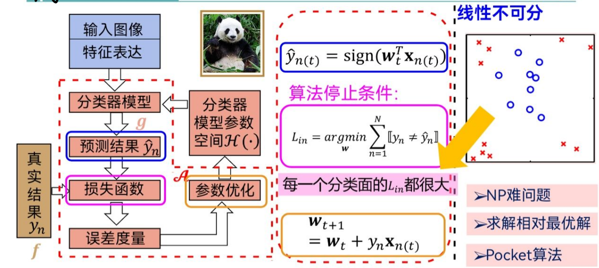
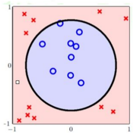
26.1 线性不可分问题



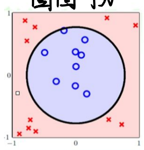


圆圈可分

$h_{sep}(\mathbf{x}) = sign(-x_1^2 - x_2^2 + 0.6)$

$$h_{sep}(\mathbf{x}) = sign(0.6\cdot 1 + (-1)\cdot x_1^2 + (-1)\cdot x_2^2)$$
$$= sign(\tilde w_0 z_0 + \tilde w_1 z_1 + \tilde w_2 z_2) = sign(\tilde{\mathbf{w}}^T \mathbf{z})$$

圆圈可分

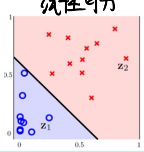


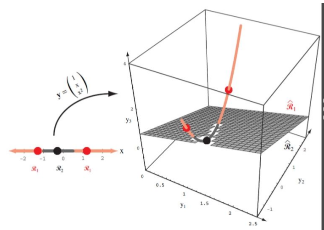
$\{(\mathbf{x}_n, y_n)\}$圆圈可分 ➡ $\{(\mathbf{z}_n, y_n)\}$线性可分

$\mathbf{x}\in\mathcal{X} \xrightarrow{\Phi} \mathbf{z}\in\mathcal{Z}$

Φ：非线性特征变换

线性可分





利用二次多项式的一般表达将样本 $\mathbf{x}$ 从 $\mathcal{X}$ 空间变换到 $\mathcal{Z}$ 空间

$$\Phi_2(\mathbf{x}) = (1, x_1, \ldots, x_d, x_1^2, x_1x_2, \ldots, x_d^2)^T$$

如果样本 $\mathbf{x}$ 是 2 维特征，则： $\Phi_2(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^T$

样本 $\mathbf{x}$ 从原来的 d 维特征空间变换到多少维特征空间?

$$\tilde d = 1 + d + d + C_d^2 = 1 + d + d + \binom{d}{2}$$

放回的组合问题

$$= 1 + d + d + \frac{d(d-1)}{2} = 1 + \frac{d(d+3)}{2} = \frac{(d+2)(d+1)}{2} = \binom{2+d}{2}$$

利用Q次多项式的一般表达将样本 $\mathbf{x}$ 从 $\mathcal{X}$ 空间变换到 $\mathcal{Z}$ 空间

$$\Phi_Q(\mathbf{x}) = (\underbrace{1, x_1, \ldots, x_d}_{一次项}, \underbrace{x_1^2, x_1x_2, \ldots, x_d^2}_{二次项}, \ldots, \underbrace{x_1^Q, x_1^{Q-1}x_2, \ldots, x_d^Q}_{Q次项})^T$$

样本 $\mathbf{x}$ 从原来的 d 维特征空间变换到多少维特征空间?

$$\tilde d = C_{Q+d}^Q = \binom{Q+d}{Q}$$

放回的组合问题

$$= \frac{(Q+d)!}{Q!\cdot d!} = \frac{(Q+d-1)(Q+d-2)\cdots(Q+1)}{d!} \Rightarrow Q^d$$

非线性变换使特征被升到高维空间

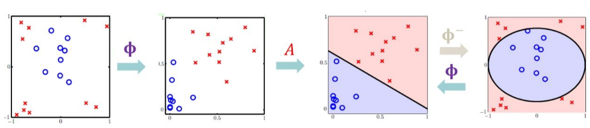
$$C_n^m = \frac{A_n^m}{A_m^m} = \frac{n!}{m!(n-m)!}$$

$$C_{Q+d}^Q = \frac{A_{Q+d}^Q}{A_Q^Q} = \frac{(Q+d)!}{Q!d!}$$

26.2 非线性变换

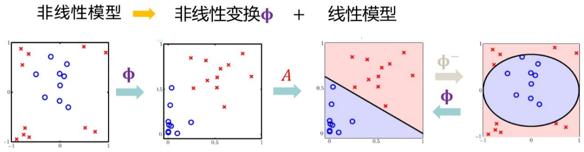
目的：通过非线性变换Φ_Q使得训练样本集$\{(z_n = \Phi_Q(x_n), y_n)\}$在Z空间找到好的分类面

非线性变换步骤



① 利用非线性变换 Φ 将原始训练样本集$\{(\mathbf{x}_n, y_n)\}$变换到$\mathcal{Z}$ 空间$\{(\mathbf{z}_n = \Phi(\mathbf{x}_n), y_n)\}$;
② 在数据集$\{(\mathbf{z}_n, y_n)\}$上选择合适的线性分类算法$\mathcal{A}$，得到最佳解 $\tilde{\mathbf{w}}^*$
③ 返回分类结果： $g(\mathbf{x}) = sign(\tilde{\mathbf{w}}^{*T}\mathbf{x})$

非线性模型 ➡ 非线性变换Φ + 线性模型



线性模型不局限于二元分类;
通过非线性变换，可以方便地实现： 二次PLA、三次PLA、更高次数多项式的PLA
二次回归、三次回归、更高次数回归。。。

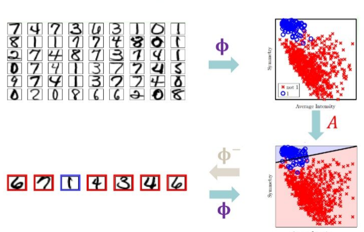
特征提取 ➡ 特征变换 Φ

非线性变换并不一定是多项式变换

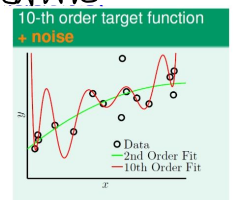
图像原始像素值
raw (pixels)

领域知识
domain
knowledge

具体特征
rete (intensity, symmetry)



26.3 泛化能力讨论



	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
L_{in}	0.050	0.034
L_{out}	0.127	9.00

	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
L_{in}	0.029	0.00001
L_{out}	0.120	7680

Vapnik-Chervonenkis (VC) Bound

For any $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$ and 'statistical' large $\mathcal{D}$, for ~~$N \geq 2$,~~ $d_{vc} \geq 2$

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

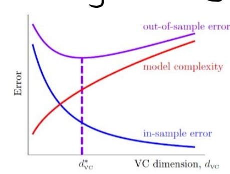
VC维度（或Vapnik-Chervonenkis维度）是衡量可以通过统计分类算法学习的函数空间的容量（复杂度，表现力，丰富度或灵活性）的度量。**它被定义为算法可以破碎（shatter）的最大点集的基数**

..., with probability $\geq 1 - \delta$, **GOOD**: $|E_{in}(g) - E_{out}(g)| \leq \epsilon$

$$\text{set} \quad \delta = 4(2N)^{d_{vc}} \exp\left(-\tfrac{1}{8}\epsilon^2 N\right)$$
$$\frac{\delta}{4(2N)^{d_{vc}}} = \exp\left(-\tfrac{1}{8}\epsilon^2 N\right)$$
$$\ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right) = \tfrac{1}{8}\epsilon^2 N$$
$$\sqrt{\tfrac{8}{N}\ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)} = \epsilon$$

$$\text{gen. error } |E_{in}(g) - E_{out}(g)| \leq \sqrt{\tfrac{8}{N}\ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)}$$
$$E_{in}(g) - \sqrt{\tfrac{8}{N}\ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)} \leq E_{out}(g) \leq E_{in}(g) + \sqrt{\tfrac{8}{N}\ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)}$$

$$\therefore E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N}\ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)}$$ 模型复杂度惩罚项$\Omega(N, \mathcal{H}, \delta)$



$d_{vc}\uparrow$: $E_{in}\downarrow$ $\Omega\uparrow$
$d_{vc}\downarrow$: $\Omega\downarrow$ $E_{in}\uparrow$
⇒ d^*_{vc}在中间取到

$\mathcal{H}$高不一定好

过拟合/泛化能力差: low E_{in} , high E_{out}

过拟合与正则化(*Overfitting and Regularization*)

$$\min_{\mathbf{w}} E_{aug}(\mathbf{w}) = E_{in}(\mathbf{w}) + \underbrace{\frac{\lambda}{N}\mathbf{w}^T\mathbf{w}}_{正则项}$$

$\|w\|^2$防止w某维度数据过大