第 12 章 非线性变换

12.1 二次假设

命题 12.1.1 (线性假设的局限与突破)

我们迄今讨论的线性假设

$$h(\mathbf{x}) = \operatorname{sign}(\mathbf{w}^\top \mathbf{x})$$

具有以下特征：

- 几何：决策边界为超平面（“直线”在高维推广）；
- 理论：对 d 维输入空间，VC 维恰为

$$d_{\mathrm{VC}} = d + 1,$$

- 数学：线性于参数 $\mathbf{w}$；
- 实践：仅对近似线性可分的数据表现良好。

局限　线性假设的表达能力受 VC 维限制，当数据分布非线性或特征维度不足时，E_{in} 往往较大，泛化性能受限。

突破路径　通过特征映射 $\phi : \mathbb{R}^d \to \mathbb{R}^D$（$D \gg d$）或核技巧，将数据嵌入高维空间，使线性超平面在原始空间表现为非线性决策边界，从而显著提升模型容量。

命题 12.1.2 (圆可分 ⇔ 线可分：二次特征映射视角)

令二次特征映射

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^6, \quad \Phi(x) = \left(1,\, x_1,\, x_2,\, x_1^2,\, x_1x_2,\, x_2^2\right)^\top,$$

记 $z = \Phi(x)$。对任意数据集 $\{(x_n, y_n)\} \subset \mathbb{R}^2 \times \{-1, +1\}$，以下等价：

i) 存在一条二次曲线（圆、椭圆、双曲线、抛物线或其退化形式）在 $\mathbb{R}^2$ 中完全分离该数据集；

ii) 存在向量 $w \in \mathbb{R}^6$ 使得在映射空间 $\mathbb{R}^6$ 中线性可分，即

$$y_n = \operatorname{sign}\left(w^\top z_n\right), \quad \forall n.$$

此时，原始空间的决策边界由

$$h(x) = \operatorname{sign}\left(w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_1x_2 + w_5x_2^2\right)$$

给出，其几何形状由系数 (w_3, w_4, w_5) 与二次型判别式共同决定。

定义 12.1.1 (通用二次假设集)

在二维输入空间 $\mathbb{R}^2$ 中，定义六维特征映射

$$\phi(x) = \left(1,\, x_1,\, x_2,\, x_1^2,\, x_1x_2,\, x_2^2\right) \in \mathbb{R}^6.$$

对应的二次假设集

$$\mathcal{H}_\phi = \left\{h : \mathbb{R}^2 \to \{-1, +1\} \mid h(x) = \operatorname{sign}(w^\top \phi(x)),\ w \in \mathbb{R}^6\right\}$$

具备如下性质：

- 与特征空间 $\mathbb{R}^6$ 中的线性感知机等价；
- 在原始空间可生成任意二次曲线边界：圆、椭圆、双曲线、抛物线等；
- 直线与常值函数为其退化特例；
- 示例：椭圆 $(x_1+x_2-3)^2+(x_1-x_2-4)^2=1$ 等价于

$$w^\top \phi(x)=0, \quad w=[33,-20,-4,3,2,3]^\top.$$

♣

12.2 非线性变换

命题 12.2.1 (从好感知机到好二次假设)

设原始输入空间为 $\mathcal{X}$，特征映射 $\phi:\mathcal{X}\to\mathcal{Z}$ 定义为

$$z=\phi(x)=\left(1,x_1,x_2,x_1^2,x_1x_2,x_2^2\right)^\top.$$

已知

- 在 $\mathcal{X}$ 空间中，我们已能利用数据 $\{(x_n,y_n)\}$ 训练出好的线性感知机；
- 对应的 $\mathcal{Z}$ 空间中的数据为 $\{(z_n=\phi(x_n),y_n)\}$。

目标

- 在 $\mathcal{Z}$ 空间中训练一条好的线性感知机

$$h(z)=\operatorname{sign}(w^\top z),$$

其在 $\mathcal{X}$ 空间即表现为一条优秀的二次分界线。

♠

算法 12.2.1: 非线性变换三步法

输入: 原始数据 $\mathcal{D}=\{(x_n,y_n)\}_{n=1}^N$；特征映射 $\Phi:\mathcal{X}\to\mathcal{Z}$；线性分类算法 $\mathcal{A}$

输出: 非线性分类器 $g:\mathcal{X}\to\{-1,+1\}$

Step 1（非线性特征映射）

将原始数据映射到特征空间

$$\mathcal{D}_\Phi \leftarrow \{(z_n=\Phi(x_n),\, y_n)\}_{n=1}^N.$$

Step 2（线性训练）

在 $\mathcal{D}_\Phi$ 上用算法 $\mathcal{A}$ 训练得到线性权重 w。

Step 3（返回分类器）

输出

$$g(x)=\operatorname{sign}\big(w^\top\Phi(x)\big).$$

定义 12.2.1 (非线性模型：非线性变换 + 线性模型)

给定特征映射

$$\Phi:\mathcal{X}\to\mathcal{Z}, \quad z=\Phi(x),$$

以及任意线性模型 $\mathcal{A}$（PLA、线性回归、逻辑回归等），可构造非线性模型

$$g(x) = \mathcal{A}\big(\Phi(x)\big).$$

两自由度

- 选择非线性变换 Φ（二次、三次、多项式⋯）；
- 选择线性模型 $\mathcal{A}$（PLA、回归、分类⋯）。

用途：二次 PLA、二次回归、三次回归、. . . 、任意阶多项式回归皆可自由实现。♣

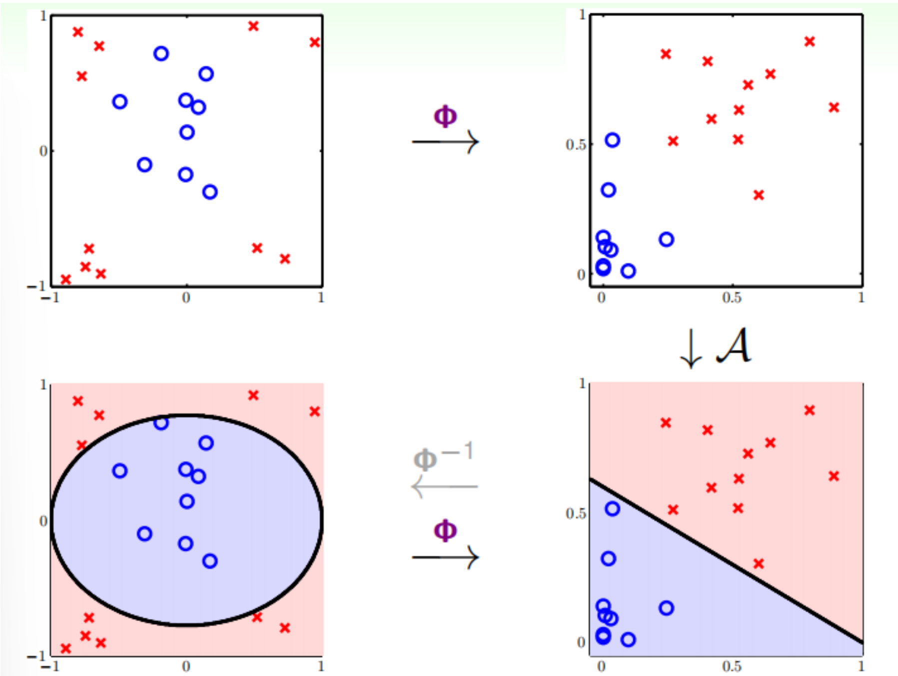

图 12.2.1: 特征变换与线性分类器结合的分类过程示意图

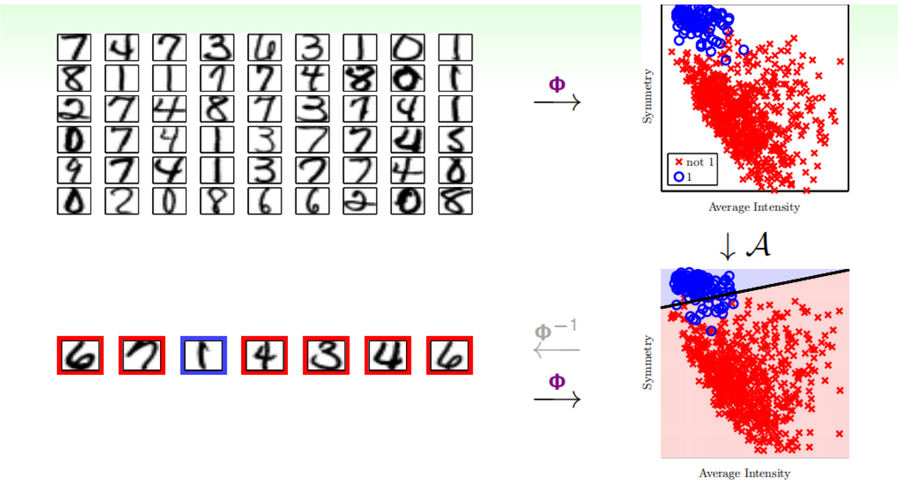

图 12.2.2: 特征变换在手写数字识别上的应用

例题 12.1 选择题：二次变换的维度计算

对于 d 维输入向量 $\boldsymbol{x} \in \mathbb{R}^d$，二次变换 $\boldsymbol{\phi}_2(\boldsymbol{x})$ 包含所有二次项 x_ix_j $(i \le j)$、线性项 x_i 及常数项 1，则其维度为：

1) d
2) $\frac{d^2}{2}+\frac{3d}{2}+1$
3) d^2+d+1
4) 2^d

解答 正确选项为 $\boxed{2}$。二次变换 $\boldsymbol{\phi}_2(\boldsymbol{x})$ 的维度由三部分构成：

- 二次项：含 x_ix_j $(i \le j)$，共 $\binom{d+1}{2}=\frac{d(d+1)}{2}$ 项；
- 线性项：含 $x_1, x_2, \ldots, x_d$，共 d 项；
- 常数项：1 项。

总维度 $=\frac{d(d+1)}{2}+d+1=\frac{d^2+d+2d+2}{2}=\frac{d^2}{2}+\frac{3d}{2}+1$。 ■

12.3 非线性变换的代价

引理 12.3.1 (Q 阶多项式映射的特征维度)

设输入维度为 d，Q 阶多项式特征映射定义为

$$\phi_Q(x)=\left(1,\, x_1,\ldots,x_d,\, x_1^2,\, x_1x_2,\ldots,x_d^2,\, \ldots,\, x_1^Q,\ldots,x_d^Q\right),$$

则该映射后的特征维度为

$$D=\binom{Q+d}{d}.$$

证明 特征维度的计算等价于统计 $\phi_Q(x)$ 中所有单项式的总数。这些单项式均具有形式 $x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}$，其中 $k_1, k_2, \ldots, k_d$ 为非负整数，且满足总次数约束：

$$0 \le k_1+k_2+\cdots+k_d \le Q.$$

为统一处理" 总次数不超过 Q" 这一条件，引入虚拟变量 $k_{d+1} \ge 0$，令其满足：

$$k_1+k_2+\cdots+k_d+k_{d+1}=Q.$$

此时，原问题中满足 $k_1+\cdots+k_d \le Q$ 的非负整数组 $(k_1,\ldots,k_d)$ 与新方程的非负整数解 $(k_1,\ldots,k_d,k_{d+1})$ 形成一一对应关系。

根据组合数学中的" 星号与竖线" 原理：对于方程 $x_1+x_2+\cdots+x_n=m$（其中 $x_i \ge 0$ 且为整数），其非负整数解的个数为 $\binom{m+n-1}{n-1}$。

在此处，变量总数为 $n=d+1$，总次数为 $m=Q$，代入上述原理可得解的个数为

$$\binom{Q+(d+1)-1}{(d+1)-1}=\binom{Q+d}{d},$$

因此特征维度 $D=\binom{Q+d}{d}$。 ■

定义 12.3.1 (Q 阶多项式映射的代价)

设输入维度为 d，定义 Q 阶多项式特征映射

$$\phi_Q(x)=\left(1,\, x_1,\ldots,x_d,\, x_1^2,\, x_1x_2,\ldots,x_d^2,\, \ldots,\, x_1^Q,\ldots,x_d^Q\right).$$

1. 计算与存储代价

映射后的特征维度

$$D = \binom{Q+d}{d} = O(Q^d).$$

需要 $O(D)$ 次运算与空间来存储 $\phi_Q(x)$ 和权重 w；当 Q 增大时，代价迅速上升。

2. 模型复杂度代价

假设集 $\mathcal{H}_{\phi_Q}$ 的 VC 维满足

$$d_{\mathrm{VC}}(\mathcal{H}_{\phi_Q}) \le D + 1 = O(Q^d).$$

直观解释：在 ϕ_Q 空间中，任意 $D+2$ 个输入无法被打散（已超出线性可分能力），回到原始 x 空间亦无法被打散。因此 Q 越大，d_{VC} 越大，模型复杂度随之升高。♣

命题 12.3.1 (泛化困境：如何选择多项式阶数 Q？)

在特征空间

$$\Phi_Q(x) \in \mathbb{R}^D \quad \text{其中} \quad D = \binom{Q+d}{d}$$

中训练得到的假设 g 满足

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \Omega\left(\frac{D \ln N}{N}\right).$$

直观权衡

- Q 过低：$E_{\text{in}}(g)$ 可能过大（欠拟合）
- Q 过高：$E_{\text{in}}(g) = 0$ 却造成 D 暴涨，$E_{\text{out}}(g)$ 远离 $E_{\text{in}}(g)$（过拟合）

如何选 Q？　凭经验“目测”往往选 $Q = 1$（原始空间）；系统做法需交叉验证或正则化，以在偏差-方差之间取得最佳折中。♠

命题 12.3.2 (“肉眼选特征”的风险)

即便在二维空间 $\mathbb{R}^2$ 中能凭直观“看出”合适的特征，当输入空间扩展到 $X = \mathbb{R}^{10}$ 时，这种“目测法”显然不再可行。

同一问题的四种特征映射示例

$$\begin{aligned}
&\text{完整二阶映射 } \Phi_2: && z = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2), d_{\mathrm{VC}} = 6;\\
&\text{视觉直观下的“简化”}: && z = (1, x_1^2, x_2^2), && d_{\mathrm{VC}} = 3;\\
&\text{进一步“简化”}: && z = (1, x_1^2 + x_2^2), && d_{\mathrm{VC}} = 2;\\
&\text{极致简化}: && z = \mathrm{sign}(0.6 - x_1^2 - x_2^2), && d_{\mathrm{VC}} = 1.
\end{aligned}$$

核心警示　人脑自带的“模型复杂度直觉”极易导致过拟合。为保证 VC 维估计具备数据无关性（data-independent），特征映射 Φ 必须在未观测数据的前提下预先确定。♠

笔记 [非线性变换的代价]

非线性映射（如多项式映射）虽可提升表达能力，但会带来维度膨胀、模型复杂度升高与泛化风险；另外，特征映射的选择应避免主观臆断，需结合交叉验证等方法进行系统优化。

12.4 结构化假设集

定义 12.4.1 (多项式变换的层级结构)

对阶数 $Q = 0, 1, 2, \dots$ 定义嵌套特征映射

$$\begin{aligned}
\Phi_0(x) &= (1), \\
\Phi_1(x) &= \big(\Phi_0(x),\ x_1, x_2, \dots, x_d\big), \\
\Phi_2(x) &= \big(\Phi_1(x),\ x_1^2, x_1x_2, \dots, x_d^2\big), \\
&\ \ \vdots \\
\Phi_Q(x) &= \big(\Phi_{Q-1}(x),\ \text{所有新增 } Q \text{ 阶单项式}\big).
\end{aligned}$$

对应假设集合呈严格嵌套：

$$\mathcal{H}_0 \subset \mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_Q.$$

层级属性

- **VC** 维单调增：$d_{\mathrm{VC}}(\mathcal{H}_0) \le d_{\mathrm{VC}}(\mathcal{H}_1) \le \cdots \le d_{\mathrm{VC}}(\mathcal{H}_Q)$.
- 训练误差单调减：$E_{\mathrm{in}}(g_0) \ge E_{\mathrm{in}}(g_1) \ge \cdots \ge E_{\mathrm{in}}(g_Q)$, 其中 $g_j = \arg\min_{h\in\mathcal{H}_j} E_{\mathrm{in}}(h)$.
- 权衡启示：选用过高阶的 $\mathcal{H}_Q$ 虽能降低 E_{in}，却因 d_{VC} 过大而升高 E_{out}（过拟合）。 ♣

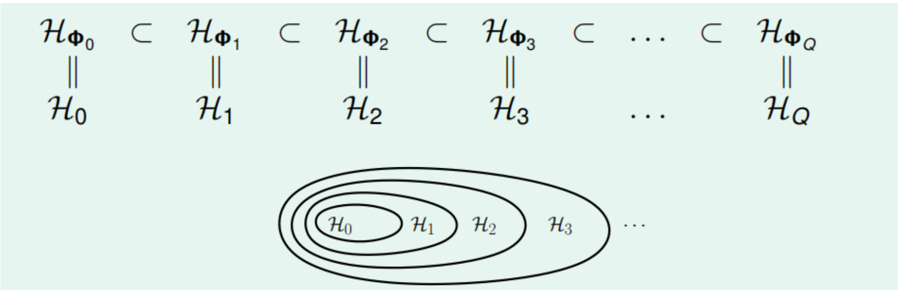

图 12.4.1: 递增假设空间序列示意图

命题 12.4.1 (线性模型优先原则)

将假设集合按嵌套关系排列为

$$\mathcal{H}_0 \subset \mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_Q,$$

其 VC 维与训练误差满足

$$d_{\mathrm{VC}}(\mathcal{H}_j)\uparrow, \quad E_{\mathrm{in}}(g_j)\downarrow, \quad \text{其中 } g_j = \arg\min_{h\in\mathcal{H}_j} E_{\mathrm{in}}(h).$$

诱人的陷阱　选用 $\mathcal{H}_{1126}$ 这类高复杂度模型，虽能将训练误差 $E_{\mathrm{in}}(g_{1126})$ 降至极低以”蒙混过关”，但会导致 VC 维急剧攀升、泛化误差 E_{out} 彻底失控，最终陷入难以挽回的境地。

稳妥策略

1) 从最简单的 $\mathcal{H}_1$（线性模型）起步；
2) 若此时 $E_{\mathrm{in}}(g_1)$ 已满足需求，即可终止并高效收尾；
3) 否则再逐步提升模型复杂度（沿假设集序列右移）；这种做法即便多耗费些许计算资源，也不会造成实质性损失。

结论 线性模型应优先选用的原则：兼具简单、高效、安全的特性，且实践中往往行之有效！

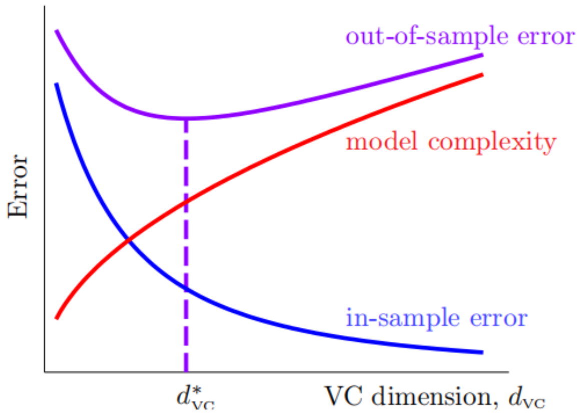

图 12.4.2: 模型复杂度、样本内误差、样本外误差与 VC 维关系图

12.5 总结

笔记 [非线性变换]

- 二次假设：在二次变换后的数据上使用线性假设。
- 非线性变换：先做 $\mathbf{z} = \Phi(\mathbf{x})$，再愉快地线性建模。
- 非线性变换的代价：计算量、存储量与模型复杂度同步上升。
- 结构化假设集：优先使用线性或更简单的模型。

笔记 [总体结论] 非线性变换通过将输入映射到高维特征空间，使得原本非线性可分的问题转化为线性可分，从而显著提升模型的表达能力。这一策略保留了线性模型在计算与训练上的优势，但也带来了维度膨胀、计算代价上升和模型复杂度增加等问题，可能导致泛化能力下降。为此，我们需在提升拟合能力与控制复杂度之间进行权衡，避免依赖主观直觉设计特征，应借助交叉验证、正则化等方法系统选择变换阶数。整体而言，非线性变换是连接简单模型与复杂任务之间的重要桥梁，是实现有效学习的关键手段之一。