六、Theory of Generalization 泛化理论（举一反三）

6.1 Restriction of Break Point 突破点的限制

回顾一下上一章中提到的成长函数$m_H(N)$的定义：假设空间在N个样本点上能产生的最大二分（dichotomy）数量，其中二分是样本点在二元分类情况下的排列组合。

上一章还介绍了突破点（break point）的概念，即不能满足完全分类情形的样本点个数，完全二分类情形（shattered）是可分出2^N 种二分类（dichotomy）的情形。

继续举例说明，假设一种分类情形最小的突破点为2，即 k=2。

容易求出在N=1 时，成长函数 $m_H(N) = 2$

在N=2时，成长函数$m_H(N) < 2_N = 4$ （突破点是2），因此最大的二分类个数不可能超过3，假设为3。

继续观察N=3时的情形，因理解稍微有些复杂，还是以图的形式表达，如图6-1所示。

如图6-1a)表示在三个样本点时随意选择一种二分的情况，这种分类没有任何问题，不会违反任意两个样本点出现四种不同的二分类情况（因为突破点是2）；

如图6-1b)表示在a)的基础上，添加不与之前重复的一种二分类，出现了两种不冲突的二分类，此时同样也没有任何问题，不会违反任意两个样本点出现四种不同的二分类情况；

如图6-1c) 表示在b)的基础上，再添加不与之前重复的一种二分类，出现了三种不冲突的二分类，此时同样也没有任何问题，不会违反任意两个样本点出现四种不同的二分类情况；

如图6-1d) 表示在c)的基础上，再添加不与之前重复的一种二分类，问题出现了，样本x_2, x_3 出现了四种不同的二分情况，和已知条件中k=2矛盾（最多只能出现三种二分类），因此将其删去。

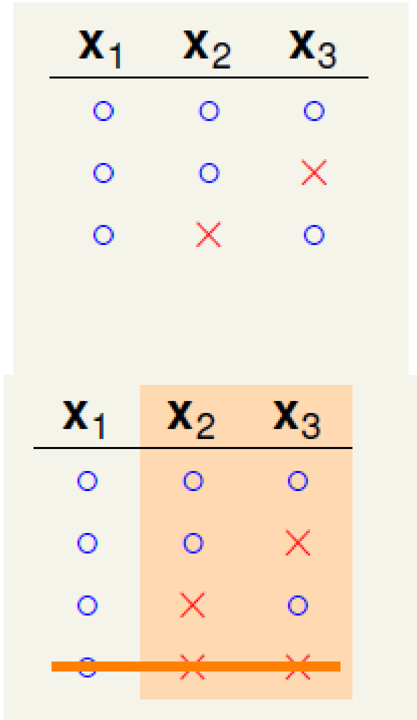
如图6-1e) 表示在c)的基础上，再添加不与之前重复的一种二分类，此时同样也没有任何问题，不会违反任意两个样本点出现四种不同的二分类情况；

如图6-1f) 表示在e)的基础上，再添加不与之前重复的一种二分类，问题又出现了，样本x_1, x_3出现了四种不同的二分情况，和已知条件中k=2的条件不符（最多只能出现三种二分类），因此将其删去。

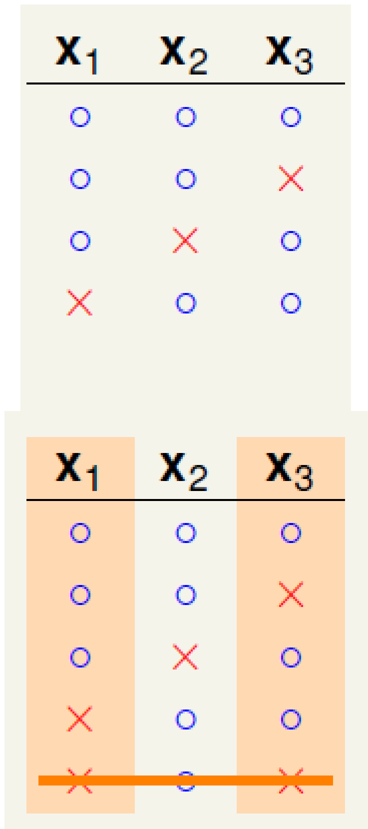
x_1	x_2	x_3
○	○	○

x_1	x_2	x_3
○	○	○
○	○	×

a)   b)



c)   d)



e)   f)

图6-1 样本数为3时的二分类情形

在突破点是2，样本点为3时，最多只能有四种二分类情况，而 $4 \ll 2^3$。

从上述情形，可以做一个猜想，成长函数$m_H(N)$ 小于等于某个与突破点**k有关的二次项**，如果可以证明这一结论，即能寻找到一个可以取代无限大M的数值，同理即能公式4-6在假设空间为无限大时也是成立，即机器是可以学习的。

假设突破点k=1，在样本数为3时，最大的二分类个数是多少？答案是1，可以想象，在样本为1时，只有一种分类情形，假设这种情形是正，则以后所有样本也为正，才能满足上述条件，所以样本N不论为多少，其最大二分类数都是1。

6.2 Bounding Function - basic function 上限函数的基本情形

根据上一小节的例子，提出一个新的概念，上限函数 $B(N,k)$（bounding function），其定义为在最小突破点为k时，表示成长函数 $m_H(N)$ **最大值**的函数。此函数将成长函数从各种假设空间的关系中解放出来，不用再去关心具体的假设，只需了解此假设的突破点，突破点相同的假设视为一类，抽象化成长函数与假设的关系。

二分类的个数或者称之为成长函数$m_H(N)$ 说白了就是二元（在图中表示为"×"或者"○ "的符号）符号在在长度为N的情况下的排列组合（每个不同的排列代表一个二分类，每种二分类可看做一个向量）个数。

在强调一遍，提出这种上限函数的好处在于，它的性质可以不用关心到底使用的是何种假设空间，只需要知道突破点便可以得到一个上限函数来对成长函数做约束。

例如：$B(N,3)$ 可以同时表示一维空间的感知器（1D perceptrons）和间隔为正（positive intervals）的两种假设空间，因为两者都满足k=3 。注意一维的成长函数为$m_H(N)=2N$，而间隔为正的成长函数为$m_H(N)=1/2N^2+1/2N+1$，$B(N,3)$ 一定比这两种情况中最大的还要大，才能约束成长函数，回忆上限函数的定义，是成长函数$m_H(N)$最大值的函数表示，从这个例子中可以理解为何还会出现"最大值"。

想要证明$B(N,k)<poly(N)$ 。

先观察已知的 $B(N,k)$如何表示，通过列表方式找出规律，如图6-2所示。

					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			15	16	16	...
	5	1				31	32	...
	6	1					63	...
	⋮	⋮						⋱

图6-2 已知的$B(N,k)$ 取值

在 N=k的这条斜线上，所有值都等于2^N-1 ，这一原因在上一节推导中已经提到过，原因是突破点代表不能完全二分类的情况，因此在此情况下最大的二分类数可以是2^N-1 。在这条斜线的右上角区域所有的点都满足完全二分类的，因此值为 2^N。

而在斜线右下角上已给出数值的点都是在上一节中已求出答案的点，空白地方的值则是下节需要介绍的内容。

当突破点等于K的时候，成长函数的上限成为上限函数，公式如下

$$B(N,k)=\max(m_H(N))$$

6.2 Bounding Function - Inductive Cases 上限函数的归纳情形

这一小节主要是将上一小节空白的内容填写上，从已有的数据上可以看出一个似乎是正确的结论：每个值等于它正上方与左上方的值相加。 如公式6-1所示。 $B(N,k)=B(N-1,k)+B(N-1,k-1)$

当然单从观察是无法证明该公式是成立的，以下从结论出发来验证它的正确性。

首先通过计算机求出N=4, k=3 的所有二分情况（自己编写程序加上限制条件，在16个二分类中找出符合条件的），其结果如图6-3所示。

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

图6-3 N=4,k=3的所有二分类

图6-3中的展示效果还是有些混乱，对这11种情况做一次重新排序，将x_4与x_1 x_3 分开观察，如图6-4所示，橙色部分为 x_1 x_3两两一致、 x_4成对（pair）出现的二分类，设橙色部分一共 α 对二分类，即 2α种二分类；紫色部分为各不相同的x_1 x_3 ，设紫色部分一共 β 种，得公式6-2。 $B(4,3)=11=2\times4+3=2\alpha+\beta$

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

图6-4 以成对和单个的形式展示

注意$k=3$，意味着在样本点为3时，不能满足完全二分的情形。需要观察在样本数为3时，这11种分类会有何变化，不妨假设这三个样本是$x_1\ x_3$，于是只剩如图6-5所示的7种二分类情形。

	x_1	x_2	x_3
α	o	o	o
	×	o	o
	o	×	o
	o	o	×
β	×	×	o
	×	o	×
	o	×	×

图6-5 三个样本时缩减之后的二分类

其中橙色部分，原本两两成对出现的二分类，在去掉x_4所属的那列样本之后，就合并成了4种二分类情况（$\alpha=4$），紫色部分不变依然为3种二分情况（$\beta=3$）。因为已知$N=4,k=3$，在样本数为3时，图6-5中即表示样本书为3 的情况，其一定不能满足完全二分，因此与α与β一定满足公式6-3。$\alpha+\beta\le B(3,3)$

继续观察橙色部分的区域，如图6-6所示。

	x_1	x_2	x_3
α	o	o	o
	×	o	o
	o	×	o
	o	o	×

图6-6 三个样本时橙色区域的二分类

以下使用反证法证明。假设$x_1\ x_3$三个样本在如上图所示的四种二分类情况下，满足任意两个样本都可完全二分类。将$x_1\ x_3$中任意两列取出，同之前被删除的x_4列相结合，一定可得到一种三个样本都满足完全二分类的情形（因为不论是哪两列与相结合都会得到8种二分类，每一行二分类都对应两个不同标记的，因此为8种。且两列四行的二分类是全完二分的，在此基础上加上不同属性的列，应该也不会重复，因此一定也是全完二分的）。但是此结论和已知任意三个样本不能完全二分冲突了，因此假设不成立，即在图6-6中一定存在某两个样本点不能完全二分的情况，因此得出如公式6-4所示的结论。$\alpha\le B$

由公式6-2~公式6-4推导出公式6-5。

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

最终还能推导出公式6-6的通用结论，也就是开始时公式6-1的猜想。

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

根据这一结论将图6-2补齐，如图6-7所示。

	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

图6-7 补齐后的上限函数表

最后可以通过如下方式证明公式6-7是成立的。

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

首先需要预先了解组合中的一个定理，如公式6-8所示。

$$C_n^k = C_{N-1}^k + C_{N-1}^{k-1}$$

很容易证明K=1 情况下公式6-7成立，如公式6-9所示。

$$B(N,1)=1=C_N^0 \le \sum_{i=0}^{1-1} C_N^i$$

上一节中已经给出了证明，不仅仅是满足不等号条件，而且满足等号。

再使用数学归纳法证明在的情况，公式6-7成立。

假设公式6-10成立，则可以得到在的情况下公式6-11成立，同时结合公式6-6的结论，公式6-12也能被推导出来。

$$B(N-1,k) \le \sum_{i=0}^{k-1} C_{N-1}^{i}$$ （公式6-10）

$$B(N-1,k-1) \le \sum_{i=0}^{k-2} C_{N-1}^{i} = \sum_{i=1}^{k-1} C_{N-1}^{i}$$ （公式6-11）

$$\begin{aligned} B(N,k) &\le B(N-1,k)+B(N-1,k-1) \\ &\le \sum_{i=0}^{k-1} C_{N-1}^{i} + \sum_{i=0}^{k-2} C_{N-1}^{i} \\ &= \sum_{i=0}^{k-1} C_{N-1}^{i} + \sum_{i=1}^{k-1} C_{N-1}^{i} \\ &= C_{N-1}^{0} + \sum_{i=1}^{k-1} (C_{N-1}^{i} + C_{N-1}^{i}) \\ &= 1 + \sum_{i=1}^{k-1} C_{N}^{i} \\ &= C_{N}^{0} + \sum_{i=1}^{k-1} C_{N}^{i} \\ &= \sum_{i=0}^{k-1} C_{N}^{i} \end{aligned}$$ （公式6-12）

这一结果意味着：成长函数 $m_H(N)$的上限函数$B(N,k)$ 的上限为N^{k-1} 。

6.4 A Pictorial Proof 一种形象化的证明

至此说明了（不敢叫证明）一个在机器学习领域很著名的理论——V-C上界制（Vapnik-Chervonenkis bound）。

2维的感知器，其突破点为4，因此其成长函数为$O(N^3)$，可以使用这个VC上界来说明在样本N足够大时候，发生坏事的情况很少，即选择一个在样本中错误率很小的g，可以得出其在整个数据上错误率也很低的结论，说明机器学习是可能的。