支持向量机

主讲人：屠恩美

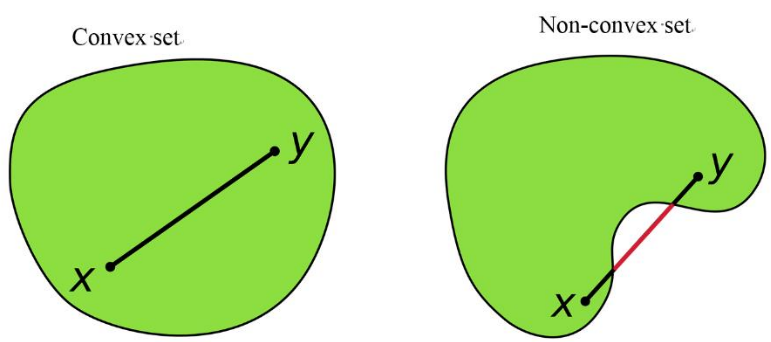
《机器学习与知识发现》

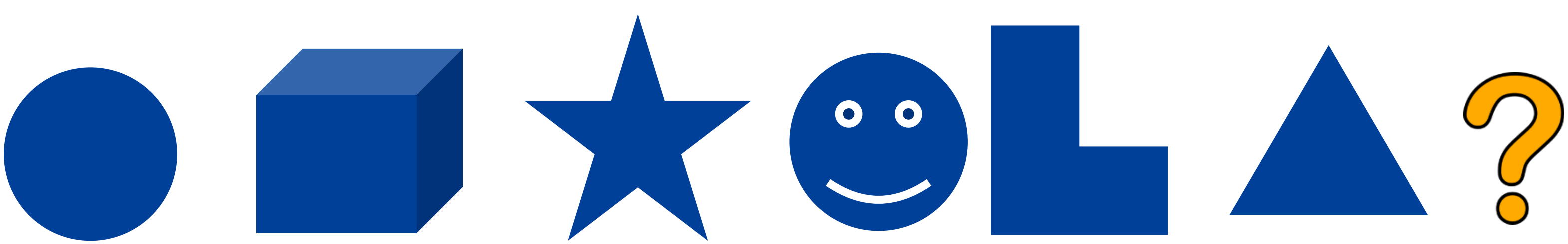
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

凸集

- 对任意的 $\sum_{i=n}^{n} \alpha_i = 1,\ \alpha_i \geq 0$， 如果集合 X 具有如下性质

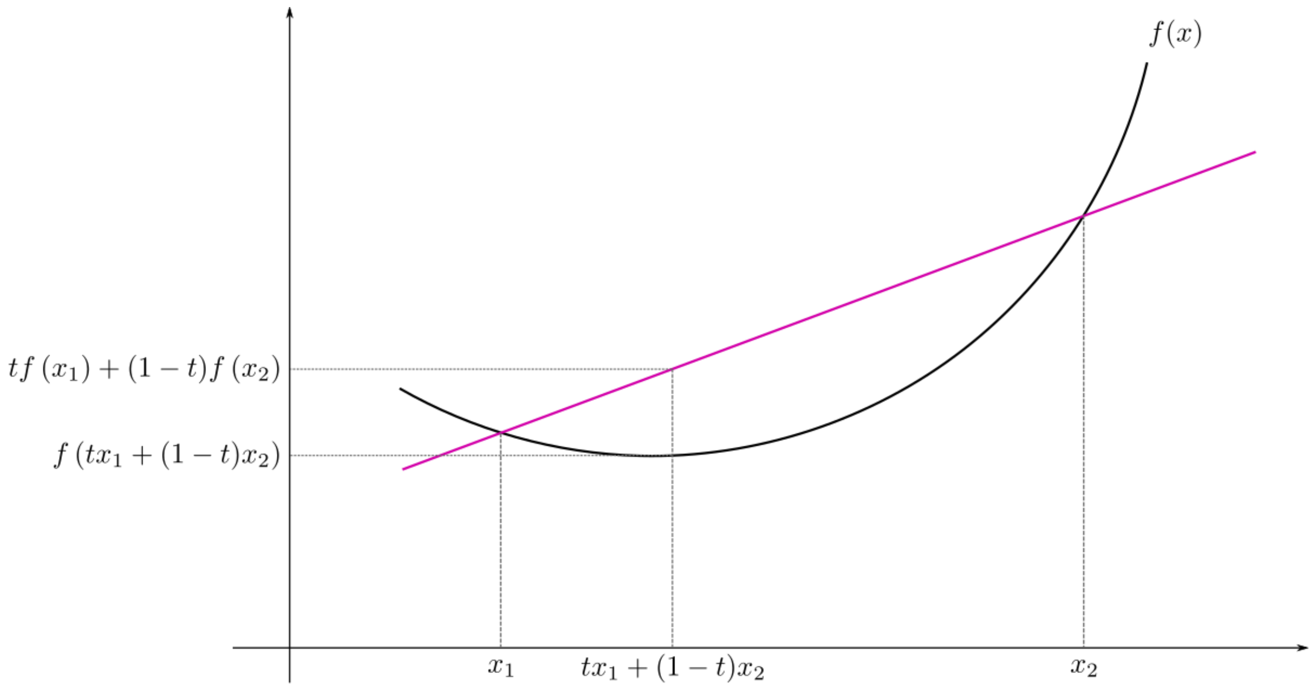
$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in X \quad \Rightarrow \quad \left(\sum_{i=1}^{n} \alpha_i \mathbf{x}_i \right) \in X$$





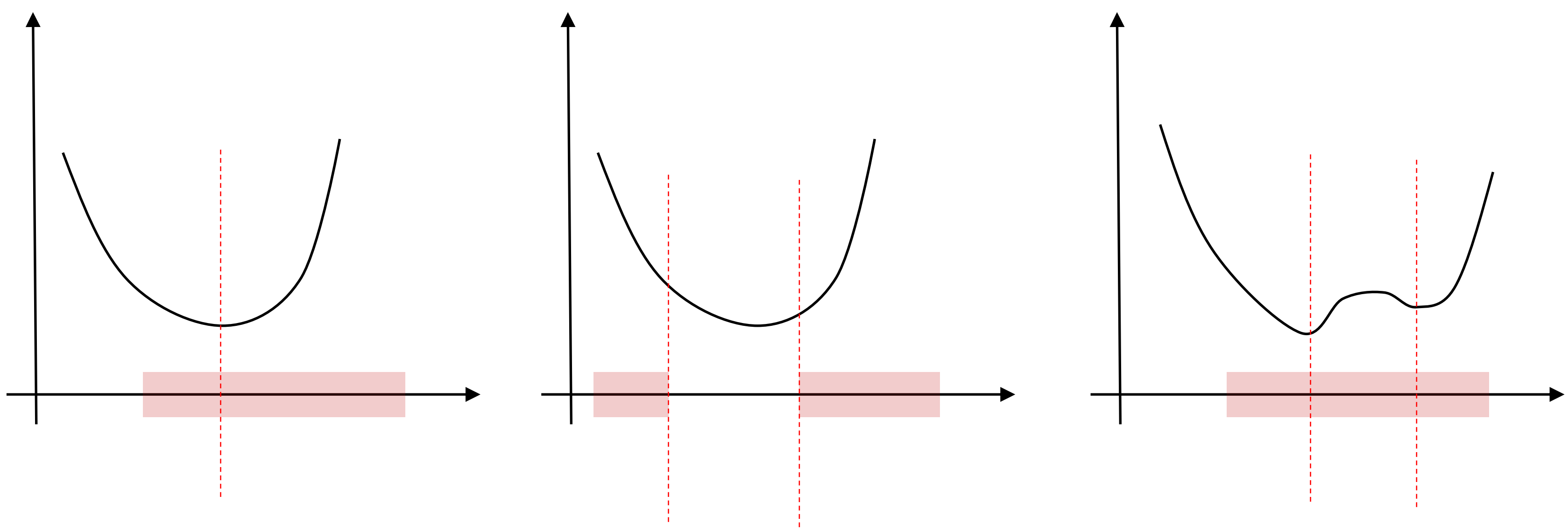
凸函数

- 定义： $t f(x_1)+(1-t) f(x_2) \geq f(tx_1+(1-t)x_2), \quad 0 \leq t \leq 1$
- 判断：如果 $f(x)$ 二阶可导，$f''(x) \geq 0$
- 常见凸函数： $\mathbf{a}^T\mathbf{x}+b, \quad \|\mathbf{x}\|_p, \quad e^{\mathbf{a}^T\mathbf{x}+b}, \quad \mathrm{tr}\left(\mathbf{A}^T\mathbf{X}\right)$



凸优化

- 凸函数在凸集上的最小值优化问题就是凸优化
- 凸优化具有全局唯一最小值（不同初始条件都能获得最优解）



凸函数 - 凸集　　凸函数 - 非凸集　　非凸函数 - 凸集

优化回顾

- 有约束优化问题

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathbb{R}^d} \quad & f_0(\mathbf{x}) \\
\text{s. t.} \quad & f_i(\mathbf{x}) \le 0, \quad i = 1,2,\dots,p \\
& g_j(\mathbf{x}) = 0, \quad j = 1,2,\dots,q
\end{aligned}
$$

目标函数

不等式约束

等式约束

- 如果 $f_i(\mathbf{x})$ 是凸函数，且 $g_i(\mathbf{x})$ 是线性函数，则该优化问题为凸优化
- 拉格朗日乘子函数

$$
\begin{aligned}
L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\tau}) &= f_0(\mathbf{x}) + \sum_{i=1}^{p} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{q} \tau_i g_i(\mathbf{x}) \\
&= f_0(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{F}(\mathbf{x}) + \boldsymbol{\tau}^T \mathbf{G}(\mathbf{x})
\end{aligned}
$$

凸优化与KKT条件

- 拉格朗日对偶函数

$$g\left(\boldsymbol{\lambda},\boldsymbol{\tau}\right)=\inf_{\mathbf{x}\in D} L\left(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\tau}\right)=\left(f_0(\mathbf{x})+\boldsymbol{\lambda}^T\mathbf{F}(\mathbf{x})+\boldsymbol{\tau}^T\mathbf{G}(\mathbf{x})\right)\Big|_{\mathbf{x}^*\leftarrow\nabla L_{\mathbf{x}}\left(\mathbf{x}^*,\boldsymbol{\lambda},\boldsymbol{\tau}\right)=0}$$

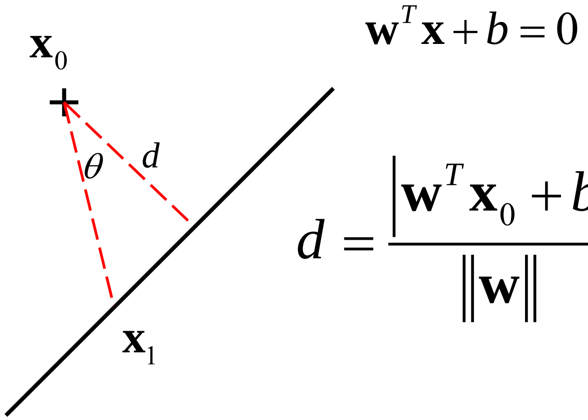
- 对偶问题

$$\max g\left(\boldsymbol{\lambda},\boldsymbol{\tau}\right)$$
$$s.t.\quad \lambda_i \ge 0,\quad i=1,\ldots,p$$

- KKT条件：最优解的必要条件(**凸优化**时也是**充分必要**条件)
  - 原始问题约束 $f_i(\mathbf{x})\le 0,\quad i=1,2,\ldots,p;\quad g_j(\mathbf{x})=0,\quad j=1,2,\ldots,q$
  - 对偶问题约束 $\lambda_i \ge 0,\quad i=1,\ldots,p$
  - 互补松弛 $\lambda_i f_i(\mathbf{x})=0,\quad i=1,\ldots,p$
  - 梯度消失 $\nabla L_{\mathbf{x}}\left(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\tau}\right)=0$

点到直线的距离

- $\mathbf{x}_0$ 到直线 $\mathbf{w}^T\mathbf{x}+b=0$ 的距离

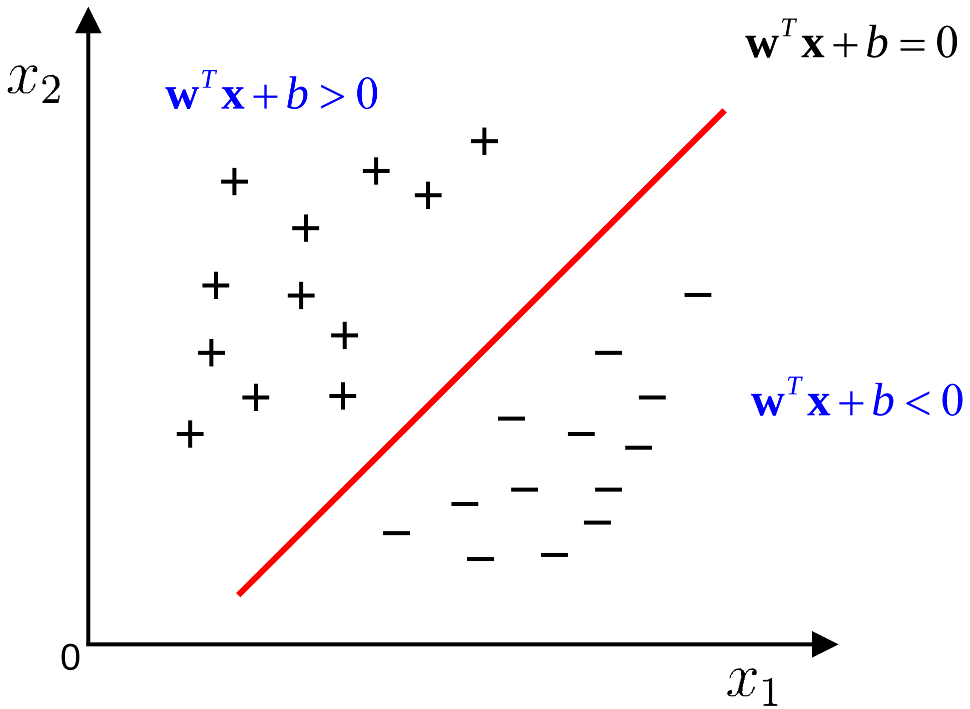


$$d=\frac{\left|\mathbf{w}^T\mathbf{x}_0+b\right|}{\|\mathbf{w}\|}$$

$$\begin{aligned}
d &= \|\mathbf{x}_0-\mathbf{x}_1\|\cos\theta \\
&= \frac{\|\mathbf{x}_0-\mathbf{x}_1\|\|\mathbf{w}\|\cos\theta}{\|\mathbf{w}\|} \\
&= \frac{\left|\mathbf{w}^T(\mathbf{x}_0-\mathbf{x}_1)\right|}{\|\mathbf{w}\|} \quad \text{(dot product definition)} \\
&= \frac{\left|\mathbf{w}^T\mathbf{x}_0-\mathbf{w}^T\mathbf{x}_1\right|}{\|\mathbf{w}\|} \\
&= \frac{\left|\mathbf{w}^T\mathbf{x}_0+b\right|}{\|\mathbf{w}\|} \quad \text{(line equation } \mathbf{w}^T\mathbf{x}_1+b=0\text{)}
\end{aligned}$$

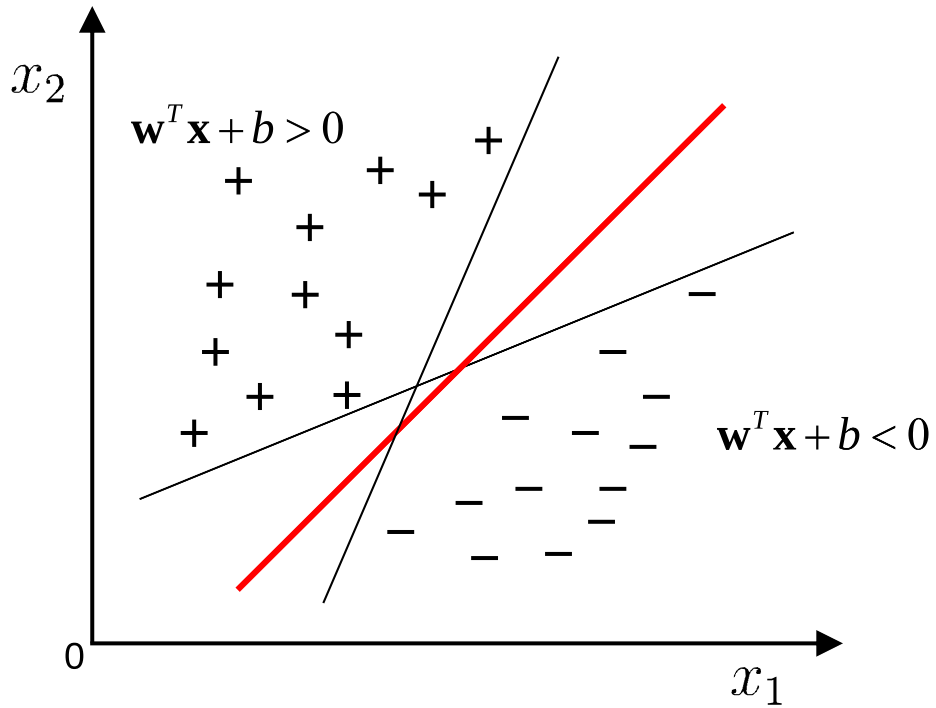
两类线性可分问题

- 考虑两类线性可分情况：在样本空间中寻找一个超平面，将不同类别的样本分开



两类线性可分问题

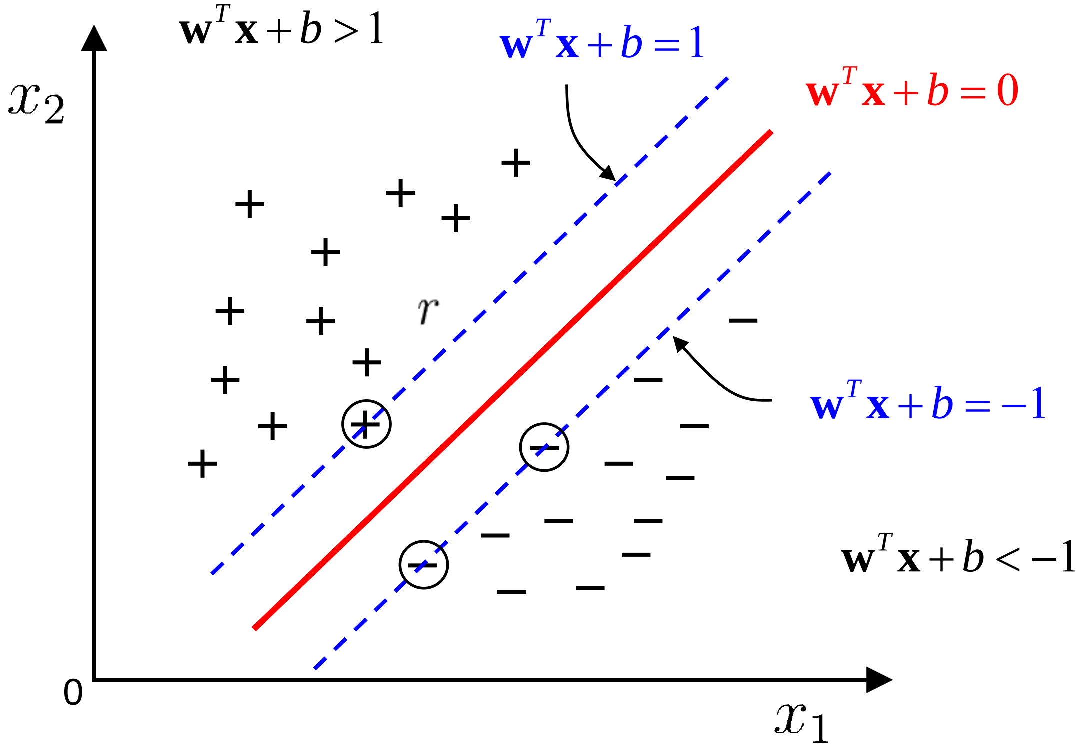
- Q: 将训练样本分开的超平面可能有很多, 哪一个好呢?



- A: 应选择"正中间", 容错性好, 稳健性高, 泛化能力最强.

两类线性可分问题

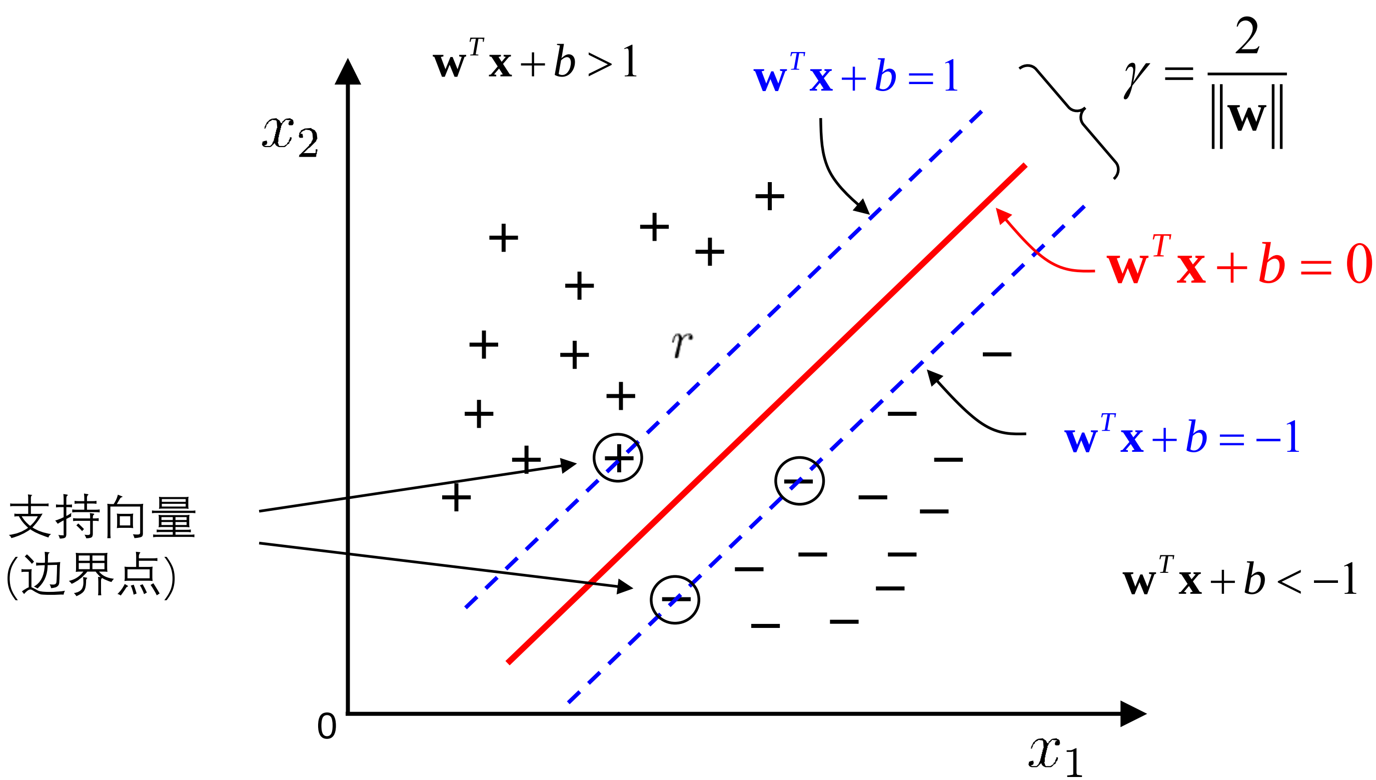
- 通过适当的比例缩放 $\mathbf{w} \to \tau\mathbf{w}, b \to \tau b$ 使正类≥1，负类≤ -1，而距离超平面最近的几个样本使等式成立



考虑正类缩放(负类同理) $\mathbf{w}^T\mathbf{x}+b>M \;\rightarrow\; \frac{1}{M}\mathbf{w}^T\mathbf{x}+\frac{1}{M}b>1 \;\rightarrow\; \tau=\frac{1}{M}$

间隔与支持向量

- 两条蓝线间隔 $\gamma = \dfrac{2}{\|\mathbf{w}\|}$ ？



支持向量机

- 假设有n个样本 $D=\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$， $y_i \in \{+1, -1\}$
- 最大间隔: 寻找参数 $\mathbf{w}$ 和 b , 使得 γ 最大，同时满足**正确分类**的约束条件.

$$\arg\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, 2, \ldots, n$$

$$\Downarrow$$

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

$$\text{s.t.} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, 2, \ldots, n$$

凸优化，存在全局唯一最小值!

问题求解

- 第一步：引入拉格朗日乘子 $\alpha_i \geq 0$ 得到拉格朗日函数

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

- 第二步：令 $L(\mathbf{w}, b, \boldsymbol{\alpha})$ 对 $\mathbf{w}$ 和 b 的偏导为零可得

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

- 第三步：回代得到一个新的优化问题

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^{n} \alpha_i$$

对偶问题

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \ldots, n$$

求解方法-SMO

- 基本思路：不断执行如下两个步骤直至收敛.
  - 第一步：选取一对需更新的变量 α_i 和 α_j .
  - 第二步：固定 α_i 和 α_j 以外的参数，求解对偶问题更新 α_i 和 α_j
- 上面第二步：仅考虑 α_i 和 α_j 时，对偶问题的约束变为

$$\alpha_i y_i + \alpha_j y_j = -\sum_{k \neq i,j} \alpha_k y_k, \quad \alpha_i \geq 0, \quad \alpha_j \geq 0$$

用一个变量表示另一个变量，再代入对偶问题中可得一个**单变量**的二次规划，该问题具有**闭式解**. （舍弃不符合约束条件的负数解）

- 偏移项 b ：通过支持向量方程来确定 $y_i f(\mathbf{x}_i) = 1$
- 最终判决 $y = \mathrm{sign}\left[f(\mathbf{x}_i) \right]$ ， $\mathrm{sign}[x]$表示取 x 符号

解得稀疏性

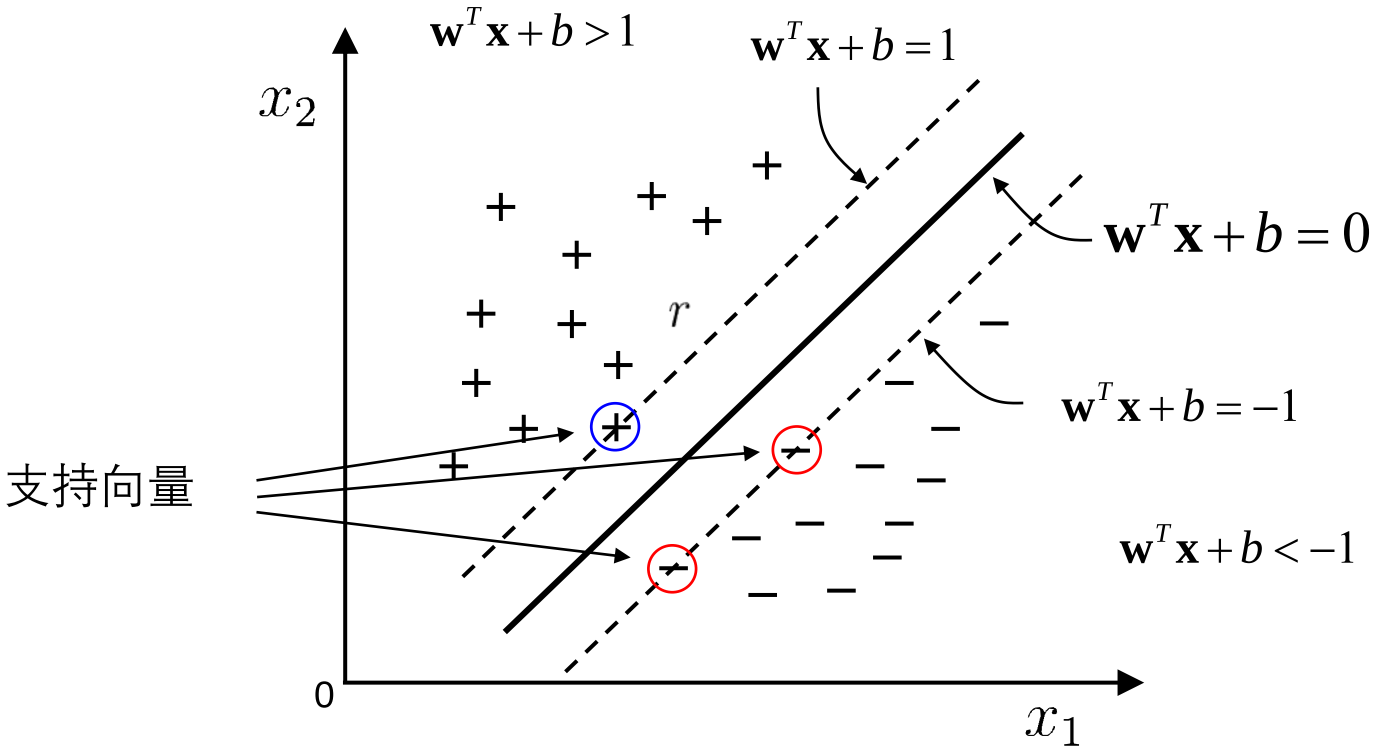
- 由KKT条件知，最优解存在的必要条件：

$$\begin{cases} y_i f(\mathbf{x}_i) - 1 \geq 0 \\ \alpha_i \geq 0 \\ \alpha_i \left(y_i f(\mathbf{x}_i) - 1 \right) = 0 \end{cases}, \quad i = 1, \ldots, n$$

- 因为大多数点都是类内点，即 $y_i f(\mathbf{x}_i) > 1$，因此多数的$\alpha_i = 0$

支持向量机解的**稀疏性**: 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


$\mathbf{w}^T\mathbf{x}+b>1$
$\mathbf{w}^T\mathbf{x}+b=1$
$\mathbf{w}^T\mathbf{x}+b=0$
$\mathbf{w}^T\mathbf{x}+b=-1$
$\mathbf{w}^T\mathbf{x}+b<-1$
x_2
x_1
r
0
支持向量

小结一下

原始问题

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

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1,2,\dots,n$$

对偶问题

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1,2,\dots,n$$

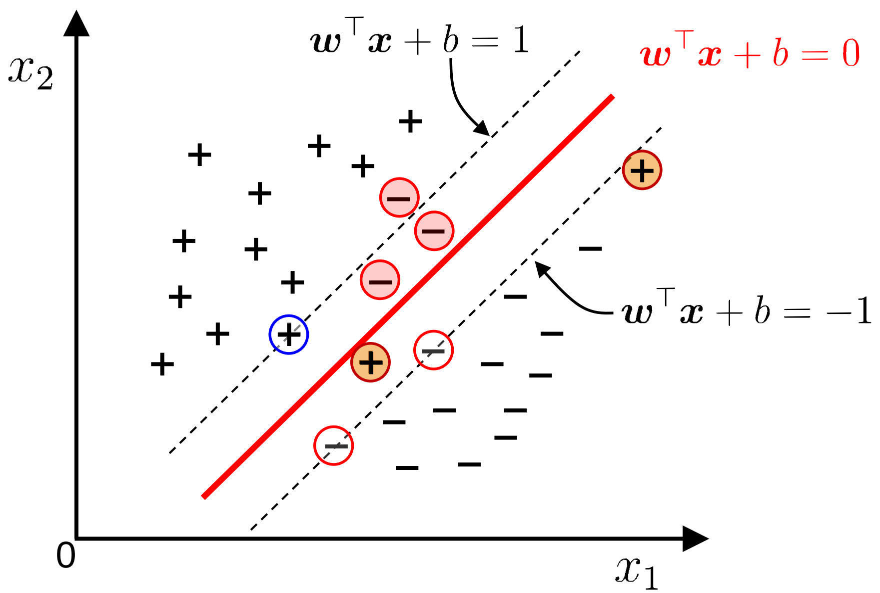
$$\boldsymbol{\alpha}^* = \left(\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*\right)$$

预测函数

$$y = \text{sign}\left(f(\mathbf{x}_i)\right) = \text{sign}\left(\mathbf{w}^T\mathbf{x} + b\right), \quad \mathbf{w} = \sum_{i=1}^{n}\alpha_i^* y_i \mathbf{x}_i$$

线性不可分情况

- 现实问题中多数情况是线性不可分的



颜色标记的为不满足分类约束的样本

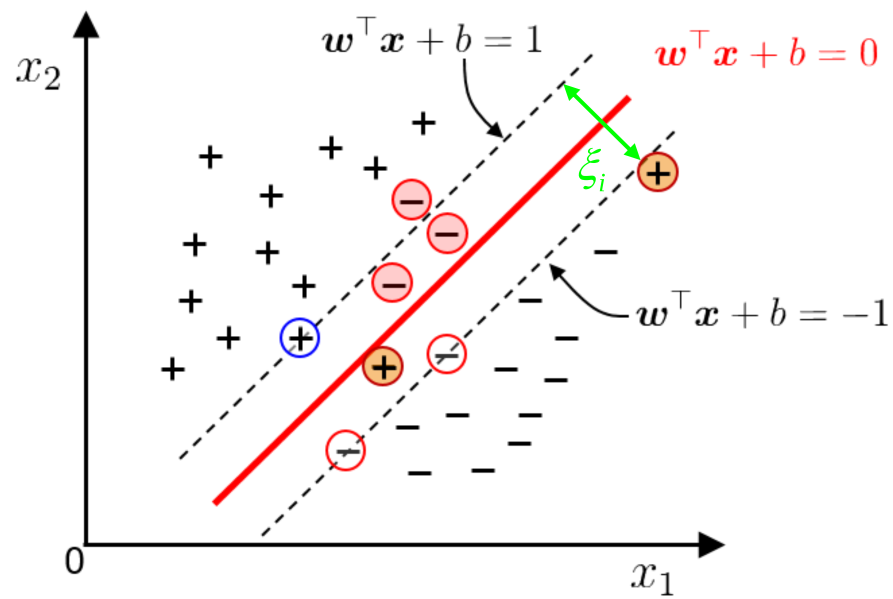
线性不可分情况

- 引入松弛变量

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

$$s.t. \quad y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i$$

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



ξ_i 表示第 i 个样本不满足

分类约束条件的程度(松弛性)

线性不可分情况

- 构建拉格朗日乘子

$$L(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i\left(y_i(\mathbf{w}^T\mathbf{x}_i+b)-1\right) - \sum_{i=1}^{n}\beta_i\xi_i$$

- 对偶问题

全内积形式出现

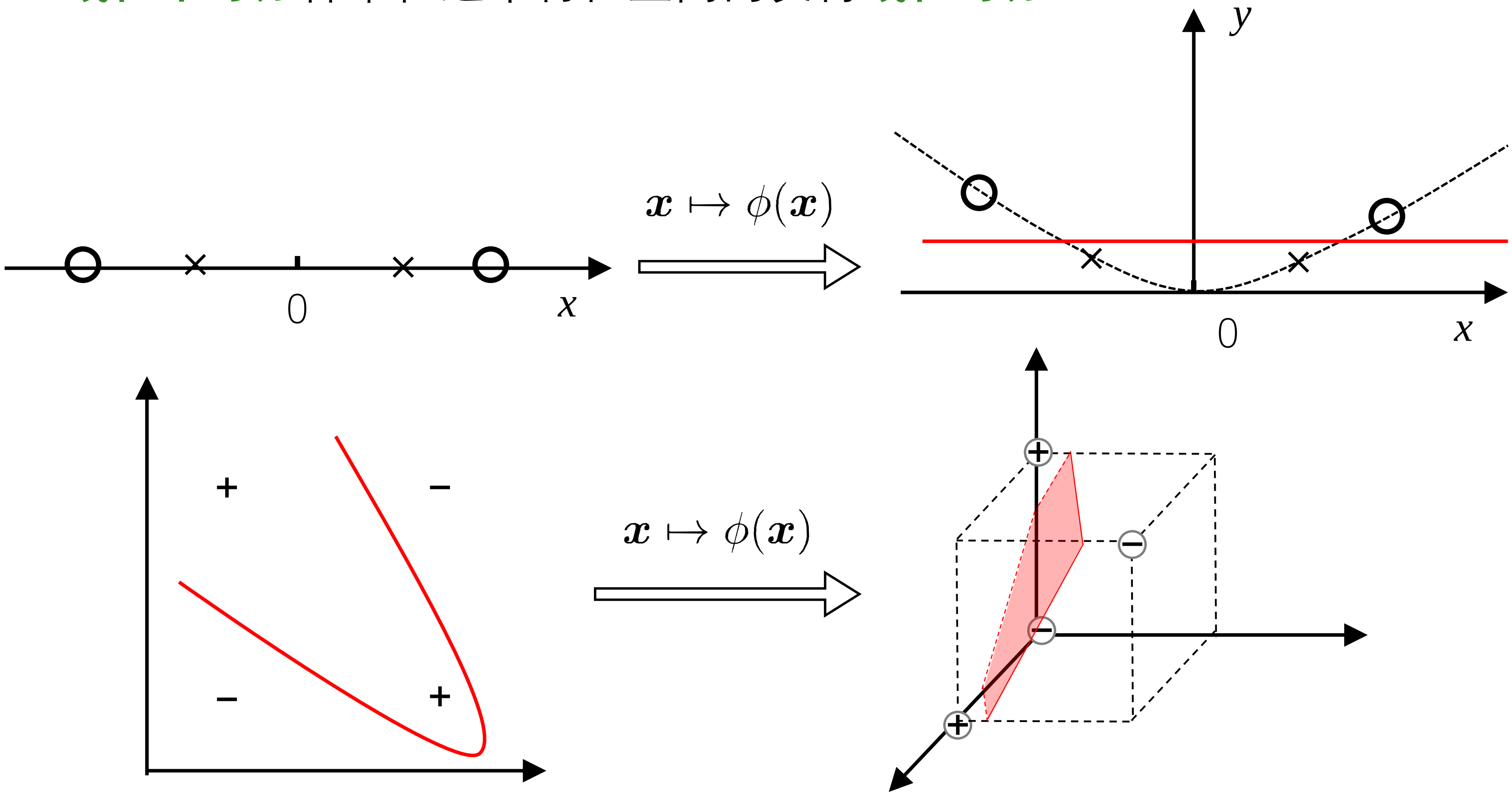
$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boxed{\mathbf{x}_i^T\mathbf{x}_j} - \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \le \alpha_i \boxed{\le C,} \quad i=1,2,\ldots,n$$

- 求解与线性可分类似，通过对偶问题求解（SMO中舍弃不满足约束的解）

特征映射

- 将样本从原始**输入空间**映射到一个更高维的**特征空间**，使得输入空间中**线性不可分**样本在这个特征空间内变得**线性可分**



特征映射后的SVM

- 设样本 $\mathbf{x}$ 映射后的向量为 $\phi(\mathbf{x})$，划分超平面为 $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$

原始问题
$$\begin{cases} \arg\max\limits_{\mathbf{w},b} \ \dfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad y_i\left(\mathbf{w}^T\phi\left(\mathbf{x}_i\right)+b\right) \geq 1, \ i=1,2,\ldots,n \end{cases}$$

对偶问题
$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} \dfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boxed{\phi\left(\mathbf{x}_i\right)^T \phi\left(\mathbf{x}_j\right)} - \sum_{i=1}^{n}\alpha_i \\ \text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i=1,2,\ldots,n \end{cases}$$

预测
$$f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{j=1}^{n}\alpha_i y_i \boxed{\phi\left(\mathbf{x}_i\right)^T \phi\left(\mathbf{x}\right)} + b$$

只以内积的形式出现，不需计算出映射，只要设计核函数 $k(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})^T\phi(\mathbf{y})$

核函数与核矩阵

- 核函数: $k:\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$
  - 对称 $k(\mathbf{x},\mathbf{y}) = k(\mathbf{y},\mathbf{x})$
  - 半正定特性 $\sum_{i=1}^{n} c_i c_j k(\mathbf{x}_i,\mathbf{x}_j) \geq 0$ ，对任意点序列 $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ 和数列 $c_1,c_2,\ldots,c_n$
- 任一核函数 k 都可以找到一个映射函数 $\phi(\mathbf{x})$，从而可以把 k 写成唯一的内积形式 $k(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = \phi(\mathbf{x})^T \phi(\mathbf{y})$
- 常见核函数

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

核函数与核矩阵

- 核矩阵：核函数 k 在点列 $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ 上的采样矩阵称为其对应的核矩阵

$$\mathbf{K}=\begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & k(\mathbf{x}_1,\mathbf{x}_2) & \cdots & k(\mathbf{x}_1,\mathbf{x}_n) \\ k(\mathbf{x}_2,\mathbf{x}_1) & k(\mathbf{x}_2,\mathbf{x}_2) & \cdots & k(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_n,\mathbf{x}_1) & k(\mathbf{x}_n,\mathbf{x}_2) & \cdots & k(\mathbf{x}_n,\mathbf{x}_n) \end{bmatrix}=\begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} & \cdots & \mathbf{K}_{1n} \\ \mathbf{K}_{21} & \mathbf{K}_{22} & \cdots & \mathbf{K}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_{n1} & \mathbf{K}_{n2} & \cdots & \mathbf{K}_{nn} \end{bmatrix}$$

- 半正定矩阵：对称且所有特征值大于等于0
- 高斯核函数为例

$$\mathbf{K}_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\frac{\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{2\sigma^2}\right)$$

基于核的SVM

原始问题

$$\begin{cases} \arg\max\limits_{\mathbf{w},b} \dfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad y_i\left(\mathbf{w}^T\phi\left(\mathbf{x}_i\right)+b\right)\geq 1,\ i=1,2,\ldots,n \end{cases}$$

对偶问题

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} \dfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{K}_{ij} - \sum_{i=1}^{n}\alpha_i \\ \text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i=1,2,\ldots,n \end{cases}$$

预测 $f(\mathbf{x})=\mathbf{w}^T\phi(\mathbf{x})+b=\sum_{j=1}^{n}\alpha_i y_i k\left(\mathbf{x}_i,\mathbf{x}\right)+b$

总结

线性可分

$$\begin{cases} \underset{\mathbf{w},b}{\arg\min} \ \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, 2, \ldots, n \end{cases}$$

线性不可分

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, n \end{cases}$$

特征映射

$$\begin{cases} \underset{\mathbf{w},b}{\arg\max} \ \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad y_i\left(\mathbf{w}^T\phi\left(\mathbf{x}_i\right) + b\right) \geq 1, \ i = 1, 2, \ldots, n \end{cases}$$

支持向量机软件包

- LIBSVM
  http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- LIBLINEAR
  http://www.csie.ntu.edu.tw/~cjlin/liblinear/
- SVMlight 、SVMperf 、SVMstruct
  http://svmlight.joachims.org/svm_struct.html
- Pegasos
  http://www.cs.huji.ac.il/~shais/code/index.html