第 6 章 泛化理论

6.1 断点的限制

定义 6.1.1 (打散（Shatter）)

设 $\mathcal{X}$ 为输入空间，$\mathcal{H} \subseteq \{h : \mathcal{X} \to \{-1, +1\}\}$ 为一假设类。给定有限子集 $S = \{x_1, \dots, x_N\} \subseteq \mathcal{X}$，记

$$\mathcal{H}_S := \{(h(x_1), \dots, h(x_N)) \mid h \in \mathcal{H}\} \subseteq \{-1, +1\}^N.$$

若 $|\mathcal{H}_S| = 2^{|S|}$，则称 $\mathcal{H}$ 打散（shatters）集合 S。进一步，若 $\mathcal{H}$ 对 $\mathcal{X}$ 中任意大小为 d 的子集皆可打散，则称 $\mathcal{H}$ 能打散任意 d 个点。♣

命题 6.1.1 (断点 $k = 2$ 时的打散（shatter）任意两个点的含义)

设假设类 $\mathcal{H}$ 的最小断点为 $k = 2$，则 $\mathcal{H}$ 不能打散任意两个点是指存在某两点 $x_1, x_2 \in \mathcal{X}$，使得

$$\left|\{(h(x_1), h(x_2)) \mid h \in \mathcal{H}\}\right| < 4.$$

因而，对任何含两个元素的子集 $S = \{x_i, x_j\} \subseteq \mathcal{X}$，必有

$$|\mathcal{H}_S| \leq 3 < 2^{|S|} = 4,$$

即 $\mathcal{H}$ 至少缺失 $(\circ, \times)$ 或 $(\times, \circ)$ 或 $(\circ, \circ)$ 或 $(\times, \times)$ 中的一种标记组合。♠

命题 6.1.2 (最小断点 $k = 2$ 的限制)

设假设集 $\mathcal{H}$ 的最小断点（break point）为 $k = 2$。则对增长函数 $m_{\mathcal{H}}(N)$ 有

1) $N = 1$：由定义必有 $m_{\mathcal{H}}(1) = 2$；
2) $N = 2$：由定义必有 $m_{\mathcal{H}}(2) < 4$，故最大可能值为 3；
3) $N = 3$：需满足假设空间 $\mathcal{H}$ 无法打散任意两点，即对于给定的三个输入点 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$，任意选取其中两点，其对应的二分法组合数必须小于 4。换言之，不存在两点能被 $\mathcal{H}$ 完全打散。根据这一限制，增长函数的上界为：

$$m_{\mathcal{H}}(3) \leq 4.$$

且若存在 5 个二分法，则必能打散某两点，与 $k = 2$ 矛盾。♠

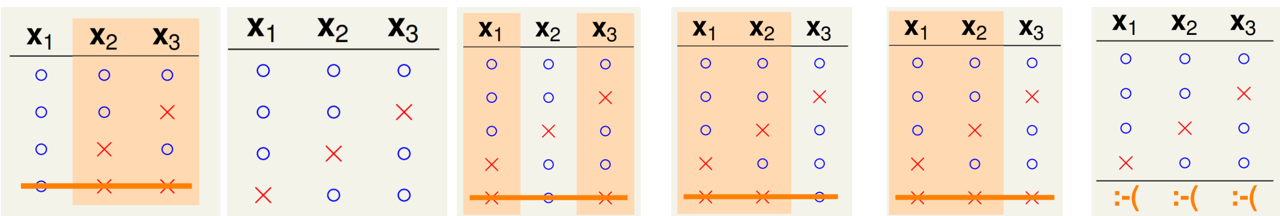


图 6.1.1: 最小断点 $k = 2$ 时二分法组合的示例图

命题 6.1.3 (断点限制（k = 2 情形）)

设假设集 $\mathcal{H}$ 的最小断点（break point）为 $k=2$，则其增长函数 $m_{\mathcal{H}}(N)$ 必须满足

$$m_{\mathcal{H}}(1)=2, \qquad m_{\mathcal{H}}(2)\le 3, \qquad m_{\mathcal{H}}(3)\le 4<2^3.$$

这表明，当 $N>k$ 时，断点 k 会显著限制 $m_{\mathcal{H}}(N)$ 的最大可能值，使其严格小于 2^N。

由此产生一个待验证的猜想：若存在最小断点 k，则对所有 N 有

$$m_{\mathcal{H}}(N)\le \text{给定}k\text{ 时的最大可能}m_{\mathcal{H}}(N)\le \text{多项式}(N).$$

从而 $m_{\mathcal{H}}(N)$ 由指数级降为多项式级，但尚需后续证明。

例题 6.1 设假设集 $\mathcal{H}$ 的最小断点 $k=1$。问：当 $N=3$ 时，$m_{\mathcal{H}}(3)$ 的最大可能值为多少？

选项 1 2 4 8

解答 由于 $k=1$，$\mathcal{H}$ 连一个点都无法打散。因此，在 3 个输入的每一列中均不能同时出现 $\circ$ 与 $\times$。包含第一种二分法后，就无法再引入任何不同的二分法，故最大可能值为 1。■

6.2 界函数（基础情形）

定义 6.2.1 (界函数（Bounding Function）)

定义界函数

$$B(N,k)\triangleq \max_{\mathcal{H}:\text{ break point}=k} m_{\mathcal{H}}(N),$$

即当最小断点为 k 时，任意假设集 $\mathcal{H}$ 在 N 个输入上所能达到的最大增长函数值。

组合意义 $B(N,k)$ 是满足下列条件的二进制向量（长度 N，元素为 $\circ,\times$）的最大数量：不存在任何长度为 k 的子向量同时包含两种符号（即无法打散任意 k 个点）。

性质与用途

- $B(N,k)$ 仅由 N,k 决定，与 $\mathcal{H}$ 的具体形式无关。
- 例如：$B(N,3)$ 同时给出正区间（$k=3$）和 1D 感知机（$k=3$）的统一上界。
- 新目标：证明 $B(N,k)\le$ 多项式 (N)。

表 6.2.1: 界函数 $B(N,k)$ 的已知值

$B(N,k)$		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			15	16	16	...
	5	1				31	32	...
	6	1					63	...
	⋮	⋮						⋱

已知规律

- 当 $N<k$ 时：$B(N,k)=2^N$（尚未触发断点条件）。

- 当 $N = k$ 时：$B(N,k) = 2^N - 1$（去掉任一单个二分法即可满足断点）。
- 当 $N > k$ 时：数值继续按组合规律递减，且总体保持多项式增长。

例题 6.2 对二维感知机（2D perceptrons），以下哪一句话正确？

选项

1) 最小断点 $k = 2$；
2) $m_{\mathcal{H}}(4) = 15$；
3) 当 $N = k =$ 最小断点时，$m_{\mathcal{H}}(N) < B(N,k)$；
4) 当 $N = k =$ 最小断点时，$m_{\mathcal{H}}(N) > B(N,k)$。

解答 已知二维感知机的最小断点为 $k = 4$，且 $m_{\mathcal{H}}(4) = 14$，而界函数给出 $B(4,4) = 15$。因此

$$m_{\mathcal{H}}(4) = 14 < 15 = B(4,4),$$

即界函数在 $N = k$ 时可能"宽松"。正确选项为 $\boxed{3}$。 ■

6.3 界函数（归纳情形）

在获得若干基础情形的界函数值后，我们可借助归纳法推导其余取值。作为示例，先通过组合分析求出 $B(4,3)$ 的具体数值，再对所得二分法进行重新归类与配对，得到如下示意图。

	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	○	×	×	○

图 6.3.1: B(4,3) 的排列组合形式

估计 $B(4,3)$ 的两部分

第一部分（$\alpha + \beta$ 的约束）

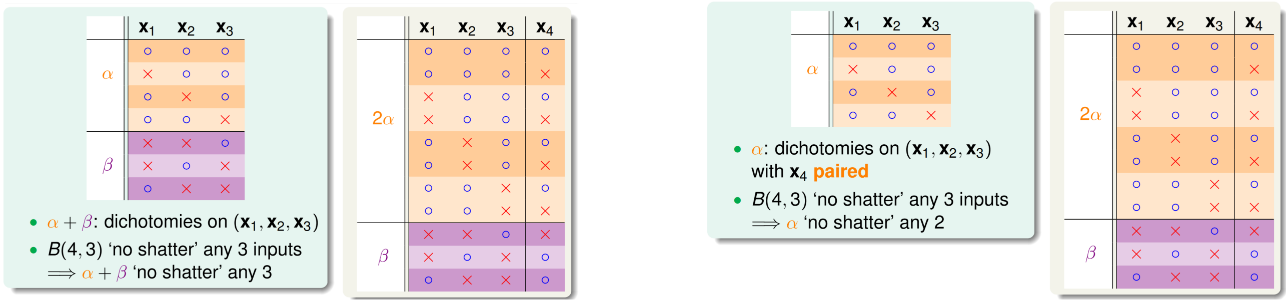
$$B(4,3) = 11 = 2\alpha + \beta, \qquad \alpha + \beta \le B(3,3).$$

其中 $\alpha + \beta$ 是在 (X_1, X_2, X_3) 上满足"不打散任意 3 点"的二分法数。

第二部分（α 的约束）

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

α 为在 (X_1, X_2, X_3) 上满足"不打散任意 2 点"的二分法数，且每个二分法与 X_4 成对出现（X_4 取相反符号）。



(a) $\alpha+\beta$ 的约束　　**(b)** α 的约束

图 6.3.2: α 和 β 在 $B(4,3)$ 上的约束

归纳上界

核心关系

$$B(4,3)=2\alpha+\beta,\quad \alpha\le B(3,2)=4,\quad \alpha+\beta\le B(3,3)=8.$$

由此得到

$$B(4,3)\le B(3,3)+B(3,2)=8+4=12,\quad \text{精确值为}11.$$

一般递推

$$B(N,k)\le B(N-1,k)+B(N-1,k-1),\quad \forall N,k\ge 1.$$

界函数上界（部分）

$B(N,k)$		$k=1$	2	3	4	5	6
N	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	15	16	16
	5	1	≤ 6	≤ 16	≤ 26	31	32
	6	1	≤ 7	≤ 22	≤ 42	≤ 57	63

定理 6.3.1 (界函数定理)

对任意正整数 N,k，界函数满足

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

其最高次项为 N^{k-1}。对固定的 k，$B(N,k)$ 被 N 的多项式上界所控制。

注 上式中的不等号“$\le$”事实上可取等号，读者可自行证明，下面给出一种证明思路。

证明 设有 N 个元素的点集 C，考虑所有 0-1 标签序列，其中最多只有 $k-1$ 个位置标记为 1，其可能的标签总数为

$$\sum_{i=0}^{k-1}\binom{N}{i}.$$

任何函数类若不能在大小为 k 的集合上构造出全部 2^k 种标记，那么它在任意 N 个点上所能构造的标签数必然不超过上述数量，因此 $B(N,k) \le \sum_{i=0}^{k-1} \binom{N}{i}$。

我们构造函数类 $\mathcal{F} = \{f_A \mid A \subseteq C, |A| \le k-1\}$，其中 $f_A(x) = 1$ 当且仅当 $x \in A$。此类函数产生的标签正好是所有 0-1 向量中"1"的个数不超过 $k-1$ 的那些，共计 $\sum_{i=0}^{k-1} \binom{N}{i}$ 种。即上界可达，因此 $B(N,k) \ge \sum_{i=0}^{k-1} \binom{N}{i}$。

因此 $B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i}$，且其最高次项为 $\binom{N}{k-1} \sim \frac{N^{k-1}}{(k-1)!}$。 ■

推论 6.3.1

若假设集 $\mathcal{H}$ 存在断点 k，则

$$m_{\mathcal{H}}(N) \le B(N,k) \le 多项式(N).$$

♡

6.4 图解证明

定理 6.4.1 (通用假设集的 PAC 上界（BAD Bound）)

对任意假设集 $\mathcal{H}$，若存在最小断点 k，则对充分大的样本量 N 和任意 $\varepsilon > 0$ 有

$$\mathbb{P}\Big[\exists h \in \mathcal{H},\ \big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \varepsilon\Big] \le 2\, m_{\mathcal{H}}(2N) \exp\Big(-\frac{\varepsilon^2 N}{8}\Big).$$

♡

为证明该定理，我们先引入必要的定义和引理：

定义 6.4.1 (经验误差与泛化误差)

设假设空间为 $\mathcal{H}$，训练集为 $S = \{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$，其中每个样本 (x_i,y_i) 独立同分布于数据生成分布 D。对任意假设 $h \in \mathcal{H}$，定义：

- 经验误差（in-sample error）：

$$E_{\text{in}}(h) \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big[h(x_i) \ne y_i\big]$$

  表示假设 h 在训练集 S 上的平均错误率，其中 $\mathbb{I}[\cdot]$ 为指示函数（条件成立时取 1，否则取 0）。

- 泛化误差（out-of-sample error）：

$$E_{\text{out}}(h) \;=\; \mathbb{E}_{(x,y)\sim D}\big[\mathbb{I}\big[h(x) \ne y\big]\big]$$

  表示假设 h 在整个数据分布 D 上的期望错误率，$\mathbb{E}_{(x,y)\sim D}[\cdot]$ 表示对分布 D 的期望。

♣

引理 6.4.1 (Symmetrization 不等式)

设 S 和 S' 是来自分布 D 的独立同分布样本集（大小均为 N），则对任意 $\epsilon > 0$：

$$\mathbb{P}_S\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \le 2 \cdot \mathbb{P}_{S,S'}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2}\right]$$

其中 $E'_{\text{in}}(h)$ 是 h 在 S' 上的经验误差。

♡

证明 [定理的证明（仅作参考，读者有兴趣可自行证明）]

步骤 **1**：固定假设的误差界

对任意固定假设 $h \in \mathcal{H}$，$E_{\text{in}}(h)$ 是 N 个独立 $[0,1]$ 变量 $\mathbb{I}[h(x_i) \neq y_i]$ 的均值，而 $E_{\text{out}}(h)$ 是其期望。由 Hoeffding 不等式：

$$\mathbb{P}\left(|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right) \leq 2\exp(-2\epsilon^2 N)$$

步骤 **2**：有限假设空间的联合界

若 $\mathcal{H}$ 含有限个假设，对所有 $h \in \mathcal{H}$ 应用联合界：

$$\mathbb{P}\left[\exists h \in \mathcal{H}, |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq \sum_{h \in \mathcal{H}} 2\exp(-2\epsilon^2 N) = 2|\mathcal{H}|\exp(-2\epsilon^2 N)$$

步骤 **3**：无限假设空间的处理

对于无限 $\mathcal{H}$，利用增长函数限制可区分数目。考虑大小为 $2N$ 的样本集 $S \cup S'$，$\mathcal{H}$ 在其上的划分数目不超过 $m_{\mathcal{H}}(2N)$，故：

$$|\{(h|_S, h|_{S'}) \mid h \in \mathcal{H}\}| \leq m_{\mathcal{H}}(2N)$$

其中 $h|_S$ 表示 h 在 S 上的预测标签。

步骤 **4**：应用 **Symmetrization** 技巧

由 Symmetrization 不等式，原概率可转化为两个经验误差差的概率：

$$\mathbb{P}\left[\exists h \in \mathcal{H}, |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2 \cdot \mathbb{P}_{S,S'}\left[\exists h \in \mathcal{H}, |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2}\right]$$

对 $S \cup S'$ 上的所有可能划分应用联合界，结合 $m_{\mathcal{H}}(2N) \leq 2^{m_{\mathcal{H}}(2N)}$ 及 Hoeffding 不等式的松弛，最终可得：

$$\mathbb{P}\left[\exists h \in \mathcal{H}, |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2 \cdot m_{\mathcal{H}}(2N) \cdot \exp\left(-\frac{\epsilon^2 N}{8}\right)$$

■

6.5 总结

笔记 [泛化理论]

- 断点的限制：一旦存在断点 k，其后的所有点都会被“截断”。
- 界函数（基础情形）：定义 $B(N,k)$，用来在断点为 k 时给出 $m_{\mathcal{H}}(N)$ 的上界。
- 界函数（归纳情形）：$B(N,k)$ 对 N 呈多项式增长，即 $\text{poly}(N)$。
- 图解证明：通过少量改动，可将原先与 M 相关的界替换为与 $m_{\mathcal{H}}(N)$ 相关的界。