Southwestern University Of Finance And Economics spring, 2021
School of Finance Lecture 5 & 6
Scribe: *Haoming Wang* Lecturer: *Prof. Xuantian Lin* @ NTU

机器学习基石 I: Theory of Generalization

Key word: Growth Function, Break Point, Bounding Function

1 泛化理论 Theory of Generalization

1.1 有效假设函数总量 Effective Number Hypothesis

上一讲讲到, 对于某个 hypothesis 函数 h 有 $\mathbb{P}\left(\text{BAD } \mathcal{D} \text{ for } h\right) \leq 2e^{-2\epsilon^2 N}$. 若 $|\mathcal{H}| = M$, 则 $\mathcal{D}$ 对于某个 $h \in \mathcal{H}$ 而言是 BAD 的概率有

$$\begin{aligned}\mathbb{P}\left(\text{BAD } \mathcal{D} \text{ for } \mathcal{H}\right) &= \mathbb{P}\left(\bigcup_{h\in\mathcal{H}} \text{BAD } \mathcal{D} \text{ for } h\right) \\ &\leq \sum_{i=1}^{M} \mathbb{P}\left(\text{BAD } \mathcal{D} \text{ for } h_i\right) \\ &\leq 2Me^{-2\epsilon^2 N}.\end{aligned}$$

这说明当样本足够大时, 我们有很大的把握认定 $E_{out}(h) = E_{in}(h)$; 再通过在 $\mathcal{H}$ 中选择 $E_{in}(h) \approx 0$ 的 hypothesis 函数 h^* 作为 g, 我们就可以推断出

$$E_{in}(h^*) \approx E_{out}(h^*) = E_{out}(g) \approx 0,$$

从而证明了学习的可行性. 即我们将学习的过程归结为两个过程:

- 使 $E_{in}(g) \approx E_{out}(g)$
- 在 $\mathcal{H}$ 找到使 $E_{in}(h)$ 足够小的 h 作为 g.

现在的问题是 $|\mathcal{H}| = M$ 如何影响学习的可行性? 当 M 很小的时候, 我们抽到 BAD 的概率上限 $2Me^{-2\epsilon^2 N}$ 也会减小, 即可以增强第一个过程的把握. 但是对于第二个过程而言, 容量过小的假设集可能使我们找不到足够使 $E_{in}(h)$ 接近于 0 的 hypothesis 函数 h.

当 M 很大的时候, 抽到 BAD 的概率上限 $2Me^{-2\epsilon^2 N}$ 也会增大 (甚至大于 1, 从而使 PAC 框架失效), 即样本 $\mathcal{D}$ 更有可能对某个 $h \in \mathcal{H}$ 是一个 BAD 样本, 从而使第一个过程的把握减小. 但是对应的, 会使我们更可能在 $\mathcal{H}$ 中找到理想的 h 作为 g.

可见 M 对机器学习的可行性是至关重要的, 但是我们的模型对于 $M = \infty$ 是无效的, 下面思考改进这一问题. 考虑霍夫丁不等式的结论:

$$\mathbb{P}\left(|E_{in}(g) - E_{out}(g) > \epsilon|\right) \leq 2Me^{-2\epsilon^2 N}.$$

对于 $h_i \in \mathcal{H}$ 而言, BAD 事件为 $\mathcal{B}_m = \{|E_{in}(h_m) - E_{out}(h_m) > \epsilon\}$, 而对于 $\mathcal{H}$ 而言 BAD 的概率为

$$\mathbb{P}\left(\mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \mathcal{B}_n \cup \cdots\right) \leq \mathbb{P}\left(\mathcal{B}_1\right) + \mathbb{P}\left(\mathcal{B}_2\right) + \cdots + \mathbb{P}\left(\mathcal{B}_n\right) + \cdots$$

需要注意, 对于 $h_i, h_j \in \mathcal{H}$, 可能 $\mathcal{B}_i \cap \mathcal{B}_j \to \mathcal{B}_i \cup \mathcal{B}_j$, 即

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}_i \cup \mathcal{B}_j\right) &= \mathbb{P}\left(\mathcal{B}_i\right) + \mathbb{P}\left(\mathcal{B}_j\right) - \mathbb{P}\left(\mathcal{B}_i \cap \mathcal{B}_j\right) \\ &\ll \mathbb{P}\left(\mathcal{B}_i\right) + \mathbb{P}\left(\mathcal{B}_j\right) < 2 \cdot 2 \cdot e^{-2\epsilon^2 N},\end{aligned}$$

这为我们将 M 替换为更小的值提供了思路.

设 $\mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}$. 考虑 $\mathcal{X} = \mathbf{x}_1 \in \mathbb{R}^2$ 的情形. 此时对于样本 $\mathbf{x}_1$ 而言, 只有两类 hypothesis 是不同的, 一是 h_1, 它将 $\mathbf{x}_1$ 推断为 $\circ$; 二是 h_2, 它将 $\mathbf{x}_1$ 推断为 $\times$.

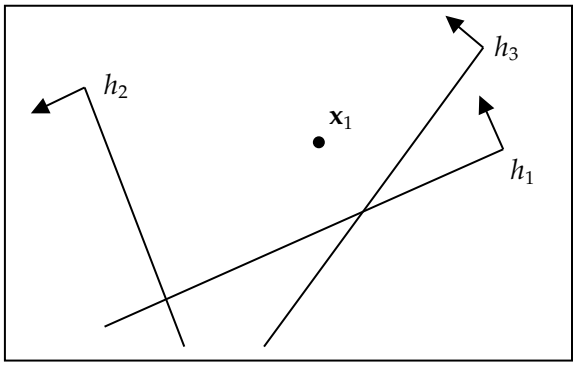

图 1: $\mathcal{X} = \mathbf{x}_1 \in \mathbb{R}^2$

而对于 h_1 与 h_3 而言, 他们对 $\mathbf{x}_1$(或者说 $\mathcal{X}$) 的推断是完全一致的, 因此若 $\mathcal{X}$ 对 h_1 而言是 BAD, 则对 h_3 而言也是 BAD. 即

$$\mathbb{P}\left(\mathcal{B}_1 \cup \mathcal{B}_2 \cup \mathcal{B}_3 \cup \cdots\right) = \mathbb{P}\left(\mathcal{B}_1 \cup \mathcal{B}_2\right) \leq 2 \cdot 2 \cdot e^{-2\epsilon^2 N}.$$

下图展示了当 $\mathcal{X} = \mathbf{x}_1$ 是 h_1 的 BAD 样本的一种可能的情况, 其中 $\mathbf{x}_1$ 以外的点表示真实的总体情况, 并不在样本中.

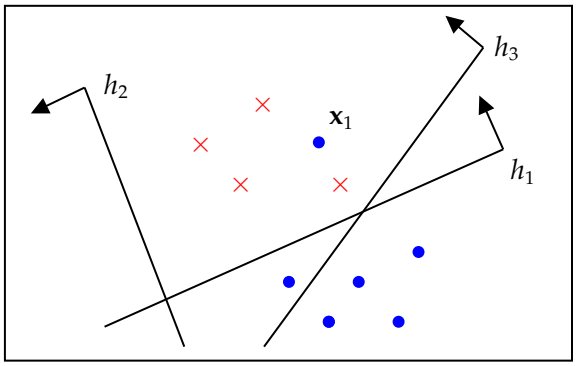

图 2: 一种可能的总体

下面考虑 $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ 的情况. 此时对于 $\mathcal{X}$ 而言, $\mathcal{H}$ 中一共有四类 hypothesis h_1, h_2, h_3, h_4. 这四类 hypothesis 将 $\mathcal{X}$ 划分为如下四类: $h_1 : \mathcal{X} \mapsto \{\circ, \circ\}$, $h_2 : \mathcal{X} \mapsto \{\times, \times\}$, $h_3 : \mathcal{X} \mapsto \{\circ, \times\}$, $h_4 : \mathcal{X} \mapsto \{\times, \circ\}$.

对于除此之外任意一个 hypothesis 如 h_5, 它对 $\mathcal{X}$ 的推断与 h_4 完全一致, 因此若 $\mathcal{X}$ 是 h_4 的一个 BAD 样本, 则它也是 h_5 的一个 BAD 样本. 因此

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}_1 \cup \mathcal{B}_2 \cup \mathcal{B}_3 \cup \mathcal{B}_4 \cup \mathcal{B}_5 \cup \cdots\right) &= \mathbb{P}\left(\mathcal{B}_1 \cup \mathcal{B}_2 \cup \mathcal{B}_3 \cup \mathcal{B}_4\right) \\ &\leq 2 \cdot 4 \cdot e^{-2\epsilon^2 N}.\end{aligned}$$

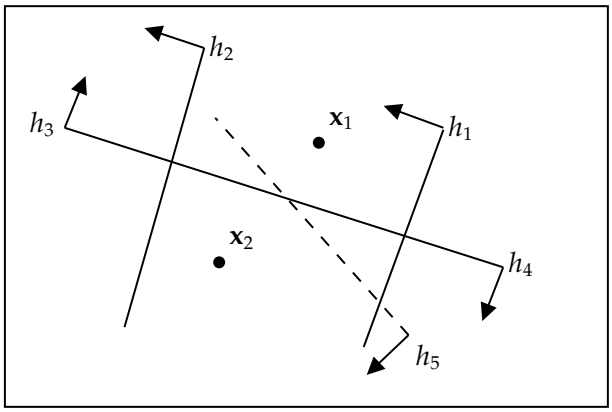

图 3: $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\} \in \mathbb{R}^2$

下图展示了一种可能的总体, 这种情况下样本 $\mathcal{D} = \mathcal{X} \times \mathcal{Y} = \{(\mathbf{x}_1, \times), (\mathbf{x}_2, \circ)\}$, 此时 $E_{in}(h_4) = E_{in}(h_5) = 0$. 但是显然 $\mathcal{D}$ 是 h_4, h_5 的 BAD 样本.

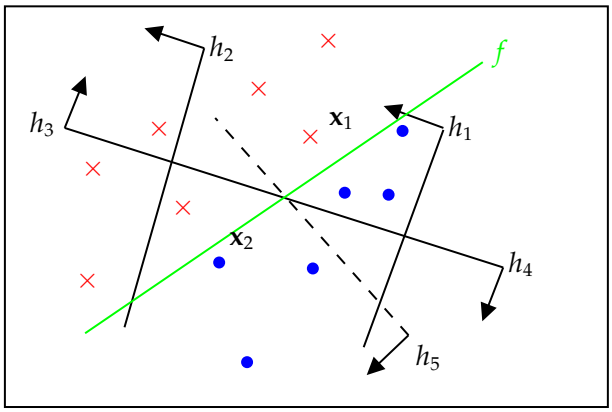

图 4: 一种可能的总体

下面我们要探索更一般的情况. 对于一个二分类的问题, 即 $\mathcal{H} = \{h_i : \mathcal{X} \mapsto \{-1, +1\}, i = 1, 2, \cdots, M\}$, 我们可以计算出若 $|\mathcal{X}| = N$, 则 $\mathcal{X}$ 最多一共有 2^N 种分类方式, 一次无论 $|\mathcal{H}| = M$ 为多少, 都只有 2^N 种 hypothesis, 因此我们可以把 M 替换成 2^N, 即

$$\mathbb{P}\left(\bigcup_i \mathcal{B}_i\right) \leq 2 \cdot 2^N \cdot e^{-2\epsilon^2 N}.$$

但是对于 $\mathcal{X} \subset \mathbb{R}^2$ 的 N 个点, 到第能够被二分类分为多少种, 是取决于假设集 $\mathcal{H}$ 的结构的. 例如当 $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} \subset \mathbb{R}^2$ 时, 虽然二分类最多可以一共分出 $2^4 = 16$ 中, 但是当我们令 $\mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}$ 时, $\{\circ, \times, \circ, \times\}$ 和 $\{\times, \circ, \times, \circ\}$ 这两种情况是无法被任意 $h \in \mathcal{H}$ 实现的. 因此对于 $\mathcal{H}$ 而言, $\mathcal{X}$ 最多只能被分为 $2^4 - 2 = 14$ 类; 换句话说, 对于 $\mathcal{X}$ 而言, $\mathcal{H}$ 中一共有 14 类 hypothesis, 因此我们可以将 M 替换为更小的某一个值, 在这个例子中:

$$\mathbb{P}\left(\bigcup_i \mathcal{B}_i\right) \leq 2 \cdot 14 \cdot e^{-2\epsilon^2 N} \leq 2 \cdot 2^4 \cdot e^{-2\epsilon^2 N} \ll 2 \cdot M \cdot e^{-2\epsilon^2 N}.$$

1.2 成长函数 Growth Function

上面讨论了当 $\mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}$ 时, 对于不同的 $\mathcal{X} = \{\mathbf{x}\}_N$, 有几种 hypothesis h. 下面一般化 $\mathcal{H}$. 设先存在一个 hypothesis set:

$$\mathcal{H} = \{\text{hypothesis } h : \mathcal{X} \mapsto \{\times, \circ\}\}.$$

我们称

$$h(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N) \coloneqq (h(\mathbf{x}_1), h(\mathbf{x}_2), \cdots, h(\mathbf{x}_N)) \in \{\times, \circ\}^N, \ \forall h \in \mathcal{H}$$

为一个 dichotomy. 将一个 hypothesis set $\mathcal{H}$ 在 $\mathcal{X} = \{\mathbf{x}_i\}_N$ 上作出的所有 dichotomy 构成一个集合记为 $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$. 比较 $\mathcal{H}$ 与 $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$.

	hypothesis set $\mathcal{H}$	dichotomies $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$
e.g.	all lines in $\mathbb{R}^2$	$\{\circ\circ\circ\circ, \circ\circ\circ\times, \circ\circ\times\times, \cdots\}$
size	M, possible ∞	$\lvert\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)\rvert \leq 2^N$

如上文分析, $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)|$ 取决于 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$. 这不仅取决于 N 还取决于 N 个点的分布. 如令 $\mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}$, $N = 4$, 若 $\mathbf{x}_1', \mathbf{x}_2', \mathbf{x}_3', \mathbf{x}_4'$ 四点共线, 则我们只能将其分为两类, 即 $|\mathcal{H}(\mathbf{x}_1', \mathbf{x}_2', \mathbf{x}_3', \mathbf{x}_4')| = 2$, 而当 $\{\mathbf{x}_i\}_4$ 分布在一个圆上时, $|\mathcal{H}(\{\mathbf{x}_i\}_4)| = 14$. 为了规避 $\{\mathbf{x}_i\}_N$ 的分布带来的影响, 我们定义

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)|$$

为 hypothesis set $\mathcal{H}$ 的成长函数 (Growth Function). 如上文分析当 $\mathcal{H} = \text{2D Perceptron}$ 时,

$$m_{\mathcal{H}}(1) = 2, m_{\mathcal{H}}(2) = 4, m_{\mathcal{H}}(3) = 8, m_{\mathcal{H}}(4) = 14.$$

对于任意二分类 hypothesis set $\mathcal{H}$, 有 $m_{\mathcal{H}}(N) \leq 2^N$.

若 $m_{\mathcal{H}}(N) = 2^N$, 则**存在** N 个 input $\{\mathbf{x}_i'\}_N$ 使得 $\mathcal{H}$ 可以将其映射到 $\{\times, \circ\}$ 中的每一点. 称 $\{\mathbf{x}_i'\}_N$ 被 $\mathcal{H}$ 击碎 (shattered), 例如, $\mathcal{H} = \text{2D Perceptron}$, 则存在 $\{\mathbf{x}_i\}_1, \{\mathbf{x}_i\}_2, \{\mathbf{x}_i\}_3$ 被 $\mathcal{H}$ shatter, 但是任意 $\{\mathbf{x}_i\}_4$ 不可被 $\mathcal{H}$ shatter.

至此, 我们可以将霍夫丁不等式中的 M 替换为 $m_{\mathcal{H}}(N)$:

$$\mathbb{P}\left(|E_{in}(g) - E_{out}(g)| > \epsilon\right) \leq 2 \cdot m_{\mathcal{H}}(N) \cdot e^{-2\epsilon^2 N}.$$

对于二分类的 $\mathcal{H}$, 有 $m_{\mathcal{H}}(N) \leq 2^N$. 下面讨论几种 $\mathcal{H}$ 的增长函数:

1. Positive Intervals $\mathcal{X} = \{\mathbf{x}_i\}_N \subset \mathbb{R}$, $\mathcal{H} = \{h(x) = \text{sign}\,(x - a), a \in \mathbb{R}\}$. 易知

$$m_{\mathcal{H}}(N) = N + 1 \ll 2^N.$$

2. Positive Intervals $\mathcal{X} = \{\mathbf{x}_i\}_N \subset \mathbb{R}$, $\mathcal{H} = \{h(x) = +1 \text{ iff } x \in [l, r); -1 \text{ otherwise}\}$. 易知

$$m_{\mathcal{H}}(N) = C_N^2 + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1 \ll 2^N.$$

3. Convex Sets $\mathcal{X} = \{\mathbf{x}_i\}_N \subset \mathbb{R}^2$, $\mathcal{H} = \{h(x) = +1 \text{ iff } x \in \text{Convex Set} \subset \mathbb{R}^2\}$. 只需要将 $\{\mathbf{x}_i\}_N$ 排列在一个圆上, 即可算得

$$m_{\mathcal{H}}(N) = 2^N.$$

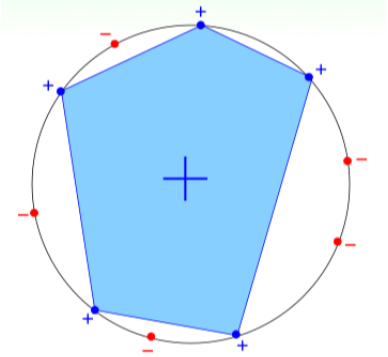

图 5: Growth Function for Convex Sets

注意, 对于任意 $N \in \mathbb{Z}^+$, 存在 $\{\mathbf{x}_i\}_N$ 被 $\mathcal{H}$ 击碎 (shattered).

1.3 突破点 Break Point

我们定义: 如果当 $N = k$ 时 $\mathcal{H}$ 不可 shatter $\{\mathbf{x}_i\}_k$, 即 $m_{\mathcal{H}}(k) < 2^k$, 则 k 为 $\mathcal{H}$ 的**突破点** (Break Point). 若 k 为 $\mathcal{H}$ 的突破点, 则

$$m_{\mathcal{H}}(k+1) \leq 2 \cdot m_{\mathcal{H}}(k) < 2 \cdot 2^k = 2^{k+1}.$$

因此 $k+1, k+2, \cdots$ 也是 $\mathcal{H}$ 的突破点. 对于 2D perceptrons 而言, 其最小突破点为 4.

比较突破点与成长函数:

$\mathcal{H}$	Min Break Point	Growth Func
Positive Rays	2	$m_{\mathcal{H}}(N) = N + 1 = O(N)$
Positive Interval	3	$m_{\mathcal{H}}(N) = 1/2N^2 + 1/2N + 1 = O(N^2)$
Convex Sets	None (∞)	$m_{\mathcal{H}}(N) = 2^N$
2D perceptrons	4	$m_{\mathcal{H}}(N) =$?

我们提出如下猜测: 若 $\mathcal{H}$ 无 min break point, 则 $m_{\mathcal{H}}(N) = 2^N$; 若 $\mathcal{H}$ 的 min break point 为 k, 则 $m_{\mathcal{H}}(N) = O(N^{k-1})$. 我们在下一讲证明这一猜想.

2 二维感知器的成长函数 2D Perceptron Growth Func

2.1 突破点的限制 Restriction of Break Point

上一讲说明 $\mathcal{H}$ 在 $\{\mathbf{x}_i\}_N$ 上产生的 dichotomy 的种类的最大值为 $m_{\mathcal{H}}(N)$. 本讲我们考察 Min Break Point 是否会对 $\mathcal{H}$ 在 $\{\mathbf{x}_i\}_N$ 上产生的 dichotomy 的种类产生更强的限制.

如果对于某个二分类 hypothesis set $\mathcal{H}$, 其 Min Break Point $k = 2$. 则有以下推论: 当 $N = 1$ 时, $m_{\mathcal{H}}(N) = 2^N = 2$. 当 $N = 2$ 时, $m_{\mathcal{H}}(N) < 2^N = 4$, 即 $\max m_{\mathcal{H}}(2) = 3$ (最大 dichotomy 最多种类为 3). 当 $N = 3$ 时, 注意: 一方面 $\max m_{\mathcal{H}}(3) \leq 7$; 另一方面, $\{\mathbf{x}_i\}_3$ 中任意两个点都不可被 shatter, 即任意两个点最多产生 3 种 dichotomy. 在这两个条件的限制下, 注意到若要 shatter 两个点, 则至少需要 4 种 dichotomy, 因此当 1 < dichotomy < 4 时都是成立的, 即我们只需要考察 4, 5, 6, 7 种 dichotomy 的情况.

	$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$	
1		No: $m_{\mathcal{H}}(1) = 2$
2		Yes
3		Yes
4		?
5		?
6		?
7		?
8		No: $\max m_{\mathcal{H}}(3) = 7$
9		No: $\max m_{\mathcal{H}}(3) = 7$

考察 4 种 dichotomy 的情况, 左图的情况下 $\mathbf{x}_2, \mathbf{x}_3$ 被 shatter; 右图的情况下, 满足条件. 因此 $\max m_{\mathcal{H}}(3) \geq 4$.

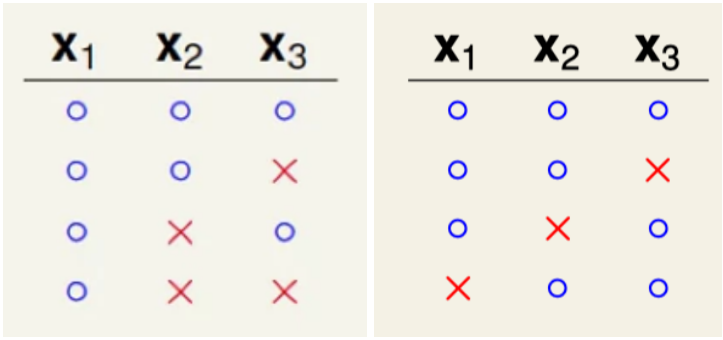

图 6: 4 种 dichotomy

考察 5 种 dichotomy 的情况. 左一图中 $\mathbf{x}_1, \mathbf{x}_3$ 被 shatter, 左二图 $\mathbf{x}_1, \mathbf{x}_2$ 被 shatter, 左三图 $\mathbf{x}_1, \mathbf{x}_2$ 被 shatter.

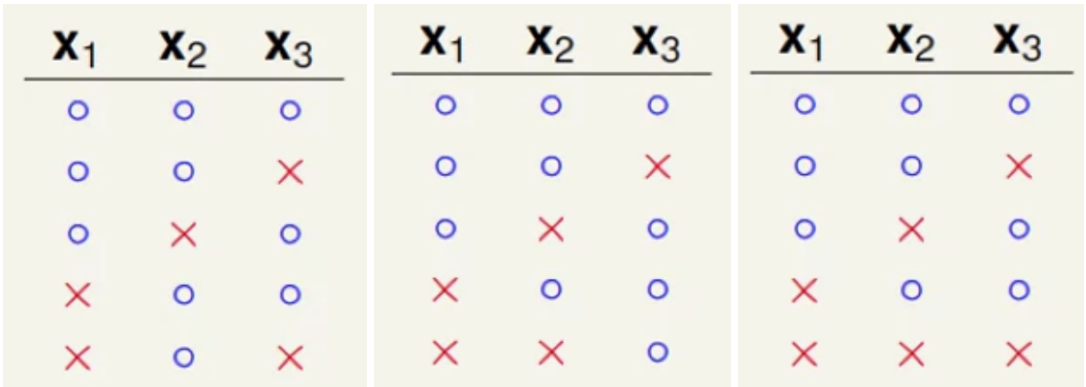

图 7: 5 种 dichotomy

任意 5 种 dichotomy 的情况下, 总有两个点被 shatter, 这违反了我们的假设, 因此 $\max m_{\mathcal{H}}(3) = 4$.

($m_{\mathcal{H}}(3)$ 可以为 2, 3, 4.) 由此可知, Min Break Point k 限制了 $\max m_{\mathcal{H}}(N)$, 对于 $N > k$.

注 2.1. *易知, 若 $\mathcal{H}$ 的 Min Break Point $k = 1$, 则 $m_{\mathcal{H}}(N) = 1, \forall N$. 因为任意两个及以上不同的 dichotomy, 必然导致某个点被 shatter.*

2.2 上限函数 Bounding Function

定义: 若 $\mathcal{H}$ 的 Min Break Point 为 k, 则 $\max m_{\mathcal{H}}(N)$ 为上限函数 $B(N,k)$ (有 N 个点, 任何 k 个点不能 shatter, 则该 N 个点的 dichotomy 最多种数的最大值为 $B(N,k)$, 注意如果 $k > N$, 则不能 shatter 的条件无效, 此时 $B(N,k) = \max m_{\mathcal{H}}(N) = 2^N$).

注 2.2. *"最多种数的最大值", "最多"是指不同分布的 N 个点在 $\mathcal{H}$ 下产生的 dichotomy 的最大值, 即 $m_{\mathcal{H}}$. "最大值"是指在所有结构的二分类 hypothesis set $\mathcal{H}$ 中产生的最大 $m_{\mathcal{H}}$.*

注意到, 我们在上一节推导 $B(3,2)$ 的过程中, 除了设定 $\mathcal{H}$ 是二分类 hypothesis set, 并没有对其结构进行其他假设, 如 Positive Interval 或 2D perceptron. 这为我们替换霍夫丁不等式中的 M 提供了方便, 因为不需要对每一类 $\mathcal{H}$ 计算其成长函数 $m_{\mathcal{H}}(N)$. 下面考察上限函数的结构:

首先, 如上文分析, $B(2,2) = 3, B(3,2) = 4, B(N,1) \equiv 1$:

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

图 8: Table of Bounding Function

如果 $k > N$, 则 N 个点中任意 k 个点不 shatter 的条件无效, 此时 $B(N,k) = \max m_{\mathcal{H}}(N) = 2^N$ (参考 Convex Set $\mathcal{H}$); 当 $N = k$ 时, 即 N 个点不能 shatter, 则 $B(N,k) = \max m_{\mathcal{H}}(N) = 2^N - 1$.

	$B(N,k)$	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	...
	⋮	⋮						⋱

图 9: Table of Bounding Function

注意, $B(N,k) = \max m_{\mathcal{H}}(N) = 2^N - 1$ 并不是说存在某个结构的 $\mathcal{H}$ 满足 $m_{\mathcal{H}}(N) = 2^N - 1$, 而只是理论上任意结构的 $\mathcal{H}$ 的 $m_{\mathcal{H}}(N)$ 上限. 例如 2D Perceptron 的 $m_{\mathcal{H}}(4) = 14 < B(4,4) = 15$.

下面考虑 $B(4,3)$ 的大小. 假设 $B(4,3)$ 种 dichotomy 由 pair 和 single 两类 dichotomy 构成. pair dichotomy 指某 output $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ 存在一个对应的仅 $\mathbf{x}_4$ 不同的 output; single dichotomy 指不存在这样的对应的 output.

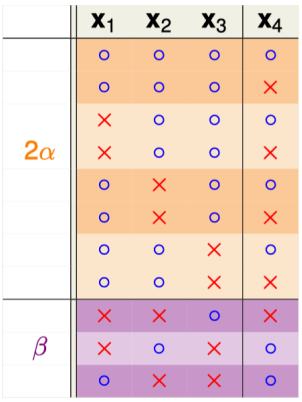

图 10: Estimating Part of $B(4,3)$

假设 pair dichotomy 与 single dichotomy 的个数分别为 2α 与 β, 则 $B(4,3) = 2\alpha + \beta$. 下面暂不考虑 $\mathbf{x}_4$ 的值, 因为 Min Break Point $k = 3$, 因此 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ 不能 shatter, 即 $\alpha + \beta \leq B(3,3)$.

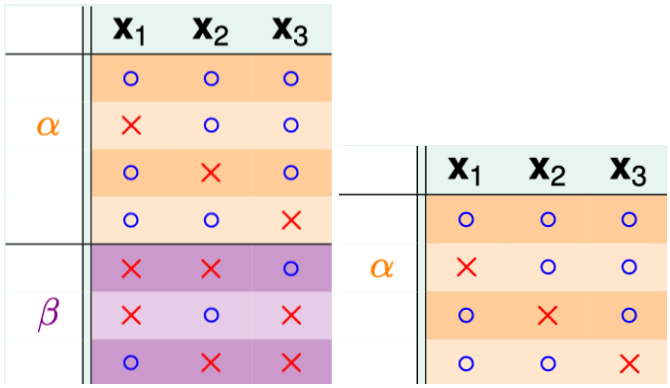

图 11: Estimating Part of $B(4,3)$

因为任意三个点不能 shatter, 因此 α 个 pair 的 dichotomy 中任意两个点不能 shatter, 否则 shatter 的两个点 (4 个 dichotomy) 加上 $\mathbf{x}_4$ 这三个点 (8 个 dichotomy) 必然 shatter. 即 $\alpha \leq B(3,2)$, 综上

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

由类似的分析方法有

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

至此我们推出了上界函数的上界 (N 个点中任意 k 个不能 shatter 下最多种类 dichotomy 的最大值的上界).

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

图 12: Table of Bounding Function

可以证明的是 $B(N,k)$ 的上界是一个多项式函数. 设 $B(N-1,k) \leq \sum_{i=1}^{k-1} C_{N-1}^i$, 则

$$\begin{aligned}
B(N,k) &\leq B(N-1,k) + B(N-1,k-1) \\
&\leq \sum_{i=0}^{k-1} C_{N-1}^i + \sum_{i=0}^{k-2} C_{N-1}^i \\
&= 1 + \sum_{i=1}^{k-1} C_{N-1}^i + \sum_{i=0}^{k-2} C_{N-1}^i \\
&= 1 + \sum_{i=0}^{k-2} C_{N-1}^{i+1} + \sum_{i=0}^{k-2} C_{N-1}^i \\
&= 1 + \sum_{i=0}^{k-2} C_N^{i+1} \\
&= 1 + \sum_{i=1}^{k-1} C_N^i \\
&= \sum_{i=0}^{k-1} C_N^i.
\end{aligned}$$

由数学归纳法可知 $B(N,k) \leq \sum_{i=1}^{k-1} C_N^i$. 因为

$$C_N^{k-1} = \frac{N!}{(k-1)!(N-k+1)!} = \frac{N \cdot N-1 \cdots N-k+2}{(k-1)!} = O(N^{k-1}),$$

所以 $B(N,k) \leq P_{k-1}(N)$(P_{k-1} 表示 $k-1$ 阶多项式). 综上若 $\mathcal{H}$ 的 Break Point 存在, 则 $m_{\mathcal{H}}(N)$ 是多项式, 对于 2D Perceptron 而言 $m_{\mathcal{H}}(N) \leq \frac{1}{6}N^3 + \frac{5}{6}N^2 + 1$.