SVM

何博士 慧科AI专家讲师

回顾LR

SVM

概念：

支持向量机其基本原理，通俗来讲，它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略便是间隔最大化线性分类器使用超平面类型的边界，非线性分类器使用超曲面。

数据： 线性可分&线性不可分

SVM两种情况

- 线性可分
- 线性不可分

情况1：样本本质上是非线性可分的

解决方法：核函数

情况2：本质上线性，非线性由噪音导致

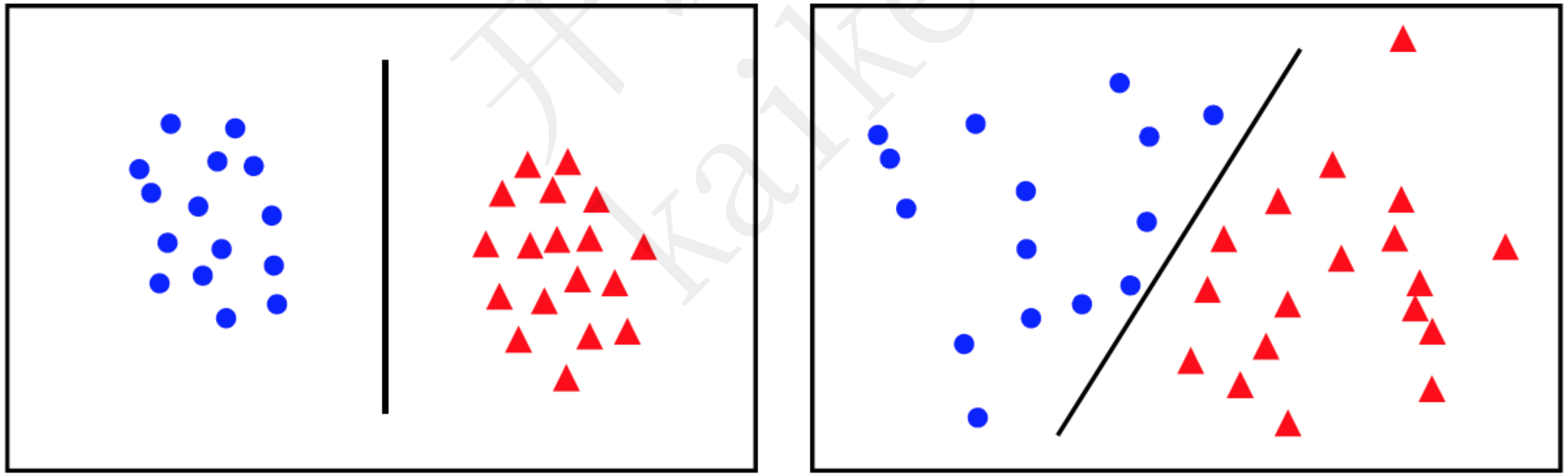
强制使用非线性函数，会导致过拟合

解决方法：软间隔

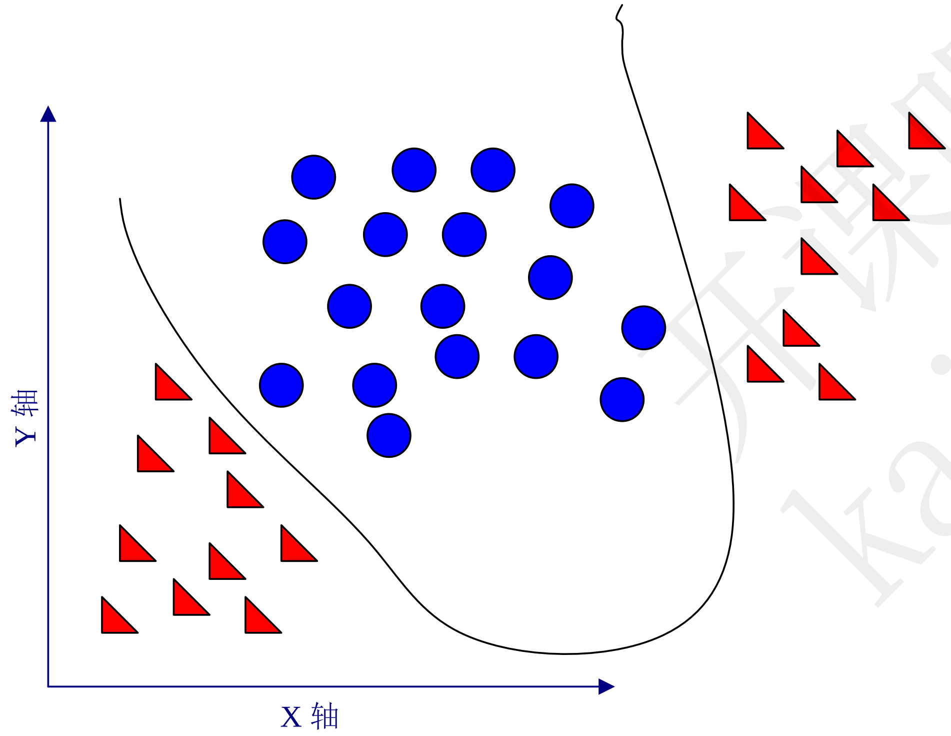
线性可分

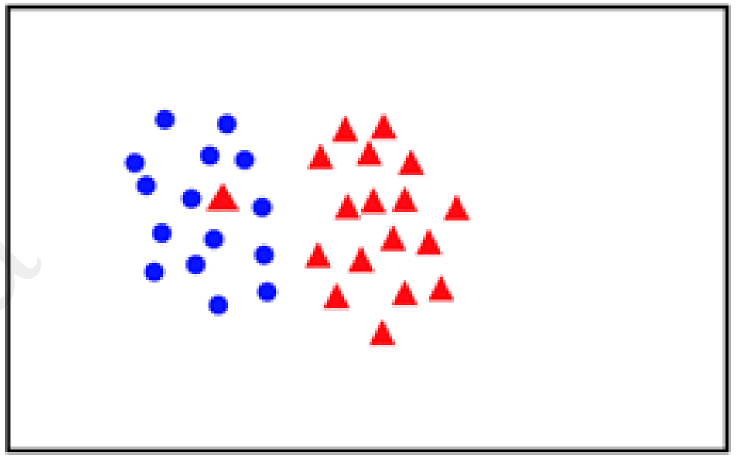
- **定义：**

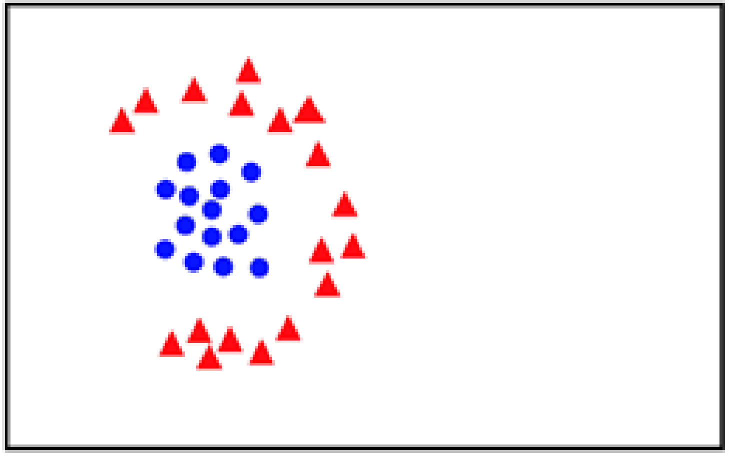
对于来自两类的一组模式能用一个线性判别函数正确分类,则称他们是线性可分的。



线性不可分



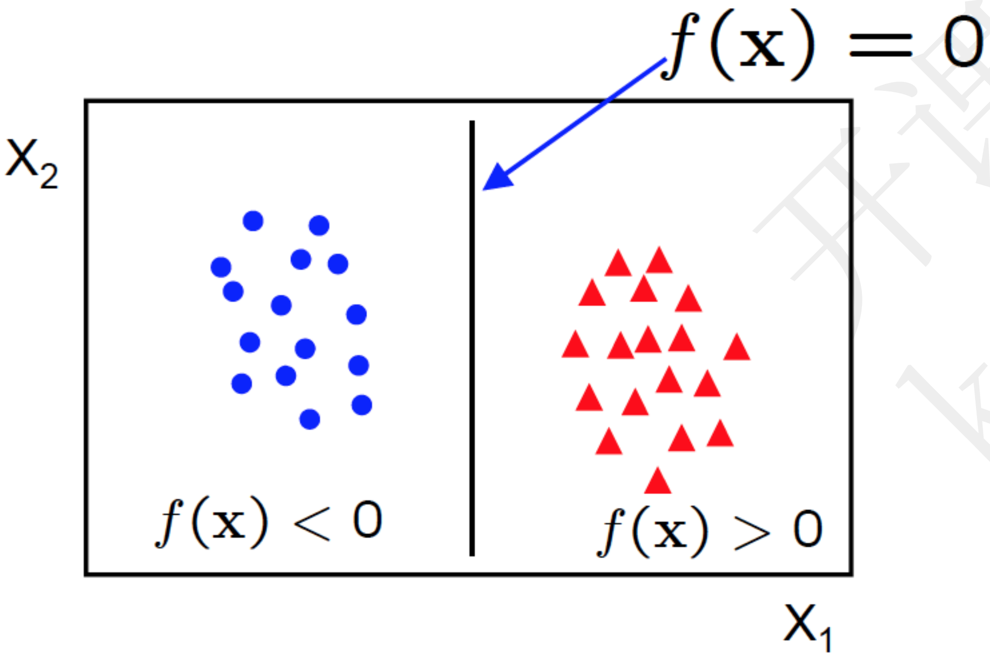




线性可分情况

线性可分情况

- 我们怎样才能取得一个最优的划分直线f(x)呢？



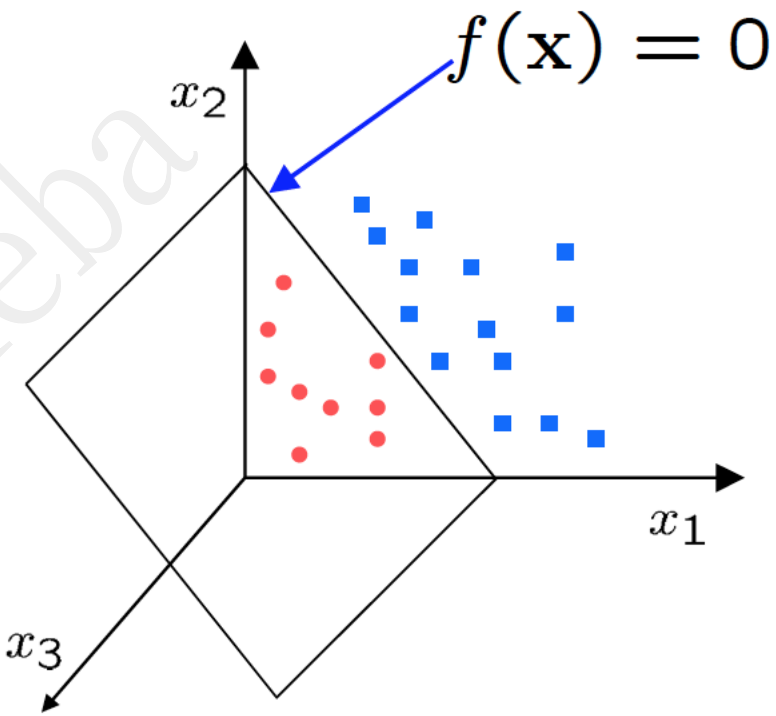
$$f(\mathbf{x}) = \mathbf{w}^{\top}\mathbf{x} + b$$

W权重向量，需要学习的参数
b为偏差，需要学习的参数

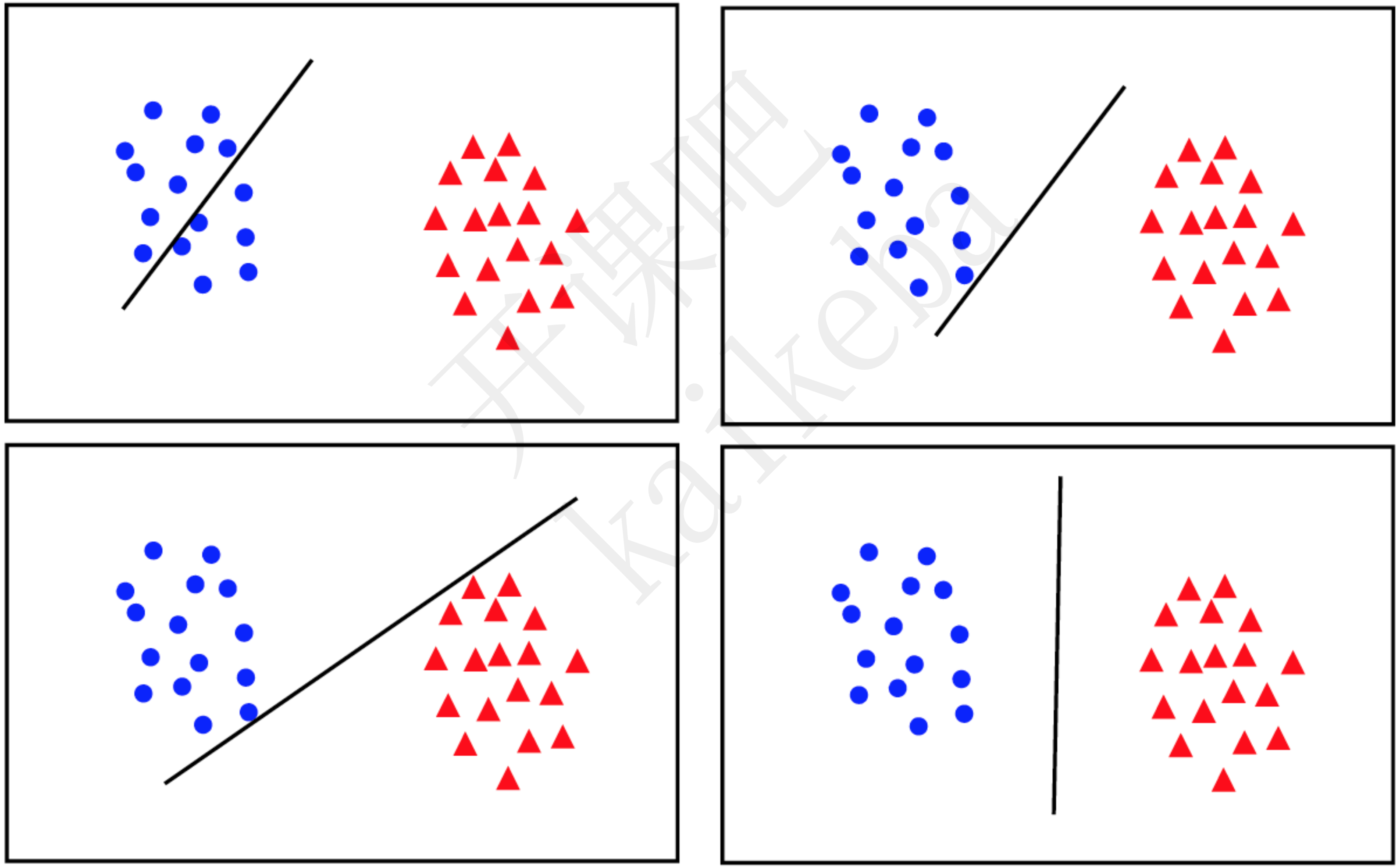
线性可分情况

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

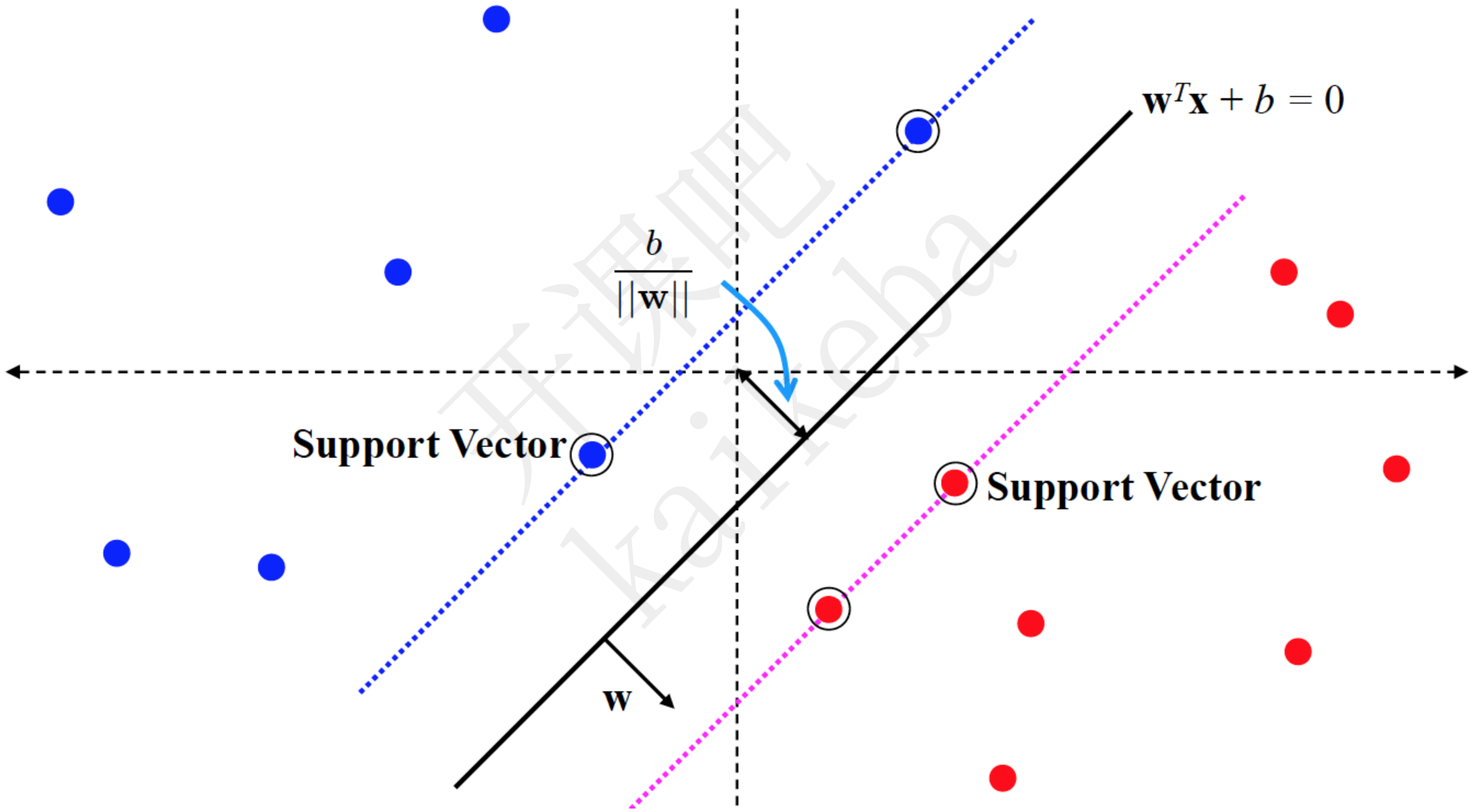
多维超平面



最大间隔



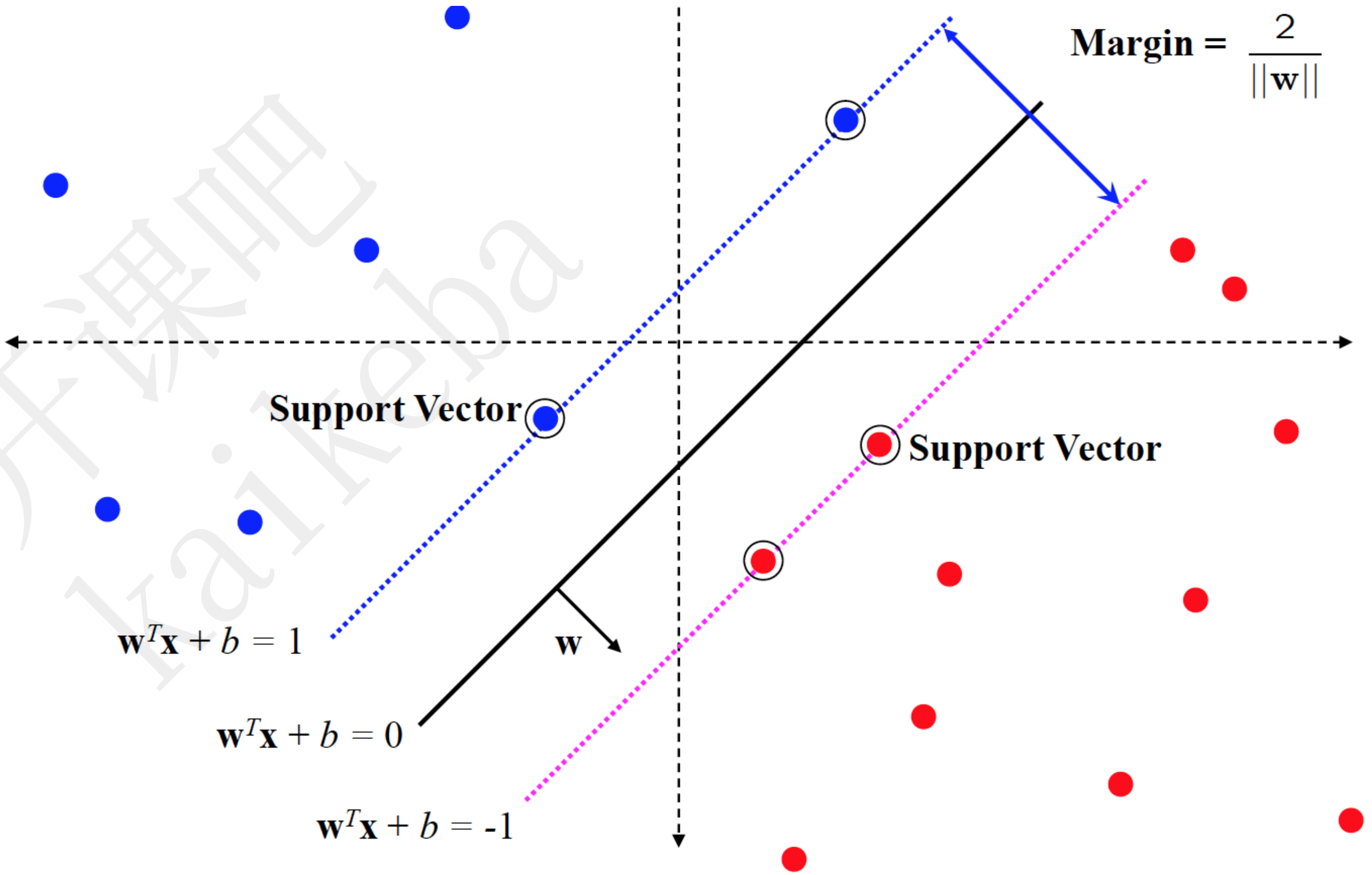
最大间隔


$\mathbf{w}^T\mathbf{x} + b = 0$
$\frac{b}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}$

最大间隔

y取{-1，+1}

$$\hat{\gamma} = y(w^T x + b) = yf(x)$$



几何间隔

- 点到直线：

$$\tilde{\gamma} = \frac{\hat{\gamma}}{\|w\|} = \frac{w^T x}{\|w\|} + \frac{b}{\|w\|} = \frac{f(x)}{\|w\|}$$

- 这时如果成比例的改变w和b，几何间隔的值不会发生改变。

最大间隔分类的目标函数

因此，我们的最大间隔分类的目标函数可以定义为：

$$\max \frac{\hat{\gamma}}{\|w\|}$$

$$s.t. \; y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}$$

我们改变优化问题的表述方式。

此时，优化问题的表达式为：

$$\min \frac{1}{2}\|w\|^2$$
$$s.t. \; y^{(i)}(w^T x^{(i)} + b) \geq 1$$

我们的优化问题转变成了一个凸优化问题

拉格朗日乘子法

- minf(w)
- s.t. gi(w)≤0 i=1,2,...,k

  hi(w)=0 i=1,2,...,l （这里0指的是零向量）

$$L(w,\alpha,\beta)=f(w)+\sum_{i=1}^{k}\alpha_i g_i(w)+\sum_{i=1}^{l}\beta_i h_i(w)$$

定义：

$$\theta_p(w)=\max_{\alpha_i\ge 0} L(w,\alpha,\beta)$$

当所有约束条件都满足时有 $\theta_p = f(w)$

$$\mathcal{L}(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i\left(y_i(w^T x_i+b)-1\right)$$

对偶问题

$$p^* = \min_{w,b} \theta(w) = \min_{w,b} \max_{\alpha_i \geq 0} L(w,b,a)$$

$$d^* = \max_{\alpha_i \geq 0} \min_{w,b} L(w,b,a)$$

一般有 $d^* \leq p^*$ ，但是在某些特定条件下（KKT），这两个最优化问题会取相同的值。

对偶问题

1.首先固定α，要让L关于w和b最小化，我们分别对w,b偏导并令其等于零，得到

$$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b}=0\Rightarrow \sum_{i=1}^{n}\alpha_i y_i=0$$

带回 $L(w,b,a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i(y_i(w^T x_i+b)-1)$得到：

$$\begin{aligned}\mathcal{L}(w,b,\alpha)&=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j-\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j-b\sum_{i=1}^{n}\alpha_i y_i+\sum_{i=1}^{n}\alpha_i\\&=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j\end{aligned}$$

对偶问题

问题转换为：

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$s.t.,\ \alpha_i\geq 0, i=1,\ldots,n$$

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

由凸二次规划的性质能保证这样最优的向量a是存在的

对偶问题

- 2.求对α的极大，即是关于对偶变量的优化问题（SMO优化算法--序列最小最优化算法）
- 然后根据

$$w^* = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$b^* = -\frac{\max_{i:y_i=-1} w^{*^T} x_i + \min_{i:y_i=1} w^{*^T} x_i}{2}$$

- 可求出最优的w和b，即最优超平面。

习题

例 7.1 数据与例 2.1 相同．已知一个如图 7.4 所示的训练数据集，其正例点是 $x_1=(3,3)^{\mathrm{T}}$， $x_2=(4,3)^{\mathrm{T}}$，负例点是 $x_3=(1,1)^{\mathrm{T}}$，试求最大间隔分离超平面．

图 7.4 间隔最大分离超平面示例

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$s.t.,\ \alpha_i\geq 0, i=1,\ldots,n$$

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

解 按照算法 7.1，根据训练数据集构造约束最优化问题：

$$\min_{w,b}\ \frac{1}{2}(w_1^2+w_2^2)$$

$$\text{s.t.}\quad 3w_1+3w_2+b\geqslant 1$$

$$4w_1+3w_2+b\geqslant 1$$

$$-w_1-w_2-b\geqslant 1$$

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

$$\alpha_1+\alpha_2-\alpha_3=0$$

$$\alpha_i\geq 0,\quad i=1,2,3$$

习题

将 $\alpha_1+\alpha_2=\alpha_3$

带入目标函数，得到关于α_1，α_2的函数：

$$-4\alpha_1^2-\frac{13}{2}\alpha_2^2-10\alpha_1\alpha_2+2\alpha_1+2\alpha_2$$

$$\alpha_i\geq 0,\quad i=1,2,3$$

$\alpha_1=1/4$

$\alpha_2=0$

$\alpha_3=1/4$

习题

代入公式：

$$w^* = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$b^* = -\frac{\max\limits_{i:y_i=-1} w^{*T} x_i + \min\limits_{i:y_i=1} w^{*T} x_i}{2}$$

求得此最优化问题的解 $w_1 = w_2 = \frac{1}{2}$，$b = -2$．于是最大间隔分离超平面为

$$\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$$

其中，$x_1 = (3,3)^T$ 与 $x_3 = (1,1)^T$ 为支持向量．

最大间隔分类的目标函数

首先给出形式化的不等式约束优化问题：

$$
\begin{aligned}
&\min_x \ f(x) \\
&s.t. \quad h_i(x) = 0, \ i = 1, 2, \ldots, m \\
&\qquad\quad g_j(x) \le 0, \ j = 1, 2, \ldots, n
\end{aligned}
$$

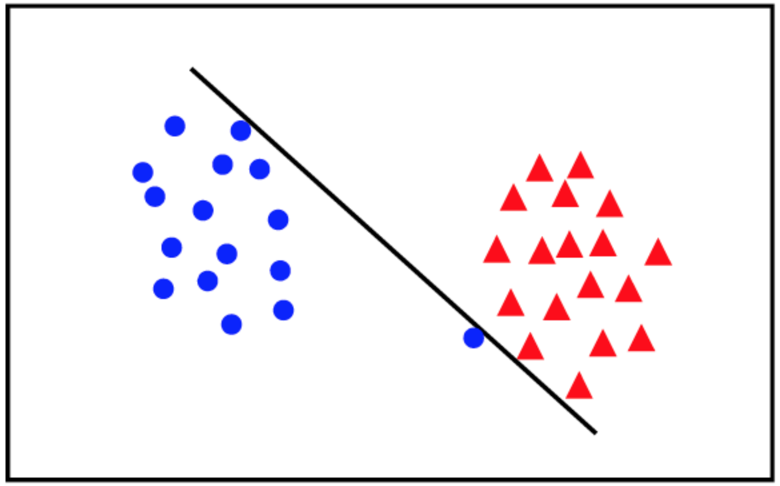
列出 Lagrangian 得到无约束优化问题：

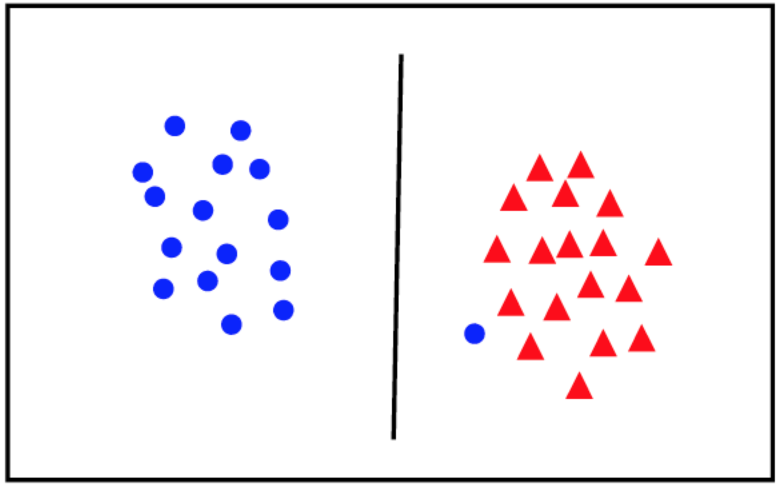
$$
L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m} \alpha_i h_i(x) + \sum_{j=1}^{n} \beta_i g_i(x)
$$

经过之前的分析，便得知加上不等式约束后可行解 x 需要满足的就是以下的 KKT 条件：

$$
\begin{aligned}
\nabla_x L(x, \alpha, \beta) &= 0 && (1) \\
\beta_j g_j(x) &= 0, \ j = 1, 2, \ldots, n && (2) \\
h_i(x) &= 0, \ i = 1, 2, \ldots, m && (3) \\
g_j(x) &\le 0, \ j = 1, 2, \ldots, n && (4) \\
\beta_j &\ge 0, \ j = 1, 2, \ldots, n && (5)
\end{aligned}
$$

软间隔





哪个划分更好？

允许一定程度犯错

$$\min \frac{1}{2}\|w\|^2$$

$$s.t.\; y^{(i)}(w^T x^{(i)} + b) \geq 1$$

开课吧
kaikeba.com
慧科集团旗下企业

软间隔

$$\min_{\mathbf{w}\in\mathbb{R}^d,\xi_i\in\mathbb{R}+} ||\mathbf{w}||^2 + C\sum_{i}^{N}\xi_i$$

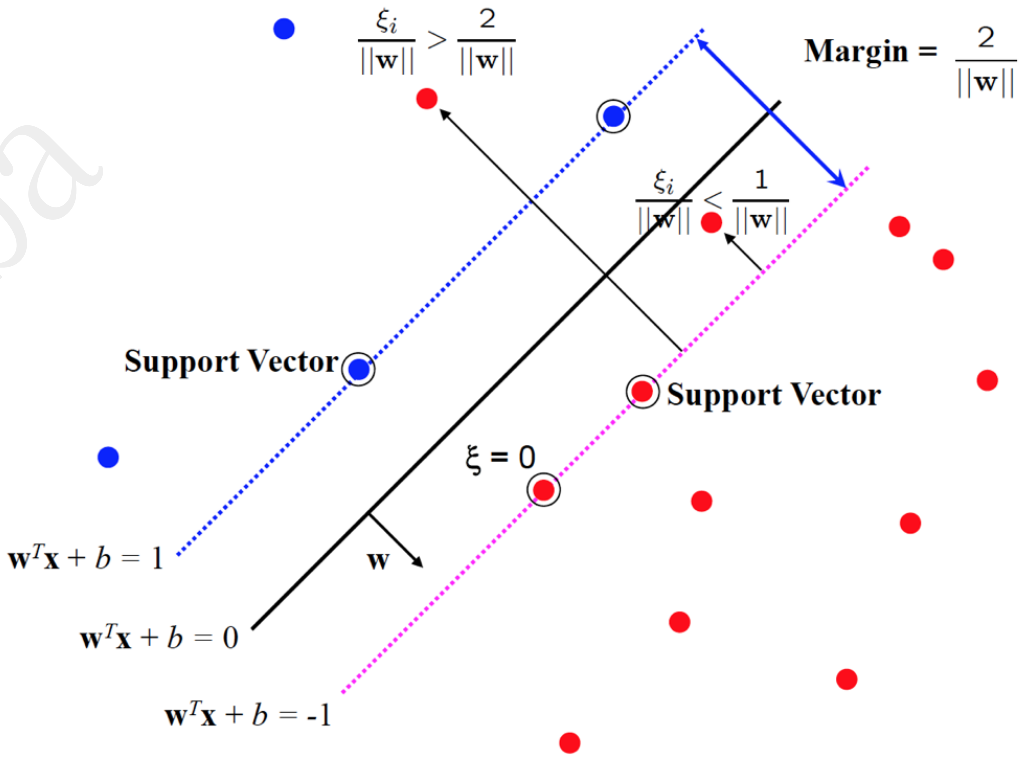
$$y_i\left(\mathbf{w}^\top\mathbf{x}_i + b\right) \geq 1 - \xi_i \text{ for } i = 1\ldots N$$

$$\xi_i \geq 0$$

$$y_i f(\mathbf{x}_i) \geq 1 - \xi_i$$

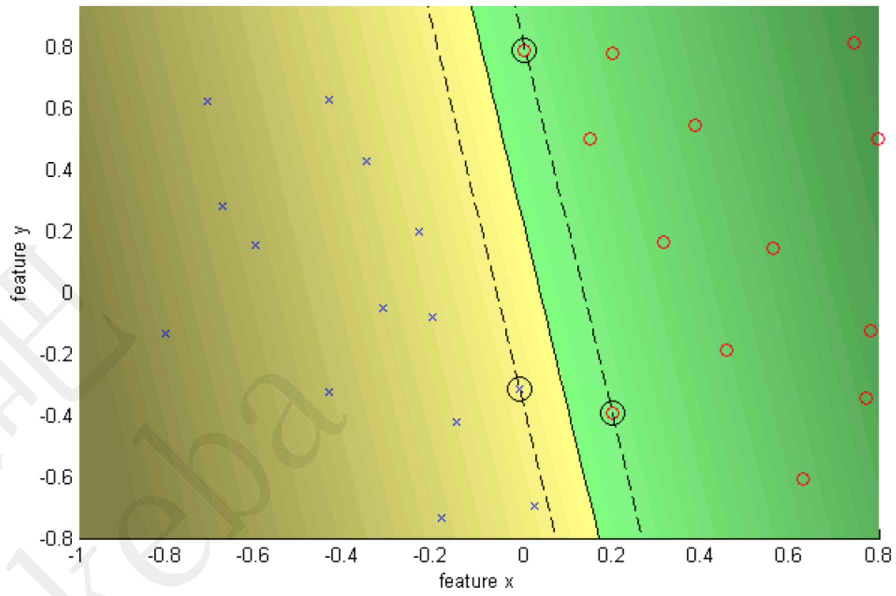
$$\xi_i = \max\left(0, 1 - y_i f(\mathbf{x}_i)\right)$$

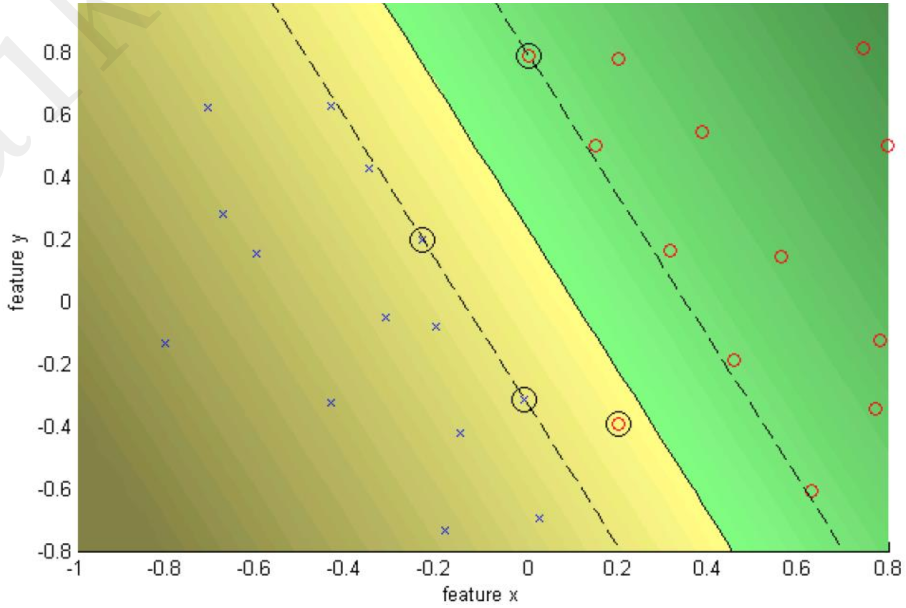
$$\min_{\mathbf{w}\in\mathbb{R}^d} \underbrace{||\mathbf{w}||^2}_{\text{regularization}} + C\sum_{i}^{N}\underbrace{\max\left(0, 1 - y_i f(\mathbf{x}_i)\right)}_{\text{loss function}}$$



C = Infinity   hard margin

C的取值和带宽





线性不可分情况下

- 根据线性可分情况下的结论：

$$w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$$

- 将分类函数变形得最终分类函数，为：

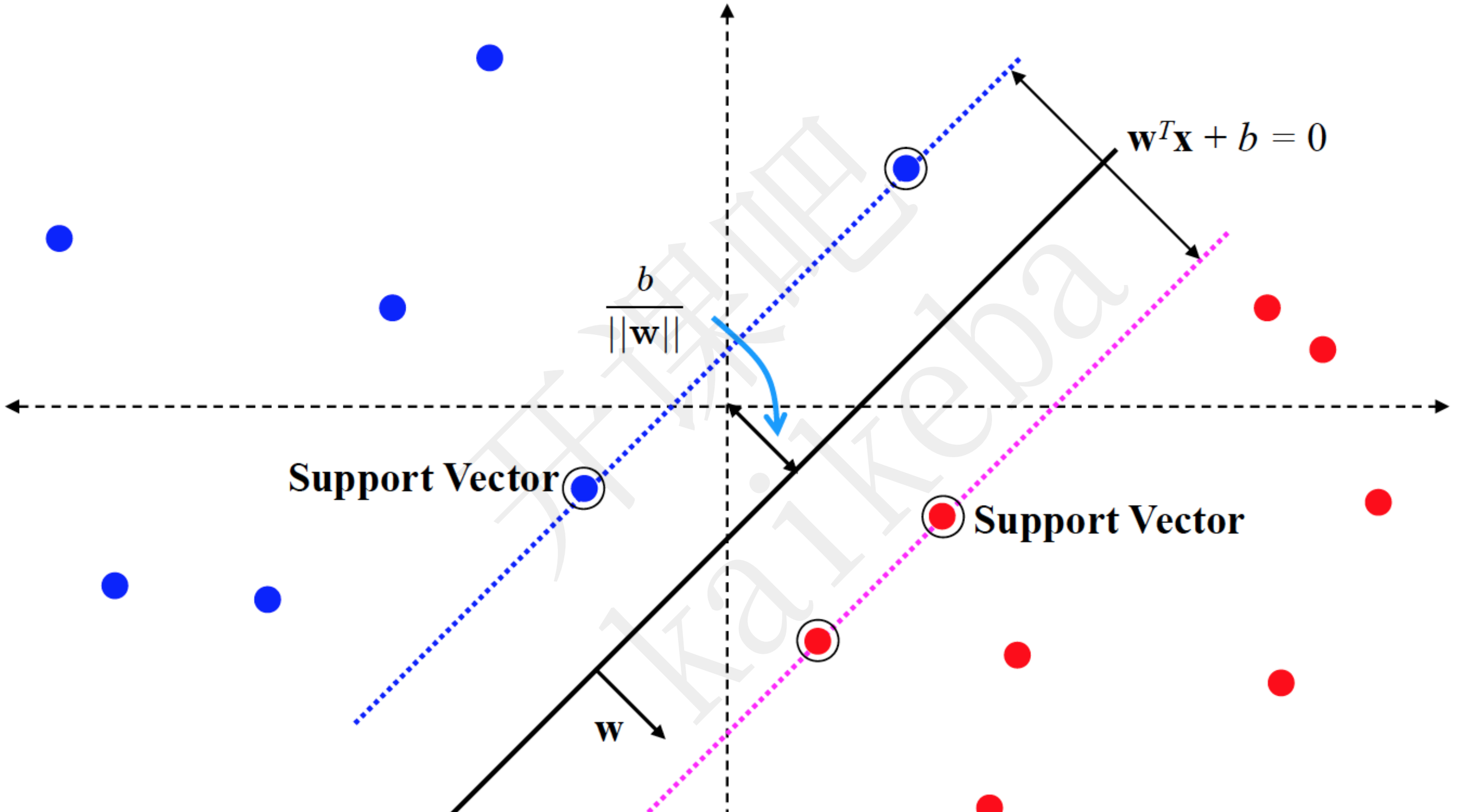
$$f(x) = \left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j \right)^{\top} \mathbf{x} + b = \sum_{j=1}^{N} \alpha_j y_j \left(\mathbf{x}_j{}^{\top} \mathbf{x} \right) + b$$

• 损失函数：

$$||\mathbf{w}||^2 = \left\{\sum_j \alpha_j y_j \mathbf{x}_j\right\}^\top \left\{\sum_k \alpha_k y_k \mathbf{x}_k\right\} = \sum_{jk} \alpha_j \alpha_k y_j y_k (\mathbf{x}_j{}^\top \mathbf{x}_k)$$

• 约束条件：

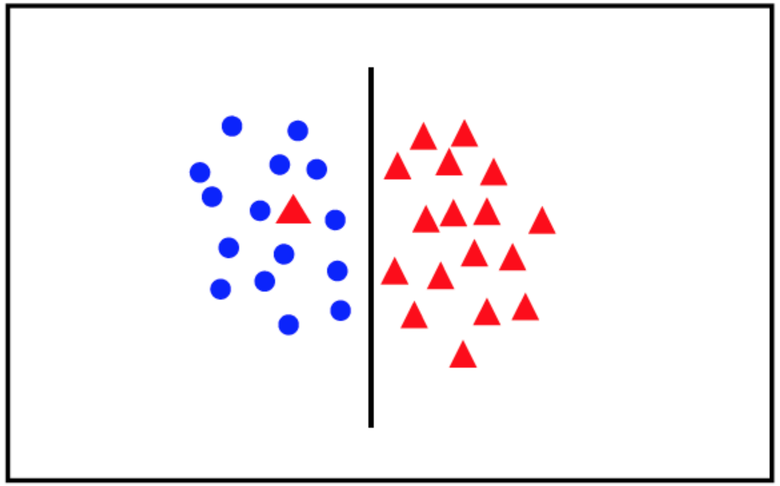
$$y_i \left(\sum_{j=1}^{N} \alpha_j y_j (\mathbf{x}_j{}^\top \mathbf{x}_i) + b\right) \geq 1$$



$$f(\mathbf{x}) = \sum_i \alpha_i y_i (\mathbf{x}_i{}^\top \mathbf{x}) + b$$

support vectors

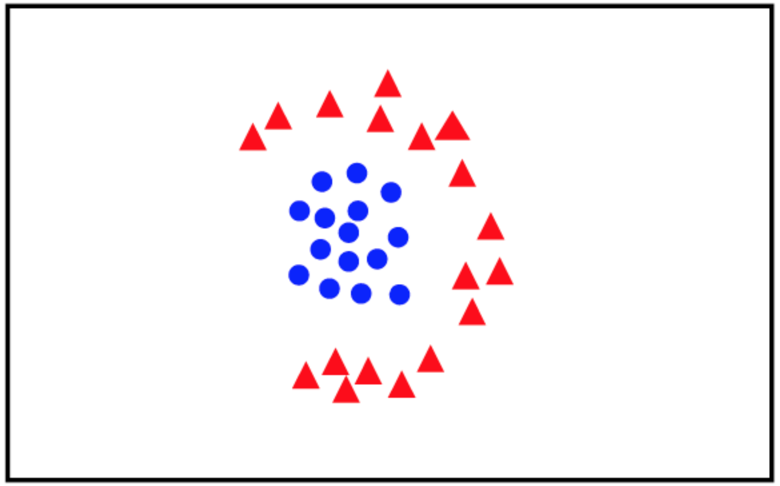
软件隔



$$\min_{\mathbf{w}\in\mathbb{R}^d,\xi_i\in\mathbb{R}^+} ||\mathbf{w}||^2 + C\sum_i^N \xi_i$$

subject to

$$y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i \text{ for } i = 1 \ldots N$$

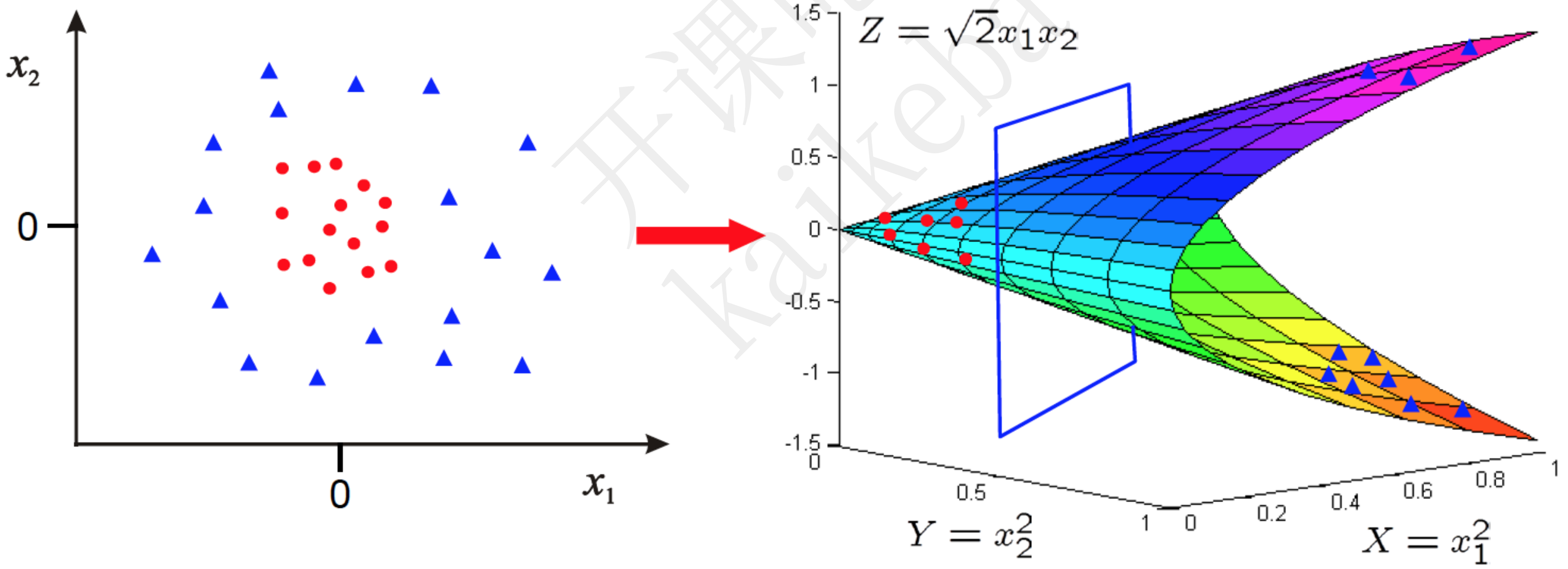


这类问题：

如何解决？？

$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \qquad \mathbb{R}^2 \to \mathbb{R}^3$$

(0,0)
(1,-1)
(1,1)
(-1,1)
(-1,-1)



- 分类问题：

$$f(\mathbf{x}) = \sum_{i}^{N} \alpha_i y_i \mathbf{x}_i{}^\top \mathbf{x} + b$$

$$\to f(\mathbf{x}) = \sum_{i}^{N} \alpha_i y_i \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}) + b$$

- 学习问题：

$$\max_{\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{jk} \alpha_j \alpha_k y_j y_k \mathbf{x}_j{}^\top \mathbf{x}_k$$

$$\to \max_{\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{jk} \alpha_j \alpha_k y_j y_k \Phi(\mathbf{x}_j)^\top \Phi(\mathbf{x}_k)$$

原来在二维空间中一个线性不可分的问题，映射到高维空间后，变成了线性可分的。因此，这也形成了我们最初想解决线性不可分问题的基本思路---向高维空间转化，使其变得线性可分。

而转化的关键的部分在于找到x到y的映射方法。

如何找到这个映射没有系统的方法，此外，在数据维度较大时，计算困难。

维度爆炸问题：

$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \quad \mathbb{R}^2 \to \mathbb{R}^3$$

$$\begin{aligned} \Phi(\mathbf{x})^\top \Phi(\mathbf{z}) &= \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right) \begin{pmatrix} z_1^2 \\ z_2^2 \\ \sqrt{2}z_1z_2 \end{pmatrix} \\ &= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \end{aligned}$$

计算次数：11次乘法和2次加法

核函数

计算次数：3次乘法和1次加法

核函数：对所有x,z属于X，满足

$$k(x,z)=\langle\phi(x)\cdot\phi(z)\rangle$$

$$\begin{aligned}\Phi(\mathbf{x}^\top\mathbf{z}) &= (\mathbf{x}^\top\mathbf{z})^2\\ &= (x_1z_1+x_2z_2)^2\\ &= x_1^2z_1^2+x_2^2z_2^2+2x_1x_2z_1z_2\end{aligned}$$

可以在特征空间中直接计算内积〈φ(xi)· φ(x)〉，就像在原始输入点的函数中一样

核函数

- 分类函数为：

$$\sum_{i=1}^{n} \alpha_i y_i \kappa(x_i, x) + b$$

- 优化问题的表达式：

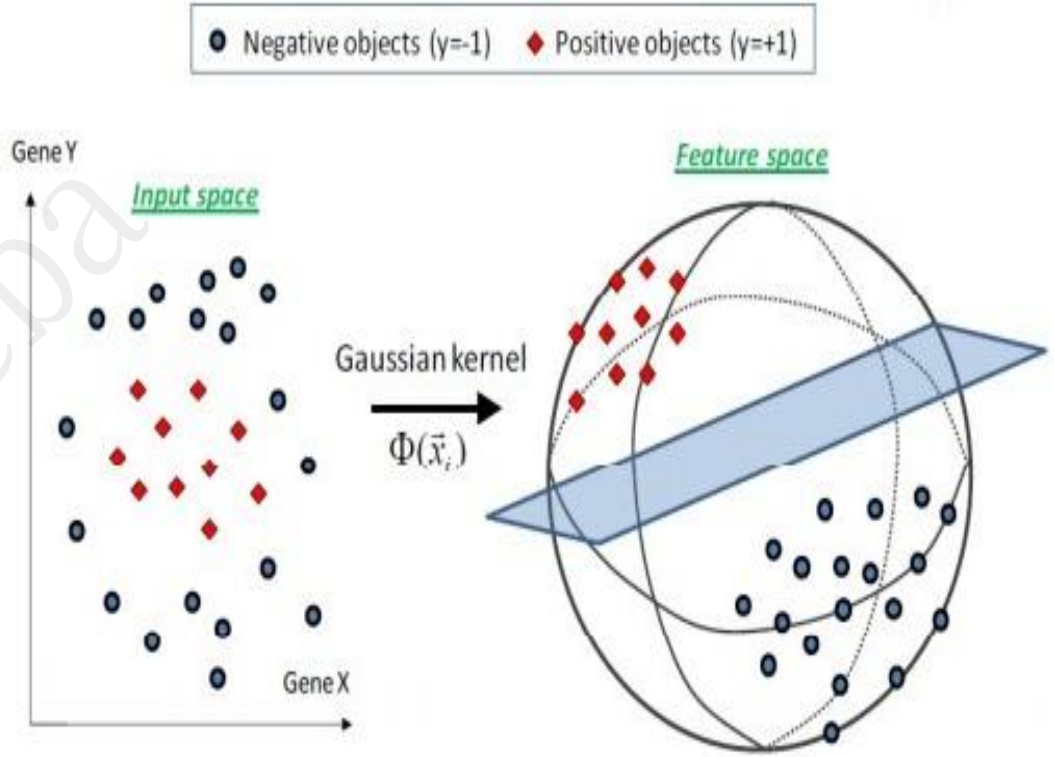
$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j)$$

$$s.t., \alpha_i \geq 0, i = 1, \ldots, n$$

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

常见核函数

- 多项式核 $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + \mathrm{c})^d$
- 线性核 $k(x, y) = \langle x \cdot y \rangle$
- 高斯径向基函数核 $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\,\sigma^2}\right)$
- **Sigmoid核** $k(\mathbf{x}, \mathbf{y}) = \tanh(\kappa(\mathbf{x} \cdot \mathbf{y}) + \Theta)$



核函数

对于核函数的选择，现在还缺乏指导原则。各种实验的观察结果表明，某些问题用某些核函数效果很好，用另一些很差，但一般来讲，径向基核函数是不会出现太大偏差的一种，一般作为首选。

无穷维-高斯核函数

$$k(x, y) = \exp(-\|x - y\|^2)$$
$$= \exp(-(x_1 - y_1)^2 - (x_2 - y_2)^2)$$
$$= \exp(-x_1^2 + 2x_1y_1 - y_1^2 - x_2^2 + 2x_2y_2 - y_2^2)$$
$$= \exp(-\|x\|^2)\exp(-\|y\|^2)\exp(2x^Ty)$$

$$k(x, y) = \exp(-\|x\|^2)\exp(-\|y\|^2)\sum_{n=0}^{\infty}\frac{(2x^Ty)^n}{n!}$$

高斯核函数的理解