公式为：

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

根据推导公式，下表给出B(N,K)值

	$B(N,k)$	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

根据递推公式，推导出B(N,K)满足下列不等式：

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

上述不等式的右边是最高阶为k-1的N多项式，也就是说成长函数$m_H(N)$的上界B(N,K)的上界满足多项式分布poly(N)，这就是我们想要得到的结果。

得到了$m_H(N)$的上界B(N,K)的上界满足多项式分布poly(N)后，我们回过头来看看之前介绍的几种类型它们的$m_H(N)$与break point的关系：

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

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

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

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

- 2D perceptrons: $m_{\mathcal{H}}(N) = ? \leq \frac{1}{6}N^3 + \frac{5}{6}N + 1$

  ×∘∘× $m_{\mathcal{H}}(4) = 14 < 2^4$: break point at 4

我们得到的结论是，对于2D perceptrons，break point为k=4，$m_H(N)$的上界是N^{k-1}。推广一下，也就是说，如果能找到一个模型的break point，且是有限大的，那么就能推断出其成长函数$m_H(N)$有界。

四、A Pictorial Proof

我们已经知道了成长函数的上界是poly(N)的，下一步，如果能将$m_H(N)$代替M，代入到Hoffding不等式中，就能得到$E_{out} \approx E_{in}$的结论：

want:

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

实际上并不是简单的替换就可以了，正确的表达式为：

actually, when N large enough,

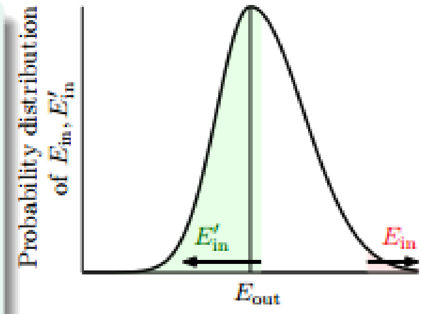
$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \leq 2 \cdot 2 m_{\mathcal{H}}(2N) \cdot \exp\left(-2 \cdot \frac{1}{16} \epsilon^2 N\right)$$

该推导的证明比较复杂，我们可以简单概括为三个步骤来证明：

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 $E_{in} - E'_{in}$

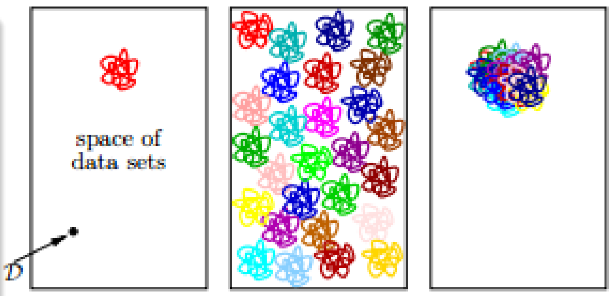


evil E_{out} removed by
verification with '**ghost data**'

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

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

- 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



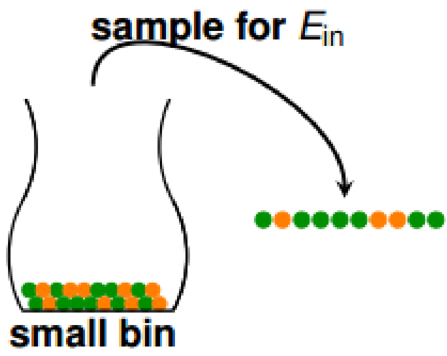
(a) Hoeffding Inequality (b) Union Bound (c) Now

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

Step 3: Use Hoeffding without Replacement

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

- 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



use Hoeffding after zooming to fixed h

这部分内容，我也只能听个大概内容，对具体的证明过程有兴趣的童鞋可以自行研究一下，研究的结果记得告诉一下我哦。

最终，我们通过引入成长函数m_H，得到了一个新的不等式，称为Vapnik-Chervonenkis(VC) bound：

Vapnik-Chervonenkis (VC) bound:

$$\begin{aligned}&\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \\ \le\ & 4m_{\mathcal{H}}(2N)\exp\left(-\frac{1}{8}\epsilon^2 N\right)\end{aligned}$$

对于2D perceptrons，它的break point是4，那么成长函数$m_H(N) = O(N^3)$。所以，我们可以说2D perceptrons是可以进行机器学习的，只要找到hypothesis能让$E_{in} \approx 0$，就能保证$E_{in} \approx E_{out}$。

五、总结

本节课我们主要介绍了只要存在break point，那么其成长函数$m_H(N)$就满足poly(N)。推导过程是先引入$m_H(N)$的上界B(N,k)，B(N,k)的上界是N的k-1阶多项式，从而得到$m_H(N)$的上界就是N的k-1阶多项式。然后，我们通过简单的三步证明，将$m_H(N)$代入了Hoffding不等式中，推导出了Vapnik-Chervonenkis(VC) bound，最终证明了只要break point存在，那么机器学习就是可行的。

注明：

文章中所有的图片均来自台湾大学林轩田《机器学习基石》课程。