Lecture 6　Theory of Generalization

整理者: LobbyBoy*　2020年2月19日

1. Restriction of Break Point

在上一章的后半段，我们针对二元分类问题下的假说集合，定义了假说集合的“增长函数”，记作$m_{\mathcal{H}}(N)$，表示该假说集合最多能够将N个样本点划分出多少种dichotomies，即“对分”。对于2D感知机，我们目前已知的信息是：$m_{\mathcal{H}}(N) < 2^N$，因为从N=4开始，无论如何它都不能将输入集合shatter掉，例如N=4时，增长函数的值便为14，而不是16。我们把N=4称为2D感知机的一个“断点(break point)”，也是最小的断点。下面我们要研究的是在(最小)断点之后增长函数的增长情况。有一点我们是明确的，那就是在(最小)断点之后，增长函数并不再以指数型的方式增长，其增长的速度会下降。

为了使结论更具一般性，我们仅假设存在一个假说集合，其(最小)断点为2，除此之外没有其他信息。我们将根据$k = 2$这条信息进行一系列推导。

首先，根据断点的定义，我们知道$m_{\mathcal{H}}(1) = 2$，$m_{\mathcal{H}}(2) < 4$。因此，$m_{\mathcal{H}}(2)$最大可能值为3，也就是说该假说集合无法shatter掉任何两个样本点。当N=3时，情况如何？我们可以采用如下方法进行演绎：每写一种dichotomy，然后回头检查已存在的dichotomies有没有违反“假说集合不能打散任何2个样本点”的前提条件：如果没有，则纳入该dichotomy；如果违反，则放弃该dichotomy。最后看看最多可以写出多少种dichotomies，将其作为N=3时增长函数的上界。

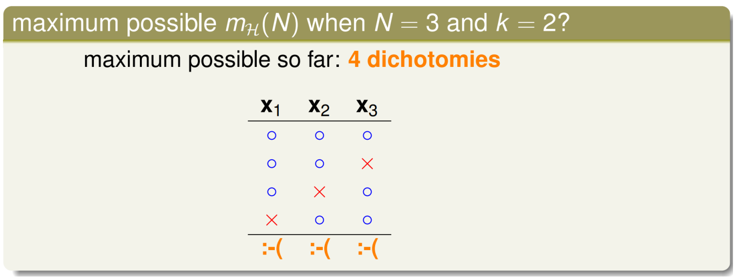

图 1: N=3, k=2

*本笔记根据台湾大学林轩田教授于线上教育平台Coursera开设的“机器学习基石”课程整理而成(课程内容见: https://www.coursera.org/learn/ntumlone-mathematicalfoundations/home/welcome)。笔记内的大多数图片来自于林老师的课程slides。感谢林老师能够将如此精彩的课程通过线上平台同所有人分享，thanks!

通过尝试我们发现，在N=3时，增长函数的上界为4，比$2^3 = 8$小了很多。当N=2时，增长函数的上界为3，只比$2^2 = 4$小1。可见发现，在(最小)断点之后，增长函数的上界被“限制”住了，且限制的力度仿佛随着N的增大而越来越显著。

最后，我们将研究方向转向增长函数的上界。理由很简单，因为研究增长函数本身，需要结合具体的某一个假说集合，不同的假说集合一般具有不同的增长函数；但研究增长函数的上界，我们可以仅仅从“断点”与“对分组合”切入，而不用关心假说集合的具体形式。例如，我们刚刚对于(最小)断点是2的假说集合的一番演绎，完全没有考虑假说集合的具体形式，但仍然推出了很多关于该假说集合的增长函数上界的信息。也就是说，增长函数上界可能仅与断点有关，而与具体的假说集合形式无关。

2. Bounding Function: Basic Cases

本节我们首先定义Bounding Function：bounding function，记作$B(N, k)$，表示当(最小)断点为k时增长函数$m_{\mathcal{H}}(N)$的最大可能值(上界)。研究bounding function可以帮助我们忽略假说集合和增长函数的具体形式，转而仅需关注组合数的问题(combinatorial quantity)。具体而言，$B(N, k)$的含义是，已知(最小)断点为k(N个样本中任何k长度的子样本都不能被打散，但k-1可以)，则N个样本最多可能产生多少种对分方式。

	$B(N,k)$	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	...
	⋮	⋮						⋱

图 2: Table of Bounding Function

首先，第一列全部是1，因为当k=1时，表示连1个点都不能打散，那么最大的可能对分数就均为1。其次，对角线部分也很容易填，因为当$N = k$时，表示不能够打散任何N个点，那么最多可能的对分就为$2^N - 1$种。最后，右上角的部分，即$N < k$，根据(最小)断点的定义，这些地方的N都能够以某种存在形式被打散，因此等于2^N。此外，$B(3, 2)$在此前计算过，为4。

3. Bounding Function: Inductive Cases

本节我们将以计算$B(4, 3)$为例，推导如何计算表格左下方的空缺。一个直观的猜想是，$B(4, 3)$可能与$B(3, ?)$有关，因为$N = 4$是在$N = 3$的基础上增加了一个样本点而得到的。因此，我们首先用电脑程序帮助我们找到B(4,3)的解：当$N = 4$时，最多最多有$2^4 = 16$种对

分，因此也就有2^{16}个“对分集(dichotomy set)”；我们用程式遍历这2^{16}个对分集，看看哪个对分集在满足任何3个样本点没有被打散的条件下，数量最多，它就是$B(4,3)$的解。结果如下，$B(4,3)=11$：

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

$\Rightarrow$

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

orange: pair; **purple**: single

图 3: Reorganized Dichotomies of $B(4,3)$

看到$B(4,3)=11$，我们猜想，$B(4,3)=11=7+4=B(3,3)+B(3,2)$。观察$B(4,3)$的解，我们大致可以将这11个对分分成两组：orange组容纳了所有在$\mathbf{x_1},\mathbf{x_2},\mathbf{x_3}$上成双成对的对分，而purple组则是形单影只的对分。我们将orange组数量的一半记为α，purple组的数量记为β，则$B(4,3)=2\alpha+\beta$。

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

图 4: $\alpha+\beta$

首先我们来看$\alpha+\beta$。$\alpha+\beta$是关于3个点：$\mathbf{x_1},\mathbf{x_2},\mathbf{x_3}$的对分的数量，而此时k=3，因此一定有：

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

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

图 5: α

接着我们单看α。α也是关于3个点：$\mathbf{x_1},\mathbf{x_2},\mathbf{x_3}$的对分的数量。但对于$\alpha$中的每一个对

分，都有另一个$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}$与之一摸一样但$\mathbf{x_4}$完全相反的对分。因此，$\alpha$中任何2个点都不能被打散，因为如果存在2个点被打散，那么配合上$\mathbf{x_4}$，则存在3个点被打散的情况，这与k=3矛盾。因此，我们会有：

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

综合上述两个推论，我们可以得到：

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

也就是说，我们得到了bounding function的上界，可以根据某个空缺的正上方及其左上方的数据填出该空缺的上界(注意，这里还不是确切值，只是上界)：

Putting It All Together

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

	$B(N,k)$	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

图 6: Putting It All Together

下面我们将利用数学归纳法及递推式：

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

证明Sauer's Lemma：

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

证明如下(补充一小点，组合数$\binom{n}{r}$中当$n < r$时也是有意义的，唯一要求是r必须为非负整数，n为整数即可。有$\binom{n}{r} = \frac{n(n-1)(n-2)\dots(n-r+1)}{r!}$)：首先，我们将验证N=1(任意k)时，上式成立。当N=1时，先考虑k=1：

$$RHS = \sum_{i=0}^{0} \binom{1}{i} = \binom{1}{0} = 1 \geq 1 = B(1,1) = LHS$$

再考虑$k \geq 2$(N=1)：

$$RHS = \sum_{i=0}^{k-1} \binom{1}{i} = \binom{1}{0} + \binom{1}{1} = 2 \geq 2 = B(1,k) = LHS$$

综上，当N=1(任意k)时，待证不等式均成立。下面假设对于$N \leq N_0$(任意k)，待证不等式均成立；我们需要推导，对于$N = N_0 + 1$(任意k)，待证不等式均成立。如下图：

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	⋮		⋱		
⋮	⋮	⋮	⋮			⋱	

图 7: 证明逻辑

由递推式可知：

$$B(N_0+1,k) \leq B(N_0,k) + B(N_0,k-1)$$

基于假设，我们可以将待证不等式带入上述递推不等式的右侧，得：

$$\begin{aligned}
LHS = B(N_0+1,k) &\leq \sum_{i=0}^{k-1} \binom{N_0}{i} + \sum_{i=0}^{k-2} \binom{N_0}{i} \\
&= \binom{N_0}{0} + \sum_{i=1}^{k-1} \binom{N_0}{i} + \sum_{i=1}^{k-1} \binom{N_0}{i-1} \\
&= 1 + \sum_{i=1}^{k-1} \left[\binom{N_0}{i} + \binom{N_0}{i-1} \right] \\
&= 1 + \sum_{i=1}^{k-1} \left[\binom{N_0+1}{i} \right] \\
&= \sum_{i=0}^{k-1} \binom{N_0+1}{i} = RHS
\end{aligned}$$

综上，基于数学归纳法，我们证明了Sauer's Lemma，即：

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

注意到，Sauer's Lemma中不等式右侧是组合数连加，且是关于N的k-1次多项式(来自$\binom{N}{k-1}$)。因此，$B(N,k) \leq ploy(N)$——当然，条件是断点k存在。

综上，对于具有(最小)断点k的某个假说集合$\mathcal{H}$，有$m_{\mathcal{H}}(N) \leq B(N,k) \leq ploy(N)$，即其增长函数被一个关于N的多项式限制住了，而非指数式。也就是说，如果某个假说集合存在(最小)断点k，那么其增长函数就可以被某个关于N的k-1次多项式bound住。

4. A Pictorial Proof

回忆multiple bins版本的Hoeffding不等式：

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

其中g表示由某个演算法A从假说集合$\mathcal{H}$中挑选并回传的hypothesis。正是由于此处的g不是固定的hypothesis，而是一个会随着演算法改变而变化的回传结果(可以看成是随机的)，因此不可以直接对它使用Hoeffding不等式。

Multiple bins版本的Hoeffding不等式有另一种写法，这也是课件第4节第一张幻灯片中的写法：

$$P\Big[\exists h \in \mathcal{H} \ s.t. \ |E_{in}(h) - E_{out}(h)| > \epsilon\Big] \leq 2 \cdot M \cdot exp(-2\epsilon^2 N)$$

记第一个不等式左侧P中的事件为B，记上式左侧P中的事件为C，易知$B \subset C$，即"回传结果g是bad"一定能够推出"在假说集合中有某个h是bad"，因为回传结果当然是从假说集合中挑选出来的。然后对$P(C)$使用union bound便得到了multiple bins版本的Hoeffding不等式的右侧。

我们本章到现在所做的工作，就是想要将上式右侧中的M换成某个其他的对假说集合大小度量的指标。现在我们有了增长函数$m_{\mathcal{H}}(N)$。那么是否能够直接将M换成$m_{\mathcal{H}}(N)$呢？可以，但是替换后还需进行一些系数的调整，最终的结果为：

$$P\Big[\exists h \in \mathcal{H} \ s.t. \ |E_{in}(h) - E_{out}(h)| > \epsilon\Big] \leq 4 \cdot m_{\mathcal{H}}(2N) \cdot exp(-\frac{1}{8}\epsilon^2 N)$$

上式被称为Vapnik-Chervonenkis bound，简称VC bound。利用上式对2D perceptron进行分析：2D perceptron的最小断点为k=4，因此其增长函数是一个最高次数为4-1=3的多项式；因此在N非常大的时候，VC bound右侧的界会趋于零；这样就保证了任何一个演算法选出的hypothesis，其in-sample error会是out-of-sample error的较好反映。