第 5 章 训练 vs 测试

5.1 回顾与展望

假设数量 M 的权衡

联合界给出

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

	小 M	大 M
$E_{\text{in}} \approx E_{\text{out}}$	√	× (上界失效)
E_{in} 足够小	× (选择少)	√

结论 M 过小或过大都会使学习失效，必须选择合适的 M（或假设集 $\mathcal{H}$）。

预告：下一步工作

已知 对有限假设集 $\mathcal{H}$ 大小为 M 时，有

$$\mathbb{P}\big[\,|E_{\text{in}}(g) - E_{\text{out}}(g)| > \varepsilon\big] \le 2M\exp(-2\varepsilon^2 N).$$

待办

- 用有限量 $m_{\mathcal{H}}$ 取代 M，得到
$$\mathbb{P}\big[\,|E_{\text{in}}(g) - E_{\text{out}}(g)| > \varepsilon\big] \le 2m_{\mathcal{H}}\exp(-2\varepsilon^2 N).$$
- 论证当 $|\mathcal{H}|$ 无限时学习仍可行。
- 研究 $m_{\mathcal{H}}$ 的性质，从而像调节 M 一样选择“正确”的 $\mathcal{H}$。
- 最终彻底揭开 PLA 的神秘面纱。

5.2 有效直线数

联合界为何失效？

已知上界

$$\mathbb{P}\big[|E_{\text{in}}(g)-E_{\text{out}}(g)|>\varepsilon\big]\le 2M\exp(-2\varepsilon^2N)$$

来自联合界 $\mathbb{P}[B_1\cup\cdots\cup B_M]\le\sum_m\mathbb{P}[B_m]$，其中 B_m 表示事件 $\{|E_{\text{in}}(h_m)-E_{\text{out}}(h_m)|>\varepsilon\}$。

联合界的不足

- 最坏情形假设所有 B_m 互斥，实际却大量重叠；
- 当 $h_1\approx h_2$ 时，$E_{\text{out}}(h_1)\approx E_{\text{out}}(h_2)$，且对大多数样本 $\mathcal{D}$ 有 $E_{\text{in}}(h_1)=E_{\text{in}}(h_2)$；
- 联合界严重高估真实概率。

下一步思路　将相似假设按“类别”合并，用有限量 $m_{\mathcal{H}}$ 取代 M，从而得到更紧致的上界。

直线到底有多少种？（二维平面）

问题设定　假设集 $\mathcal{H}$ 为 $\mathbb{R}^2$ 中所有直线。

从输入点看种类数

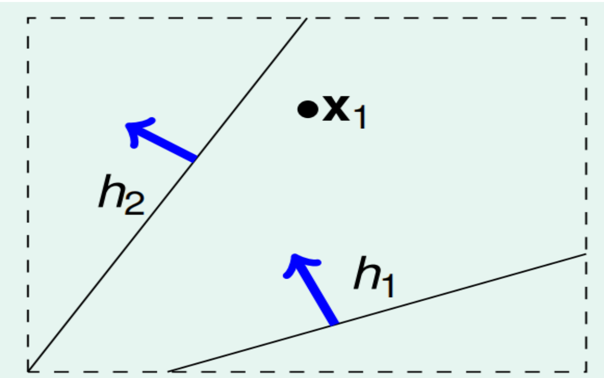
- 若仅看单个输入 $\mathbf{x}_1$，直线可把 $\mathbf{x}_1$ 分成两类：

  $$\text{正类}\circ\quad\text{或}\quad\text{负类}\times.$$

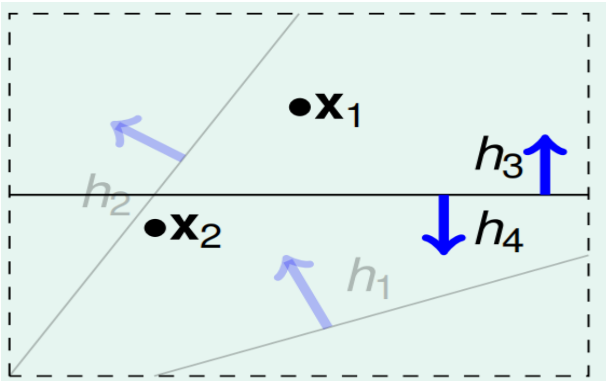
  故 **2** 种。
- 若看两个输入 $\mathbf{x}_1,\mathbf{x}_2$，直线可把 $(\mathbf{x}_1,\mathbf{x}_2)$ 分成

  $$(\circ,\circ),\ (\circ,\times),\ (\times,\circ),\ (\times,\times)$$

  共 **4** 种。



(a) 一个输入的分类



(b) 两个输入的分类

图 5.2.1: 二维平面输出的分类

三个输入时直线有多少种划分？

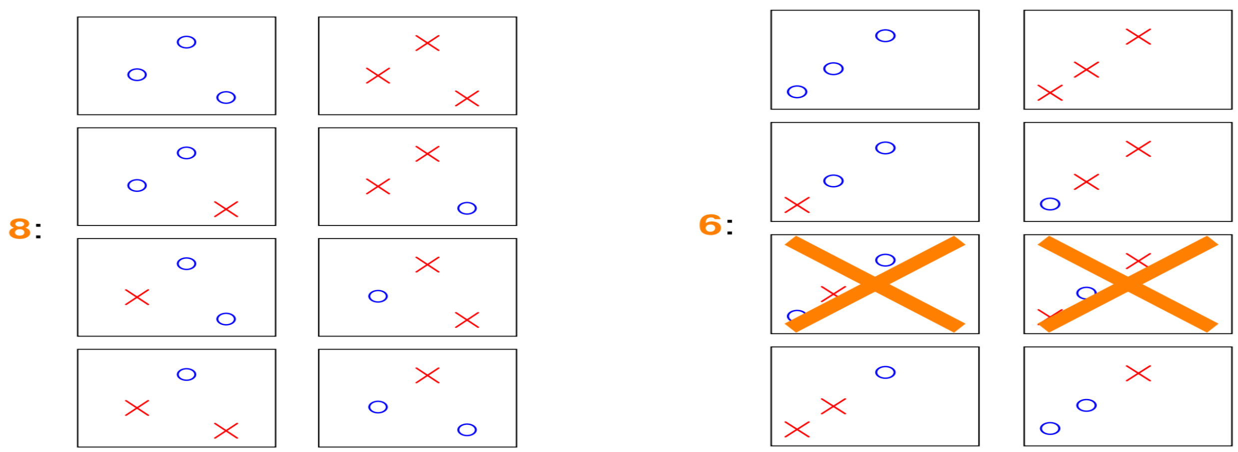
一般情形 当输入 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ 处于一般位置（不共线）时，平面中的直线最多可将这 3 个点分成

$$2^3 = 8$$

种标签组合，每种组合均可由某条直线实现。

退化情形 若三点共线或出现重复输入，则某些标签组合无法由直线实现，实际可实现的划分种类数 **少于 8 种**（例如只有 6 种）。即

$$m_{\mathcal{H}}(3) = 6.$$



(a) 一般情形 (b) 退化情形

图 5.2.2: 三个输入时直线的划分情况

四条输入时直线有多少种划分？

设定假设集 $\mathcal{H}$ 为二维平面 $\mathbb{R}^2$ 中的所有直线，给定任意四个输入点 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$。

最大可实现划分 无论四点如何摆放（一般位置或退化），平面直线最多可将这 4 个点分成 14 种不同的标签组合（而非 $2^4 = 16$ 种）。即

$$m_{\mathcal{H}}(4) = 14.$$

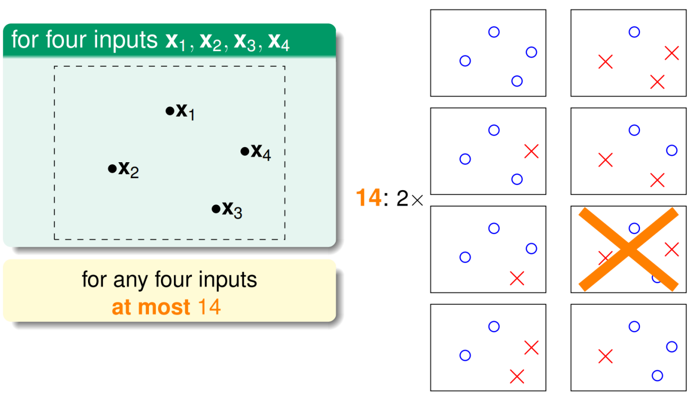


图 5.2.3: 四个输入时直线的划分情况

命题 5.2.1 (有效直线数)

令 $\mathcal{H}$ 为二维平面所有直线组成的假设集，$m_{\mathcal{H}}(N)$ 表示对任意 N 个输入点 $\mathbf{x}_1,\dots,\mathbf{x}_N$ 的最大可实现划分种类数，则有

$$m_{\mathcal{H}}(N) \le 2^N \quad \text{且} \quad \begin{cases} m_{\mathcal{H}}(1) = 2, \\ m_{\mathcal{H}}(2) = 4, \\ m_{\mathcal{H}}(3) = 8, \\ m_{\mathcal{H}}(4) = 14. \end{cases}$$

利用 $m_{\mathcal{H}}(N)$ 替代原先的 M，可得

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

由于 $m_{\mathcal{H}}(N) \ll 2^N$ 且随 N 多项式增长，即使 $\mathcal{H}$ 包含无限多条直线，学习仍然可行。♠

例题 5.1 五条输入的有效直线数 在二维平面 $\mathbb{R}^2$ 中，给定五条输入点 $\mathbf{x}_1,\dots,\mathbf{x}_5$。问：平面直线最多可把这五点分成多少种不同的标签组合？

选项 14 16 22 32

解答 将五条输入大致均匀摆放在一个圆周上。对圆周上任意一段“连续”的点，可画一条直线使其落在同一侧，其余点在另一侧。按此方式枚举，可得 22 种不同的划分，远小于理论上限 $2^5 = 32$。后续课程将给出形式化证明。■

5.3 有效假设数

二分法（Dichotomies）：迷你假设

定义 设假设集 $\mathcal{H} = \{h : \mathcal{X} \to \{\times, \circ\}\}$。给定 N 个输入点 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$，定义

$$\mathcal{H}(\mathbf{x}_1,\dots,\mathbf{x}_N) = \big\{(h(\mathbf{x}_1),\dots,h(\mathbf{x}_N)) \mid h \in \mathcal{H}\big\}$$

为 $\mathcal{H}$ 在这些点上“实现”的所有 **二分法**（dichotomies）。

示例 $\mathcal{H}$ = 平面所有直线，则

$$\mathcal{H}(\mathbf{x}_1,\dots,\mathbf{x}_N) = \{\circ \circ \circ\circ, \circ \circ \circ\times, \circ \circ \times\times, \dots\}$$

其大小可能无限，但受限于 2^N。

用途 $|\mathcal{H}(\mathbf{x}_1,\dots,\mathbf{x}_N)|$ 可作为有限量替代原先的 M，用于建立 PAC 界。

定义 5.3.1 (增长函数（Growth Function）)

设假设集 $\mathcal{H} \subseteq \{h : \mathcal{X} \to \{\times, \circ\}\}$。定义其增长函数为

$$m_{\mathcal{H}}(N) \triangleq \max_{(\mathbf{x}_1,\dots,\mathbf{x}_N) \in \mathcal{X}^N} \big|\mathcal{H}(\mathbf{x}_1,\dots,\mathbf{x}_N)\big|,$$

其中 $\mathcal{H}(\mathbf{x}_1,\dots,\mathbf{x}_N)$ 表示 $\mathcal{H}$ 在 N 个输入点上可实现的所有二分法集合。$m_{\mathcal{H}}(N)$ 满足：

- 有限且 $m_{\mathcal{H}}(N) \le 2^N$；
- 用于 PAC 界：$\mathbb{P}\big[|E_{\text{in}} - E_{\text{out}}| > \varepsilon\big] \le 2 m_{\mathcal{H}}(N) e^{-2\varepsilon^2 N}$；

- 对二维直线，$m_{\mathcal{H}}(N)$ 的前几项为

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

♣

正射线（Positive Rays）的增长函数

设定　设输入空间 $\mathcal{X}=\mathbb{R}$（一维），假设集

$$\mathcal{H}=\{h \mid h(x)=\operatorname{sign}(x-a),\, a\in\mathbb{R}\},$$

即所有以阈值 a 为界的“正射线”，输出左侧为 -1，右侧为 $+1$。

二分法计数　对任意 N 个有序输入 $x_1<x_2<\cdots<x_N$，阈值 a 只能在 $(-\infty,x_1],[x_1,x_2],\ldots,[x_N,+\infty)$ 这 $N+1$ 个区间取值，每个区间对应一种不同的二分法，故

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

性质

$$N+1<2^N \quad \text{当}N\text{ 较大时成立}.$$

因此即使 $\mathcal{H}$ 无限，仍可用 $m_{\mathcal{H}}(N)=N+1$ 取代原先的 M，保证 PAC 上界有效。

正区间（Positive Intervals）的增长函数

设定　输入空间 $\mathcal{X}=\mathbb{R}$（一维），假设集

$$\mathcal{H}=\{h \mid h(x)=\mathbb{I}(x\in[a,b))\}.$$

其中 $\mathbb{I}[\cdot]$ 为示性函数：区间内输出 $+1$，其余输出 -1。

二分法计数　对任意 N 个有序输入 $x_1<x_2<\cdots<x_N$，区间端点 a,b 只能落在

$$(-\infty,x_1],\ [x_1,x_2],\ \ldots,\ [x_N,+\infty)$$

共 $N+1$ 个位置。选取两个端点（可重合）决定区间，因此

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

（常简记为 $\mathcal{O}(N^2)$）

性质

$$\frac{N^2+N+2}{2}\ <\ 2^N \quad \text{当}N\text{ 较大时成立},$$

故 $\mathcal{H}$ 虽无限，仍可用多项式增长的 $m_{\mathcal{H}}(N)$ 替代原先的 M，PAC 学习可行。

凸集（Convex Sets）的增长函数

设定　输入空间 $\mathcal{X}=\mathbb{R}^2$，假设集

$$\mathcal{H}=\{h \mid h(\mathbf{x})=\mathbb{I}(\mathbf{x}\in C),\ C\subseteq\mathbb{R}^2\text{ 为凸集}\},$$

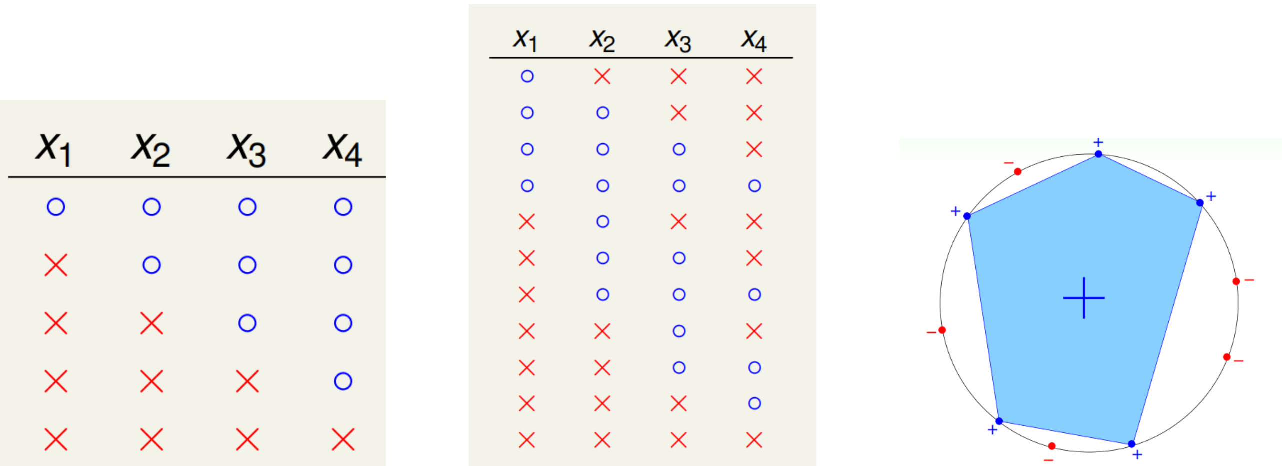
其中 $\mathbb{I}(\cdot)$ 为示性函数：在凸集 C 内输出 $+1$，否则 -1。

增长函数把 N 个点均匀放在一个大圆周上，则

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

并称这 N 个点被 $\mathcal{H}$ **打散**（shattered）。

结论　凸集假设集满足 $m_{\mathcal{H}}(N)=2^N$，VC 维无限，需额外技术（如正则化）控制复杂度。



(a) 正射线的增长函数　(b) 正区间的增长函数　(c) 凸集的增长函数

图 5.3.1: 三种假设集的增长函数对比

5.4 断点

四种假设集的增长函数

假设集	增长函数 $m_{\mathcal{H}}(N)$
正射线（positive rays）	$N+1$
正区间（positive intervals）	$\dfrac{N^2+N+2}{2}$
凸集（convex sets）	2^N
二维感知机（2D perceptrons）	多项式上界 $<2^N$

PAC 界　用 $m_{\mathcal{H}}(N)$ 替代原先的 M 得

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

结论　多项式 $m_{\mathcal{H}}(N)$：好；指数 $m_{\mathcal{H}}(N)$：坏。

笔记 [断点分类与理论猜想（The Four Break Points）]

四种典型的 Break Point 情况与理论猜想

- 正射线（**Positive Rays**）：
  - 断点出现在 2
  - 增长函数：$\mathcal{H}(N) = N + 1 = \mathcal{O}(N)$
- 正区间（**Positive Intervals**）：
  - 断点出现在 3
  - 增长函数：$\mathcal{H}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1 = \mathcal{O}(N^2)$
- 凸集（**Convex Sets**）：
  - 无断点
  - 增长函数：$\mathcal{H}(N) = 2^N$
- 二维感知机（**2D Perceptrons**）：
  - 断点出现在 4
  - 增长函数：$\mathcal{H}(N) < 2^N$ （在某些情况下）

理论猜想（**Conjecture**）：

- 若没有断点，则 $\mathcal{H}(N) = 2^N$
- 若存在断点 k，则 $\mathcal{H}(N) = \mathcal{O}(N^{k-1})$

5.5 总结

笔记 [训练 vs 测试]

- 回顾与展望：两个核心问题：$E_{\text{out}}(g) \approx E_{\text{in}}(g)$ 与 $E_{\text{in}}(g) \approx 0$。
- 有效直线数：在 4 个输入点看来，最多只有 14 条有效直线。
- 有效假设数：在 N 个输入点看来，有效假设数至多为增长函数 $m_{\mathcal{H}}(N)$。
- 断点（break point）：当 $m_{\mathcal{H}}(N)$ 不再呈指数增长时的关键转折点。

笔记 [总体结论] 为了解决在假设集无限时原有联合界失效的问题，引入了增长函数 $m_{\mathcal{H}}(N)$，用于衡量在 N 个样本点上，假设集 $\mathcal{H}$ 实际能实现的不同划分数量（即有效假设数）。尽管 $\mathcal{H}$ 可能是无限的，但其在有限数据上能产生的划分是有限的，例如二维平面中直线在 4 个点上最多只能划分出 14 种情况。增长函数提供了更精确的复杂度度量，使我们能够用更紧致的泛化上界替代原有依赖假设数 M 的不等式，从而在控制模型复杂度的同时，确保学习的可行性。