第 7 章 VC 维

7.1 VC 维的定义

命题 7.1.1 (VC 界回顾)

设假设集 $\mathcal{H}$ 具有最小断点 k。对任意学习算法 $\mathcal{A}$ 输出的假设 $g=\mathcal{A}(\mathcal{D})$，当样本量 N 足够大时，对任意 $\varepsilon>0$ 有

$$\mathbb{P}\Big[\exists h\in\mathcal{H},\ \big|E_{\text{in}}(h)-E_{\text{out}}(h)\big|>\varepsilon\Big]\le 4\,(2N)^{k-1}\exp\Big(-\frac{\varepsilon^2 N}{8}\Big).$$

若同时满足：

1) 假设集存在断点 k，即拥有好假设集；
2) 数据量 N 足够大，即拥有好数据；
3) 算法 $\mathcal{A}$ 选出具有较小 E_{in} 的假设 g，即拥有好算法；

则大概率成立 $E_{\text{out}}(g)\approx E_{\text{in}}(g)$，即学习成功。

定义 7.1.1 (VC 维)

假设集 $\mathcal{H}$ 的 VC 维（记作 $d_{\text{VC}}(\mathcal{H})$）定义为

$$d_{\text{VC}}(\mathcal{H})=\max\{N\in\mathbb{N}\mid m_{\mathcal{H}}(N)=2^N\},$$

即能被打散的最多输入点数。等价地，

$$d_{\text{VC}}(\mathcal{H})=\text{最小断点}k-1.$$

性质

- 若 $N\le d_{\text{VC}}$，则 $\mathcal{H}$ 能够打散某 N 个输入；
- 若 $k>d_{\text{VC}}$，则 k 是 $\mathcal{H}$ 的断点；
- 当 $N\ge 2$ 且 $d_{\text{VC}}\ge 2$ 时，$m_{\mathcal{H}}(N)\le N^{d_{\text{VC}}}$.

四种假设集的 VC 维及其学习保证

假设集	$m_{\mathcal{H}}(N)$	$d_{\text{VC}}(\mathcal{H})$
正射线（positive rays）	$N+1$	1
正区间（positive intervals）	$\dfrac{N^2+N+2}{2}$	2
凸集（Convex Sets）	2^N	∞
二维感知机（2D perceptrons）	$\le N^3\ (N\ge 2)$	3

学习保证 只要 $d_{\text{VC}}(\mathcal{H})<\infty$，学习所得假设 g 必以高概率满足 $E_{\text{out}}(g)\approx E_{\text{in}}(g)$ 且该结论

- 与具体学习算法 $\mathcal{A}$ 无关；
- 与输入分布 $\mathcal{P}$ 无关；
- 与真实目标函数 f 无关。

7.2 线性分类器的 VC 维

再探二维感知机学习算法（PLA）

设定

- 数据集 $\mathcal{D}$ 线性可分，$X_n \sim \mathcal{P}$，$y_n = f(X_n)$；
- PLA 保证收敛。

VC 维保证

$$\mathbb{P}\big[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \varepsilon\big] \le 4\,(2N)^{d_{\text{VC}}} \exp\Big(-\frac{\varepsilon^2 N}{8}\Big) \quad (\text{由 } d_{\text{VC}} = 3 \text{ 控制}).$$

当迭代次数 T 与样本量 N 足够大时

$$E_{\text{in}}(g) = 0 \quad \Longrightarrow \quad E_{\text{out}}(g) \approx E_{\text{in}}(g) \approx 0.$$

猜想　对于特征维度 $d > 2$ 的一般 PLA，上述结论同样成立，只需将 d_{VC} 替换为 $d+1$。

例题 7.1 选择题：证明 $d_{\text{VC}} \ge d+1$　下列哪一句话最能说明 $d_{\text{VC}} \ge d+1$?

1) 存在一组 $d+1$ 个输入可以被我们打散。
2) 任意一组 $d+1$ 个输入都可以被我们打散。
3) 存在一组 $d+2$ 个输入我们无法打散。
4) 任意一组 $d+2$ 个输入我们都无法打散。

解答　正确选项为 $\boxed{1}$。根据 VC 维定义，d_{VC} 是使得 $m_{\mathcal{H}}(N) = 2^N$ 的最大 N。因此若存在一组 $d+1$ 个输入可以被我们打散，则能在某 $d+1$ 个输入上构造出全部 2^{d+1} 种二分法，则 $m_{\mathcal{H}}(d+1) = 2^{d+1}$，从而 $d_{\text{VC}} \ge d+1$。■

命题 7.2.1 ($d_{\text{VC}} \ge d+1$（一般 d 维情形）)

在 d 维空间 $\mathbb{R}^d$（含偏置项）中，存在 $d+1$ 个输入点可以被线性假设集打散。构造“平凡”输入矩阵

$$X = \begin{bmatrix} -x_1^T- \\ -x_2^T- \\ -x_3^T- \\ \vdots \\ -x_{d+1}^T- \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}_{(d+1)\times(d+1)}$$

则 X 可逆。对任意标签向量 $\mathbf{y} \in \{-1,+1\}^{d+1}$，令

$$\mathbf{w} = X^{-1}\mathbf{y} \Leftrightarrow X\mathbf{w} = \mathbf{y}$$

便有

$$\text{sign}(X\mathbf{w}) = \mathbf{y},$$

因而这 $d+1$ 个点可以被线性分类器打散，故

$$d_{\text{VC}} \ge d+1.$$

♠

例题 7.2 选择题：证明 $d_{\text{VC}} < d+2$　下列哪一句话最能说明 $d_{\text{VC}} < d+2$?

1) 存在一组 $d+1$ 个输入可以被我们打散。

2) 任意一组 $d+1$ 个输入都可以被我们打散。
3) 存在一组 $d+2$ 个输入我们无法打散。
4) 任意一组 $d+2$ 个输入我们都无法打散。

解答 正确选项为 $\boxed{4}$。根据 VC 维定义，d_{VC} 为使得 $m_{\mathcal{H}}(N)=2^N$ 的最大 N。
若任意一组 $d+2$ 个输入我们都无法打散，则对任意 $d+2$ 个输入都无法产生全部 2^{d+2} 种二分法，即 $m_{\mathcal{H}}(d+2)<2^{d+2}$，则 $d+2$ 是一个断点，从而 $d_{\text{VC}}<d+2$，亦即 $d_{\text{VC}}\le d+1$。 ■

命题 7.2.2 ($d_{\text{VC}}\le d+1$（一般 d 维情形）)

设输入空间为 $\mathbb{R}^d$（含偏置项），令 $X=[\mathbf{x}_1,\dots,\mathbf{x}_{d+2}]\in\mathbb{R}^{(d+1)\times(d+2)}$ 为任意 $d+2$ 个输入组成的矩阵。由于列数多于行数，X 的列必然线性相关，即存在不全为零的系数 $a_1,\dots,a_{d+1}$ 使得

$$\mathbf{x}_{d+2}=a_1\mathbf{x}_1+a_2\mathbf{x}_2+\cdots+a_{d+1}\mathbf{x}_{d+1}.$$

假设线性分类器能打散这 $d+2$ 个点，则对任意标签分配 $y_1,\dots,y_{d+2}\in\{-1,+1\}$，都存在 $\mathbf{w}\in\mathbb{R}^{d+1}$ 满足：

$$\operatorname{sign}(\mathbf{w}^\top\mathbf{x}_i)=y_i\quad(\forall i=1,\dots,d+2).$$

特别地，取 $y_i=\operatorname{sign}(a_i)$（若 $a_i=0$ 则 y_i 任意），且 $y_{d+2}=-1$。此时：

$$\mathbf{w}^\top\mathbf{x}_{d+2}=\sum_{i=1}^{d+1}a_i(\mathbf{w}^\top\mathbf{x}_i)=\sum_{i=1}^{d+1}a_iy_i=\sum_{i=1}^{d+1}|a_i|>0,$$

但根据标签定义要求 $\mathbf{w}^\top\mathbf{x}_{d+2}=y_{d+2}<0$，矛盾。因此，任意 $d+2$ 个输入均无法被线性假设集打散，从而

$$d_{\text{VC}}\le d+1.$$

♠

定理 7.2.1 (线性分类器的 VC 维)

设输入空间为 $\mathbb{R}^d$（即包含偏置项的 $(d+1)$ 维增广空间），则线性假设集

$$\mathcal{H}=\{h\mid h(\mathbf{x})=\operatorname{sign}(\mathbf{w}^\top\mathbf{x}),\ \mathbf{w}\in\mathbb{R}^{d+1}\}$$

的 VC 维满足

$$d_{\text{VC}}(\mathcal{H})=d+1.$$

♡

7.3 VC 维的直观物理意义

自由度（Degrees of Freedom）

参数自由度	$\mathbf{w}=(w_0,w_1,\dots,w_d)\in\mathbb{R}^{d+1}$
假设数量	$M=\lvert\mathcal{H}\rvert$ （“模拟”自由度)
假设“威力”	$d_{\text{VC}}(\mathcal{H})=d+1$ （“有效二进制”自由度)
结论	$d_{\text{VC}}(\mathcal{H})$ 衡量假设集的表达能力

VC 维与自由参数的例子

正射线 $(d_{\text{VC}} = 1)$

$$h(x) = \text{sign}(x - a), \qquad \text{自由参数：阈值} a \in \mathbb{R}.$$

正区间 $(d_{\text{VC}} = 2)$

$$h(x) = \text{sign}\big[(x - l)(r - x)\big], \qquad \text{自由参数：左端} l, \text{右端} r \in \mathbb{R}.$$

经验法则

$$d_{\text{VC}} \approx \text{自由参数个数} \quad (\text{通常成立，但并非绝对})$$

例题 7.3 过原点超平面的 VC 维 在 $\mathbb{R}^d$ 中，仅考虑通过原点的超平面（即固定偏置项 $w_0 = 0$ 的感知机），其 VC 维为多少？

选项 1 d d+1 ∞

解答 证明思路与普通感知机几乎相同，与普通感知机的区别在于通过原点的超平面的偏置项为 0，即输入空间 $\mathbb{R}^d$ 的维度与 VC 维大小相同，因此 $d_{\text{VC}} = d$。

但可直接利用 $d_{\text{VC}} \approx$ 自由参数个数这一直觉：通过原点的超平面仅含 d 个自由参数 $(w_1, \ldots, w_d)$，因此 $d_{\text{VC}} = d$。 ■

7.4 VC 维的解读

命题 7.4.1 (VC 界的详细重述：模型复杂度的惩罚项（Penalty for Model Complexity）)

设

- $\mathcal{H}$ 为任意假设集，其 VC 维 $d_{\text{VC}} \geq 2$；
- $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ 为独立同分布样本，$N \geq 2$；
- $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$ 为算法 $\mathcal{A}$ 输出的假设。

$$\mathbb{P}_{\mathcal{D}}\Big[\underbrace{|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon}_{\text{BAD}}\Big] \leq \underbrace{4(2N)^{d_{\text{VC}}} \exp\left(-\frac{1}{8}\epsilon^2 N\right)}_{\delta}$$

则对任意 $\delta > 0$，以概率至少 $1 - \delta$ 成立

$$\big|E_{\text{in}}(g) - E_{\text{out}}(g)\big| \leq \underbrace{\sqrt{\frac{8}{N}\left(\frac{4(2N)^{d_{\text{VC}}}}{\delta}\right)}}_{\Omega(N, d_{\text{VC}}, \delta)}.$$

其中

$$\Omega(N, d_{\text{VC}}, \delta) = \sqrt{\frac{8}{N} \ln\left(\frac{4(2N)^{d_{\text{VC}}}}{\delta}\right)}$$

♠

VC 维相关信息

以高概率成立

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \underbrace{\sqrt{\frac{8}{N}\left(\frac{4(2N)^{d_{\text{VC}}}}{\delta}\right)}}_{\Omega(N, d_{\text{VC}}, \delta)}$$

- 模型复杂度 $d_{\text{VC}} \uparrow$
  $\Longrightarrow$ 泛化界 $\Omega \uparrow$，但训练误差 $E_{\text{in}} \downarrow$。
- 模型复杂度 $d_{\text{VC}} \downarrow$
  $\Longrightarrow$ 泛化界 $\Omega \downarrow$，但训练误差 $E_{\text{in}} \uparrow$。
- 最优 d_{VC}^* 位于中间地带。

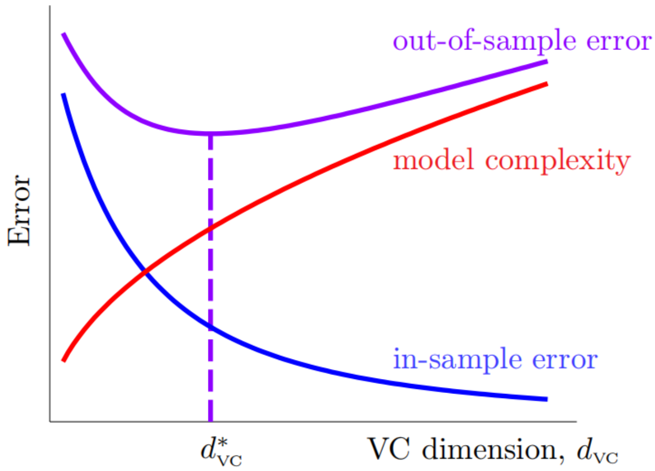

图 7.4.1: 模型复杂度、样本内误差、样本外误差与 VC 维关系图

命题 7.4.2 (VC 界的重新表述：样本复杂度（Sample Complexity）)

给定精度 $\varepsilon = 0.1$、置信参数 $\delta = 0.1$ 与 VC 维 $d_{\text{VC}} = 3$，若要求

$$4\,(2N)^{d_{\text{VC}}} \exp\left(-\frac{\varepsilon^2 N}{8}\right) \le \delta,$$

则最小样本量 N 的理论上界为

$$N \approx 29{,}300 \approx 10^4\, d_{\text{VC}}.$$

实际经验法则常取

$$N \approx 10\, d_{\text{VC}}$$

即可满足泛化需求。

命题 7.4.3 (VC 界的松弛性（Looseness of the VC Bound))

$$\mathbb{P}\big[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \varepsilon\big] \le 4\,(2N)^{d_{\text{VC}}} \exp\left(-\frac{\varepsilon^2 N}{8}\right).$$

理论与实践的差距

理论要求：$N \sim 10^4\, d_{\mathrm{VC}}$；实际经验：$N \sim 10\, d_{\mathrm{VC}}$ 即够。为何如此松弛？

四大保守来源

1) 使用 Hoeffding 不等式处理未知 E_{out}，需对任意分布、任意目标函数均成立；
2) 用增长函数 $m_{\mathcal{H}}(N)$ 代替条件集合 $\mathcal{H}(X_1, \ldots, X_N)$ 的规模；
3) 进一步用 $N^{d_{\mathrm{VC}}}$ 代替 $m_{\mathcal{H}}(N)$，对所有同 d_{VC} 的假设集通用；
4) 采用联合界（union bound）覆盖最坏情形，同时保护算法 $\mathcal{A}$ 的任意选择。

结论　虽然 VC 界在数值上非常宽松，但它对所有模型"同样宽松"，因此仍能有效指导机器学习改进方向——这正是 VC 界的重要哲学意义。

7.5 总结

笔记 [VC 维]

- VC 维的定义：最大的非突破点，即能够被完全打散（shatter）的最大点数。
- 线性分类器的 VC 维：$d_{\mathrm{VC}}(\mathcal{H}) = d + 1$。
- VC 维的直观物理意义：d_{VC} 约等于模型中自由参数的个数。
- VC 维的解读：粗略地反映了模型复杂度与所需样本复杂度。

笔记 [总体结论] VC 维衡量的是假设集能打散的最大样本数，是模型复杂度的核心指标。它决定了在多大样本量下学习算法能有效泛化，即训练误差接近于测试误差。VC 界给出了泛化误差的概率上界，虽数值上偏保守，但为不同模型提供了统一的理论保障。整体上，VC 维帮助我们在模型复杂度与泛化能力之间取得平衡，是理解机器学习本质的重要工具。