MACHINE LEARNING

机器学习

Support Vector Machine

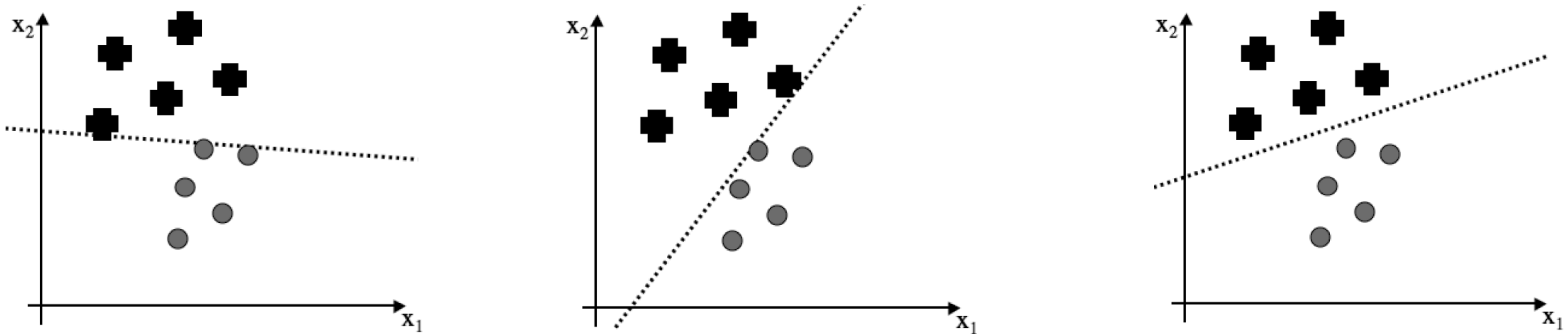
支持向量机

参考：
《机器学习》

Machine Learning Course

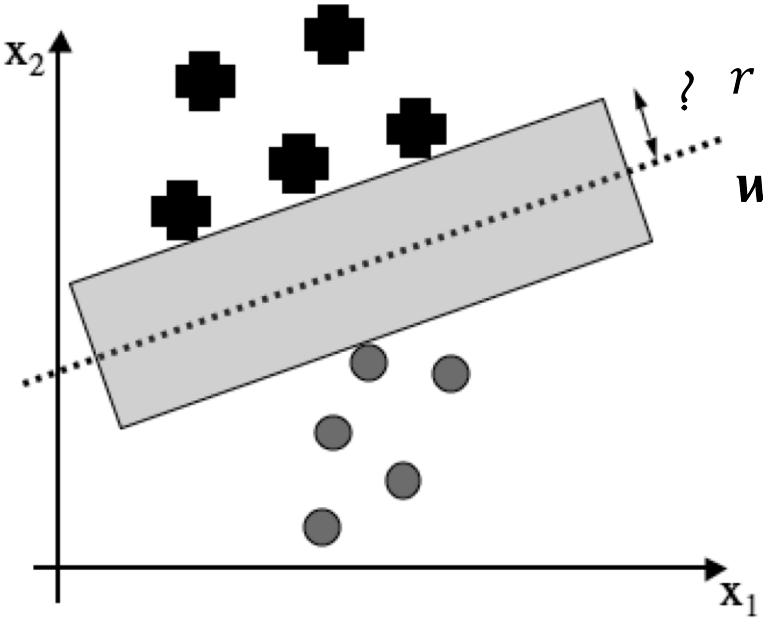
间隔与线性可分支持向量机

- 分类超平面的优劣



间隔与线性可分支持向量机

- 间隔

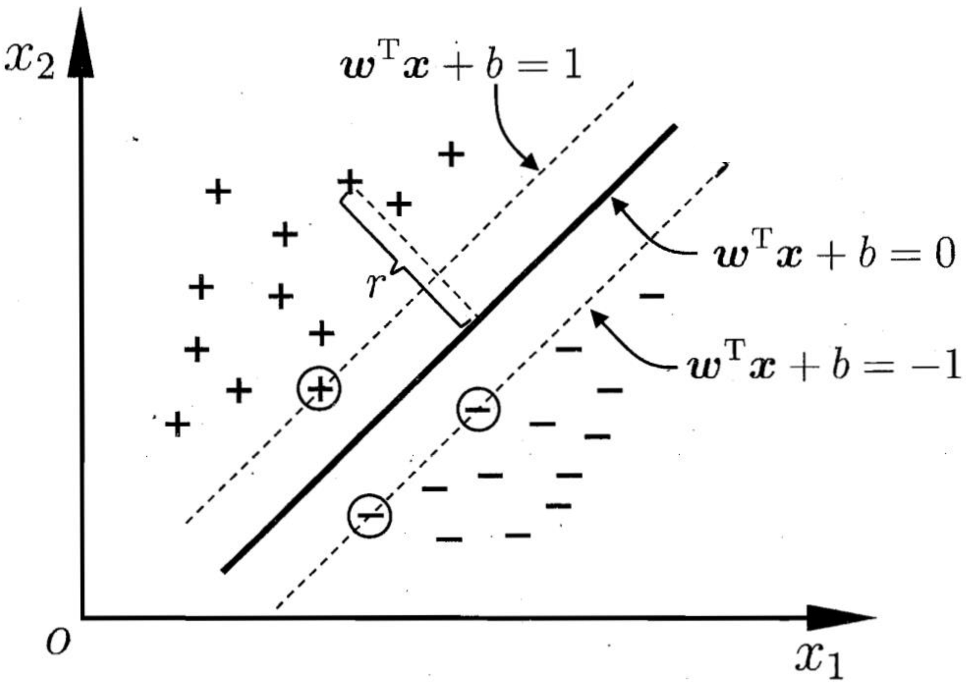


$$r = min\frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b|}{\|\boldsymbol{w}\|}$$

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

间隔与线性可分支持向量机

- 间隔



$$r = min\frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b|}{\|\boldsymbol{w}\|} \longrightarrow r=\frac{1}{\|\boldsymbol{w}\|}$$

分类全分对：$\begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b \geqslant 0, & y_i=+1 \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b \leqslant 0, & y_i=-1 \end{cases}$

$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b) \geqslant 0$

$|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b| = y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)$

约束 $\boldsymbol{w}$ 刚好满足：

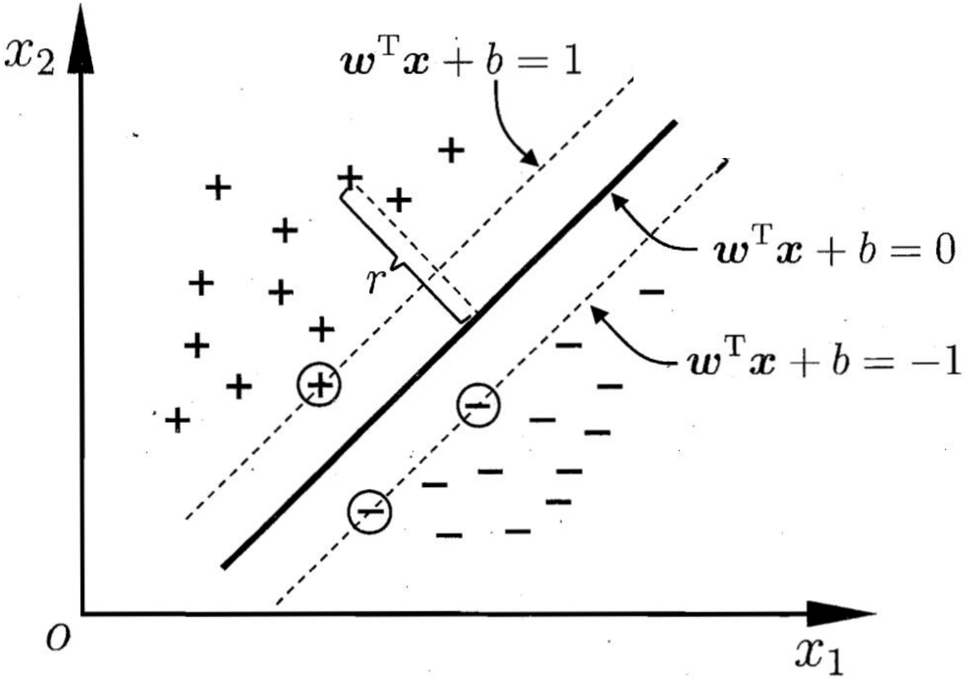
$y_s(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_s+b)=1$

$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b) \geqslant 1$

离分类超平面最近的点

间隔与线性可分支持向量机

- 间隔



$$\max_{\boldsymbol{w},b} \frac{1}{\|\boldsymbol{w}\|}$$

$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b) \geqslant 1, i=1,2,\ldots,m$$

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

$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b) \geqslant 1, i=1,2,\ldots,m$$

间隔与线性可分支持向量机

- 对偶问题

原问题

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

$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1, i = 1,2,\ldots,m$$

对偶问题 $L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + \sum_{i=1}^{m} \alpha_i(1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b))$

令L对w和b的偏导为零：$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, 0 = \sum_{i=1}^{m} \alpha_i y_i$

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

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

$$\alpha_i \geqslant 0, i = 1,2,\ldots,m$$

解出最优的α后

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

间隔与线性可分支持向量机

- 求解技术

> 二次规划 quadratic programming (QP)

> 序列最小化优化 Platt‘s sequential minimal optimization (SMO)

> 次梯度 Sub-gradient descent, 坐标下降法coordinate descent

- 工具包

```
from sklearn.svm import SVC
from sklearn.linear_model import SGDClassifier
from sklearn.svm import LinearSVC
```

间隔与线性可分支持向量机

- Karush - Kuhn - Tucker条件与支持向量

对于非线性规划（Nonlinear Programming）问题能有最优化解法的一个必要和充分条件是KKT条件：

$$\begin{cases} \alpha_i \geqslant 0 \\ y_i f(\boldsymbol{x}_i) - 1 \geqslant 0 \\ \alpha_i (y_i f(\boldsymbol{x}_i) - 1) = 0 \end{cases}$$

可以发现，若样本的α满足$\alpha_i > 0$ 则

(1), $y_i f(\boldsymbol{x}_i) = 1$ 即它是一个离分类超平面最近的点，即是$\boldsymbol{x}_s$，称为特征向量

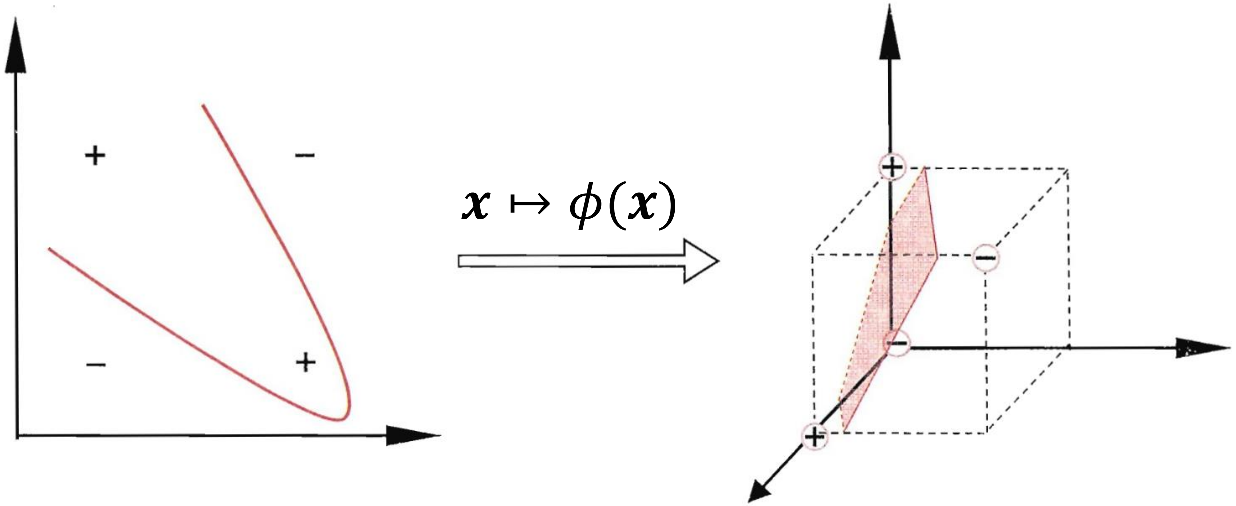
(2), 可利用支持向量 $S = \{i \mid \alpha_i > 0, i = 1,2, \dots, m\}$ 得到偏置项b

$$y_s \left(\sum_{i \in S} \alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_s + b \right) = 1, \quad b = \frac{1}{|S|} \sum_{s \in S} (y_s - \sum_{i \in S} \alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_s)$$

(3), 训练完成后，大部分的训练样本都不需保留，最终模型仅与支持向量有关。支持向量机这个名字强调了此类学习器的关键是如何从支持向量构建出解；同时也暗示着其复杂度主要与支持向量的数目有关.

核支持向量机

- 空间变换的力量



如果原始空间是有限维，即属性数有限，那么一定存在一个高维特征空间使样本可分.

核支持向量机

- 空间变换下的支持向量机

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

$$\min_{w,b} \frac{1}{2} \| w \|^2$$
$$\text{s.t. } y_i(w^T\phi(x_i) + b) \geqslant 1, i = 1,2,\dots,m$$

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$$
$$\text{s.t. } \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geqslant 0, i = 1,2,\dots,m$$

整个过程的求解式总是涉及到$\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$，而直接计算高维空间下的样本内积可能是代价高昂的

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i y_i \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}) + \frac{1}{|S|}\sum_{s\in S}\left(y_s - \sum_{i\in S} \alpha_i y_i \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_s)\right)$$

核支持向量机

- 核函数

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}_j)$$

- 示例

$$\begin{aligned}
\mathbf{K}(\mathbf{x}, \mathbf{y}) &= (\mathbf{x}^T \mathbf{y})^2 \\
&= x_1^2 y_1^2 + 2x_1 x_2 y_1 y_2 + x_2^2 y_2^2 \\
&= (x_1^2, \sqrt{2} x_1 x_2, x_2^2)(y_1^2, \sqrt{2} y_1 y_2, y_2^2) \\
&= \Phi(\mathbf{x})^T \Phi(\mathbf{y})
\end{aligned}$$

核支持向量机

- 核函数的选择

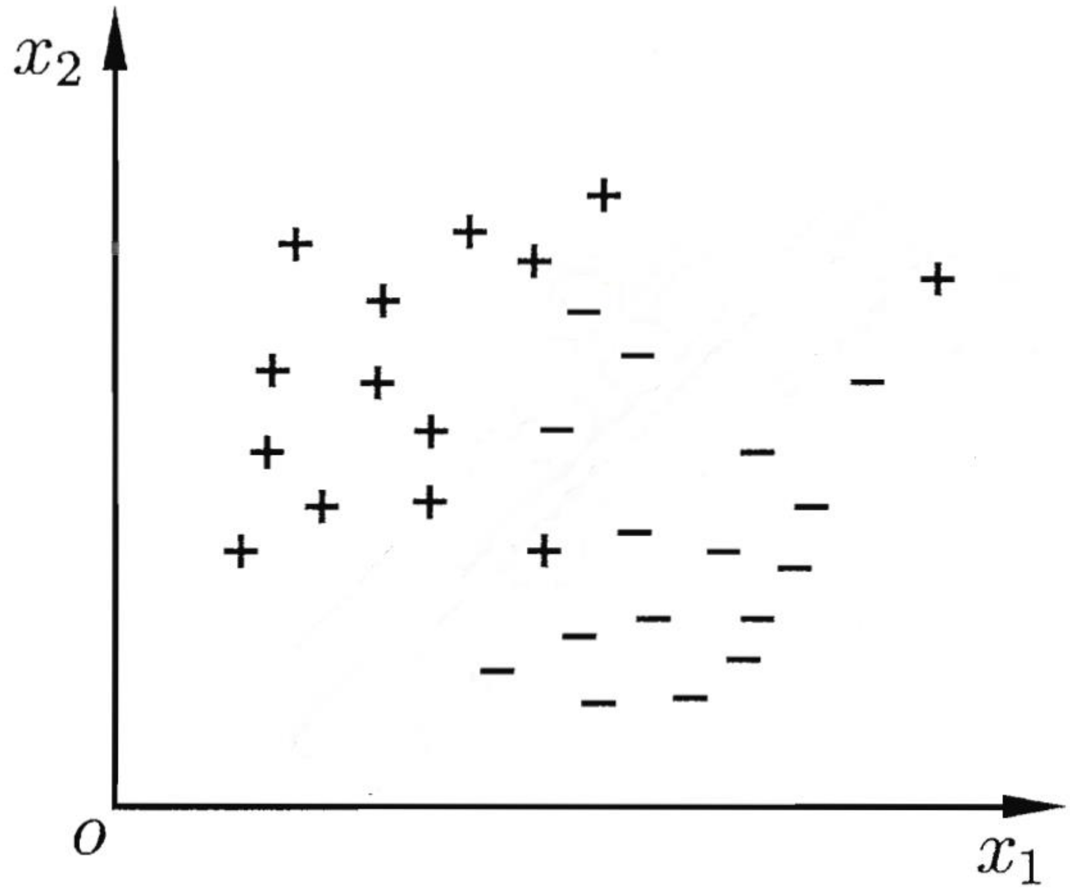
名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j)^d$	$d \geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert}{\sigma}\right)$	$\sigma > 0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\beta \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + \theta)$	tanh 为双曲正切函数, $\beta > 0, \theta < 0$

$$\gamma_1 \kappa_1 + \gamma_2 \kappa_2$$
$$\kappa_1 \otimes \kappa_2(\boldsymbol{x}, \boldsymbol{z}) = \kappa_1(\boldsymbol{x}, \boldsymbol{z}) \kappa_2(\boldsymbol{x}, \boldsymbol{z})$$
$$\kappa(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}) \kappa_1(\boldsymbol{x}, \boldsymbol{z}) g(\boldsymbol{z})$$

软间隔支持向量机

- 软间隔 *soft margin*

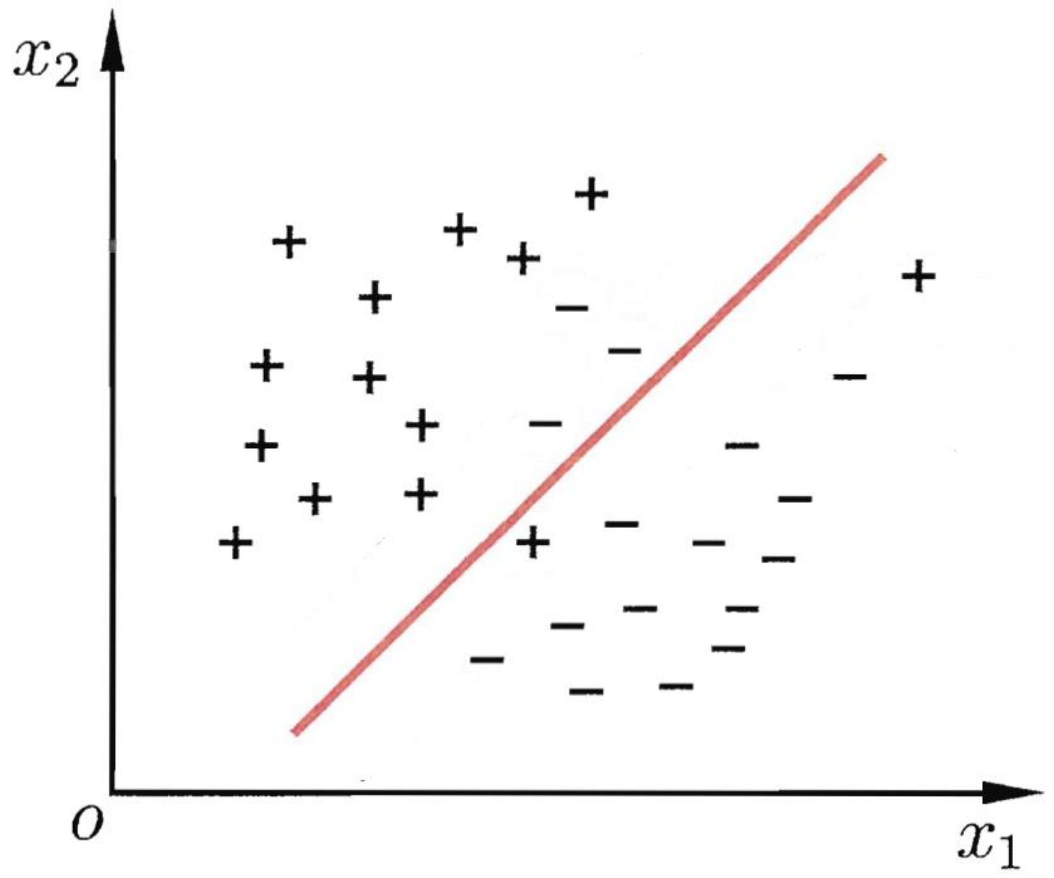
- 基础支持向量机假定训练样本在样本空间或特征空间中是线性可分的
- 一些离群点或噪音点可能会造成训练样本仅仅是“接近线性可分的”



软间隔支持向量机

- 软间隔 *soft margin*

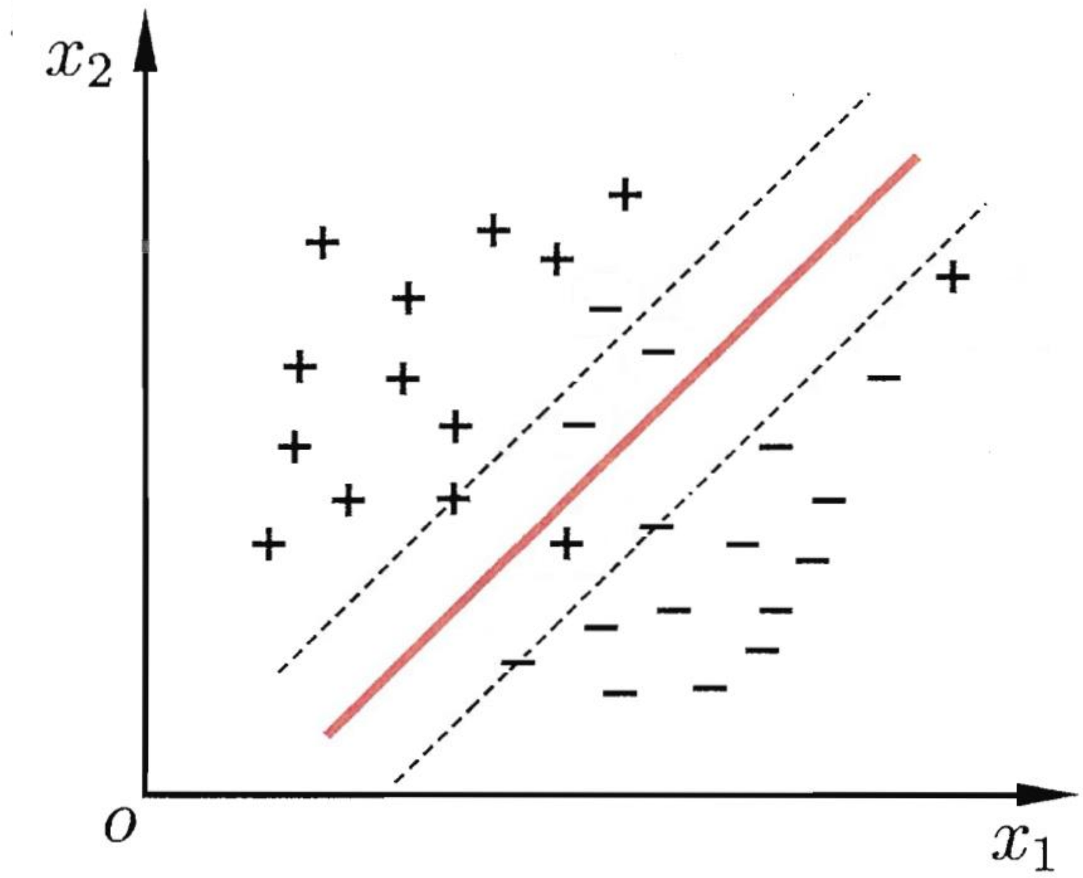
- 基础支持向量机假定训练样本在样本空间或特征空间中是线性可分的
- 一些离群点或噪音点可能会造成训练样本仅仅是“接近线性可分的”



软间隔支持向量机

- 软间隔 *soft margin*

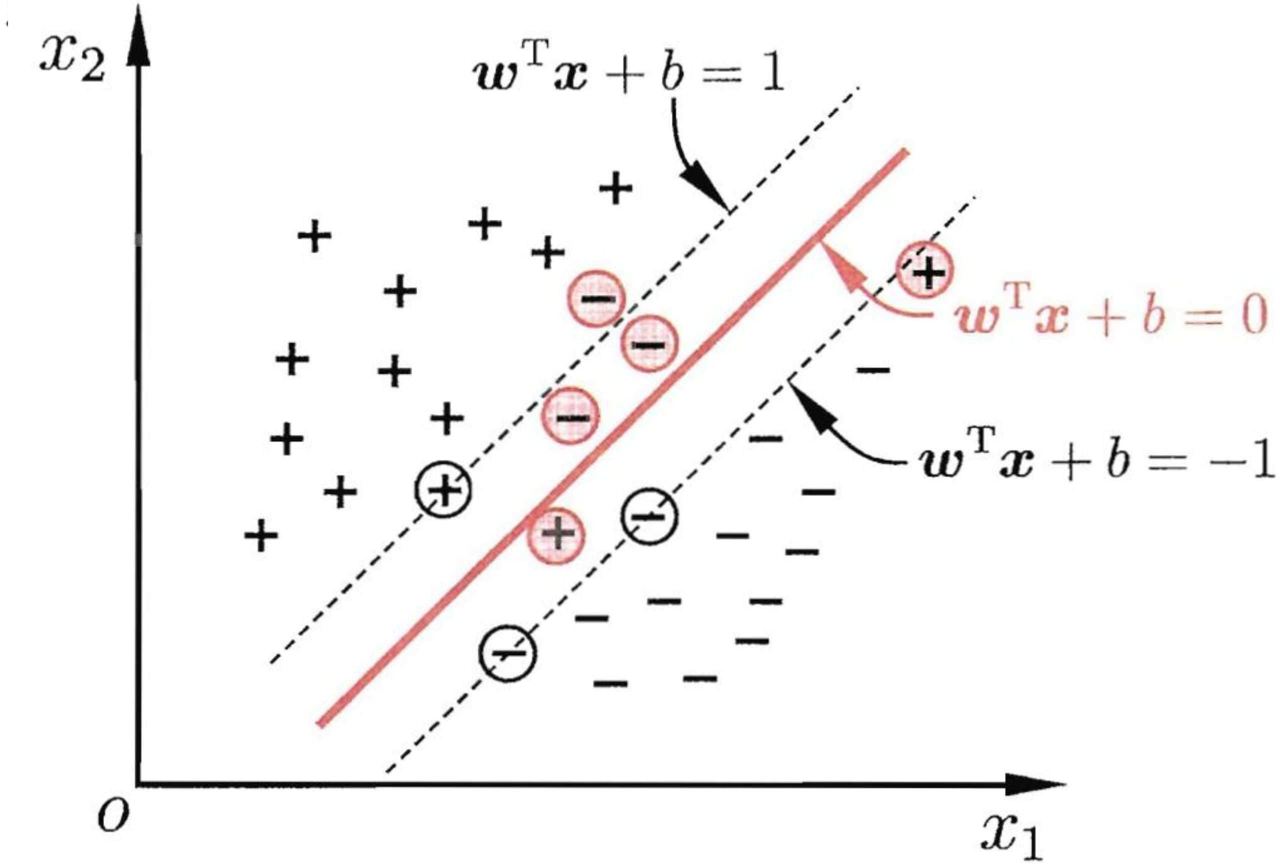
- 基础支持向量机假定训练样本在样本空间或特征空间中是线性可分的
- 一些离群点或噪音点可能会造成训练样本仅仅是“接近线性可分的”



软间隔支持向量机

- 软间隔 *soft margin*

- 基础支持向量机假定训练样本在样本空间或特征空间中是线性可分的
- 一些离群点或噪音点可能会造成训练样本仅仅是“接近线性可分的”
- 软间隔允许某些样本不满足$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1$



软间隔支持向量机

- 软间隔 *soft margin*

\- 硬间隔约束

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1$$

\- 破坏硬间隔的0-1损失

$$\ell_{0/1}\left(y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) - 1\right),\ \ell_{0/1}(z) = \begin{cases} 1, \text{if } z < 0 \\ 0, \text{otherwise} \end{cases}$$

\- hinge损失

$$\ell_{hinge}(z) = max(0,1-z)$$

$$\min_{\boldsymbol{w},b} \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + C \sum_{i=1}^{m} max(0,1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b))$$

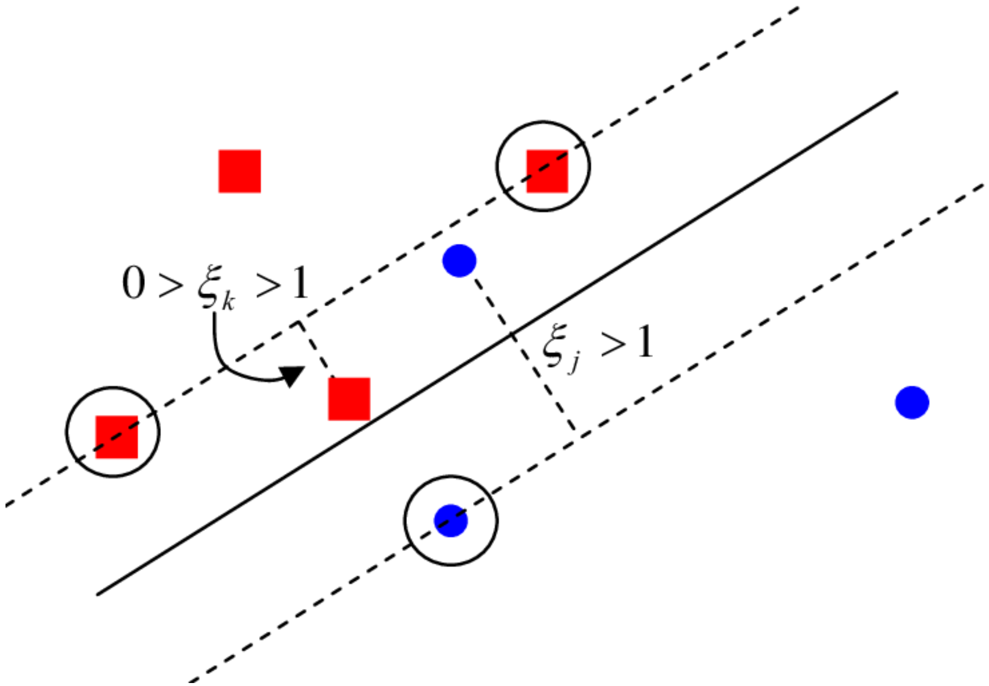
软间隔支持向量机

- 软间隔 *soft margin*

- hinge损失与松弛变量

$$\min_{\boldsymbol{w},b} \frac{1}{2}\parallel \boldsymbol{w} \parallel^2 + C\sum_{i=1}^{m} max(0,1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b))$$

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi_i} \quad & \frac{1}{2}\parallel \boldsymbol{w} \parallel^2 + C\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b) \geqslant 1-\xi_i \\ & \xi_i \geqslant 0, i=1,2,\dots,m \end{aligned}$$



软间隔支持向量机

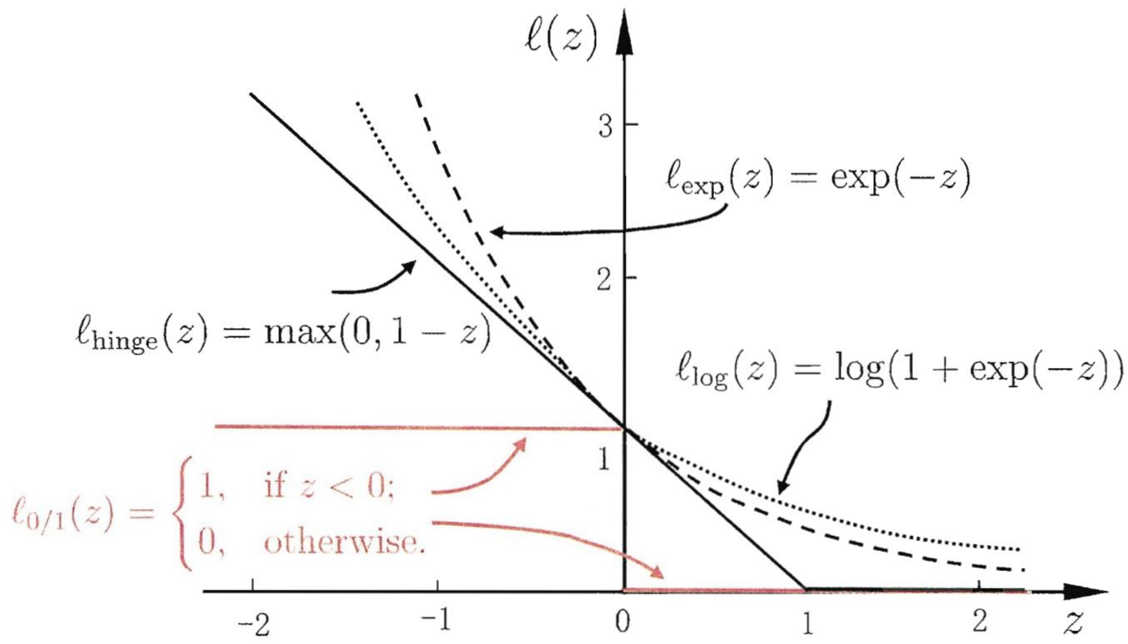
- 常用的分类损失

  - 函数间隔为关注对象 $z = y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)$

hinge 损失: $\ell_{\text{hinge}}(z) = \max(0, 1 - z)$

指数损失(exponential loss): $\ell_{\exp}(z) = \exp(-z)$

对率损失(logistic loss): $\ell_{\log}(z) = \log(1 + \exp(-z))$

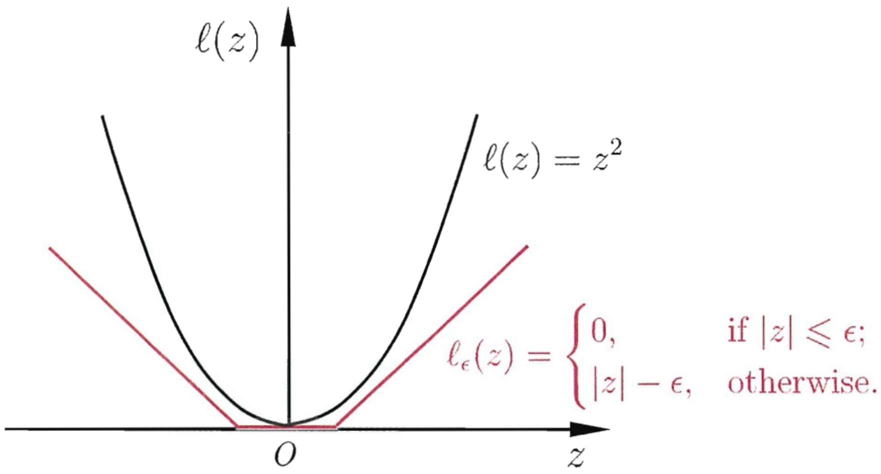


支持向量回归

- ε不敏感损失

$$\min_{\boldsymbol{w},b} \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{m} \ell_\epsilon(f(\boldsymbol{x}_i) - y_i)$$

$$\ell_\epsilon(z) = \begin{cases} 0, & \text{if } |z| \leqslant \epsilon \\ |z| - \epsilon, & \text{otherwise} \end{cases}$$



$$\min_{\boldsymbol{w},b,\xi_i,\hat{\xi}_i} \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{m} (\xi_i + \hat{\xi}_i)$$

s.t. $f(\boldsymbol{x}_i) - y_i \leqslant \epsilon + \xi_i$

$y_i - f(\boldsymbol{x}_i) \leqslant \epsilon + \hat{\xi}_i$

$\xi_i \geqslant 0, \hat{\xi}_i \geqslant 0, i = 1,2,\dots,m$