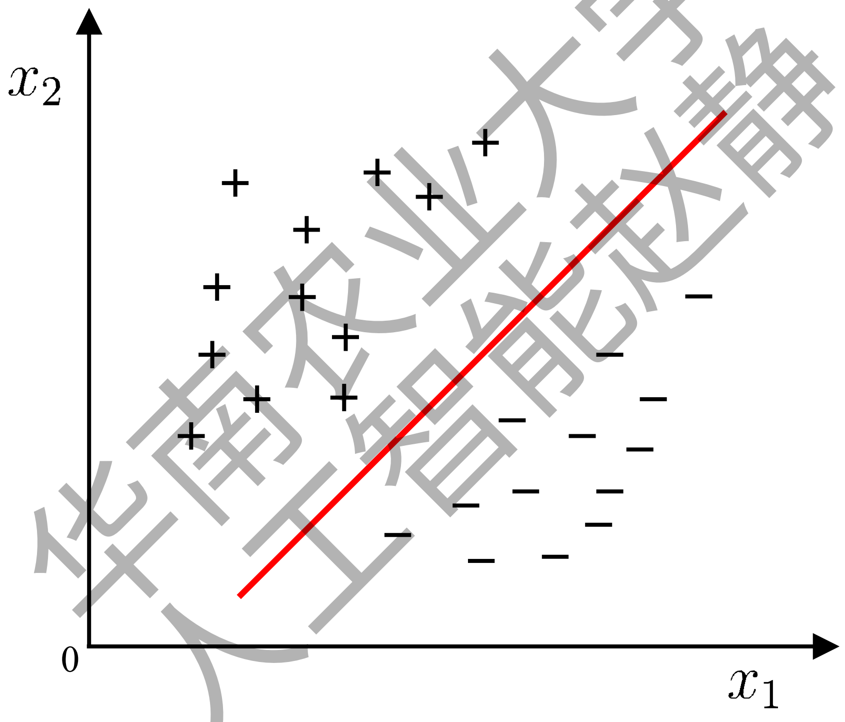
明理 精工 笃学 致远

第5章 支持向量机

Support Vector Machine

电子工程学院、人工智能学院
college of Electronic Engineering , college of Artificial Intelligence

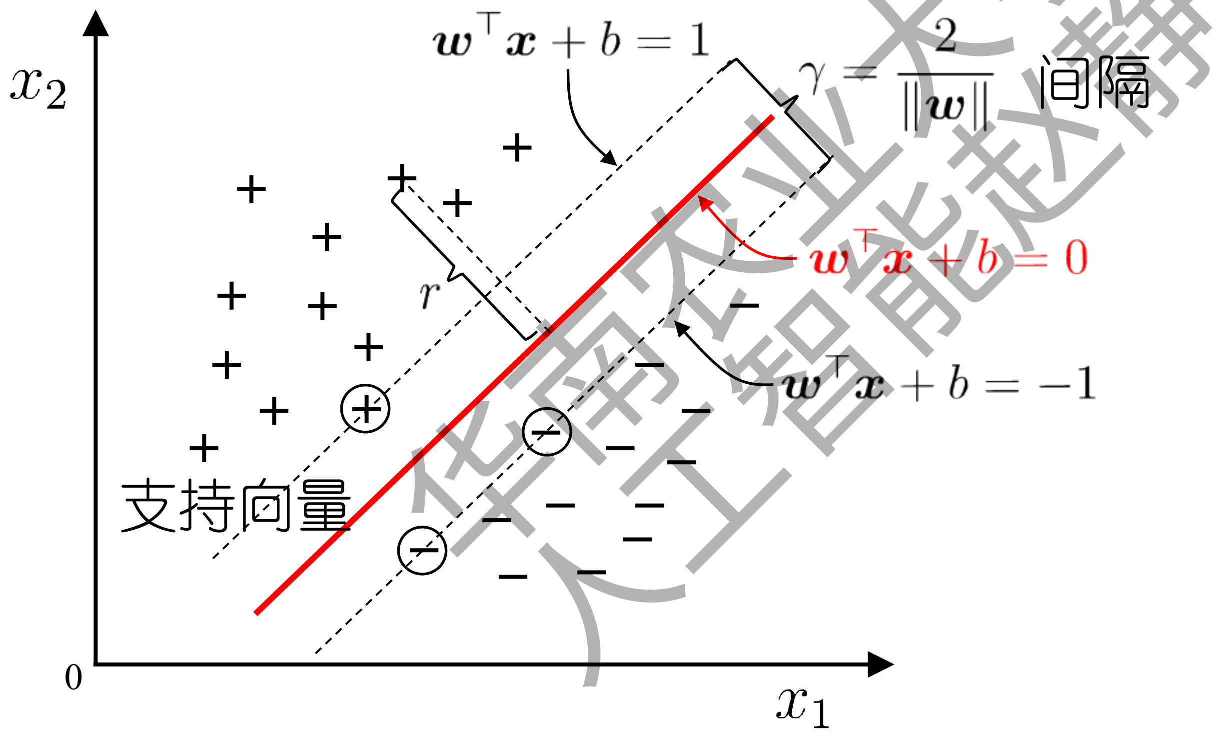
线性模型：在样本空间中寻找一个超平面, 将不同类别的样本分开.



- 线性SVM
- 带松弛变量的SVM
- 合页损失函数
- SVM的对偶问题
- 核化SVM模型
- SVM回归（SVR）

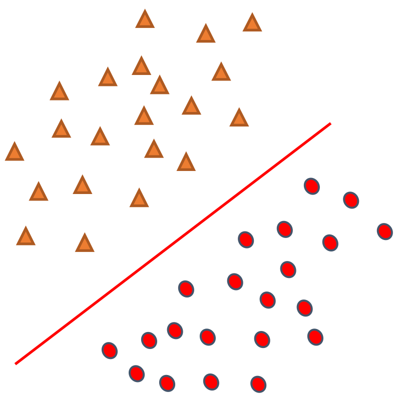
1. 线性SVM

$$\boldsymbol{w}^\top \boldsymbol{x} + b = 0$$

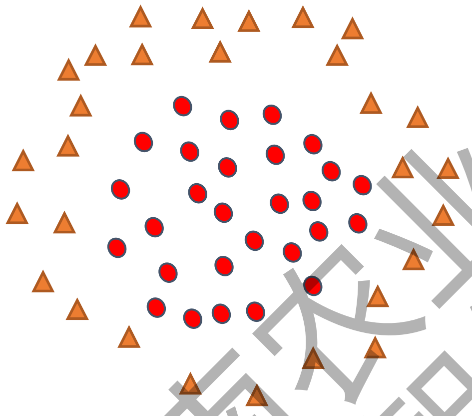


Vladimir Vapnik

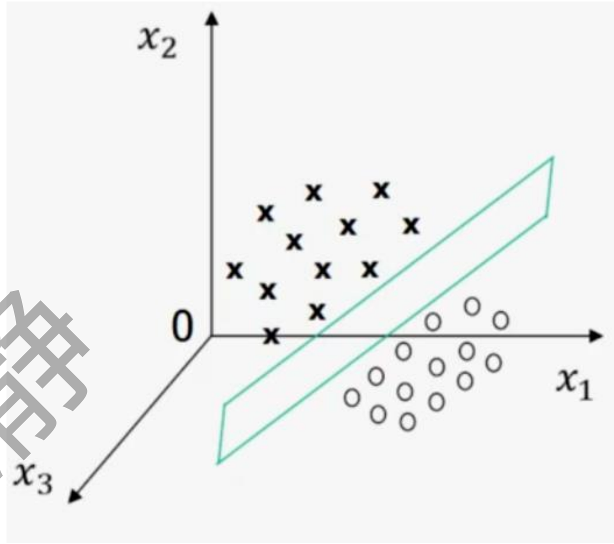
➢ 线性可分与线性不可分

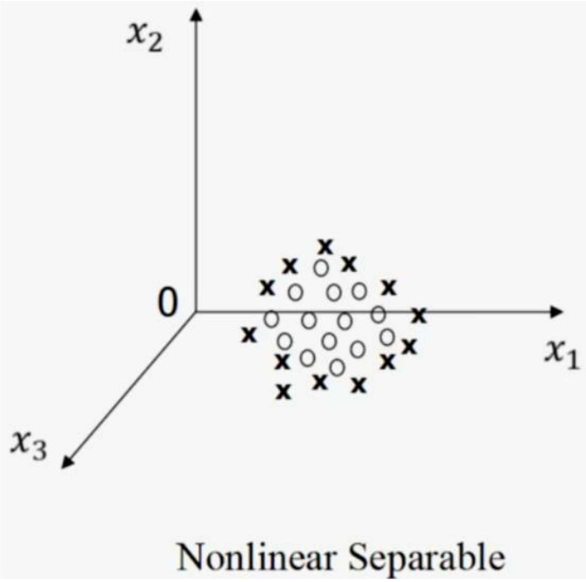


线性可分



线性不可分





➢ 定义线性可分

$$D=\{(x_1,y_1),(x_2,y_2),......(x_m,y_m)\}$$

$$y\in\{+1,-1\}$$

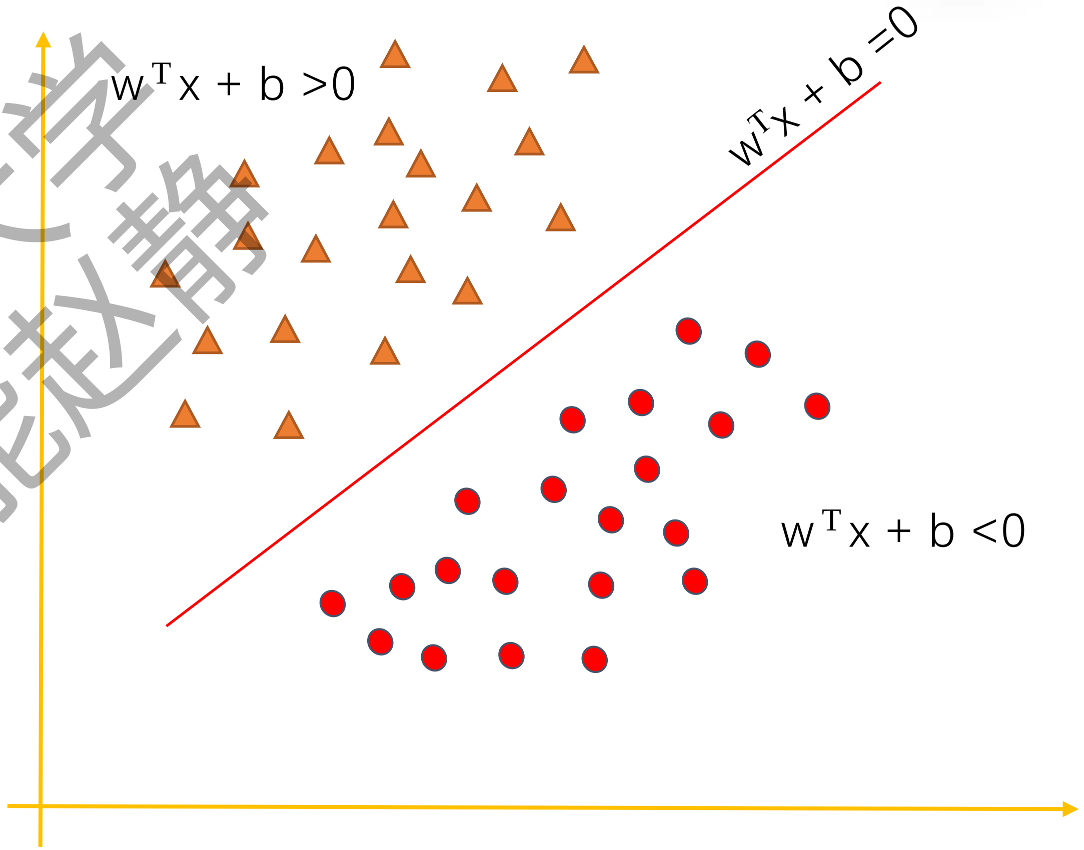
- 线性分类器$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$
- 若存在w，使得：

所有满足$f(\mathbf{x})<0$ 的点，其对应的y等于-1

所有满足$f(\mathbf{x})>0$ 的点，其对应的y等于1

则数据线性可分

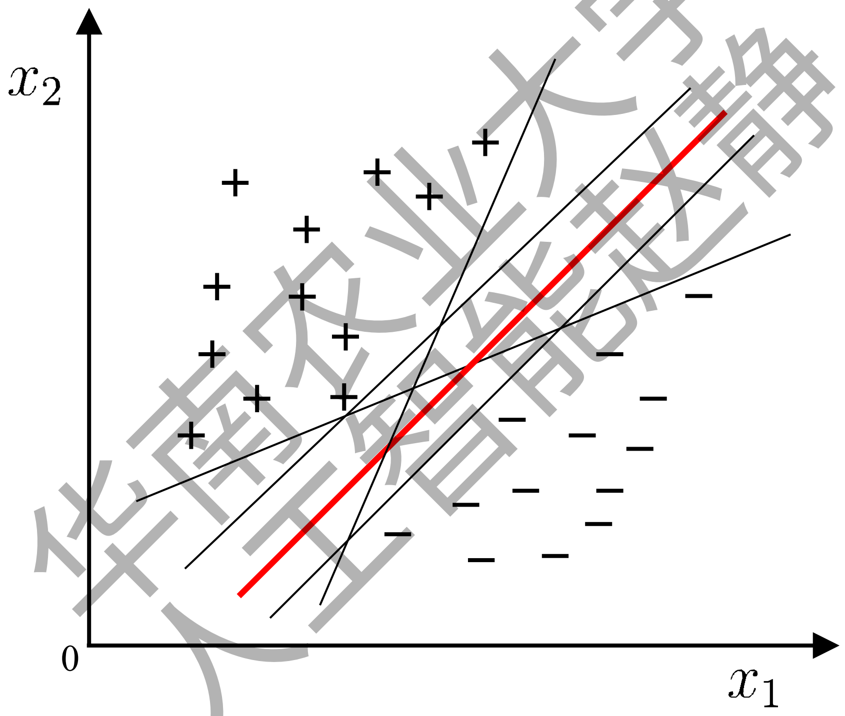
- 线性判别函数$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$，$\mathbf{x}$是位于超平面面上的点



超平面：$w^Tx+b=0$

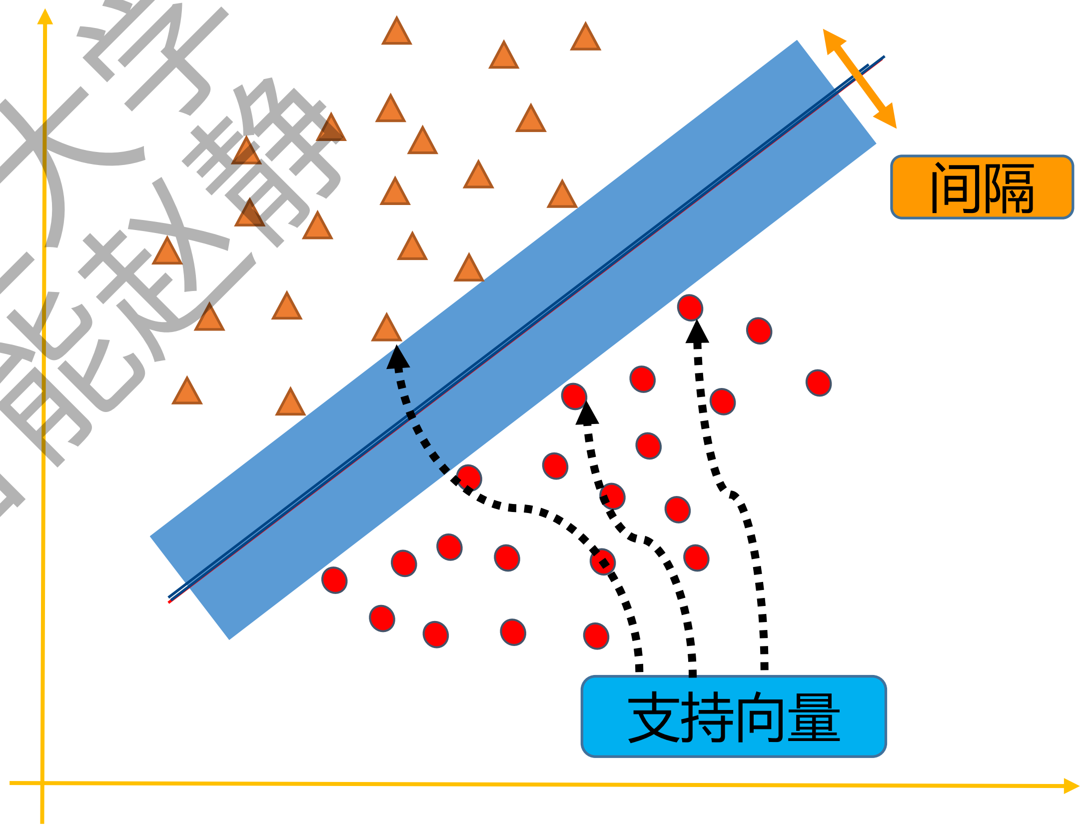
➢ 线性可分支持向量机

-Q:将训练样本分开的超平面可能有很多, 哪一个好呢?

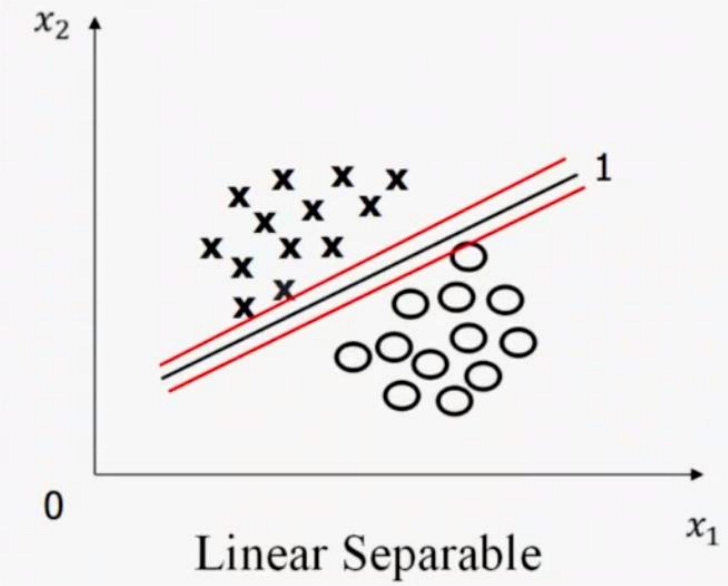


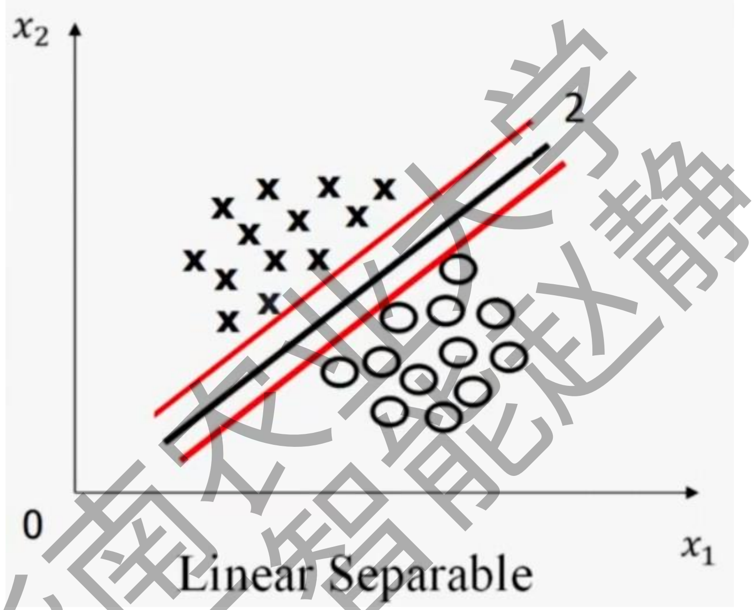
-A:应选择"正中间", 容忍性好, 鲁棒性高, 泛化能力最强.

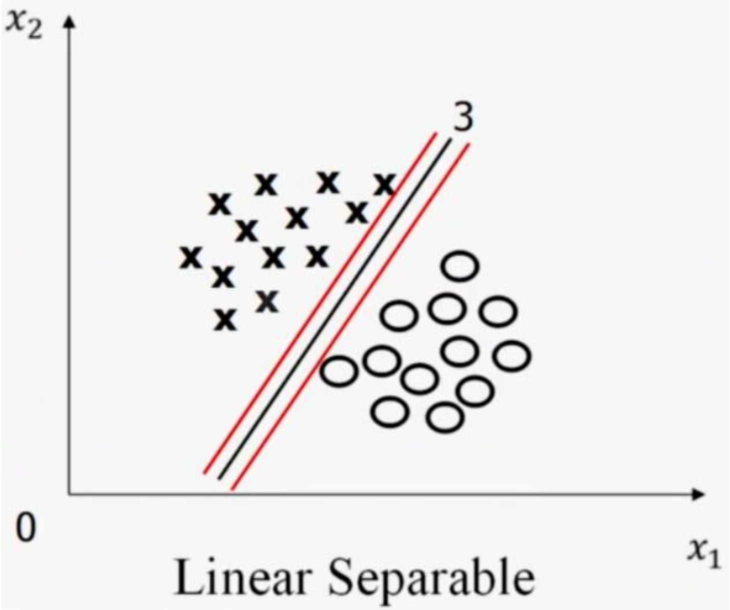
找到集合边缘上的若干数据（称为支持向量（Support Vector）），用这些点找出一个平面（称为决策面），使得支持向量到该平面的距离最大。



➢ 间隔（Margins）最大





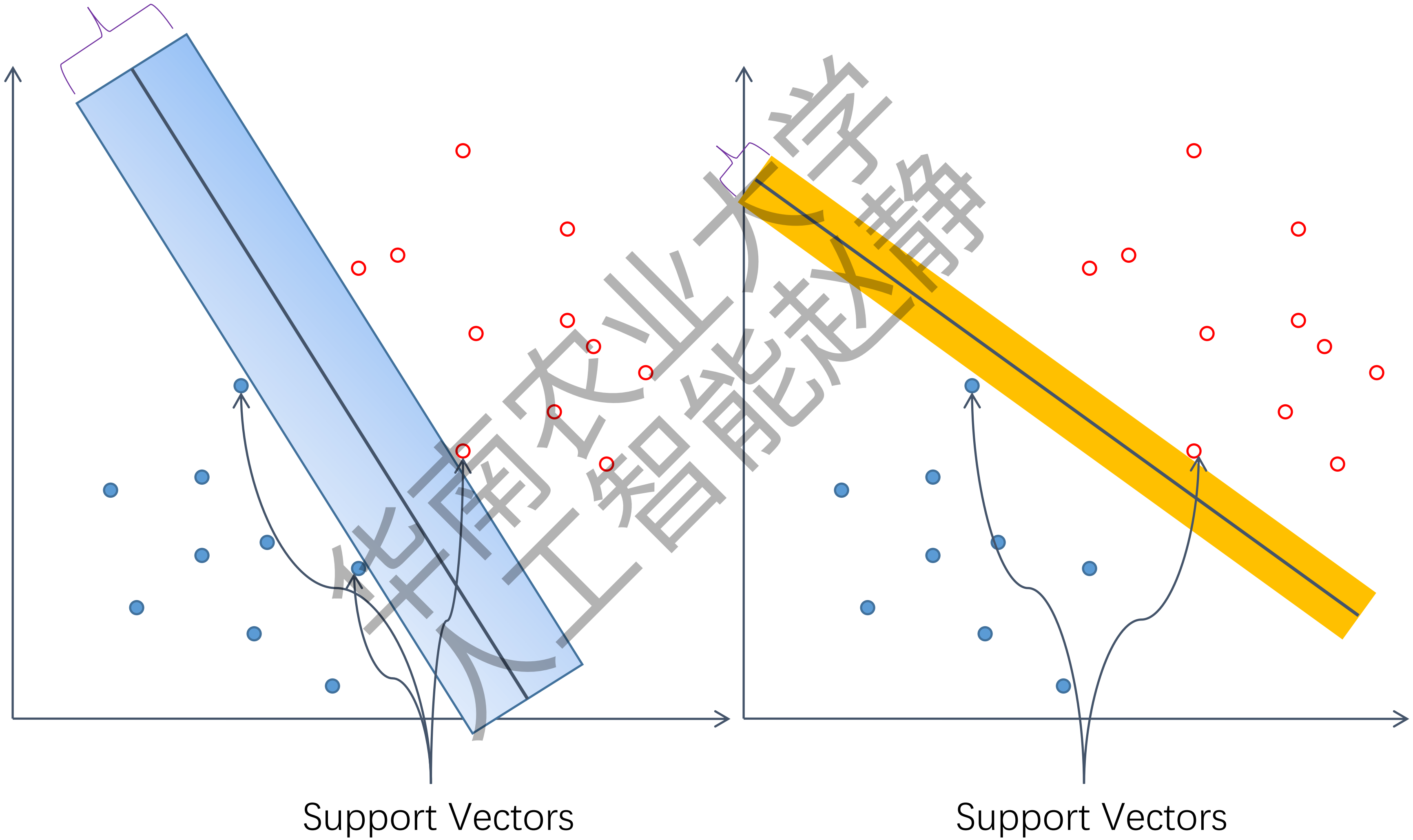


直线能分开两类
寻找间隔最大的直线
直线在间隔的正中间

$$\gamma = \frac{2}{\|\boldsymbol{w}\|}$$

超平面能分开两类
寻找间隔最大的超平面
超平面在间隔的正中间

$$\gamma = \frac{2}{\|\boldsymbol{w}\|}$$



➢ 线性可分支持向量机最优化问题

- 最大间隔: 寻找参数 $\boldsymbol{w}$ 和 b, 使得 γ 最大.

$$\underset{w,b}{argmax} \frac{1}{||w||}$$

$$s.t.\ y_i(w^T x_i + b) \geq 1,\ i = 1,2,..,m$$

$$\underset{w,b}{argmin}\ \frac{1}{2}||w||^2$$

$$s.t.\ y_i(w^T x_i + b) \geq 1,\ i = 1,2,..,m$$

$$\omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{bmatrix}$$

$$\|\omega\|^2 = \omega_1^{\ 2} + \omega_2^{\ 2} \ldots + \omega_m^{\ 2} = \sum_{i=1}^{m} \omega_i^{\ 2}$$

- 凸二次规划(convex quadratic programming)

点到平面的距离：

$$w^T x + b = 0$$

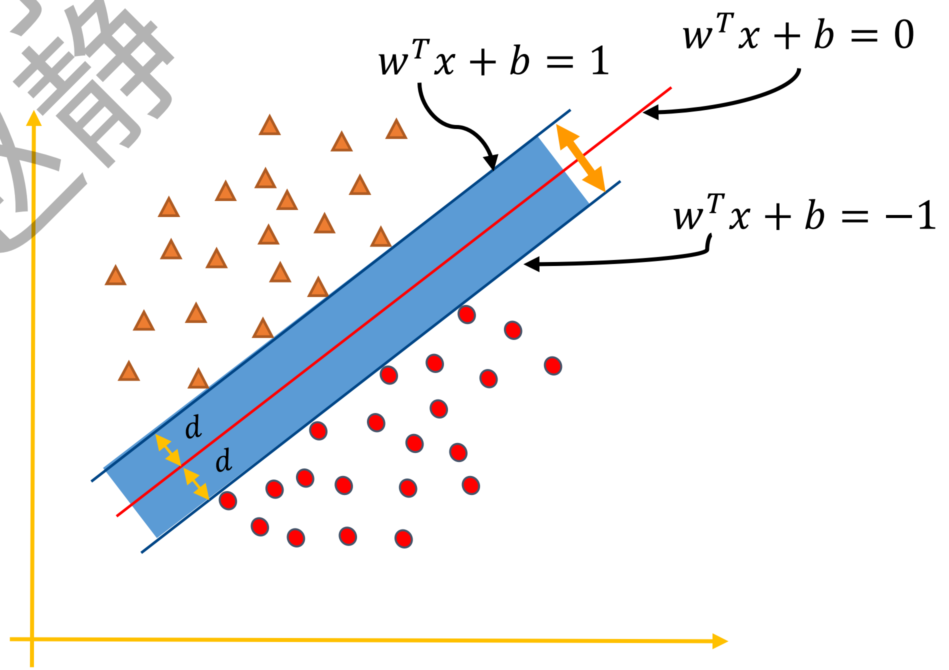
二维空间点 (x, y)到直线 $Ax + By + C = 0$的距离公式是：

$$\frac{|Ax + By + C|}{\sqrt{A^2 + B^2}}$$

扩展到 n 维空间后，点 $x = (x_1, x_2 \ldots x_n)$ 到超平面

$w^T x + b = 0$ 的距离为：$d = \frac{|w^T x + b|}{||w||}$

其中 $||w|| = \sqrt{w_1^2 + \cdots w_n^2}$



决策超平面：

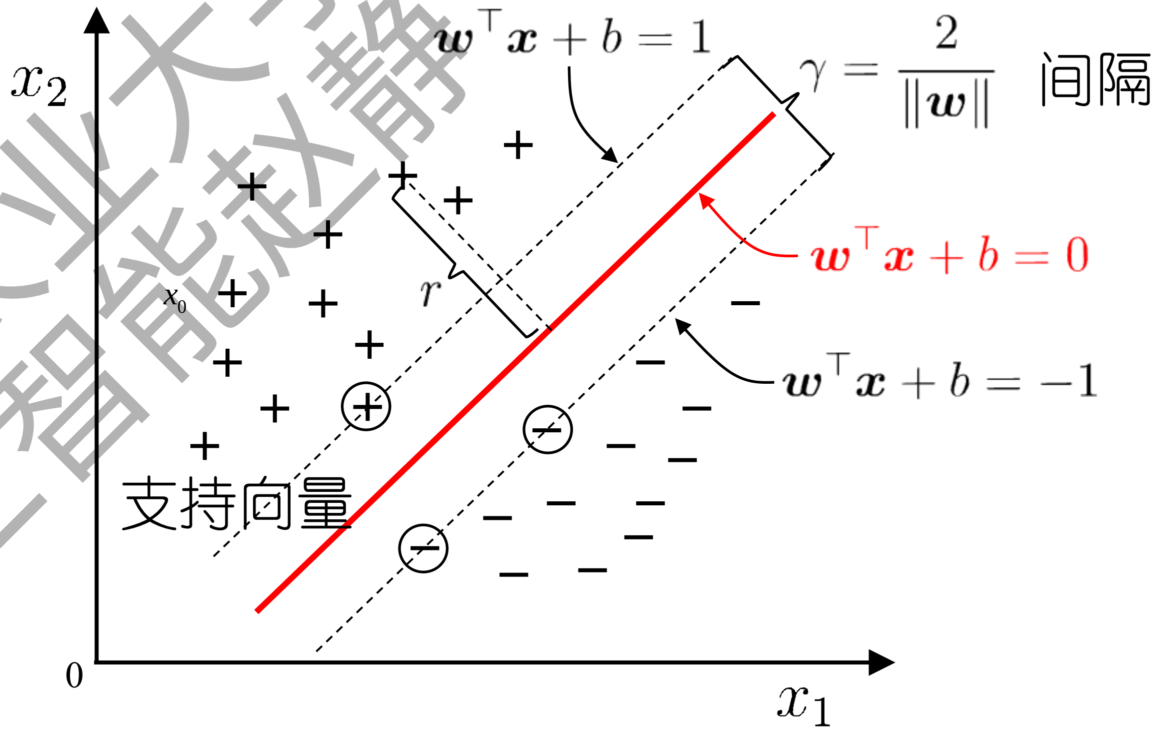
$w^T x + b = 0$ 与 $(aw^T)x + ab = 0$ $(a \neq 0)$ 是同一个超平面

$(w, b) \longrightarrow (aw, ab)$

支持向量决定的超平面：

支持向量 x_0 $\quad w^T x_0 + b = 1$

非支持向量x_0 $\quad w^T x_0 + b > 1$



支持向量 x到超平面的距离

$$d = \frac{|w^T x + b|}{||w||} = \frac{1}{||w||}$$

- 支持向量 $\boldsymbol{x}$ 到超平面的距离

$$d = \frac{|w^T x + b|}{||w||} = \frac{1}{||w||}$$

最大化 $\frac{1}{||w||}$，即最小化$||w||$ $\Rightarrow$ $\underset{w,b}{argmin}\ \frac{1}{2}||w||^2$

- 非支持向量 $\boldsymbol{x}$ 到超平面的距离

$$\begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \geqslant +1, & y_i = +1\,; \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \leqslant -1, & y_i = -1\,. \end{cases}$$

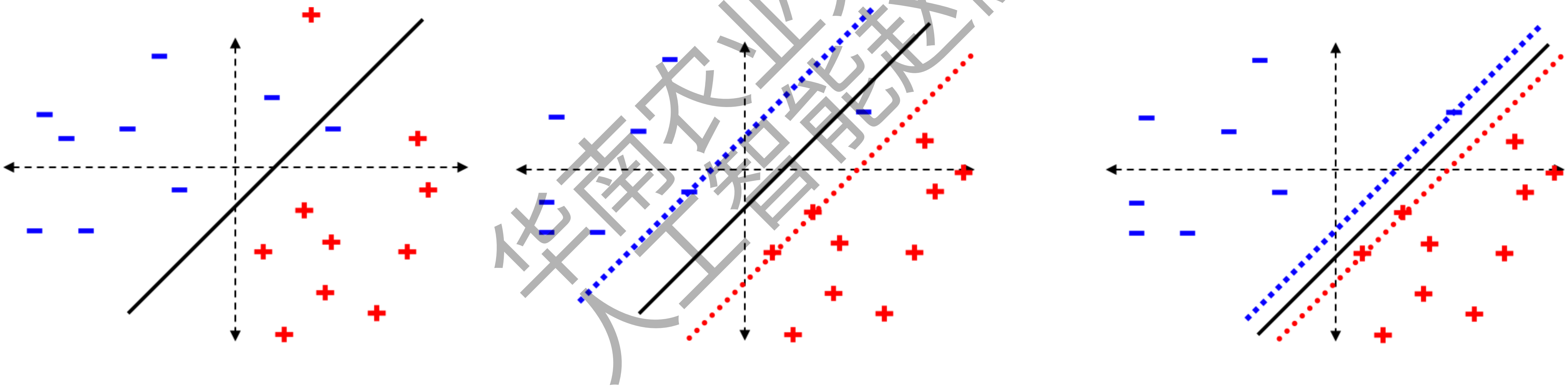
$\Downarrow$

$$s.t.\ y_i(w^T x_i + b) \geq 1,\ i = 1,2,..,m$$

$$\boxed{\begin{gathered} \underset{w,b}{argmin}\ \frac{1}{2}||w||^2 \\ s.t.\ y_i(w^T x_i + b) \geq 1,\ i = 1,2,..,m \end{gathered}}$$

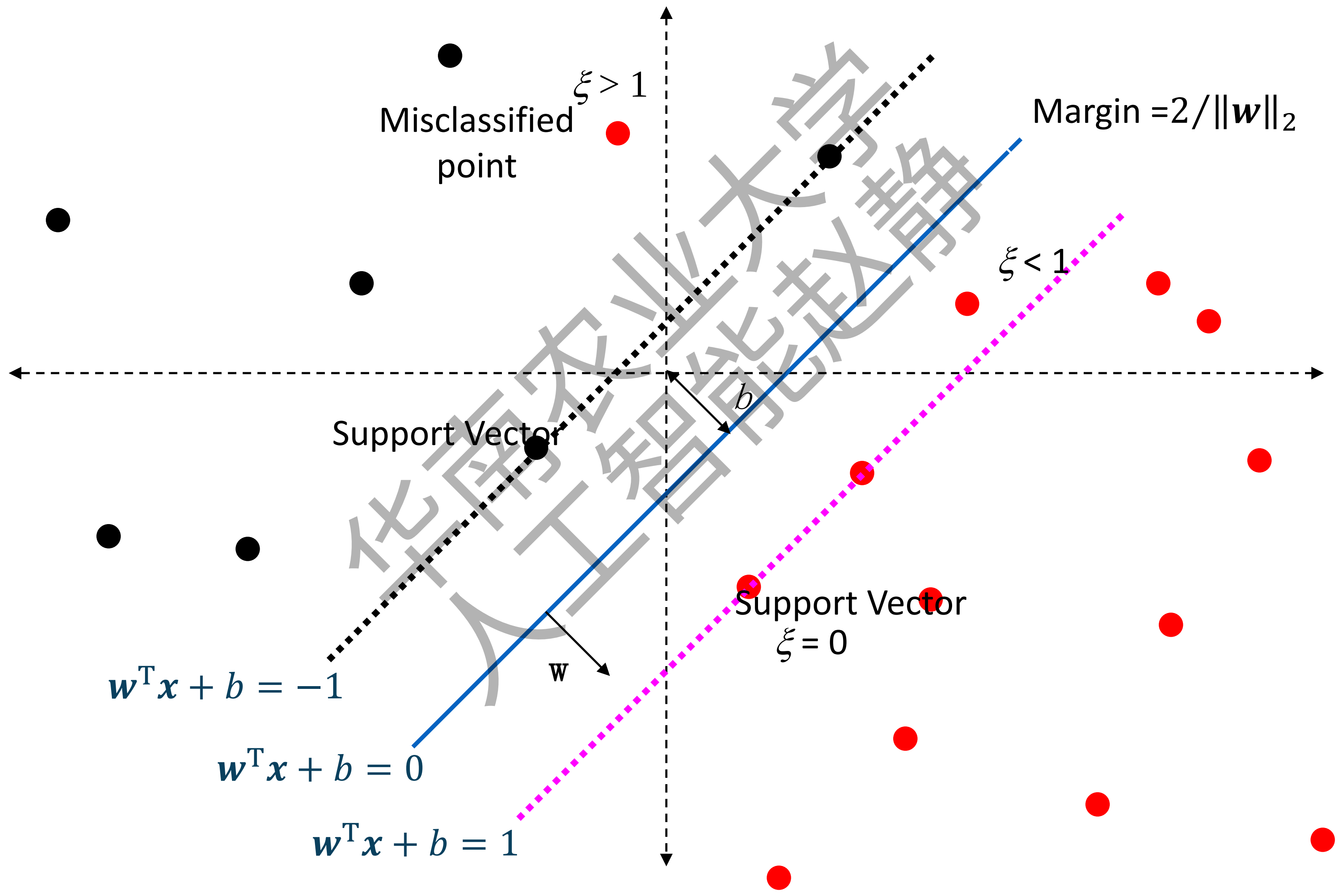
- 线性SVM
- **带松弛因子的SVM**
- 合页损失函数
- SVM的对偶问题
- 核化SVM模型
- SVM回归（SVR）

2.带松弛因子的SVM



少量样本被错分，但间隔大　　样本被完全分对，但间隔小

数据不完全线性可分：松弛变量ξ



➢ C-SVM

若数据线性不可分，则可以引入松弛变量(slack variable)$\xi \geq 0$，使函数间隔加上“**松弛变量**”大于等于1

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

则软间隔最大化SVM（C-SVM）的**目标函数**

$$J(\boldsymbol{w}, b, C) = C \sum_{i=1}^{N} \xi_i + \frac{1}{2} \|\boldsymbol{w}\|_2^2$$

$$\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b\right) \geq 1 - \xi_i, \qquad i = 1,2, \dots, N$$

$$\xi_i \geq 0, \quad i = 1,2, \dots, N$$

➤ C-SVM目标函数

$$J(\boldsymbol{w}, b, C) = C\sum_{i=1}^{N}\xi_i + \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \geq 1 - \xi_i, \qquad i = 1,2,\ldots,N$$

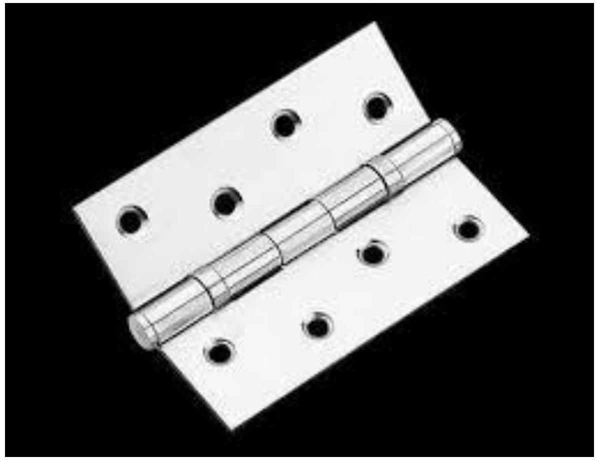
$$\xi_i \geq 0, \quad i = 1,2,\ldots,N$$

形式与带正则的线性回归或Logistic回归的目标函数类似

$$J(\boldsymbol{w}, \lambda) = C\sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i, \boldsymbol{w})) + R(\boldsymbol{w})$$

- 线性SVM
- 带松弛因子的SVM
- 合页损失函数
- SVM的对偶问题
- 核化SVM模型
- SVM回归（SVR）

3. 合页损失函数（Hinge Loss）

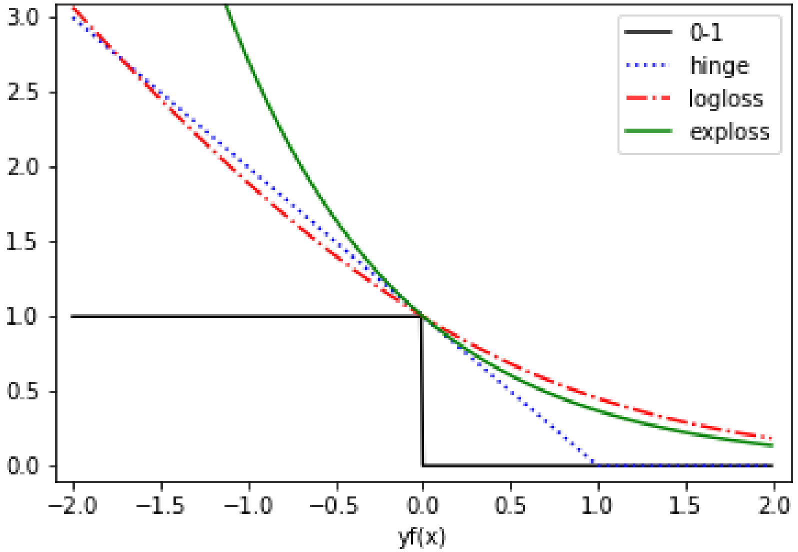


在C-SVM中

- 当$y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \geq 1$，$\xi_i = 0$
- 其他点：$\xi_i = 1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)$

合页损失(Hinge Loss)

$$\xi = L_{Hinge}(y, \hat{y}) = \begin{cases} 0 & y\hat{y} \geq 1 \\ 1 - y\hat{y} & otherwise \end{cases}$$



将合页损失代入C-SVM的目标函数

$$J(\boldsymbol{w}, b, C) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i$$

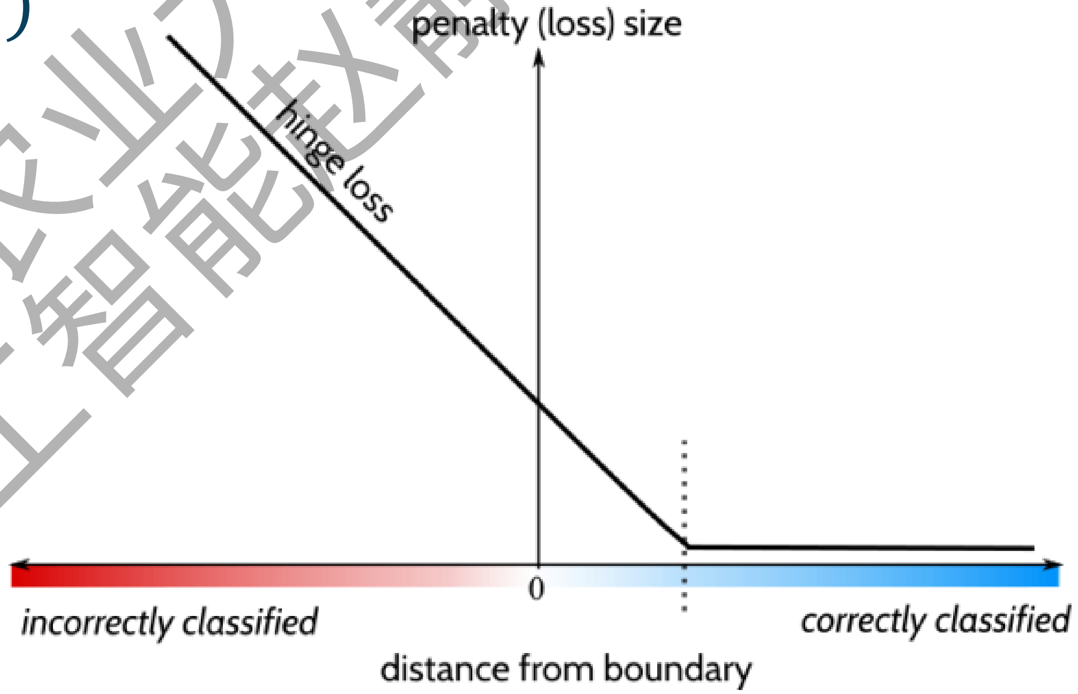
$$= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}L_{Hinge}(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}, b))$$

对比一般机器学习模型的目标函数

$$J(\boldsymbol{\theta}, \lambda) = C\sum_{i=1}^{N}L(y_i, f(\boldsymbol{x}_i, \boldsymbol{\theta})) + R(\boldsymbol{\theta})$$

$$\xi = L_{Hinge}(y, \hat{y}) = \begin{cases} 0 & y\hat{y} \geq 1 \\ 1 - y\hat{y} & otherwise \end{cases}$$

$$\xi_i = \max(0, 1 - y_i(w^T x_i + b))$$

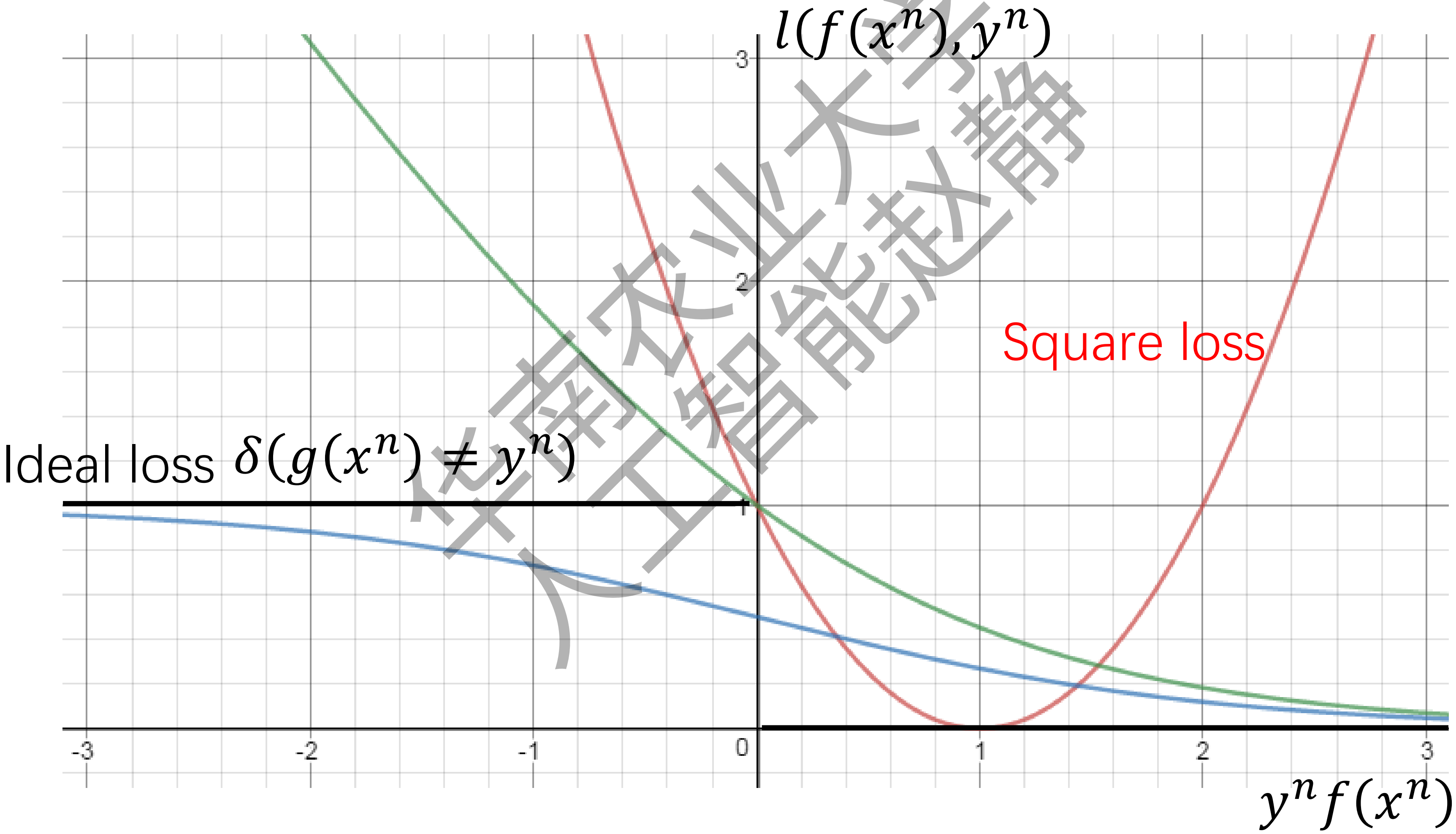


合页函数 hinge loss

✓ 总结：损失函数

- 平方差

$$\hat{y} = \begin{cases} w^T x_i + b = f(x) > 0 & y = +1 \\ w^T x_i + b = f(x) < 0 & y = -1 \end{cases}$$

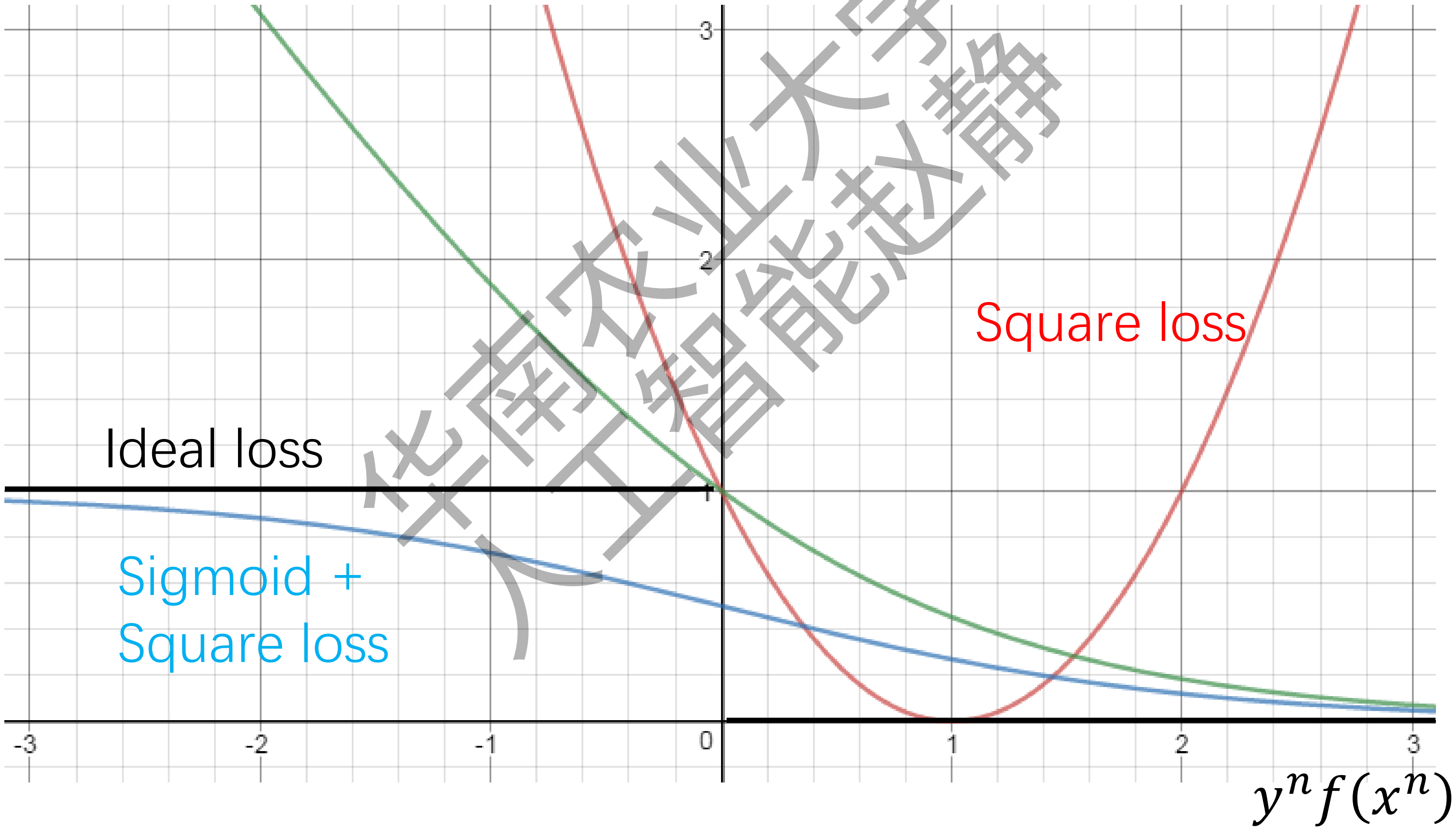


- **Sigmoid + Square Loss:**

If $y^n = 1$, $\sigma(f(x))$ close to 1

If $y^n = -1$, $\sigma(f(x))$ close to 0

$$l(f(x^n), \hat{y}^n) = \left(\sigma\left(\hat{y}^n f(x^n)\right) - 1\right)^2$$

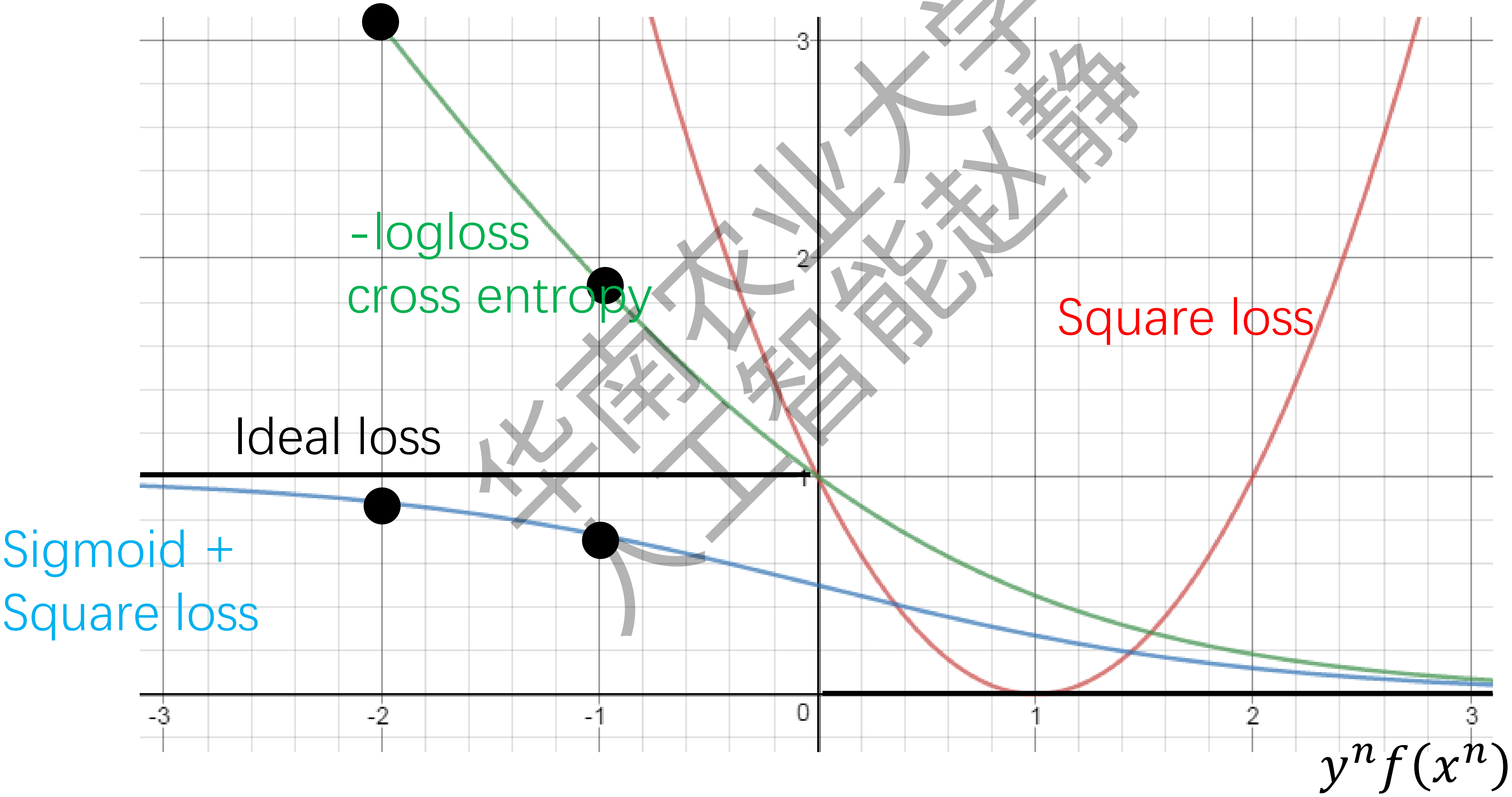


- **-logloss**

cross entropy

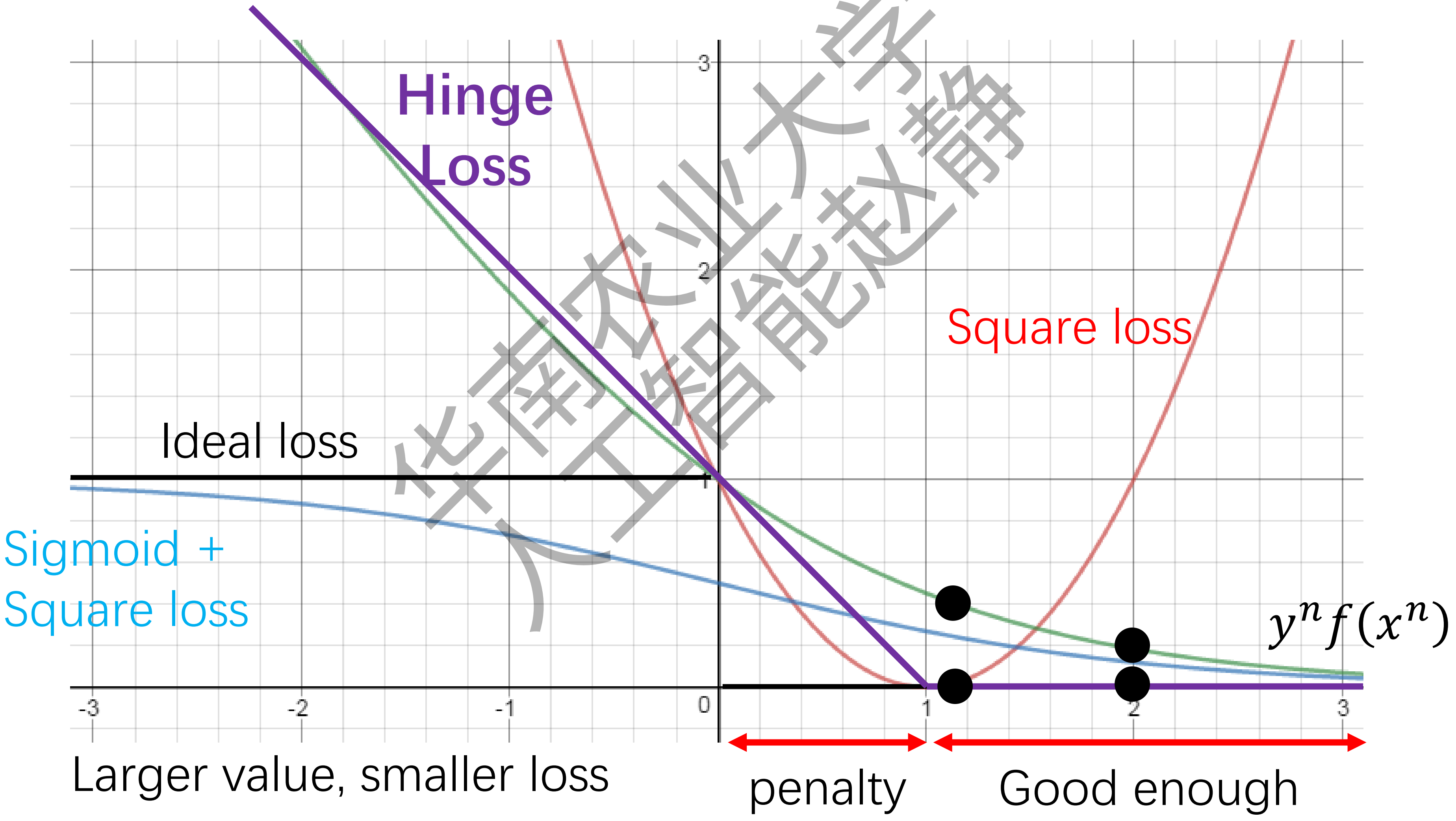
$$\ell(\mu) = -\ln p(\mathcal{D})$$

$$= -\sum_{i=1}^{N} \ln p(y_i|\boldsymbol{x}_i) = -\sum_{i=1}^{N} \ln\left(\mu(\boldsymbol{x}_i)^{y_i}(1-\mu(\boldsymbol{x}_i))^{(1-y_i)}\right)$$



- **Hinge Loss**

$$l(f(x^n), y^n) = max(0, 1 - y^n f(x^n))$$



- 线性SVM
- 带松弛因子的SVM
- 合页损失函数
- **SVM的对偶问题**
- 核化SVM模型
- SVM回归（SVR）

4. 对偶问题

- C-SVM的目标函数为

$$J(\boldsymbol{w}, b, C) = C\sum_{i=1}^{N}\xi_i + \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\text{s.t.}\ y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1 - \xi_i, \qquad i = 1,2,\dots,N$$

$$\xi_i \geq 0, \quad i = 1,2,\dots,N$$

➢ 拉格朗日乘子法

- C-SVM原问题目标函数：

$$J(\boldsymbol{w}, b, C) = C\sum_{i=1}^{N}\xi_i + \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \geq 1 - \xi_i, \qquad i = 1,2,\ldots,N$$

$$\xi_i \geq 0, \quad i = 1,2,\ldots,N$$

- 写成标准的不等式约束问题：

$$\min C\sum_{i=1}^{N}\xi_i + \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\text{s.t. } 1 - \xi_i - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \leq 0, \qquad i = 1,2,\ldots,N$$

$$-\xi_i \leq 0, \quad i = 1,2,\ldots,N$$

- 对应的广义拉格朗日函数：

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{N}\mu_i\xi_i$$

$$\text{s.t. } \alpha_i \geq 0, \quad \mu_i \geq 0, \quad \xi_i \geq 0, \quad i = 1,2,\ldots,N$$

➢ SVM的对偶问题

- 拉格朗日函数$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{N}\mu_i\,\xi_i$$

- 原问题(Primal)的最优解: $p^* = \min_{\boldsymbol{w},\boldsymbol{b}} \theta_P(\boldsymbol{w}, \boldsymbol{b}) = \min_{\boldsymbol{w},\boldsymbol{b}} \max_{\alpha_i,\mu_i \geq 0} L(w, b, \boldsymbol{\alpha}, \mu)$
- 对偶问题(Dual)的最优解: $d^* = \max_{\alpha_i,\mu_i \geq 0} \min_{\boldsymbol{w},\boldsymbol{b}} L(w, b, \boldsymbol{\alpha}, \mu)$
- 二者关系$d^* \leq p^*$
- 满足KKT条件时，$d^* = p^*$

注意：和原问题相比，对偶问题交换了max和min的顺序。

- 对$\boldsymbol{w}, b, \xi_i$求导，令一阶导数为0，求 $\min\limits_{\boldsymbol{w},\boldsymbol{b}} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{N}\mu_i\,\xi_i$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})}{\partial \boldsymbol{w}} = 0 \Longrightarrow \boldsymbol{w} = \sum_{i=1}^{N}\alpha_i\, y_i \boldsymbol{x}_i$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{N}\alpha_i\, y_i = 0$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})}{\partial \xi_i} = 0 \Longrightarrow C = \alpha_i + \mu_i$$

- 将上述结论代入拉格朗日函数$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{N}\mu_i\xi_i$$

$$C = \alpha_i + \mu_i$$

$$= \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{N}\alpha_i\left(y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i\right) + \sum_{i=1}^{N}\alpha_i\xi_i$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$= \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} - \sum_{i=1}^{N}\alpha_i y_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - \sum_{i=1}^{N}\alpha_i y_i b + \sum_{i=1}^{N}\alpha_i$$

$$\boldsymbol{w} = \sum_{i=1}^{N}\alpha_i y_i\boldsymbol{x}_i$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j - b\sum_{i=1}^{N}\alpha_i y_i + \sum_{i=1}^{N}\alpha_i$$

$$= \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$$

- 从而得到对偶变量$\boldsymbol{\alpha}$的优化问题：

$$\max\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j\right)$$

$$\text{s.t.}\sum_{i=1}^{N}\alpha_i\,y_i=0$$

$$0\le\alpha_i\le C,\qquad i=1,2,\ldots,N$$

w, b的计算

得到最佳的$\boldsymbol{\alpha}$后，可计算$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$

用任意一个支持向量即可求得b: $b = y_i - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i$

为了得到更稳定的解，对所有支持向量求平均

$$b = \frac{1}{N_{\mathcal{S}}} \sum_{m \in \mathcal{S}} \left(y_m - \sum_{m' \in \mathcal{S}} \alpha_{m'} y_{m'} \langle \boldsymbol{x}_m, \boldsymbol{x}_{m'} \rangle \right)$$

得到$\boldsymbol{\alpha}, \mathbf{w}, b$ 后，可计算SVM模型：

$$
\begin{aligned}
f(\mathbf{x}) &= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \\
&= \left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i\right)^{\mathrm{T}} \boldsymbol{x} + b \\
&= \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x} + b \\
&= \sum_{i=1}^{N} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b
\end{aligned}
$$

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

点$\boldsymbol{x}$ 的标签可根据$f(\boldsymbol{x})$的符号得到：$\hat{y} = \mathrm{sign}(f(\boldsymbol{x}))$

➢ α的稀疏性

SVM目标函数对应的KKT条件中，每个训练样本点：

$$\alpha_i\left(1-\xi_i-y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\right)=0$$

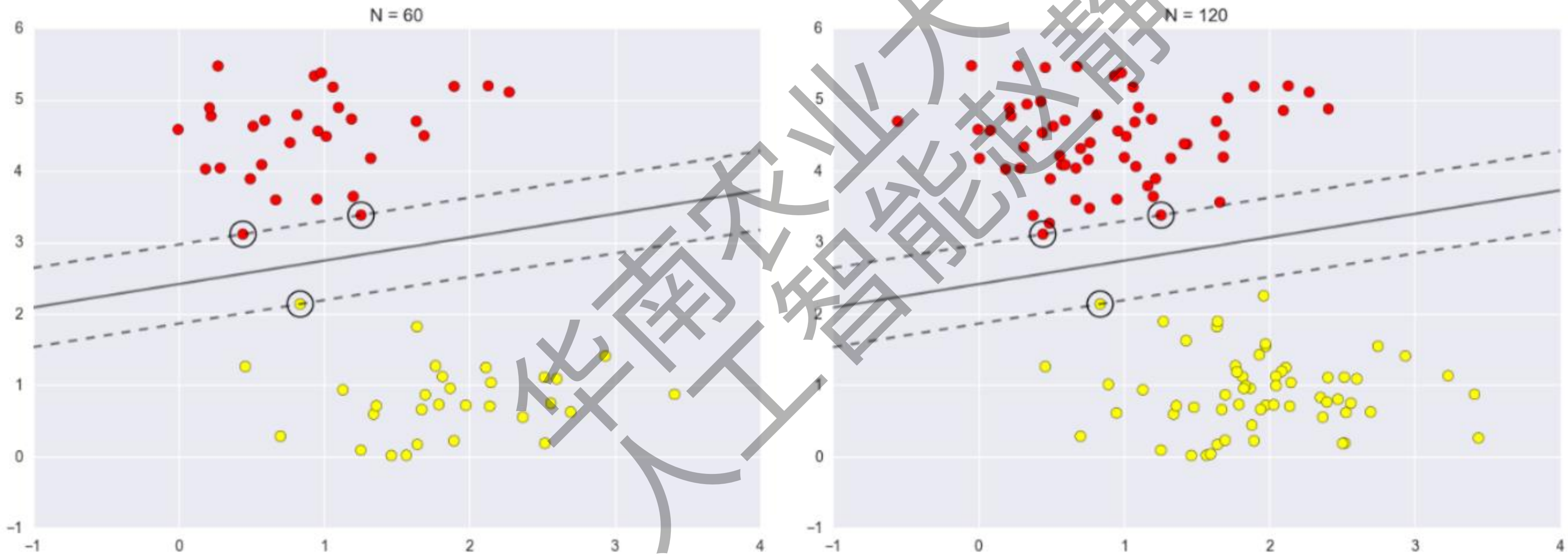
因此，　$\alpha_i=0$　或　$1-\xi_i-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)=0$

- 当$\alpha_i=0$ 时，该点在决策函数中不起作用（非支持向量）

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=\sum_{i=1}^{N}\alpha_i y_i\langle\boldsymbol{x}_i,\boldsymbol{x}\rangle+b$$

- 当$\alpha_i\neq 0$时，$1-\xi_i-y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)$=0，这些样本点被称为支持向量。

支持向量：真正发挥作用的数据点，α值不为0的点



例：数据为3个样本，其中正例 $X_1(3,3)$，$X_2(4,3)$，负例$X_3(1,1)$

解： $\min\limits_{\alpha} \left(\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^{\mathrm{T}} x_j - \sum_{i=1}^{m} \alpha_i \right)$

$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0,$$

$$\alpha_i \geq 0$$

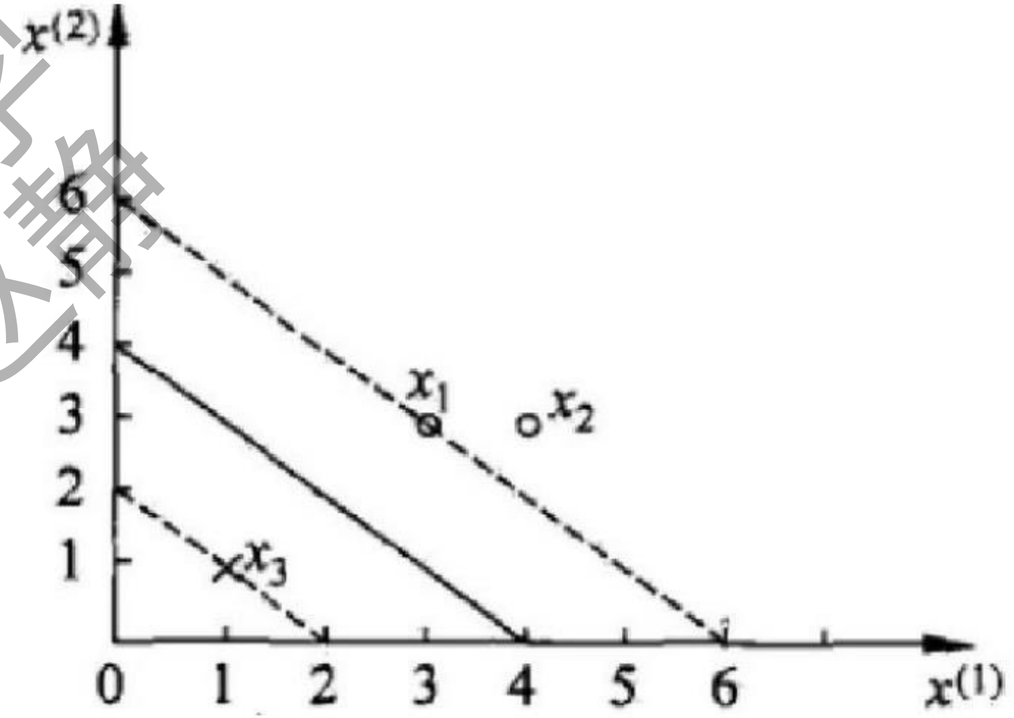
$$\frac{1}{2}\left(18\alpha_1^2 + 25\alpha_2^2 + 2\alpha_3^2 + 42\alpha_1\alpha_2 - 12\alpha_1\alpha_3 - 14\alpha_2\alpha_3\right) - \alpha_1 - \alpha_2 - \alpha_3$$

由于：$\alpha_1 + \alpha_2 = \alpha_3$ 化简可得：$4\alpha_1^2 + \frac{13}{2}\alpha_2^2 + 10\alpha_1\alpha_2 - 2\alpha_1 - 2\alpha_2$

分别对α_1和α_2求偏导，偏导等于0可得，α_1 =1.5 α_2 =-1

所以解应在边界上α_1 =0 α_2 =-2/13

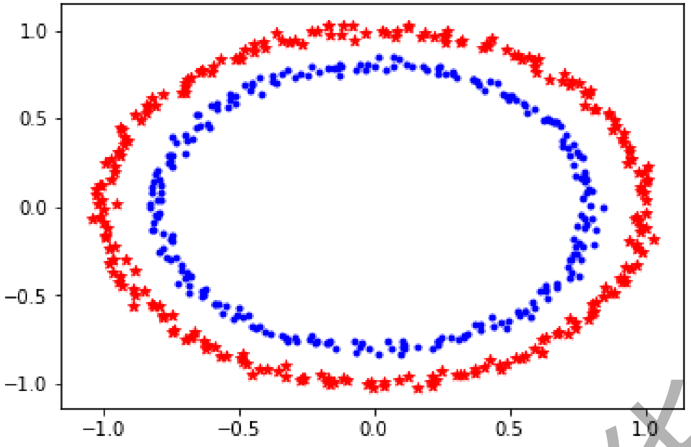
α_1 =0.25 α_2 =0, α_3 =0.25 √



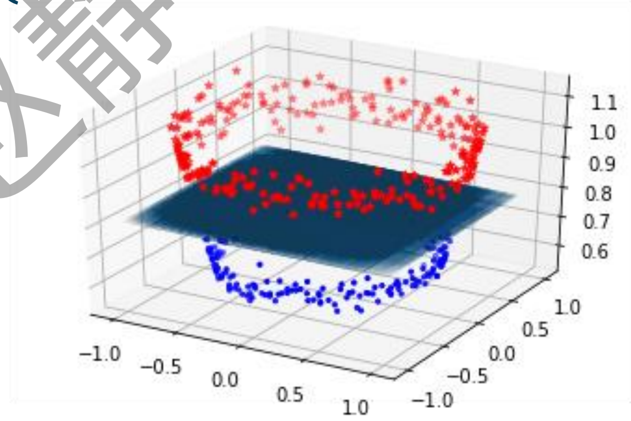
$$\left.\begin{aligned} w^* &= \sum_{i=1}^{m} \alpha_i^* y_i x_i = 1/4 * 1 * (3,3) + 1/4 * (-1) * (1,1) = (1/2,1/2) \\ b^* &= y - w^{*\,T} x = 1 - (1/2,1/2)^{\mathrm{T}} (3,3) = -2 \end{aligned}\right\} \Rightarrow \quad 1/2\, x_1 + 1/2\, x_2 - 2 = 0$$

- 线性SVM
- 带松弛因子的SVM
- 合页损失函数
- SVM的对偶问题
- **核化SVM模型**
- SVM回归（SVR）

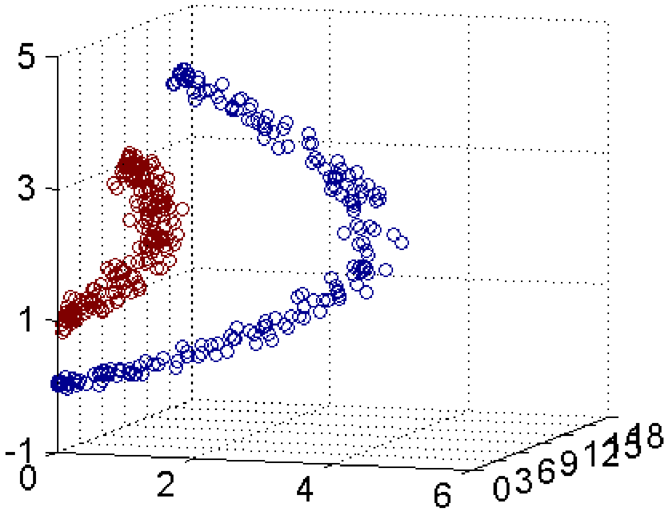
5. 核方法



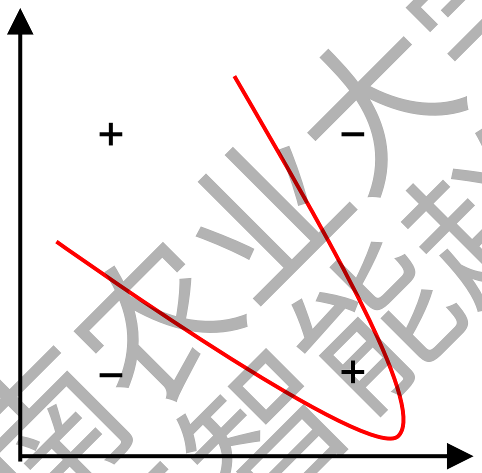
$$(z_1, z_2, z_3) = (x, y, x^2 + y^2)$$



将原始空间映射到一个更高维特征空间，使得在这个特征空间数据线性可分。

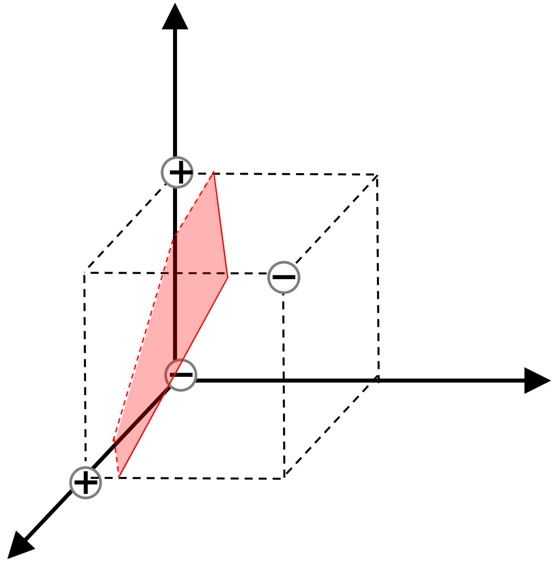


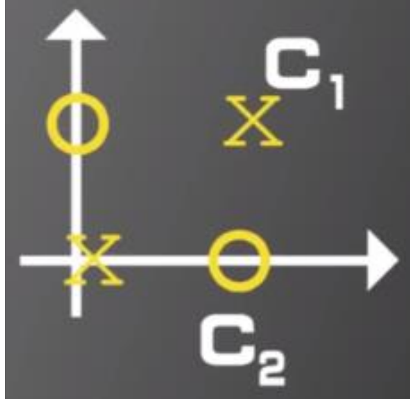
高维下线性可分



XOR数据集

$$\boldsymbol{x} \mapsto \phi(\boldsymbol{x})$$





即：

$$x_1=\begin{bmatrix}0\\0\end{bmatrix}\in C_1 \quad x_2=\begin{bmatrix}1\\1\end{bmatrix}\in C_1 \quad x_3=\begin{bmatrix}1\\0\end{bmatrix}\in C_2 \quad x_4=\begin{bmatrix}0\\1\end{bmatrix}\in C_2$$

构造一个5维函数

$$\varphi(x): \quad x=\begin{bmatrix}a\\b\end{bmatrix} \longrightarrow \varphi(x)=\begin{bmatrix}a^2\\b^2\\a\\b\\ab\end{bmatrix}$$

$$\varphi(x_1)=\begin{bmatrix}0\\0\\0\\0\\0\end{bmatrix} \quad \varphi(x_2)=\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$$

$$\varphi(x_3)=\begin{bmatrix}1\\0\\1\\0\\0\end{bmatrix} \quad \varphi(x_4)=\begin{bmatrix}0\\1\\0\\1\\0\end{bmatrix}$$

设：

$$\omega=\begin{bmatrix}-1\\-1\\-1\\-1\\6\end{bmatrix} \qquad b=1$$

$$\omega^T\varphi(x_1)+b=1\geqslant 0 \qquad \omega^T\varphi(x_2)+b=3\geqslant 0$$

$$\omega^T\varphi(x_3)+b=-1<0 \qquad \omega^T\varphi(x_4)+b=-1<0$$

➢ 目标函数

- 令$\phi(\boldsymbol{x})$ 表示将$\boldsymbol{x}$映射后的特征向量，则在特征空间中划分超平面对应的模型可表示为

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})+b$$

- 根据之前SVM的推导，得到特征映射后的SVM目标函数

$$\min\ C\sum_{i=1}^{N}\xi_i+\frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\mathrm{s.t.}\ 1-\xi_i\le y_i\left(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b\right),\qquad i=1,2,\ldots,N$$

$$\xi_i\ge 0,\quad i=1,2,\ldots,N$$

$$\boldsymbol{x} \mapsto \phi(\boldsymbol{x})$$

$$\min_{w,b,\xi} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i$$

$s.t.$ ① $\xi_i \geq 0$

② $y_i(w^T x_i + b) \geq 1 - \xi_i$

$$\min_{w,b,\xi} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i$$

$s.t.$ ① $\xi_i \geq 0$

② $y_i(w^T \emptyset(x_i) + b) \geq 1 - \xi_i$

➢ 核方法——对偶

- 相应的对偶问题为：

$$\max\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j y_i y_j\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle\right)$$

$$\mathrm{s.t.}\sum_{i=1}^{N}\alpha_i\,y_i=0$$

$$0\le\alpha_i\le C,\qquad i=1,2,\ldots,N$$

- 求得对偶问题的解$\boldsymbol{\alpha}$后，可计算$\boldsymbol{w},b$，从而得到分类判别函数：

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})+b$$

$$=\sum_{i=1}^{N}\alpha_i\,y_i\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x})\rangle+b$$

➢ 核技巧（Kernel Trick）

- 判别函数为：$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b = \sum_{i=1}^{N} \alpha_i \, y_i \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle + b$
- **核函数**：高维空间中的点积可写成核(*kernel*)的形式

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$$

- SVM核化目标函数为

$$W(\boldsymbol{\alpha}) = \sum_{i=0}^{N} \alpha_i - \frac{1}{2} \sum_{i=0}^{N} \sum_{j=0}^{N} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- 判别函数为：$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b = \sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b$

Kernel Trick

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\phi(x) = \begin{bmatrix} {x_1}^2 \\ \sqrt{2}x_1x_2 \\ {x_2}^2 \end{bmatrix}$$

$$K(x,z) = \phi(x)\cdot\phi(z) = \begin{bmatrix} {x_1}^2 \\ \sqrt{2}x_1x_2 \\ {x_2}^2 \end{bmatrix} \cdot \begin{bmatrix} {z_1}^2 \\ \sqrt{2}z_1z_2 \\ {z_2}^2 \end{bmatrix}$$

$$= {x_1}^2{z_1}^2 + 2x_1x_2z_1z_2 + {x_2}^2{z_2}^2$$

$$= (x_1z_1 + x_2z_2)^2 = \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\right)^2$$

$$= (x\cdot z)^2$$

Kernel Trick

$$K(x,z) = (x \cdot z)^2$$

$$= (x_1 z_1 + x_2 z_2 + \cdots + x_k z_k)^2$$

$$= \underline{x_1}^2 \underline{z_1}^2 + \underline{x_2}^2 \underline{z_2}^2 + \cdots + \underline{x_k}^2 \underline{z_k}^2$$

$$+2\underline{x_1 x_2}\,\underline{z_1 z_2} + 2\underline{x_1 x_3}\,\underline{z_1 z_3} + \cdots$$

$$+2\underline{x_2 x_3}\,\underline{z_2 z_3} + 2\underline{x_2 x_4}\,\underline{z_2 z_4} + \cdots$$

$$= \phi(x) \cdot \phi(z)$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \quad z = \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix}$$

$$\phi(x) = \begin{bmatrix} x_1{}^2 \\ \vdots \\ x_k{}^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1x_3 \\ \vdots \\ \sqrt{2}x_2x_3 \\ \vdots \end{bmatrix}$$

➢ 构造核函数

核函数K和映射ϕ是一一对应关系

核函数的形式不能随意取

满足一定的条件

两个ϕ内积的形式

- 令$\mathcal{X}$为输入空间， $k(.,.)$是定义在$\mathcal{X}*\mathcal{X}$上的对称函数，则k是核函数的充要条件是对任意数据$\mathcal{D}=\{\boldsymbol{x}_i\}_{i=1}^{N}$，核矩阵$\boldsymbol{K}$总是半正定的：

$$\boldsymbol{K}=\begin{bmatrix} \kappa(\boldsymbol{x}_1,\boldsymbol{x}_1) & \kappa(\boldsymbol{x}_1,\boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_1,\boldsymbol{x}_N) \\ \kappa(\boldsymbol{x}_2,\boldsymbol{x}_1) & \kappa(\boldsymbol{x}_1,\boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_2,\boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\boldsymbol{x}_N,\boldsymbol{x}_1) & \kappa(\boldsymbol{x}_N,\boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_N,\boldsymbol{x}_N) \end{bmatrix}$$

✓ 多项式核

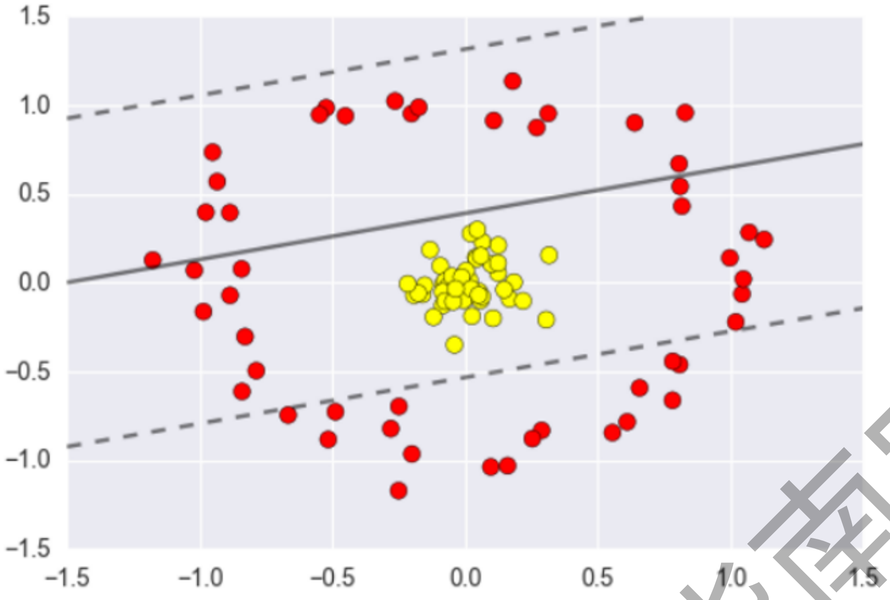
- 多项式核： $k(\boldsymbol{x},\boldsymbol{x}')=\left(\gamma\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'+r\right)^{M}$
- 当$M=1$，$\gamma=1$，$r=0$时，为线性核，

$$\phi(\boldsymbol{x})=\boldsymbol{x}$$

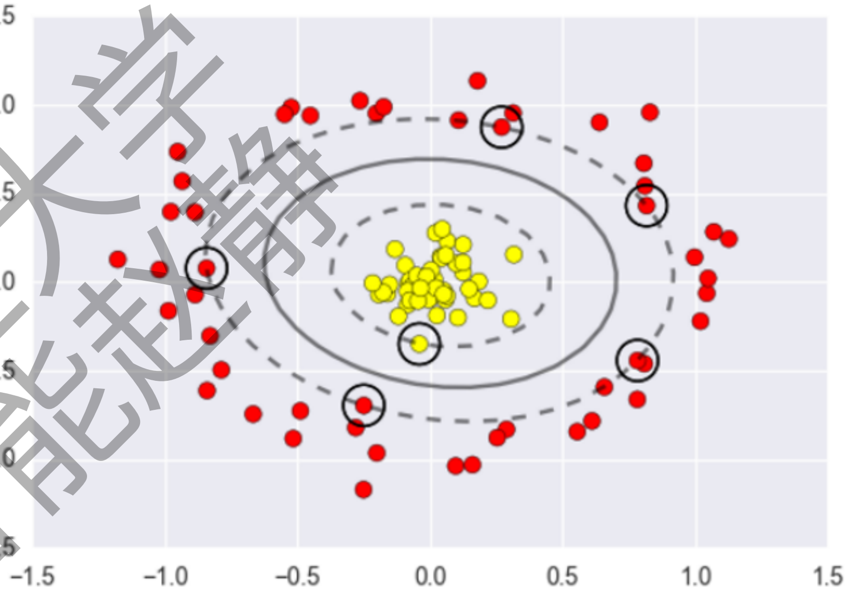
$$k(\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'$$

✓ 高斯核函数（径向基核函数 RBF）

$$k(\boldsymbol{x},\boldsymbol{x}')=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^{2}}{2\sigma^{2}}\right)=\exp\left(-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^{2}\right)$$


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

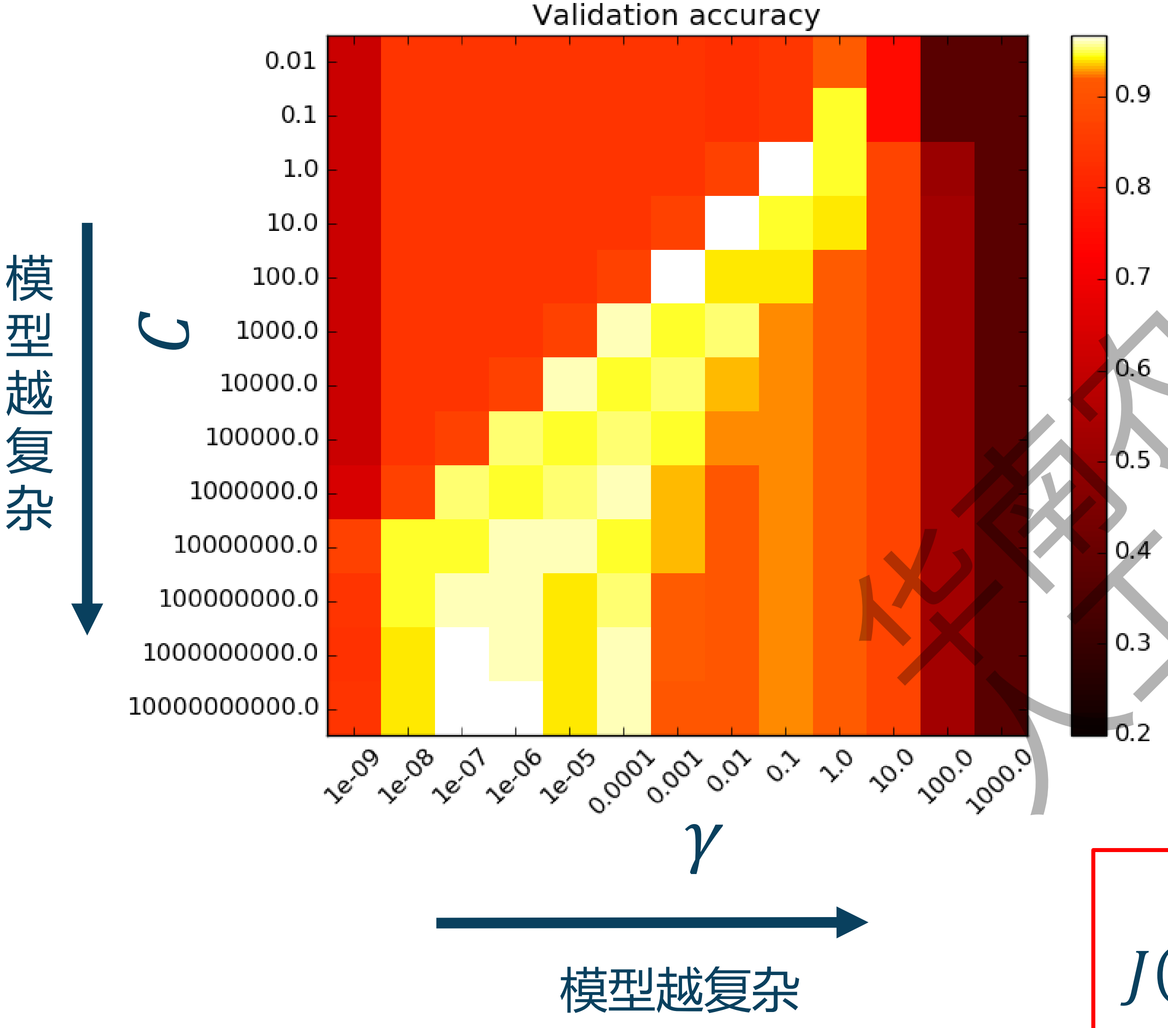
线性核函数


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

高斯核函数

例： RBF核

在鸢尾花分类任务上（只取了前2维特征），不同参数值的RBF核SVM分类器的交叉验证精度



C： 训练样本有多重要

越大，越看重训练样本，模型越复杂

γ： RBF核宽度的倒数，

表示每个训练样本的影响范围

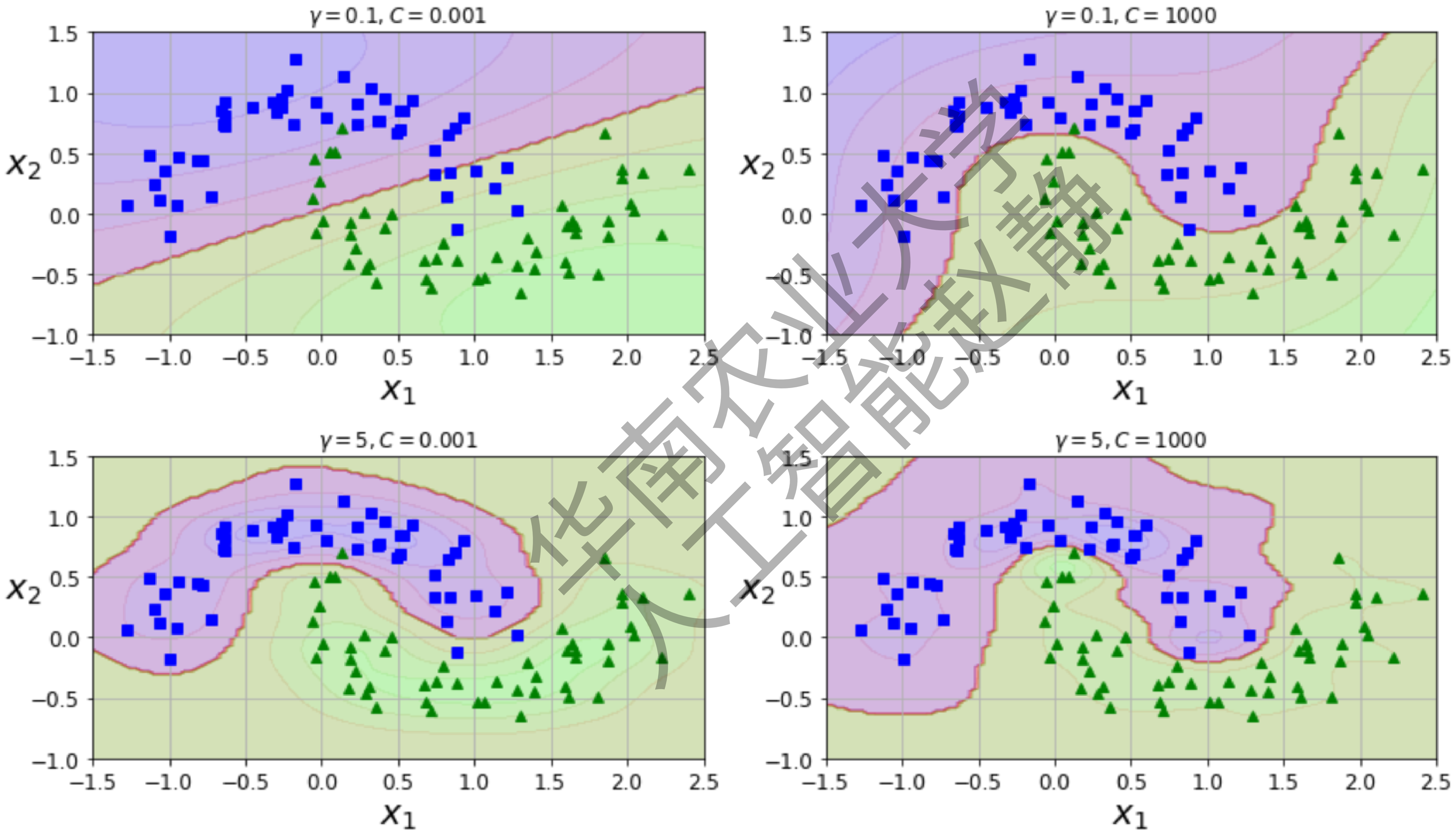
越大， RBF核宽度越小，

每个训练样本/支持向量的范围越小，

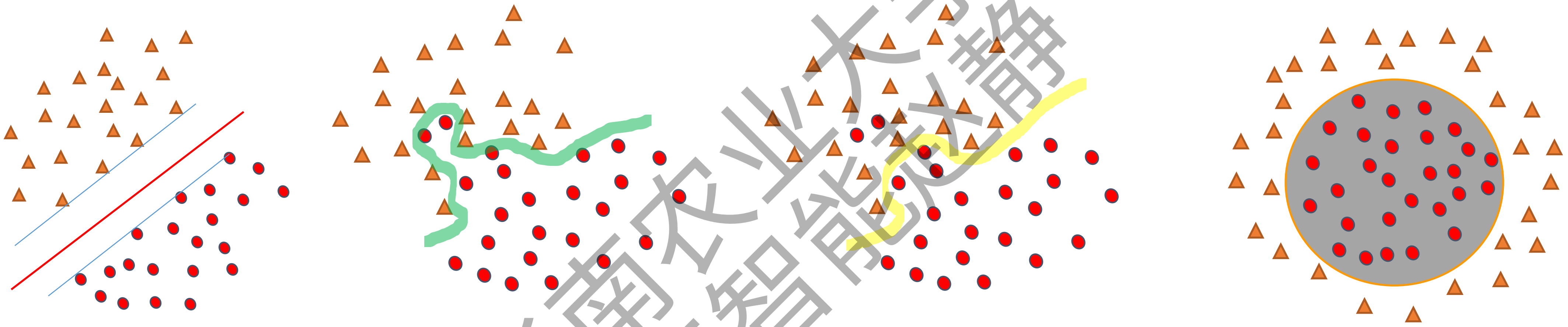
模型越复杂

$$J(\mathbf{w}, b; C) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N} L_{Hinge}(y_i, f(\mathbf{x}_i; \mathbf{w}, b))$$

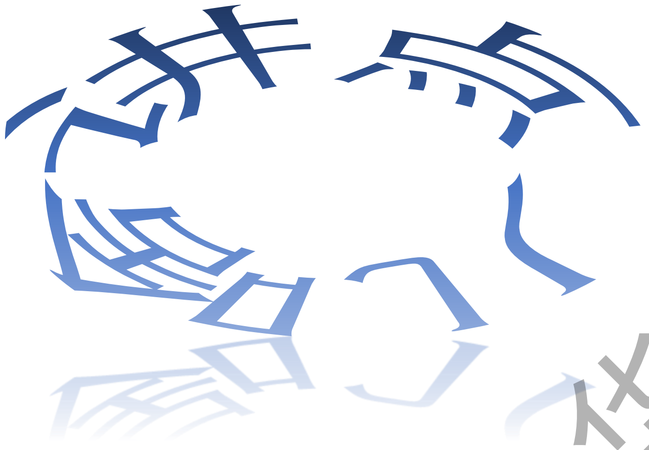
例： RBF核



硬间隔、软间隔和非线性 SVM



硬间隔　　软间隔　　软间隔　　线性不可分

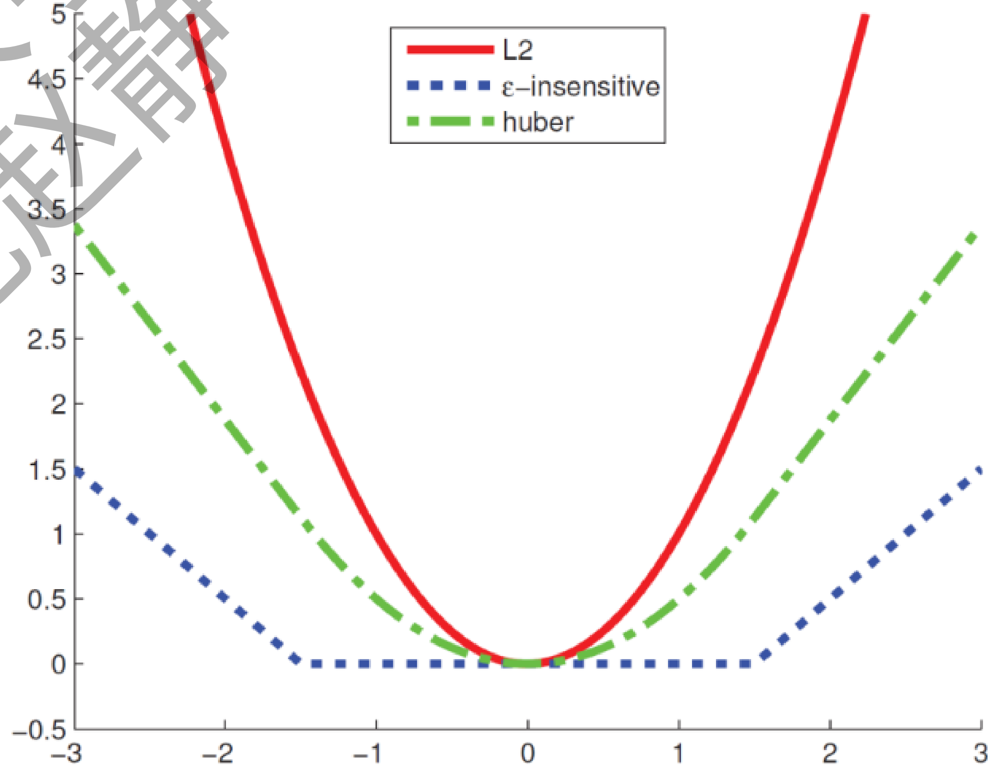


- 线性SVM
- 带松弛因子的SVM
- 合页损失函数
- SVM的对偶问题
- 核化SVM模型
- SVM回归（SVR）

- ϵ不敏感损失函数（ϵ insensitive loss ）

$$L_{\epsilon}(y,\hat{y}) = \begin{cases} 0 & |y-\hat{y}| \leq \epsilon \\ |y-\hat{y}|-\epsilon & otherwise \end{cases}$$

不惩罚小的损失



➢ 支持向量回归（Support Vector Regression， SVR）

- 假设回归函数为线性模型：$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$，SVR的目标函数为

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

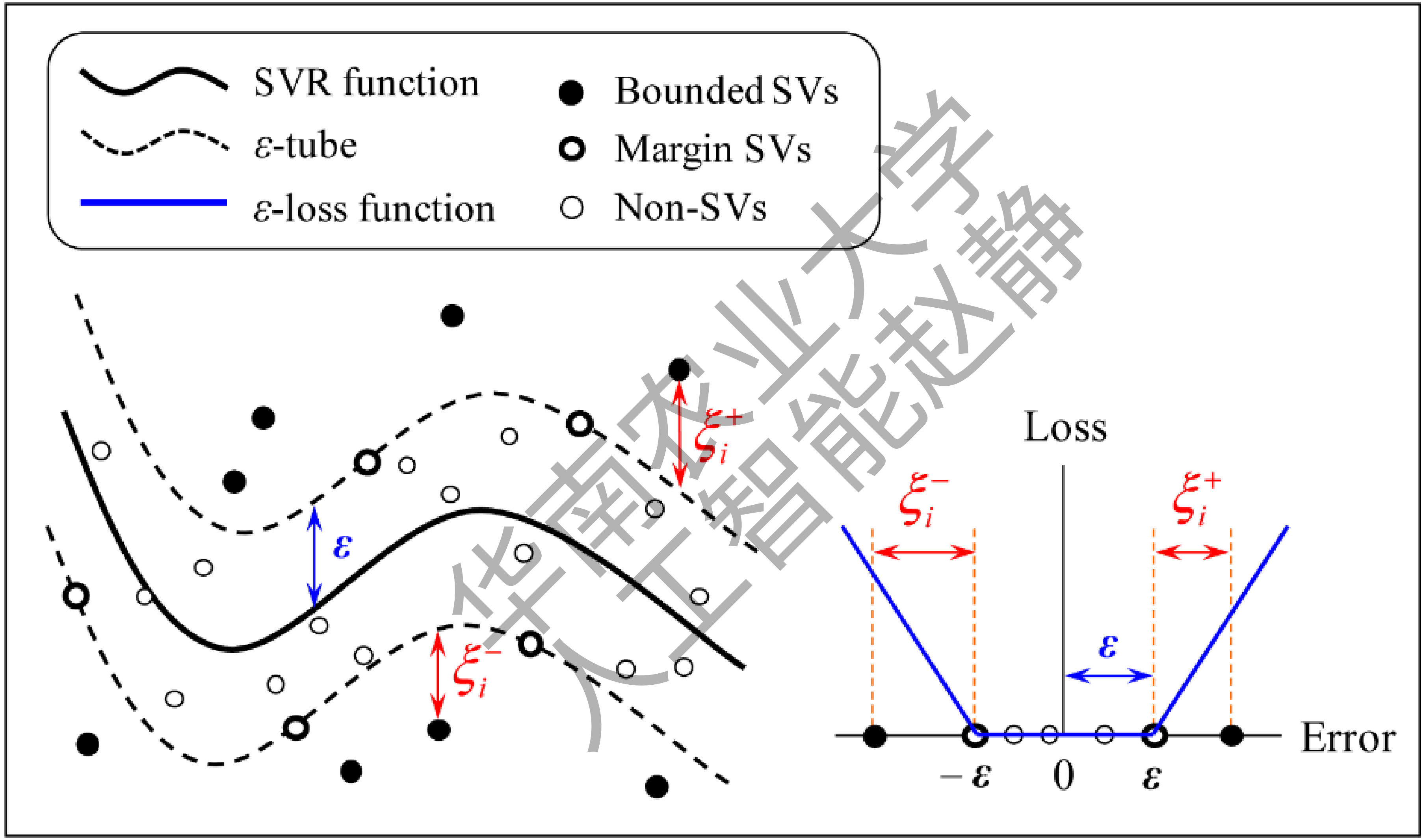
$$\text{s.t.}\ \left|y_i - \left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)\right| \le \epsilon, \quad i = 1,2,\dots,N$$

- 加入松弛变量$\xi_i \ge 0$，用于表示每个点在ϵ管道外的程度，SVR模型的目标函数变为

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\left(\xi_i^{\vee} + \xi_i^{\wedge}\right)$$

$$\text{s.t.}\ -\epsilon - \xi_i^{\vee} \le y_i - \left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \le \epsilon + \xi_i^{\wedge}, \qquad i = 1,2,\dots,N$$

$$\xi_i^{\wedge} \ge 0, \qquad \xi_i^{\vee} \ge 0, \qquad i = 1,2,\dots,N$$


SVR function
ε-tube
ε-loss function
Bounded SVs
Margin SVs
Non-SVs
ε
ξ_i^+
ξ_i^-
Loss
Error
$-\varepsilon$
0
ε

➢ 拉格朗日函数

- 与SVM类似，SVR的拉格朗日函数为：

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}^{\vee}, \boldsymbol{\alpha}^{\wedge}, \boldsymbol{\xi}^{\vee}, \boldsymbol{\xi}^{\wedge}, \boldsymbol{\mu}^{\vee}, \boldsymbol{\mu}^{\wedge}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}(\xi_i^{\vee} + \xi_i^{\wedge})$$

$$+\sum_{i=1}^{N}\alpha_i^{\vee}(-\epsilon - \xi_i^{\vee} - y_i + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)$$

$$+\sum_{i=1}^{N}\alpha_i^{\wedge}(y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - b - \epsilon - \xi_i^{\wedge})$$

$$-\sum_{i=1}^{N}\mu_i^{\vee}\xi_i^{\vee} - \sum_{i=1}^{m}\mu_i^{\wedge}\xi_i^{\wedge}$$

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\mathrm{s.t.} -\epsilon - \xi_i^{\vee} \le y_i - (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \le \epsilon + \xi_i^{\wedge}$$

$$\xi_i^{\wedge} \ge 0, \qquad \xi_i^{\vee} \ge 0$$

➢ SVR的对偶问题

- 拉格朗日函数 $L(\boldsymbol{w}, b, \boldsymbol{\alpha}^{\vee}, \boldsymbol{\alpha}^{\wedge}, \boldsymbol{\xi}^{\vee}, \boldsymbol{\xi}^{\wedge}, \boldsymbol{\mu}^{\vee}, \boldsymbol{\mu}^{\wedge})$ 对 $\mathbf{w}, b, \xi_i^{\wedge}, \xi_i^{\vee}$ 的一阶导数:

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha}^{\vee}, \boldsymbol{\alpha}^{\wedge}, \boldsymbol{\xi}^{\vee}, \boldsymbol{\xi}^{\wedge}, \boldsymbol{\mu}^{\vee}, \boldsymbol{\mu}^{\wedge})}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N} (\alpha_i^{\wedge} - \alpha_i^{\vee})\, \boldsymbol{x}_i$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha}^{\vee}, \boldsymbol{\alpha}^{\wedge}, \boldsymbol{\xi}^{\vee}, \boldsymbol{\xi}^{\wedge}, \boldsymbol{\mu}^{\vee}, \boldsymbol{\mu}^{\wedge})}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} (\alpha_i^{\wedge} - \alpha_i^{\vee}) = 0$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha}^{\vee}, \boldsymbol{\alpha}^{\wedge}, \boldsymbol{\xi}^{\vee}, \boldsymbol{\xi}^{\wedge}, \boldsymbol{\mu}^{\vee}, \boldsymbol{\mu}^{\wedge})}{\partial \xi_i^{\wedge}} = 0 \Rightarrow C - \alpha_i^{\wedge} + \mu_i^{\wedge}$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha}^{\vee}, \boldsymbol{\alpha}^{\wedge}, \boldsymbol{\xi}^{\vee}, \boldsymbol{\xi}^{\wedge}, \boldsymbol{\mu}^{\vee}, \boldsymbol{\mu}^{\wedge})}{\partial \xi_i^{\vee}} = 0 \Rightarrow C - \alpha_i^{\vee} + \mu_i^{\vee}$$

- 将上述结论代入拉格朗日函数 $L(\boldsymbol{w}, b, \boldsymbol{\alpha}^{\vee}, \boldsymbol{\alpha}^{\wedge}, \boldsymbol{\xi}^{\vee}, \boldsymbol{\xi}^{\wedge}, \boldsymbol{\mu}^{\vee}, \boldsymbol{\mu}^{\wedge})$， 得到对偶问题：

$$\max_{\boldsymbol{\alpha}^{\vee},\boldsymbol{\alpha}^{\wedge}} \left(-\sum_{i=1}^{N} \left((\epsilon - y_i)\alpha_i^{\wedge} + (\epsilon + y_i)\alpha_i^{\vee} \right) - \sum_{i=1}^{N} \frac{1}{2}(\alpha_i^{\wedge} - \alpha_i^{\vee})(\alpha_j^{\wedge} - \alpha_j^{\vee})\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \right)$$

$$\mathrm{s.t.} \sum_{i=1}^{N} (\alpha_i^{\wedge} - \alpha_i^{\vee}) = 0$$

$$0 \le \alpha_i^{\wedge} \le C, \qquad i = 1,2,\ldots,N$$

$$0 \le \alpha_i^{\vee} \le C, \qquad i = 1,2,\ldots,N$$

- 去掉负号，将上述目标函数中的max换成min，得到等价问题：

$$\min_{\boldsymbol{\alpha}^{\vee},\boldsymbol{\alpha}^{\wedge}} \left(\sum_{i=1}^{N} \left((\epsilon - y_i)\alpha_i^{\wedge} + (\epsilon + y_i)\alpha_i^{\vee} \right) + \sum_{i=1}^{N} \frac{1}{2}(\alpha_i^{\wedge} - \alpha_i^{\vee})(\alpha_j^{\wedge} - \alpha_j^{\vee})\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \right)$$

➢ SVR模型

- 求得$\boldsymbol{\alpha}$的最优值后，可计算$\boldsymbol{w}, b$，从而得到SVR模型

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = \sum_{i=1}^{N}(\alpha_i^{\wedge} - \alpha_i^{\vee})\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x} + b$$
$$= \sum_{i=1}^{N}(\alpha_i^{\wedge} - \alpha_i^{\vee})\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + b$$

- $\boldsymbol{x}$ 与训练数据的点积$\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle$换成核函数$k(\boldsymbol{x}_i, \boldsymbol{x})$，得到核化SVR模型

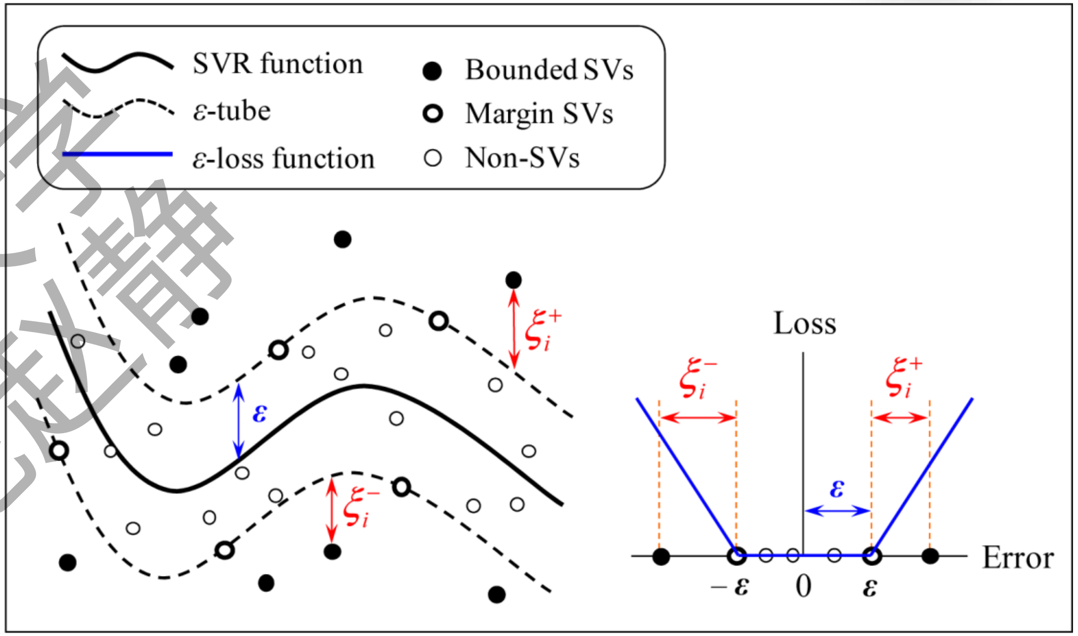
$$f(\boldsymbol{x}) = \sum_{i=1}^{N}(\alpha_i^{\wedge} - \alpha_i^{\vee})k(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

➢ SVR的支持向量

- 根据KKT条件：

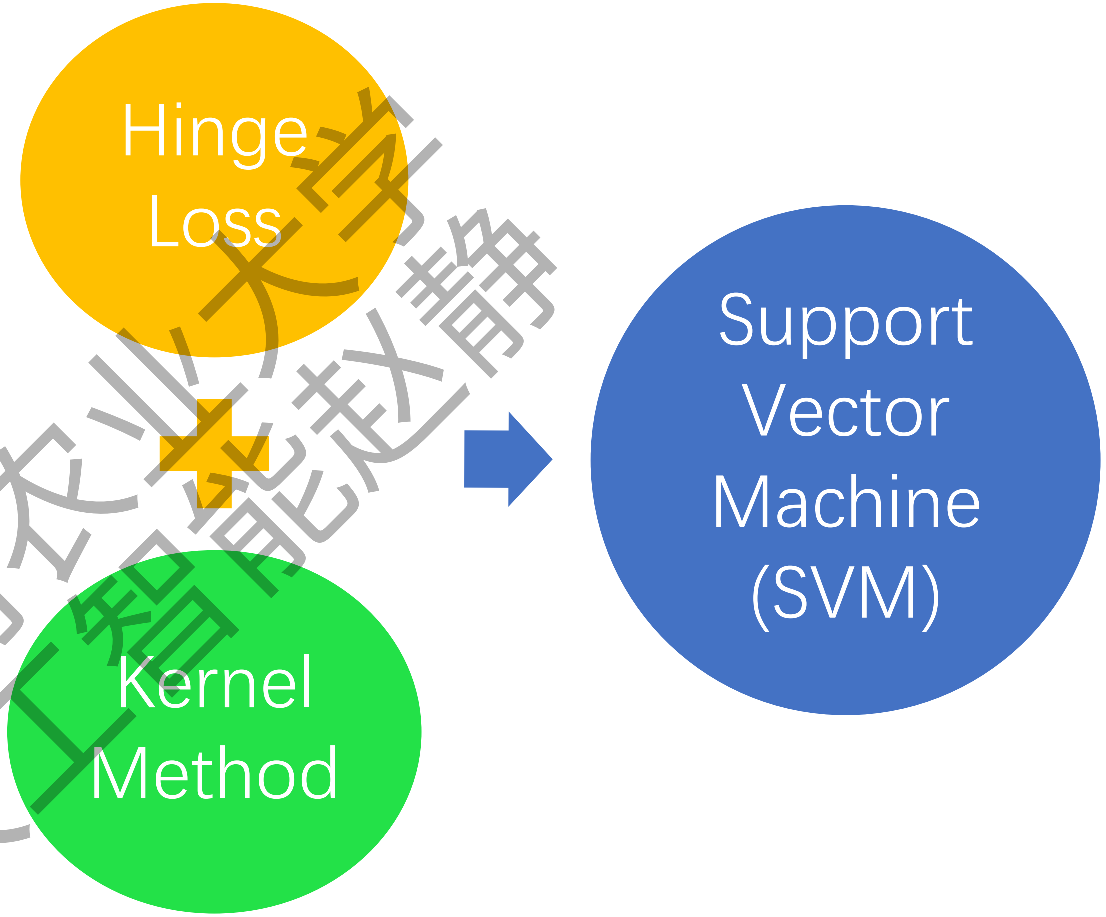
$$\alpha_i^{\vee}\left(\epsilon+\xi_i^{\vee}+y_i-\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\right)=0$$

$$\alpha_i^{\wedge}\left(\epsilon+\xi_i^{\wedge}-y_i+\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\right)=0$$



支持向量：

仅当样本$(\boldsymbol{x}_i, y_i)$落在ϵ间隔中，$\alpha_i^{\vee}$、$\alpha_i^{\wedge}$才取非0值。


Hinge Loss
Kernel Method
Support Vector Machine (SVM)