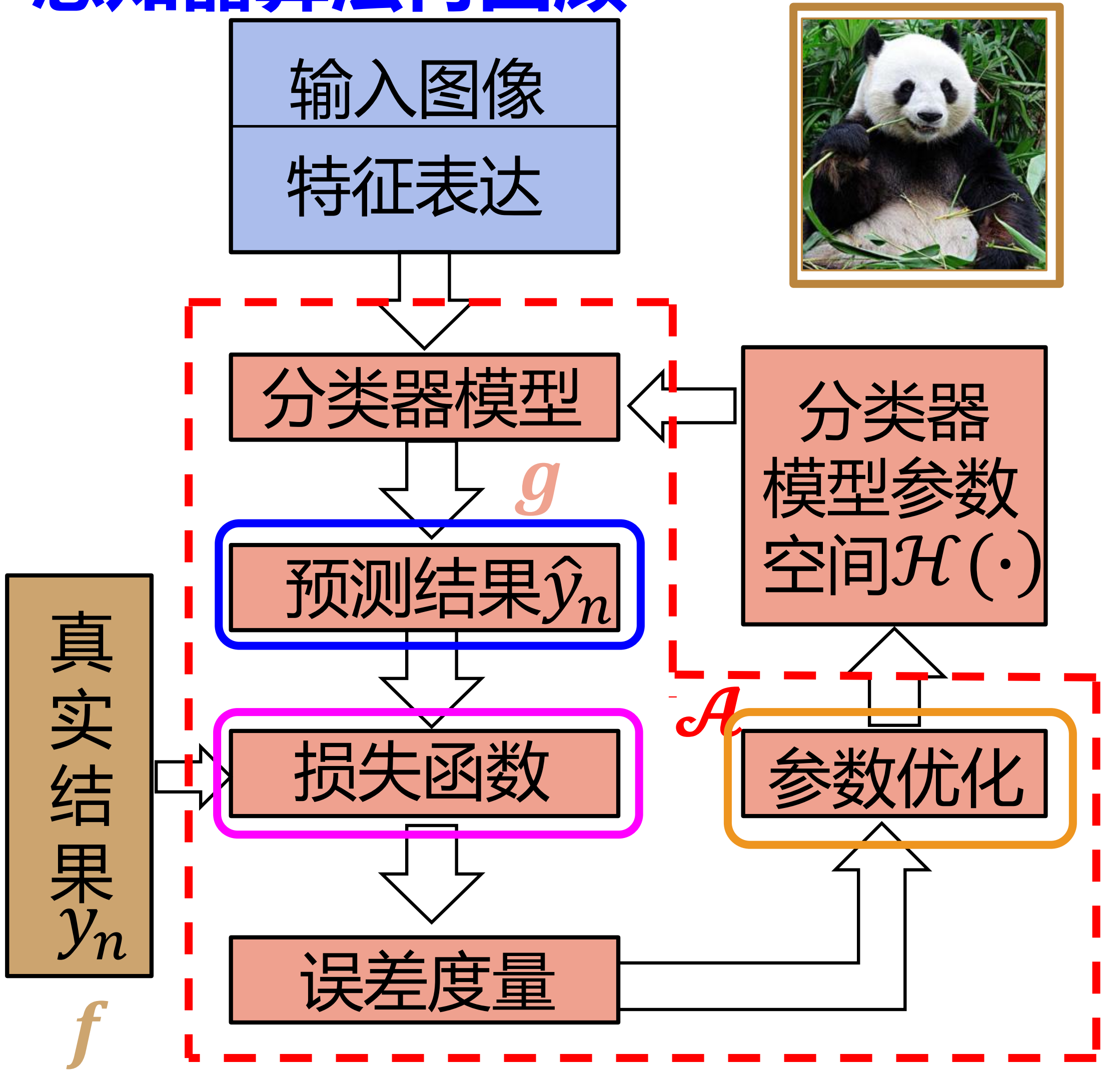
第六讲　非线性变换（*Nonlinear Transformation*）

6.1 线性不可分问题

感知器算法再回顾



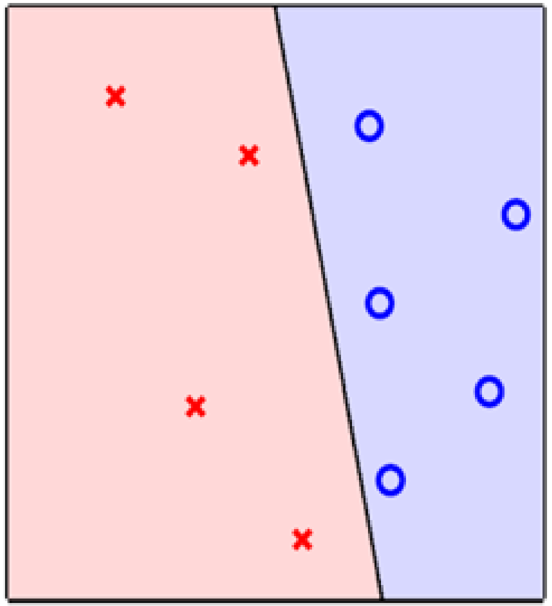
$$\hat{y}_{n(t)} = \text{sign}(\mathbf{w}_t^T \mathbf{x}_{n(t)})$$

算法收敛：

$$L_{in} = \sum_{n=1}^{N} [\![y_n \neq \hat{y}_n]\!] = 0$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + y_n \mathbf{x}_{n(t)}$$

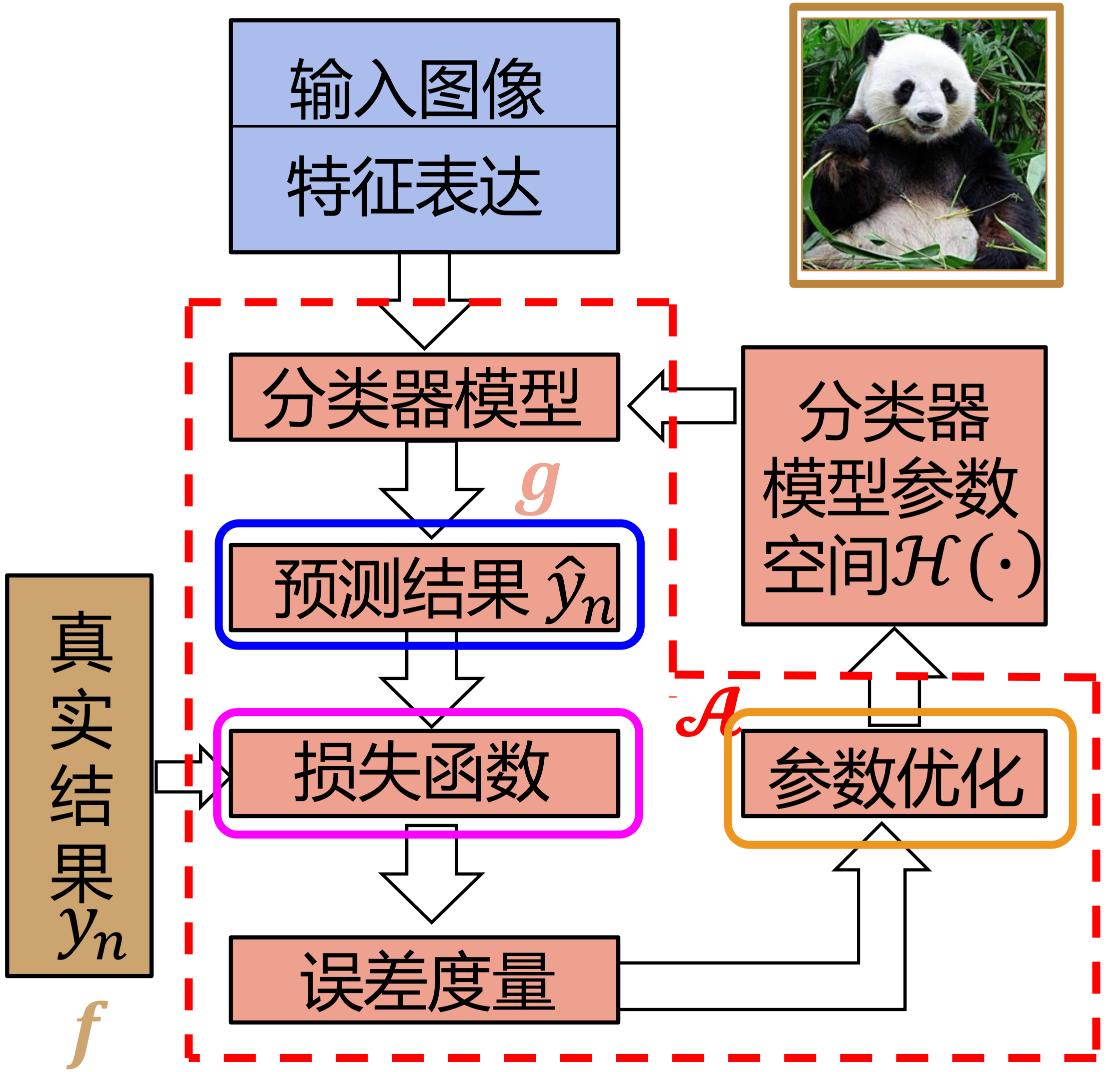
线性可分



- 设置初始分类面(权重)$\mathbf{w}_0$
- 如果有样本分错，就修正权重

Ref.: NTU-LIN

6. 1 线性不可分问题



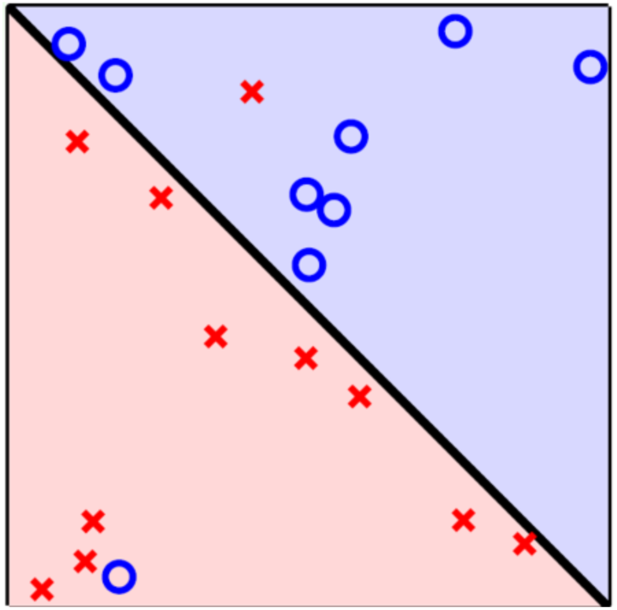
$$\hat{y}_{n(t)} = \mathrm{sign}(\boldsymbol{w}_t^T \mathbf{x}_{n(t)})$$

算法停止条件：

$$L_{in} = \underset{\boldsymbol{w}}{argmin} \sum_{n=1}^{N} [\![y_n \neq \hat{y}_n]\!]$$

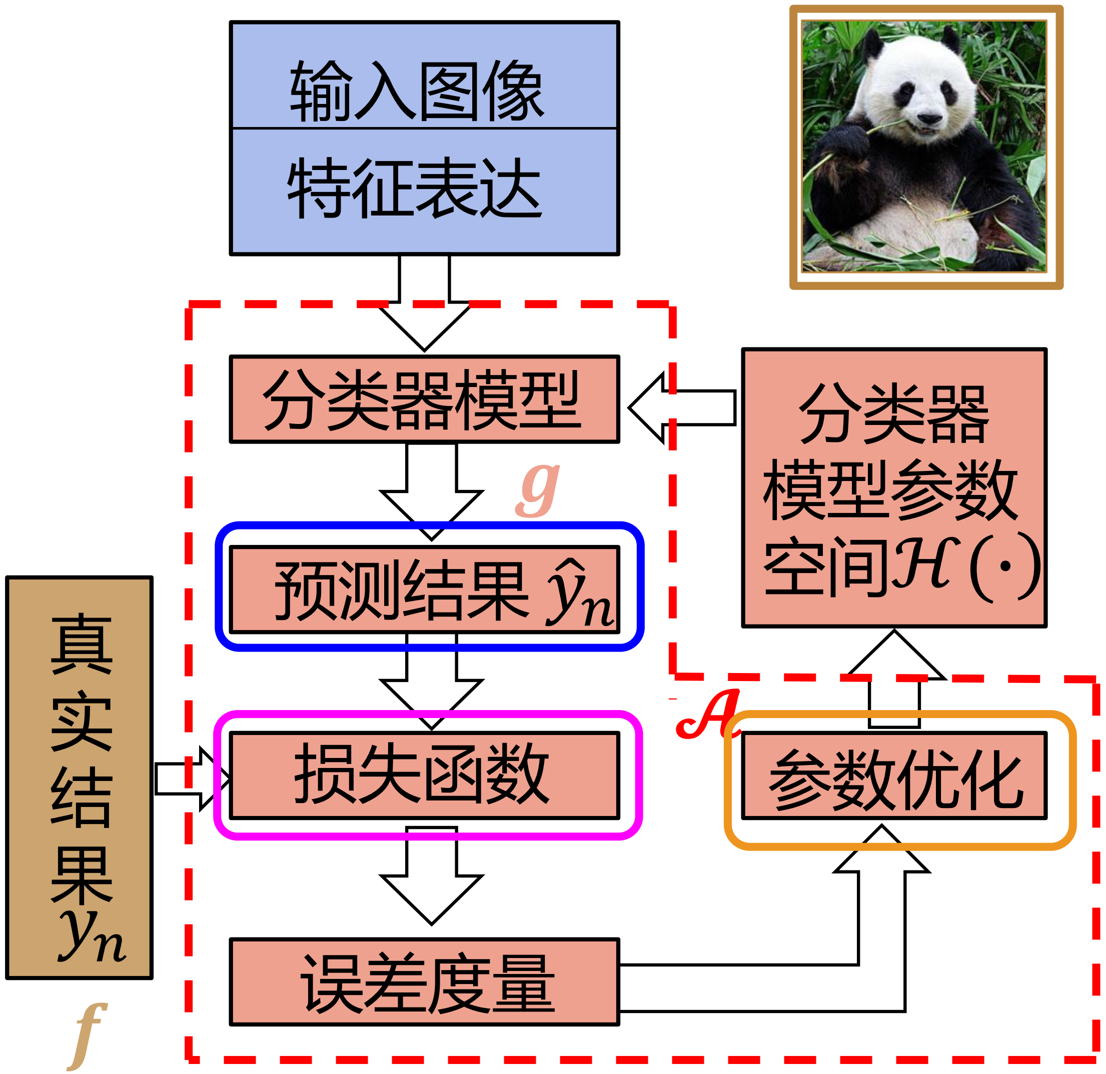
$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + y_n \mathbf{x}_{n(t)}$$

线性不可分



- NP难问题
- 求解相对最优解
- Pocket算法

Ref.: NTU-LIN



$$\hat{y}_{n(t)} = \mathrm{sign}(\mathbf{w}_t^T \mathbf{x}_{n(t)})$$

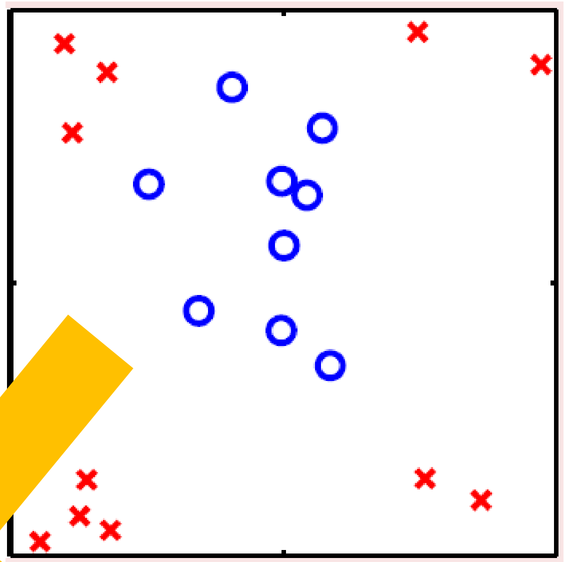
算法停止条件：

$$L_{in} = \underset{\mathbf{w}}{argmin} \sum_{n=1}^{N} [\![y_n \neq \hat{y}_n]\!]$$

每一个分类面的L_{in}都很大

$$\mathbf{w}_{t+1} = \mathbf{w}_t + y_n \mathbf{x}_{n(t)}$$

线性不可分

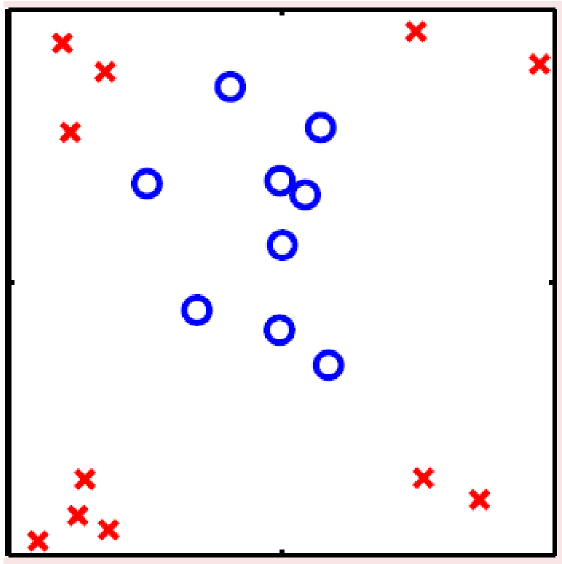


- NP难问题
- 求解相对最优解
- Pocket算法

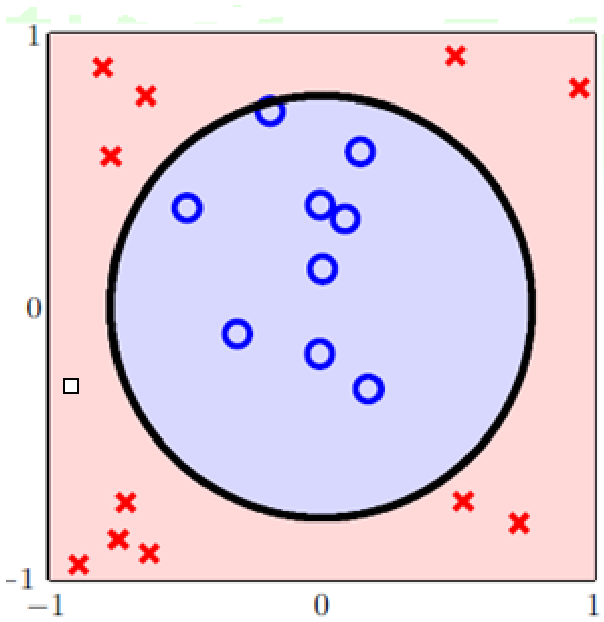
Ref.: NTU-LIN

6.1 线性不可分问题

如何突破线性分类限制



线性不可分



圆圈可分

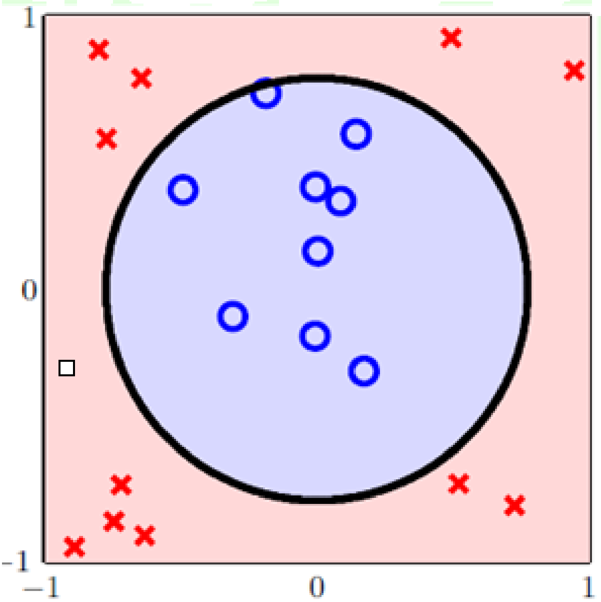
$$h_{sep}(\mathbf{x}) = sign(-x_1^2 - x_2^2 + 0.6)$$

Ref.: NTU-LIN

圆圈可分与线性可分

$$h_{sep}(\mathbf{x}) = sign(0.6 \cdot 1 + (-1) \cdot x_1^2 + (-1) \cdot x_2^2)$$

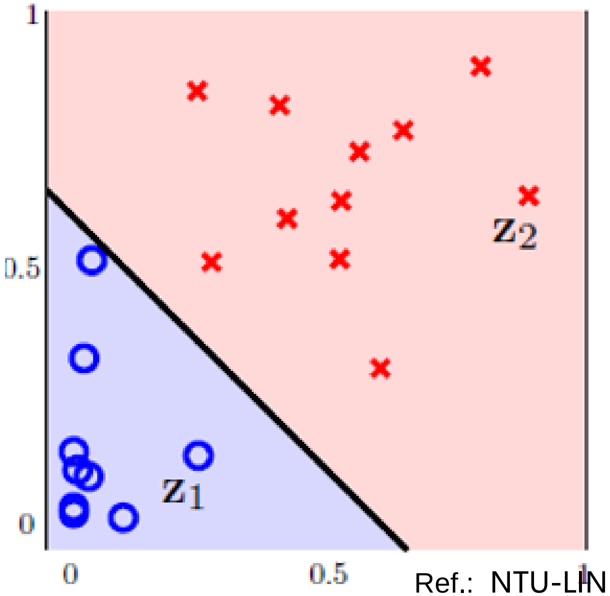
$$= sign(\tilde{w}_0 z_0 + \tilde{w}_1 z_1 + \tilde{w}_2 z_2) = sign(\tilde{\mathbf{w}}^T \mathbf{z})$$



$\{(\mathbf{x}_n, y_n)\}$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



Ref.: NTU-LIN

6.1 线性不可分问题

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

$$\boldsymbol{\phi}_2(\mathbf{x}) = (1, x_1, \dots, x_d, x_1^2, x_1x_2, \dots, x_d^2)^T$$

如果样本 $\mathbf{x}$ 是 2 维特征，则： $\boldsymbol{\phi}_2(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^T$

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

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

不放回的组合问题

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

6.1 线性不可分问题

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

$$\boldsymbol{\phi}_2(\mathbf{x}) = (1, x_1, \dots, x_d, x_1^2, x_1x_2, \dots, x_d^2)^T$$

如果样本 $\mathbf{x}$ 是 2 维特征，则： $\boldsymbol{\phi}_2(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^T$

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

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

6.1 线性不可分问题

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

$$\boldsymbol{\phi}_Q(\mathbf{x}) = (1, \underbrace{x_1, \dots, x_d}_{\text{一次项}}, \underbrace{x_1^2, x_1x_2, \dots, x_d^2}_{\text{二次项}}, \quad \dots \quad , \underbrace{x_1^Q, x_1^{Q-1}x_2, \dots, x_d^Q}_{\text{Q次项}})^T$$

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

非线性变换使特征被升到高维空间

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

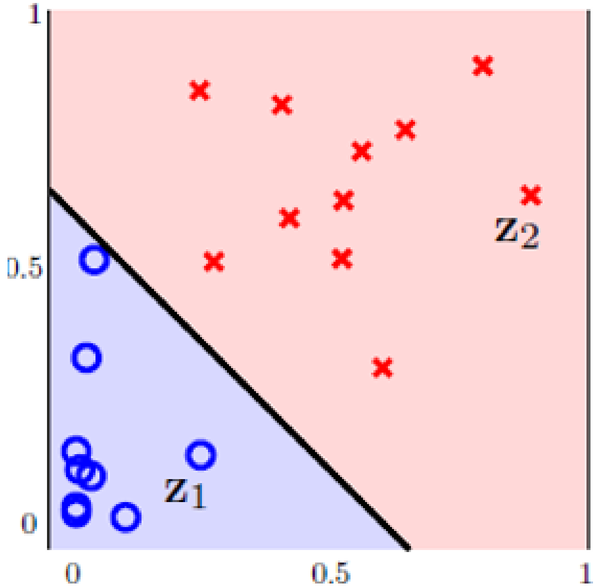
6.1 线性不可分问题（*Nonlinear Data Problem*）

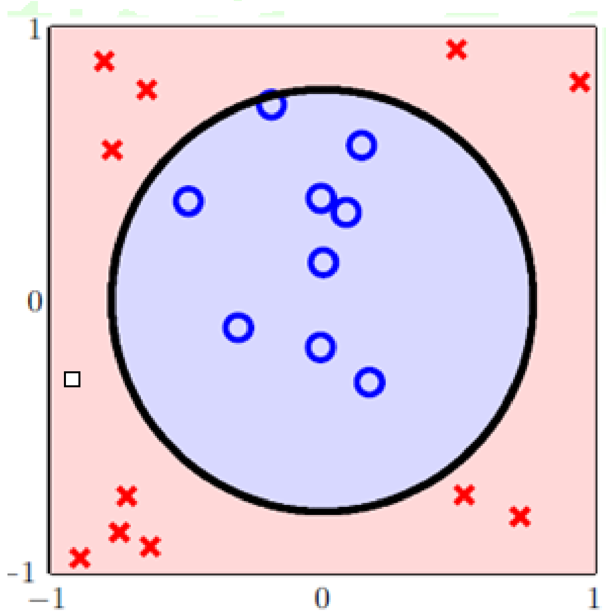
6.2 非线性变换（*Nonlinear Transform*）

6.3 知识拓展（*Knowledge Extension* ）

6.2 非线性变换

非线性变换的目的



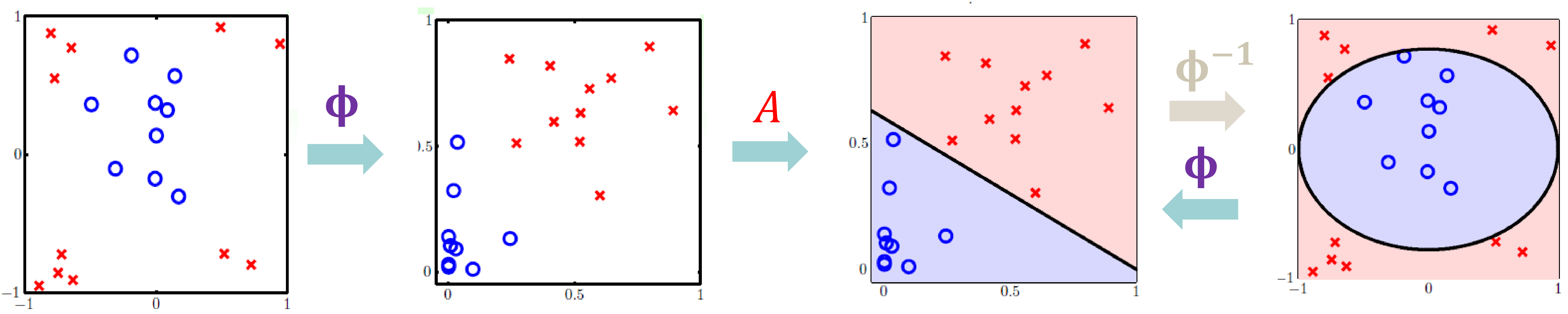


通过非线性变换 Φ_Q 使得训练样本集$\{(z_n = \Phi_Q(\mathbf{x}_n), y_n)\}$在$\mathcal{Z}$ 空间找到好的分类面

Ref.: NTU-LIN

6.2 非线性变换

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$\{(\mathbf{z}_n = \Phi(\mathbf{x}_n), y_n)\}$;

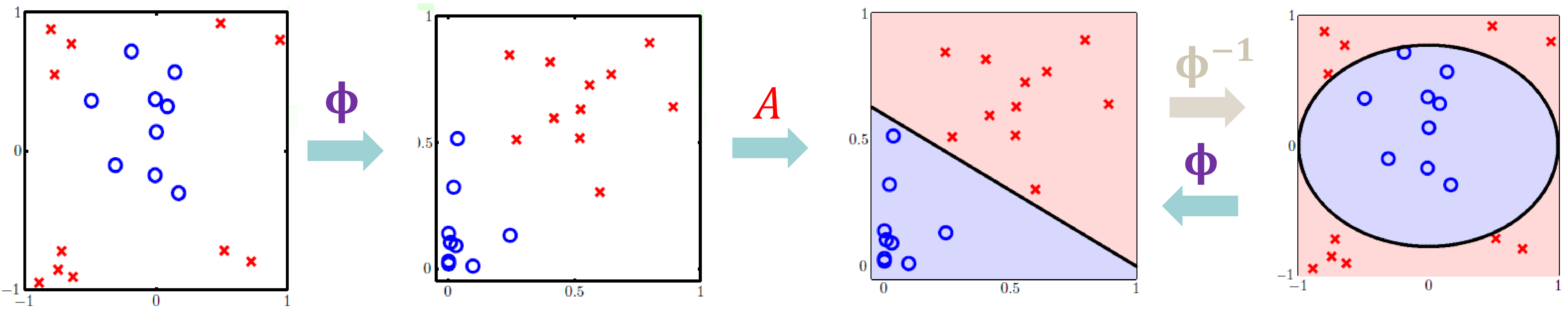
② 在数据集$\{(\mathbf{z}_n, y_n)\}$上选择合适的线性分类算法$\mathcal{A}$，得到最佳解 $\tilde{\mathbf{w}}^*$

③ 返回分类结果： $g(\mathbf{x}) = sign(\tilde{\mathbf{w}}^{*T}\mathbf{x})$

Ref.: NTU-LIN

6.2 非线性变换

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。。。

Ref.: NTU-LIN

6. 2 非线性变换

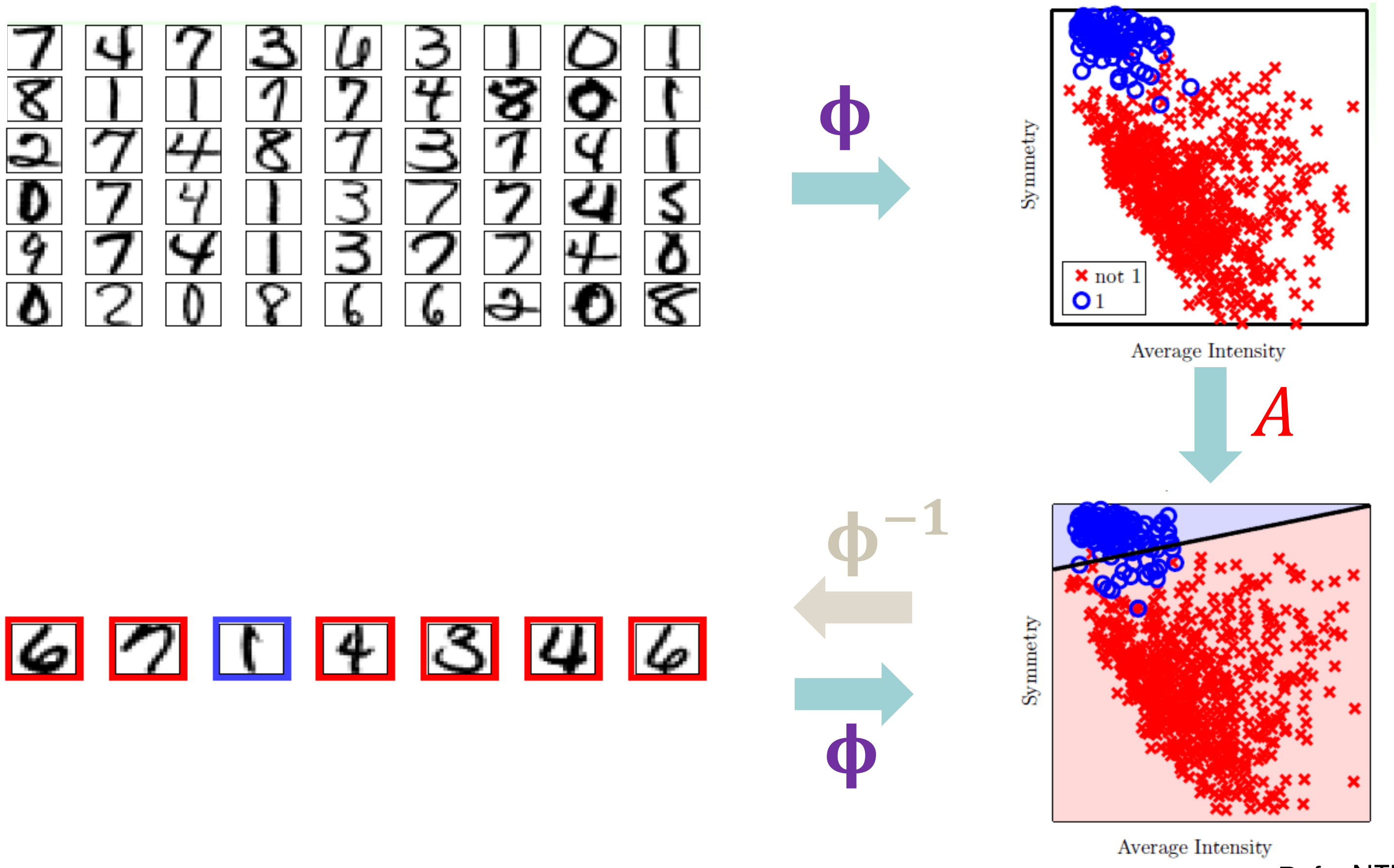
特征提取 → 特征变换 ϕ

非线性变换并不一定是多项式变换

图像原始像素值
raw (pixels)

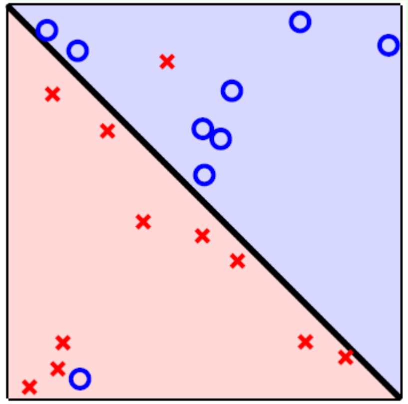
领域知识
domain knowledge

具体特征
concrete (intensity, symmetry)



Ref.: NTU-LIN

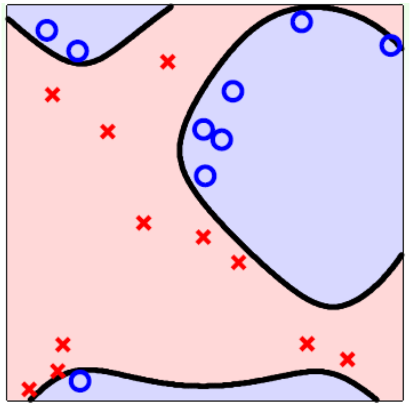
模型泛化能力讨论(Generalization Issue)



$\boldsymbol{\Phi}_1$ (original $\mathbf{x}$)

$L_{in} \neq 0$

你认为哪个分类面更好?



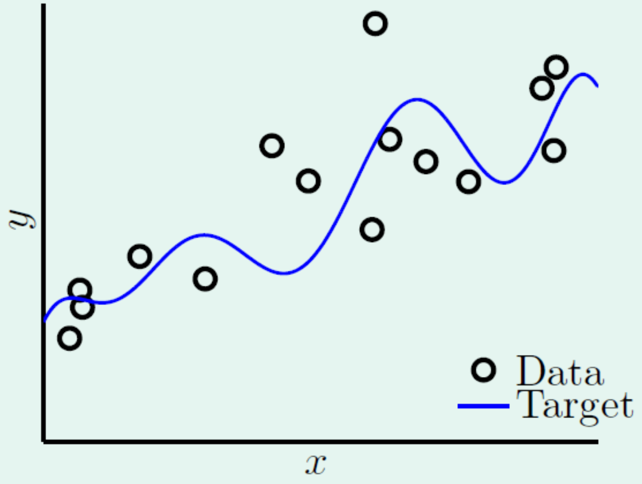
$\boldsymbol{\Phi}_4$

$L_{in} = 0$

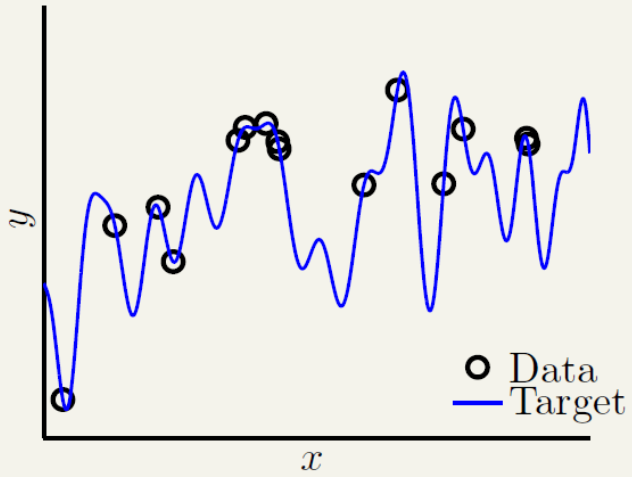
Ref.: NTU-LIN

模型泛化能力讨论(Generalization Issue)

10-th order target function
+ noise



50-th order target function
noiselessly



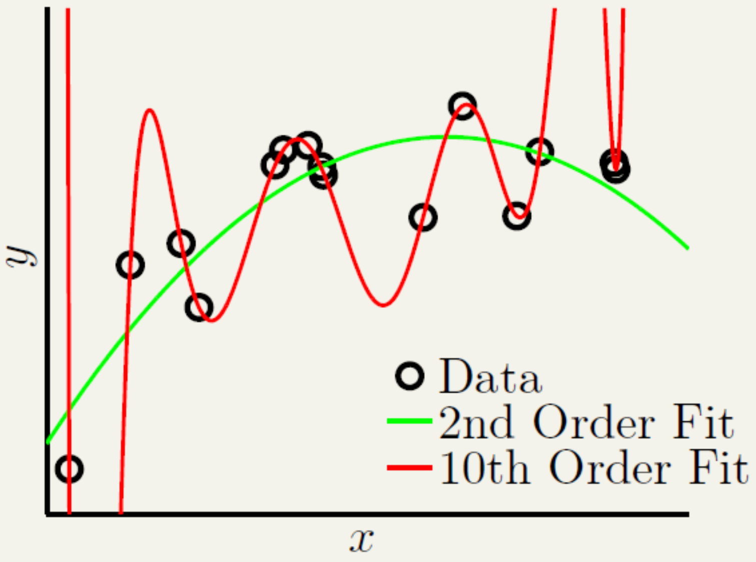
Ref.: NTU-LIN

模型泛化能力讨论(Generalization Issue)

10-th order target function
+ noise

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

50-th order target function
noiselessly



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

Ref.: NTU-LIN

模型泛化能力讨论(Generalization Issue)

Vapnik-Chervonenkis (VC) Bound

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

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

Ref.: NTU-LIN

Vapnik-Chervonenkis (VC) Bound

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

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

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

$$\begin{aligned}
\text{set} \quad \delta &= 4(2N)^{d_{\mathrm{VC}}} \exp\left(-\tfrac{1}{8}\epsilon^2 N\right) \\
\frac{\delta}{4(2N)^{d_{\mathrm{VC}}}} &= \exp\left(-\tfrac{1}{8}\epsilon^2 N\right) \\
\ln\left(\frac{4(2N)^{d_{\mathrm{VC}}}}{\delta}\right) &= \tfrac{1}{8}\epsilon^2 N \\
\sqrt{\frac{8}{N}\ln\left(\frac{4(2N)^{d_{\mathrm{VC}}}}{\delta}\right)} &= \epsilon
\end{aligned}$$

Ref.: NTU-LIN

Vapnik-Chervonenkis (VC) Bound

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

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

..., with probability $\geq 1 - \delta$, **GOOD**!

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

$$E_{\text{in}}(g) - \sqrt{\frac{8}{N}\ln\left(\frac{4(2N)^{d_{\text{VC}}}}{\delta}\right)} \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N}\ln\left(\frac{4(2N)^{d_{\text{VC}}}}{\delta}\right)}$$

Ref.: NTU-LIN

模型复杂度

with a high probability,

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

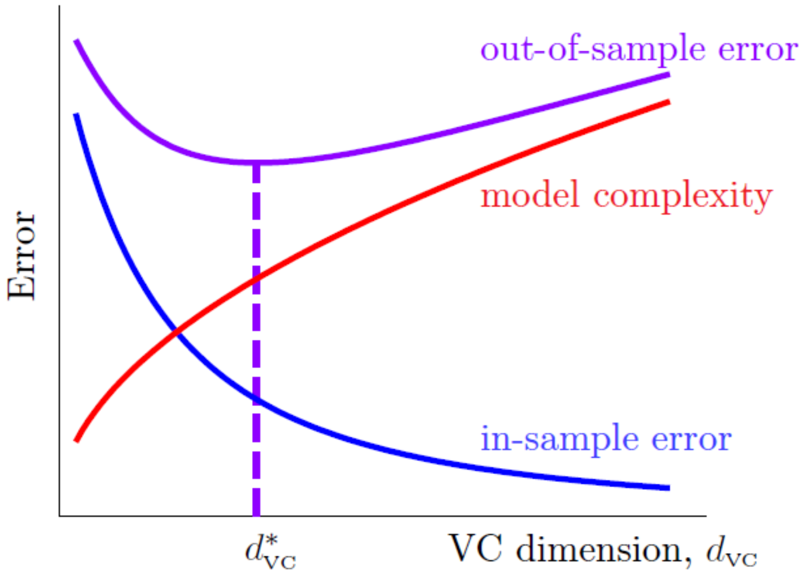
$\underbrace{\sqrt{\cdots}}_{\Omega(N,\mathcal{H},\delta)}$: penalty for **model complexity**

Ref.: NTU-LIN

模型复杂度

with a high probability,

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

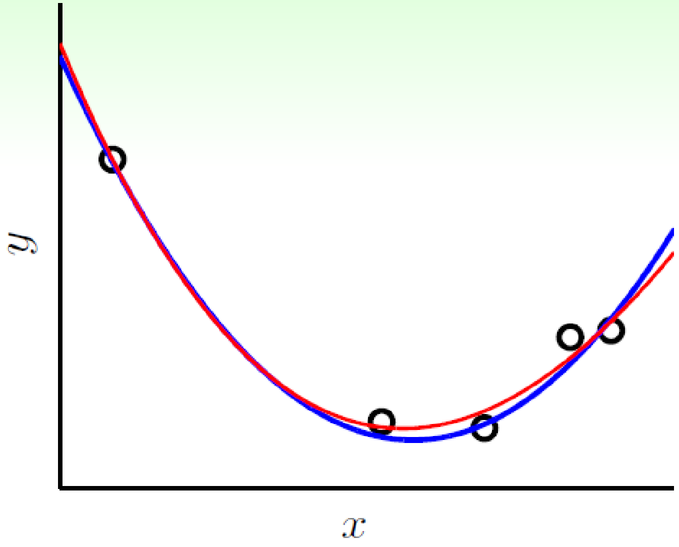


- $d_{VC}\uparrow$: $E_{in}\downarrow$ but $\Omega\uparrow$
- $d_{VC}\downarrow$: $\Omega\downarrow$ but $E_{in}\uparrow$
- best d^*_{VC} in the middle

powerful $\mathcal{H}$ not always good!

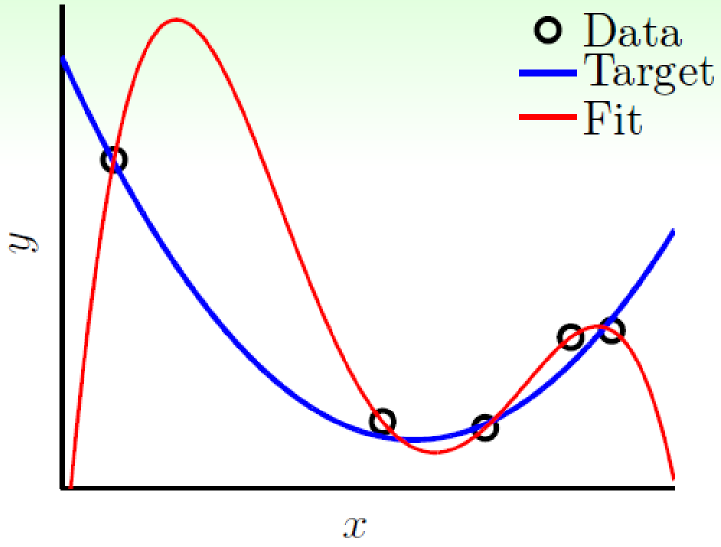
Ref.: NTU-LIN

泛化性能不好与过拟合(*Bad Generalization and Overfitting*)



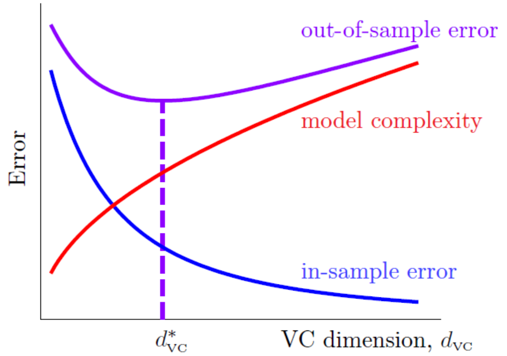
‘good fit’

$\Longrightarrow$



overfit

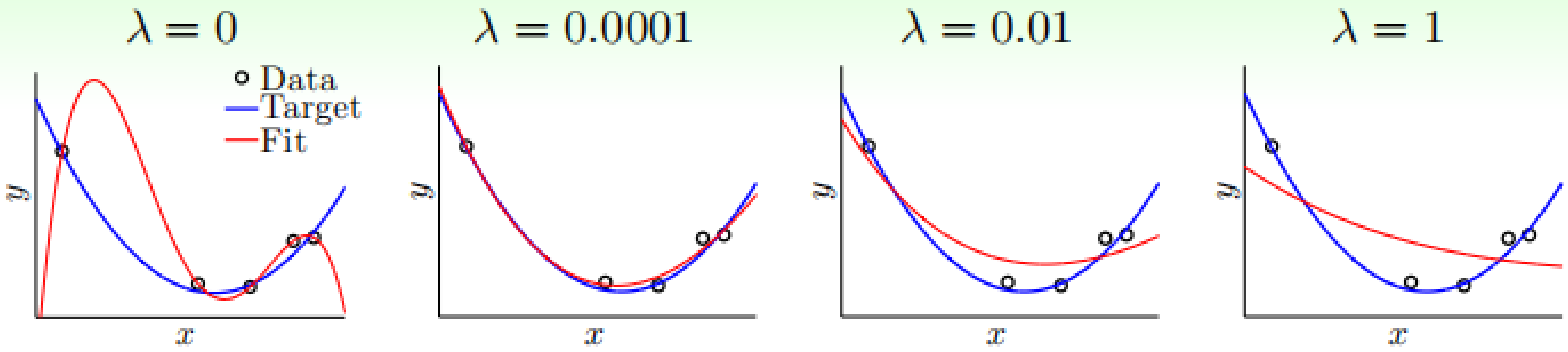
bad generalization: low E_{in}, high E_{out};
overfitting: low**er** E_{in}, high**er** E_{out}



Ref.: NTU-LIN

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

$$\min_{\mathbf{w}} E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$$



overfitting $\Longrightarrow$ $\Longrightarrow$ $\Longrightarrow$ underfitting

Ref.: NTU-LIN

第六讲　非线性变换（*Nonlinear Transformation*）

6.1线性不可分问题（Nonlinear Data Problem）

通过多项式变换后的数据集符合线性模型特点

6.2 非线性变换

利用$Z=\Phi(X)$变换后，可以方便地使用线性模型处理

6.3 知识拓展

***VC Bound**、模型复杂度、泛化能力、过拟合、正则化*