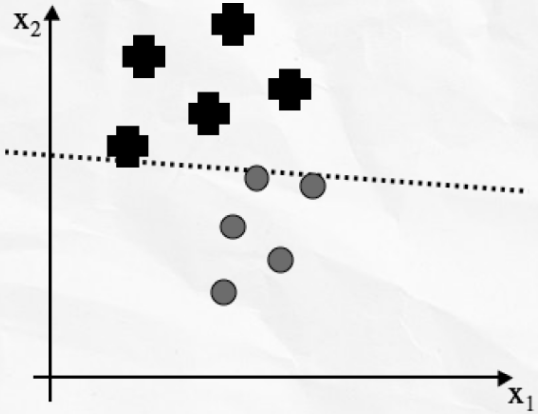
机器学习

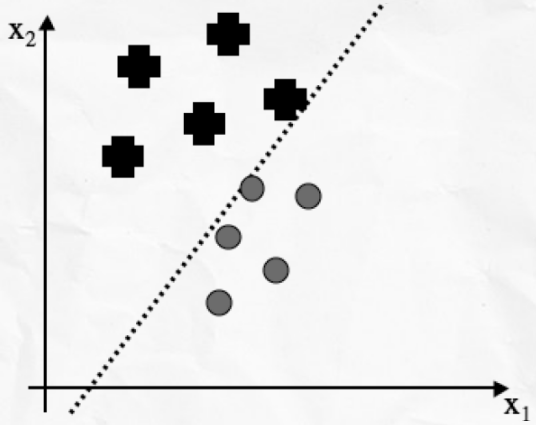
支持向量机

涂文婷
tu.wenting@mail.shufe.edu.cn

间隔

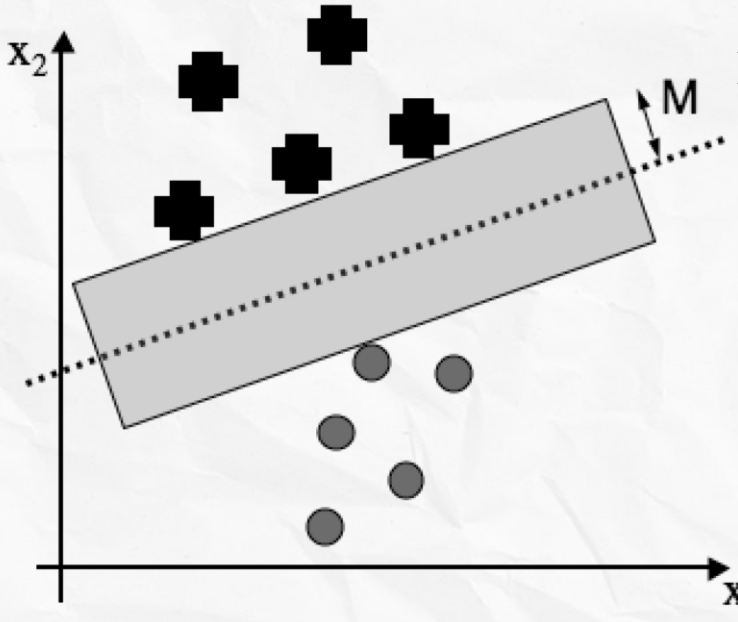
- 分类超平面的优劣





间隔

- 空间间隔

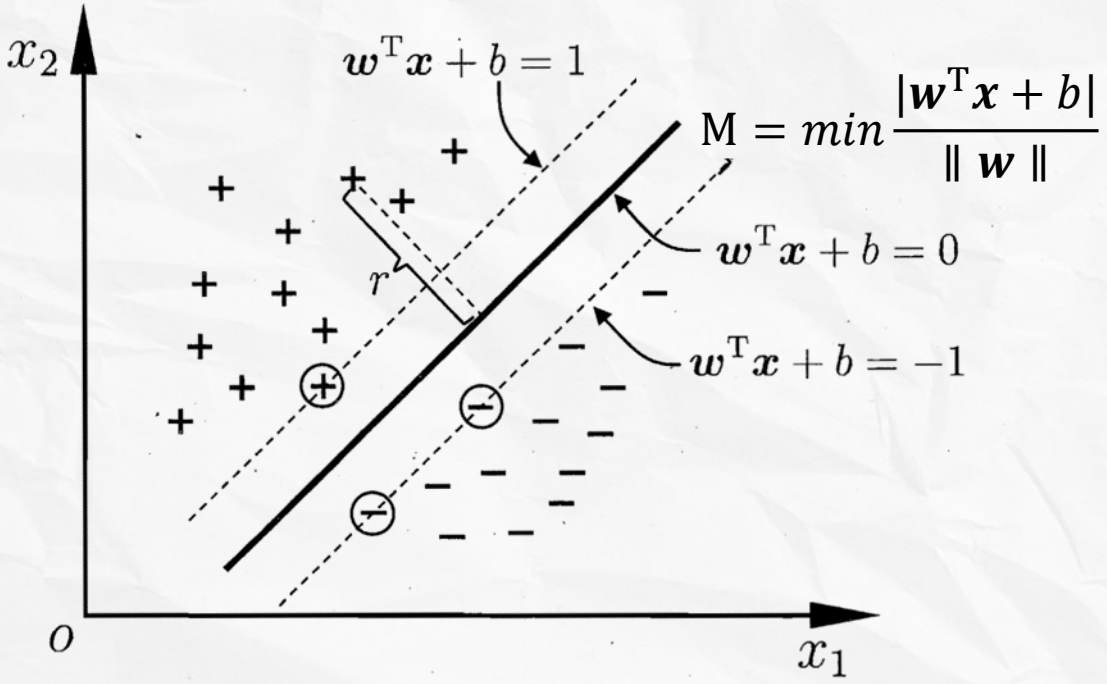


$$\mathrm{M} = min\frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b|}{\| \boldsymbol{w} \|}$$

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

线性可分支持向量机

- 最大空间间隔

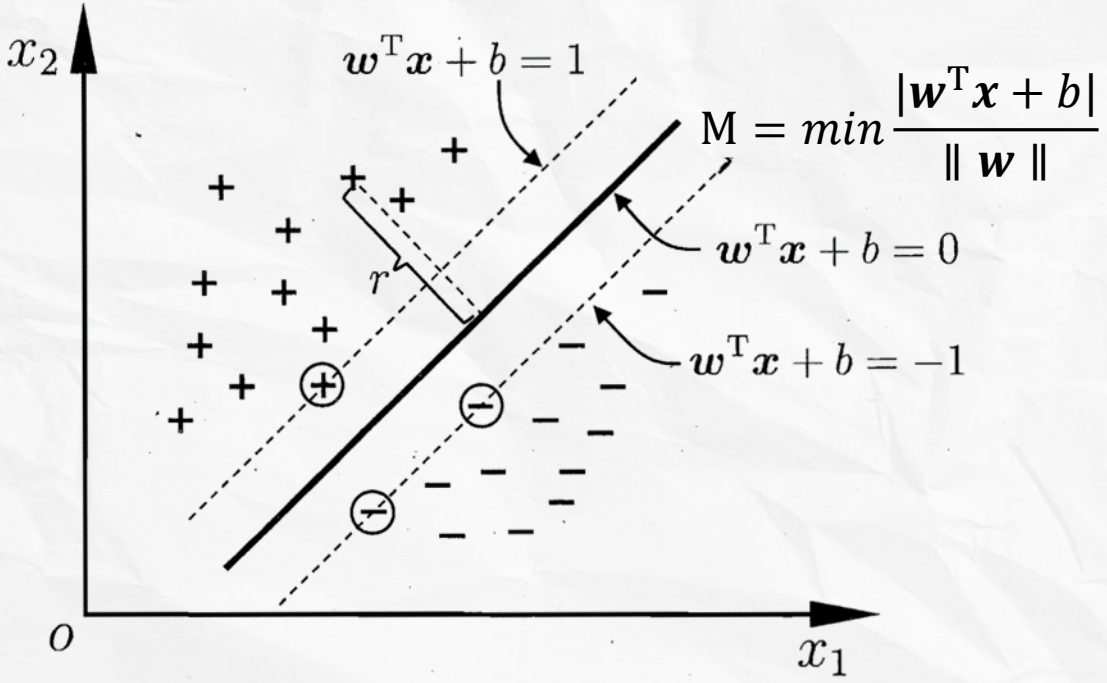


线性可分支持向量机

- 最大空间间隔

$$\max_{\boldsymbol{w},b} \mathrm{M}$$

$$\text{s.t.} \begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \geqslant 0, & y_i = +1 \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \leqslant 0, & y_i = -1 \end{cases}$$

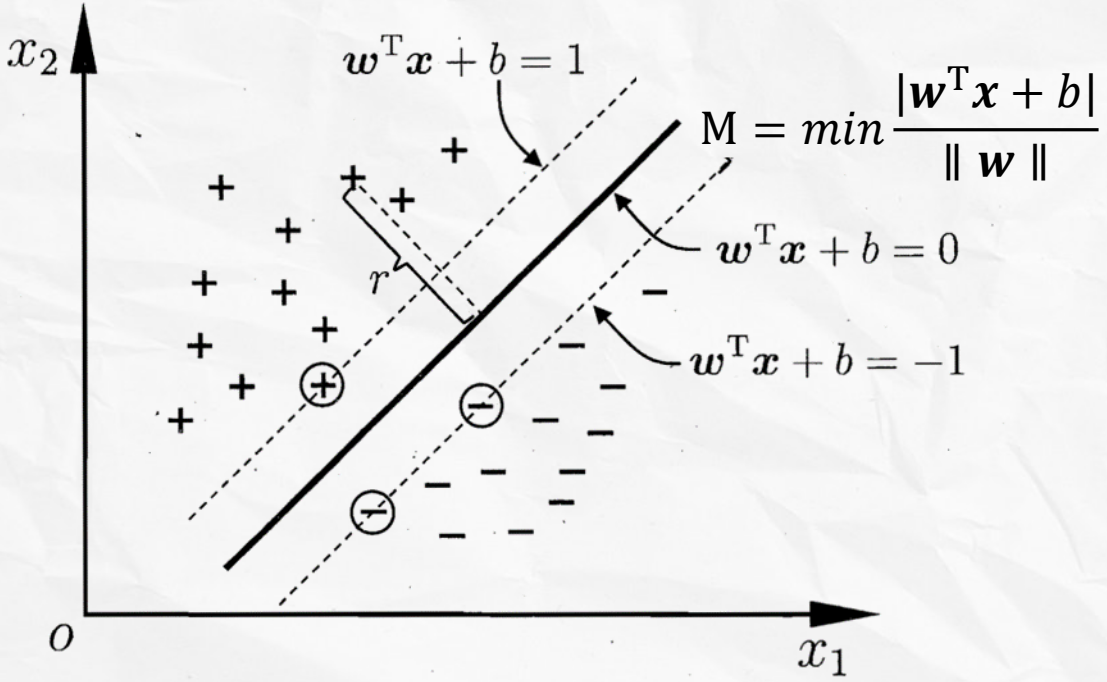


线性可分支持向量机

- 最大空间间隔

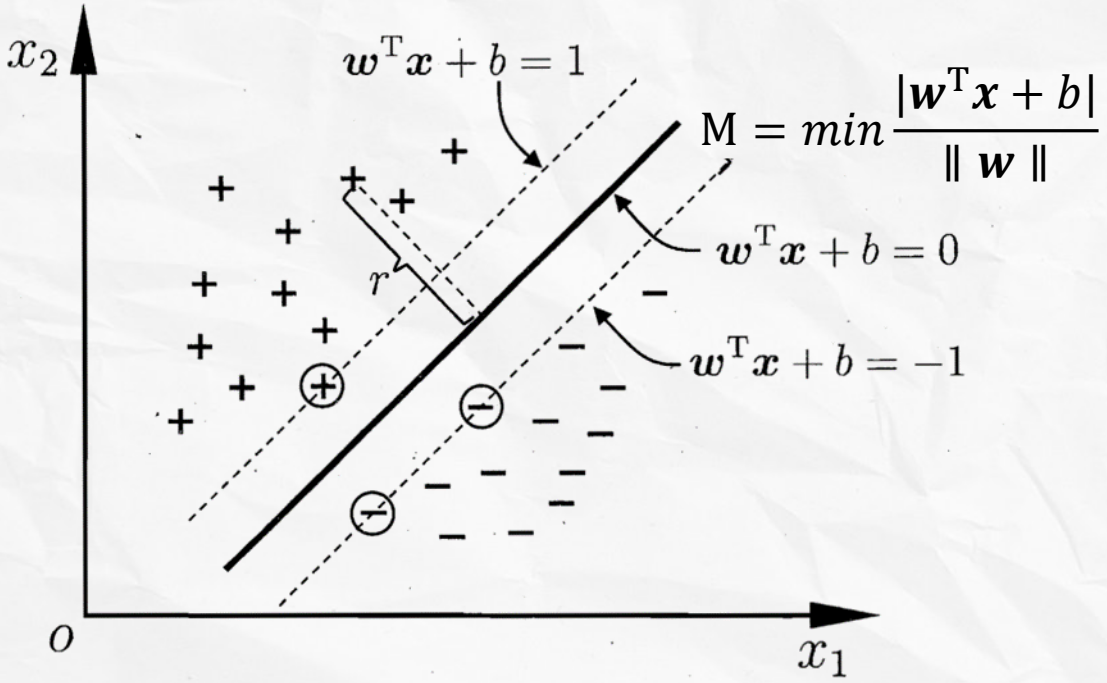
$$\max_{\boldsymbol{w},b} \mathrm{M}$$

$$\text{s.t.} \begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \geqslant 0, & y_i = +1 \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \leqslant 0, & y_i = -1 \end{cases} \quad <=> \quad y_i\,(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 0$$



线性可分支持向量机

- 最大空间间隔



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

$$\max_{w,b} \mathrm{M}$$

$$\text{s.t.} \begin{matrix} w^T x_i + b \geqslant 0, & y_i = +1 \\ w^T x_i + b \leqslant 0, & y_i = -1 \end{matrix} \quad <=> \; y_i \, (w^T x_i + b) \geqslant 0$$

上述优化存在无穷多个最优解，加入限制

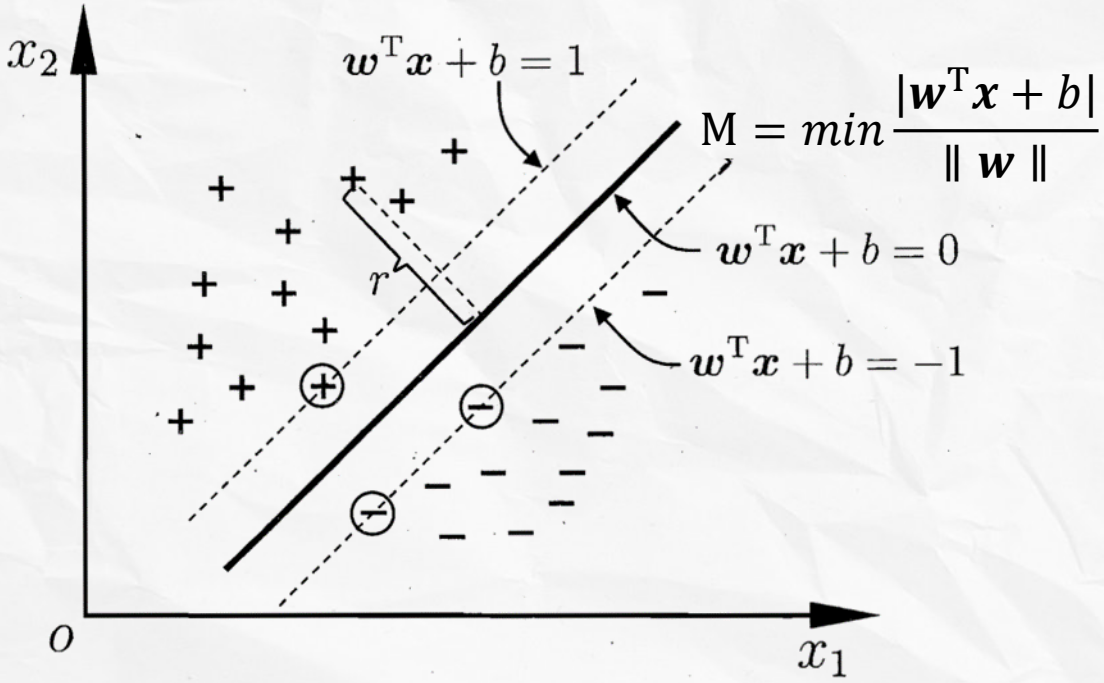
$$y_s(w^T x_s + b) = 1$$

离分类超平面最近的点

$$<=> \; y_i \, (w^T x_i + b) \geqslant 1$$

线性可分支持向量机

- 最大空间间隔



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

$$\max_{\boldsymbol{w},b}\ \mathrm{M}$$

$$\text{s.t.}\begin{matrix}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\geqslant 0, & y_i=+1\\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\leqslant 0, & y_i=-1\end{matrix}\quad <=>\ y_i\,(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 0$$

上述优化存在无穷多个最优解，加入限制

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

↖ 离分类超平面最近的点

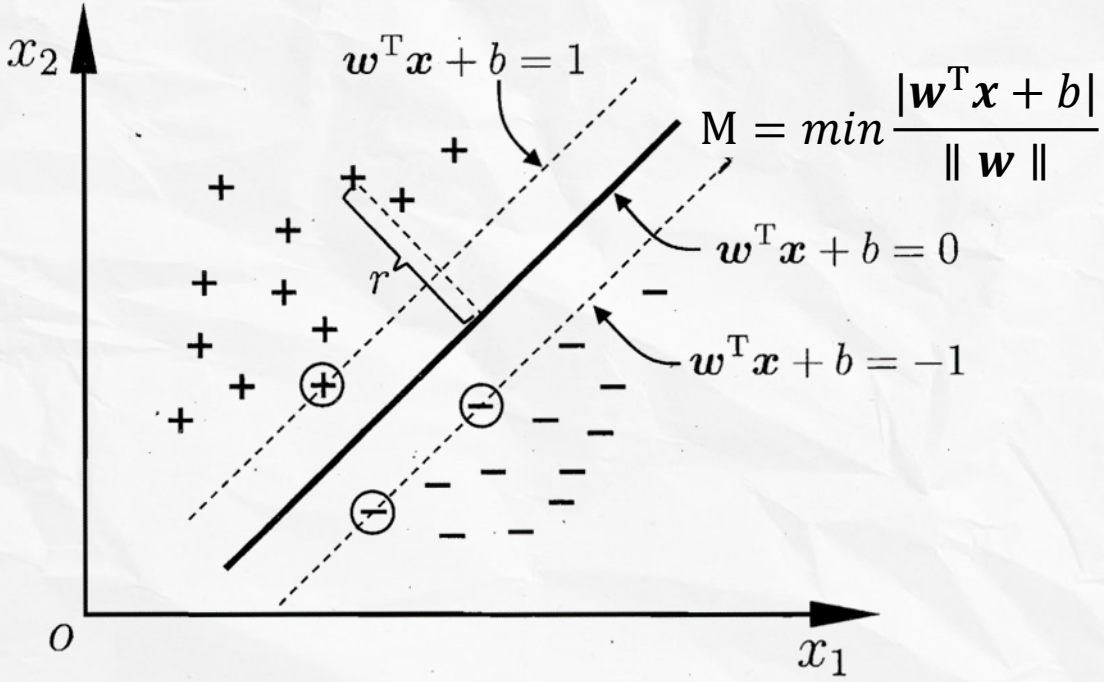
$$<=>\ y_i\,(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1$$

由于此时$min\ \left|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right|=min\quad y\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right)=1$

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

线性可分支持向量机

- 最大空间间隔



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

$$\max_{w,b} \mathrm{M}$$

$$\text{s.t.} \begin{matrix} w^T x_i + b \geqslant 0, & y_i = +1 \\ w^T x_i + b \leqslant 0, & y_i = -1 \end{matrix} \quad <=> \; y_i \, (w^T x_i + b) \geqslant 0$$

上述优化存在无穷多个最优解，加入限制

$$y_s (w^T x_s + b) = 1$$

↖ 离分类超平面最近的点

$$<=> \; y_i \, (w^T x_i + b) \geqslant 1$$

由于此时$min \; |w^T x + b| = min \quad y(w^T x + b) = 1$

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

优化任务变为：

$$\max_{w,b} \frac{1}{\| w \|} <=> \min_{w,b} \frac{1}{2} \| w \|^2$$

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

线性可分支持向量机

- 利用对偶问题求解

原问题

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

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

对偶问题

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

↓ 令L对$\boldsymbol{w}$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 →

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

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：

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

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$b$

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

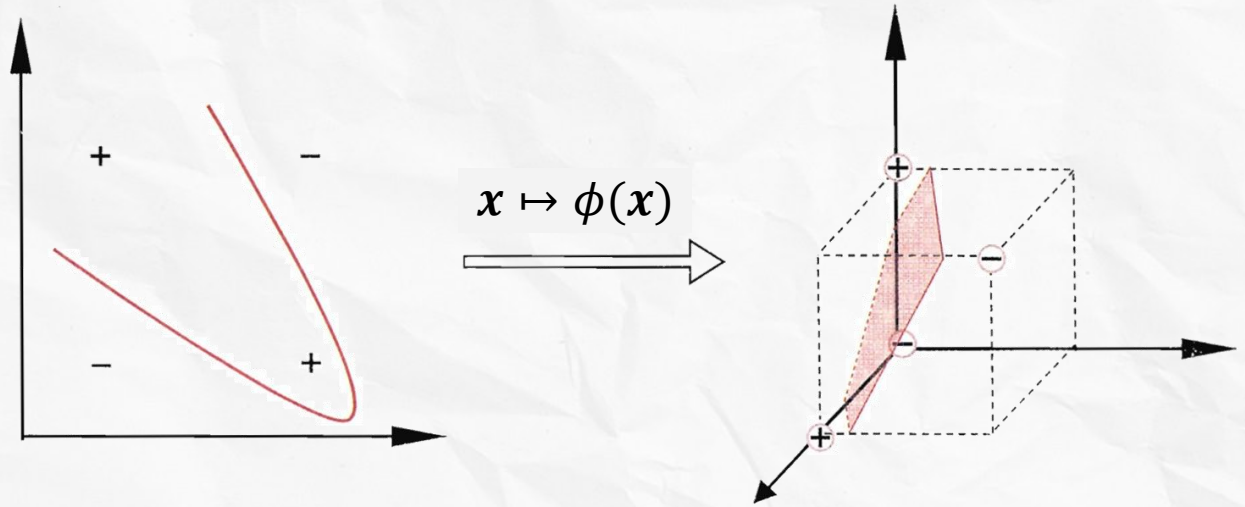
(3)，训练完成后，大部分的训练样本都不需保留，最终模型仅与支持向量有关。

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

- 核技术
  - 空间变换



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

- 核技术
  - 空间变换下的支持向量机

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

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

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

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

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

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

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

核支持向量机

- 核技术
  - 核函数

$$\kappa(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{y})$$

e. g. ,

$$\kappa(\mathbf{x},\mathbf{y}) = \left(\mathbf{x}^{T}\mathbf{y}\right)^{2}$$

$$= x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2$$

$$= \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)\left(y_1^2, \sqrt{2}y_1y_2, y_2^2\right)$$

$$= \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{y}),\ \phi(\mathbf{v}) = \left(v_1^2, \sqrt{2}v_1v_2, v_2^2\right)$$

核支持向量机

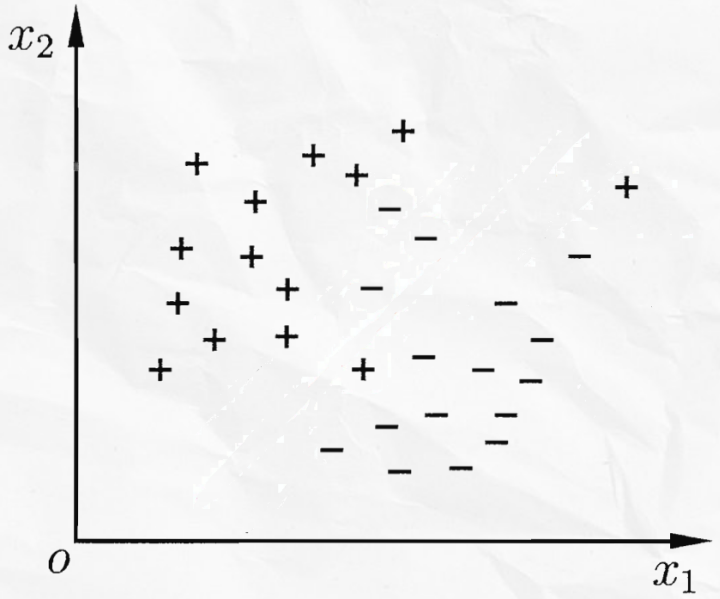
核技术

常用核函数

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

软间隔支持向量机

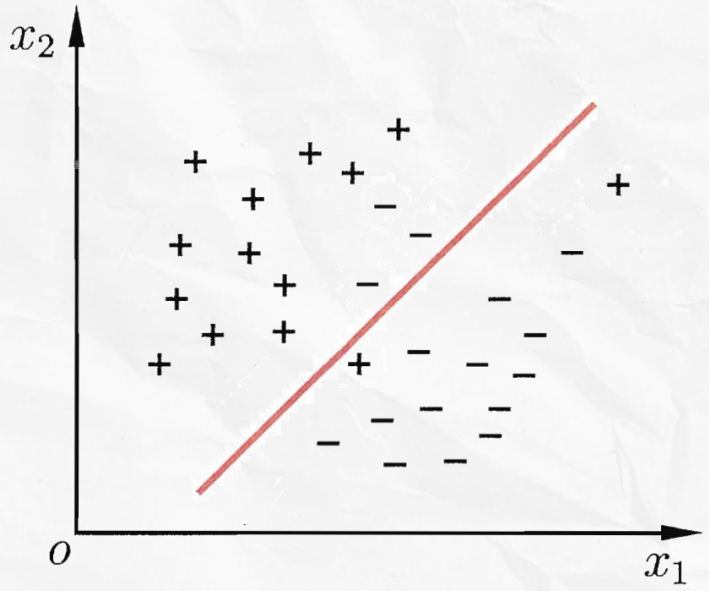
- 背景



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

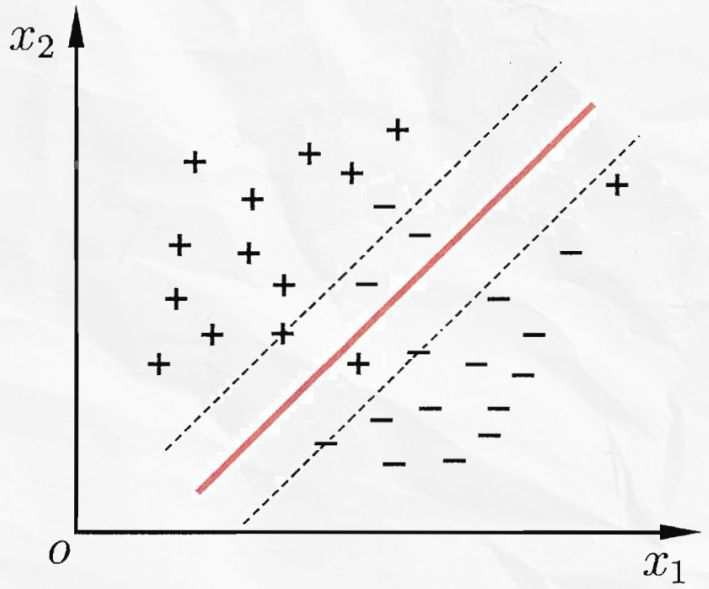
- 背景



当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”，
如何学习到一个可以在一定程度接受破坏间隔约束的数据点存在的分类超平面？

软间隔支持向量机

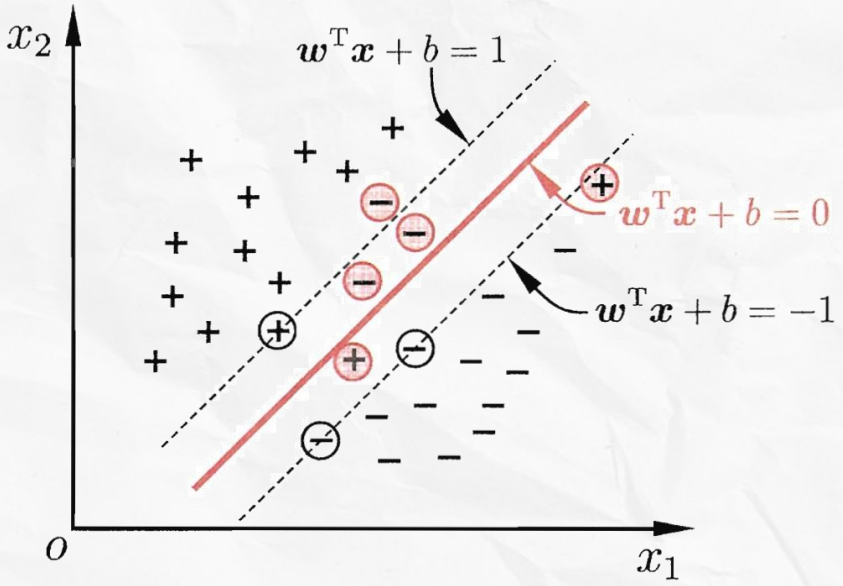
- 背景



当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”，
如何学习到一个可以在一定程度接受破坏间隔约束的数据点存在的分类超平面？

软间隔支持向量机

- 软间隔



硬间隔要求所有样本满足$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1$

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

- 从正则化角度看支持向量机算法
  - 硬间隔支持向量机

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

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

<=>

$$\min_{\boldsymbol{w},b} \frac{1}{2} \| \boldsymbol{w} \|^2 + C\sum_{i=1}^{m} \ell_{0/1}\left(y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) - 1\right), C = +\infty, \quad \ell_{0/1}(z) = \begin{cases} 1, & z < 0 \\ 0, & \text{otherwise} \end{cases}$$

软间隔支持向量机

- 从正则化角度看支持向量机算法
  - 软间隔支持向量机
    - hinge损失

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

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

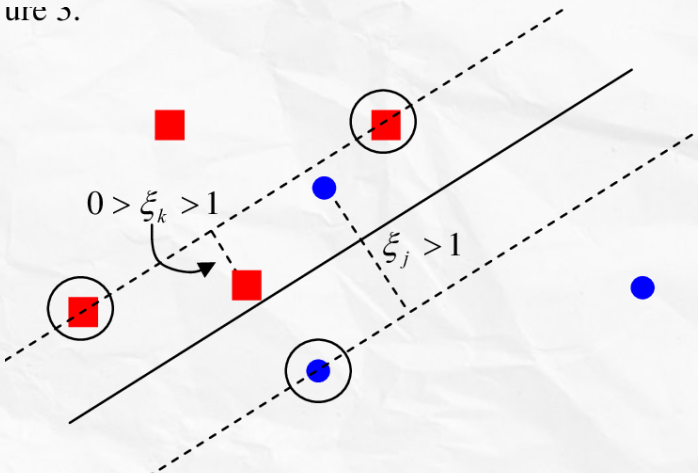
    - 松弛变量

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

$$\text{s.t.} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1 - \xi_i$$

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

ure 3.



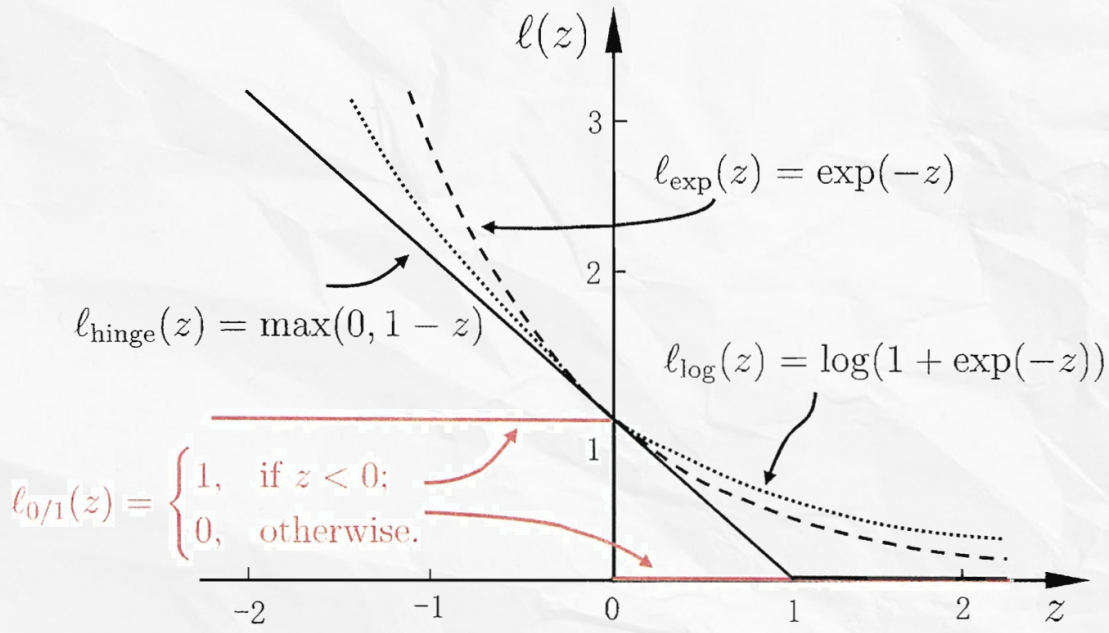
软间隔支持向量机

- 常用的分类损失
  - 分类损失总与函数间隔$z = y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)$相关

    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**): $\ell_{\log}(z) = \log(1 + \exp(-z))$

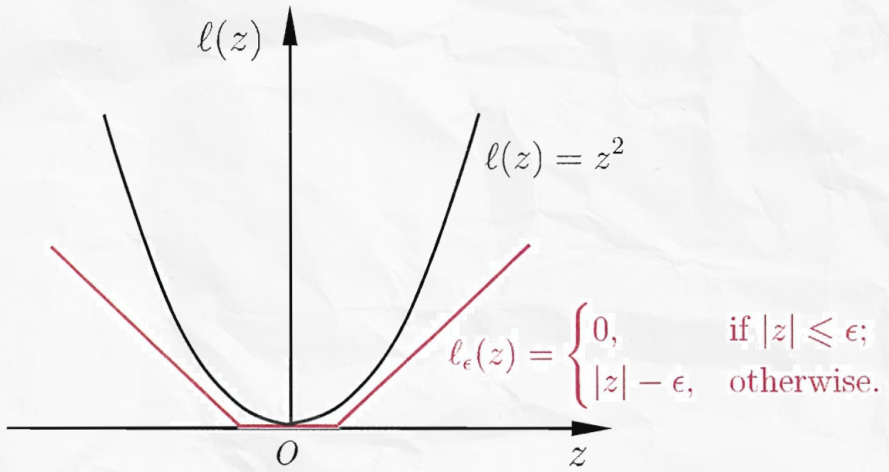


支持向量回归

- ε不敏感损失

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

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



支持向量机的实现

- 求解技术
  - 二次规划 *quadratic programming (QP)*
  - 序列最小化优化 *Platt's sequential minimal optimization (SMO)*
  - 坐标下降法 *coordinate descent*

- 工具包
  - sklearn.svm.LinearSVC
  - sklearn.svm.SVC
  - sklearn.linear_model.SGDClassifier (loss='hinge')