SICHUAN UNIVERSITY

机器学习-第十四章 非线性支持向量机

教师：胡俊杰 副教授

邮箱：hujunjie@scu.edu.cn

1.凸优化概述

■ 优化问题的一般形式

$$minimize\ f_0(x)$$

待优化的目标函数

$$subject\ to\ f_i(x) \boxed{\leq} 0, i = 1,2, \dots, m$$
$$(s.t.)$$

m个不等式约束

$$h_i(x) = 0, i = 1,2, \dots, p$$

p个等式约束

1.凸优化概述

- 凸集（Convex set）：如果连接集合C中任意两点的线段都在C内，则C为凸集，即

  对于$x_1, x_2 \in C$,且$0 \le \theta \le 1$

  $\theta x_1 + (1-\theta)x_2 \in C$

- 凸函数（Convex function）： $\boldsymbol{dom}f$是凸集，对于所有$x, y \in \boldsymbol{dom}f$,且$0 \le \theta \le 1$，以下不等式均成立

  $$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y)$$

- 凸函数 vs 非凸函数

  简单　　复杂

1.凸优化概述

$$minimize\ f_0(x)$$

$$s.t.\ f_i(x) \le 0, i = 1,2,\dots,m$$

$$h_i(x) = 0, i = 1,2,\dots,p$$

凸优化

- $\boldsymbol{dom}f$是凸集， 目标函数$f_0(x)$和不等式约束$f_i(x)$为凸函数，等式约束$h_i(x)$ 为仿射函数，凸优化的目标在于找到全局最优解$x^* \in \boldsymbol{dom}f$，使得对任意$x \in \boldsymbol{dom}f$，$f(x^*) \le f(x)$均成立

- 可行解（feasible solution）：满足所有约束条件的解
- 最优解（optimal solution）：满足所有约束条件，且对任意$x \in \boldsymbol{dom}f$，$f(x^*) \le f(x)$均成立

1.凸优化概述

$$minimize\ f_0(x)$$

$$s.t.\ f_i(x) \le 0, i = 1,2,\dots,m$$

$$h_i(x) = 0, i = 1,2,\dots,p$$

- 拉格朗日函数
  - 为每个约束指定一个拉格朗日乘子，以乘子为加权系数将约束增加到目标函数中

$$L(x,\lambda,v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)$$

$$L(x,\lambda,v): R^n \times R_+^m \times R^p \to R$$

$$x \in R^n \quad \lambda \in R_+^m, v \in R^p$$

1.凸优化概述

$$minimize\ f_0(x)$$
$$s.t.\ f_i(x) \le 0, i = 1,2,\dots,m$$
$$h_i(x) = 0, i = 1,2,\dots,p$$

- 拉格朗日函数

$$L(x,\lambda,v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)$$

- 拉格朗日对偶函数
  - 对拉格朗日函数$L(x,\lambda,v)$中的x取下确界可定义拉格朗日对偶函数

$$g(\lambda,v) = \inf_{x\in R^n} L(x,\lambda,v) = \inf_{x\in R^n}\left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)\right)$$

- 拉格朗日对偶函数 $g(\lambda,v): R_+^m \times R^p \to R$

inf (infimum): 下确界，数学分析中的概念，小于等于集合中的所有成员的最大实数

$$\inf\{x \in R: 0 < x < 1\} = 0$$

sup (supremum) : 上确界，大于等于集合中所有成员的最小实数

$$\sup\{x \in R: 0 < x < 1\} = 1$$

1.凸优化概述

■ 拉格朗日对偶函数

$$g(\lambda, v) = \inf_{x \in R^n} L(x, \lambda, v) = \inf_{x \in R^n} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x) \right)$$

- 拉格朗日对偶函数是凹函数，无论原问题是否为凸问题
- 拉格朗日对偶函数给出了原问题最优值的下界：$g(\lambda, v) \leq p^*$， p^*：原问题（primal problem）的最优值（optimal value）

1.凸优化概述

拉格朗日对偶函数

$$g(\lambda, v) = \inf_{x \in R^n} L(x, \lambda, v) = \inf_{x \in R^n} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x) \right)$$

从拉格朗日对偶函数获得的下界中，哪个是最优的?

当$g(\lambda, v) = -\infty$，则其提供的下界无实际意义

拉格朗日对偶问题

$$\max_{\lambda \geq 0, v} g(\lambda, v) = \max_{\lambda \geq 0, v} \inf_{x \in R^n} L(x, \lambda, v)$$

- 假设拉格朗日对偶问题（dual problem）的最优值为d^*

1.凸优化概述

■ 弱对偶

$$d^* \leq p^*$$

- 拉格朗日对偶函数给出了原问题最优值的下界：$g(\lambda, v) \leq p^*$
- 拉格朗日对偶问题：$\max\limits_{\lambda \geq 0, v} g(\lambda, v)$

■ 强对偶

$$d^* = p^*$$

$$\begin{aligned} & minimize\ f_0(x) \\ & s.t.\ f_i(x) \leq 0, i = 1,2, \ldots, m \\ & \qquad h_i(x) = 0, i = 1,2, \ldots, p \end{aligned}$$

- 如果$f_0(x), \ldots, f_m(x)$为凸函数，通常情况下强对偶成立
- 强对偶成立的一般条件：Slater条件

■ 对偶间隙

$$p^* - d^*$$

1.凸优化概述

$$minimize\ f_0(x)$$

$$s.t.\ f_i(x) \le 0, i = 1,2,\dots,m$$

$$h_i(x) = 0, i = 1,2,\dots,p$$

■ 拉格朗日函数

$$L(x,\lambda,v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)$$

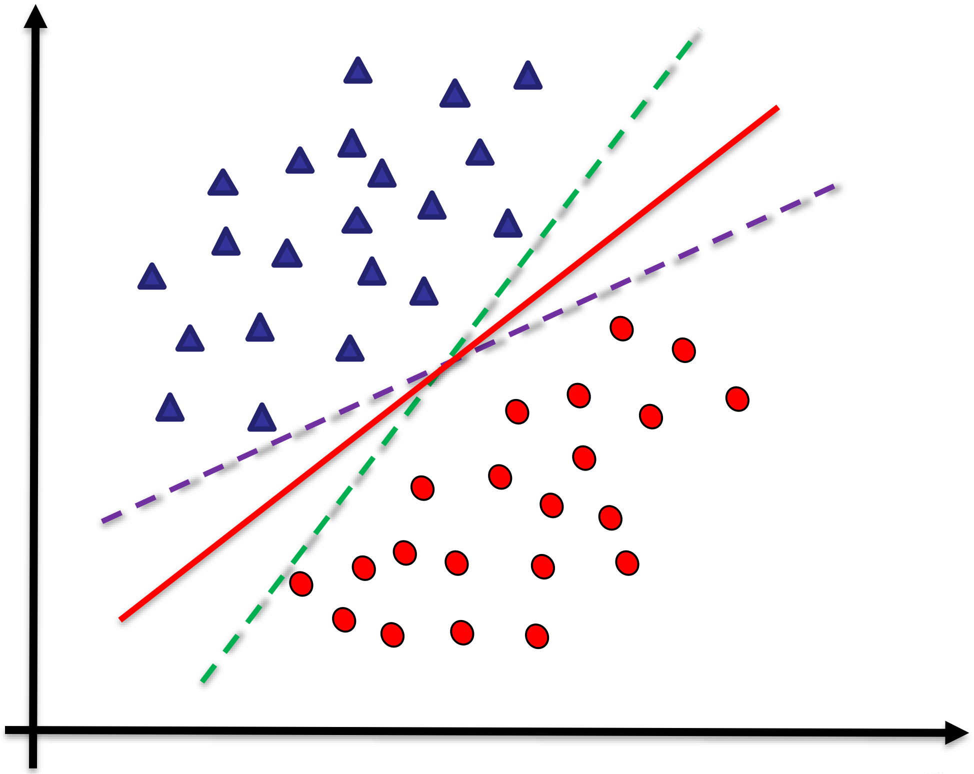
■ KKT条件

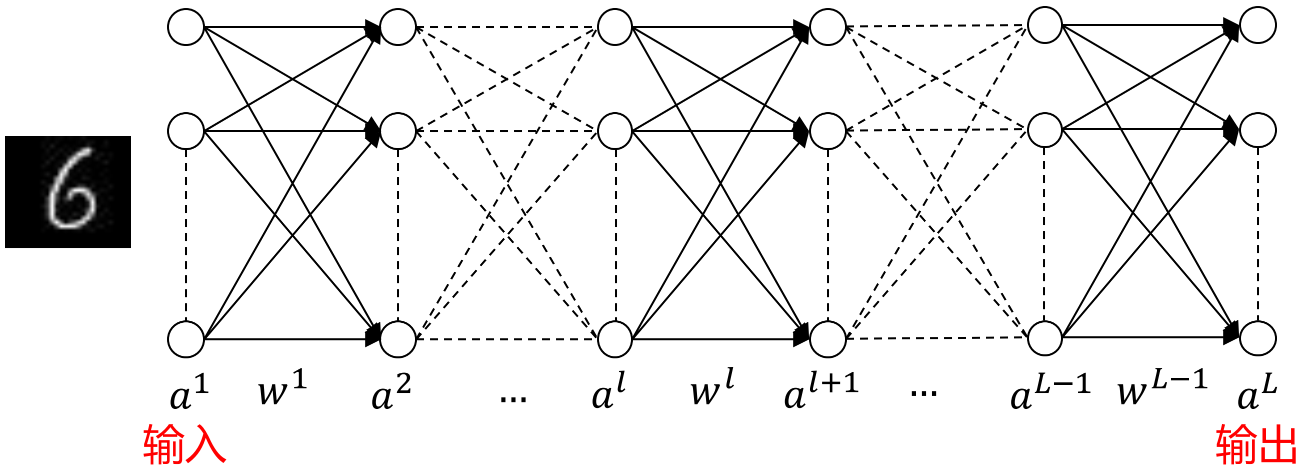
x^*和(λ^*, v^*)分别是原问题和对偶问题的最优解，且对偶间隙为0，则

$$\begin{cases} \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(x^*) = 0 & \text{稳定性条件} \\ f_i(x^*) \le 0, i = 1,\dots,m & \text{原始可行性条件} \\ h_i(x^*) = 0, i = 1,\dots,p & \text{原始可行性条件} \\ \lambda_i^* \ge 0, i = 1,\dots,m & \text{对偶可行性条件} \\ \lambda_i^* f_i(x^*) = 0, i = 1,\dots,m & \text{互补松弛条件} \end{cases}$$

- 当原问题为凸问题时，且不等式约束为凸函数，等式约束为仿射变换，则KKT条件为充要条件，且对偶间隙为0 。即满足KKT条件的解为最优解，最优解一定满足KKT条件

2.支持向量机概述





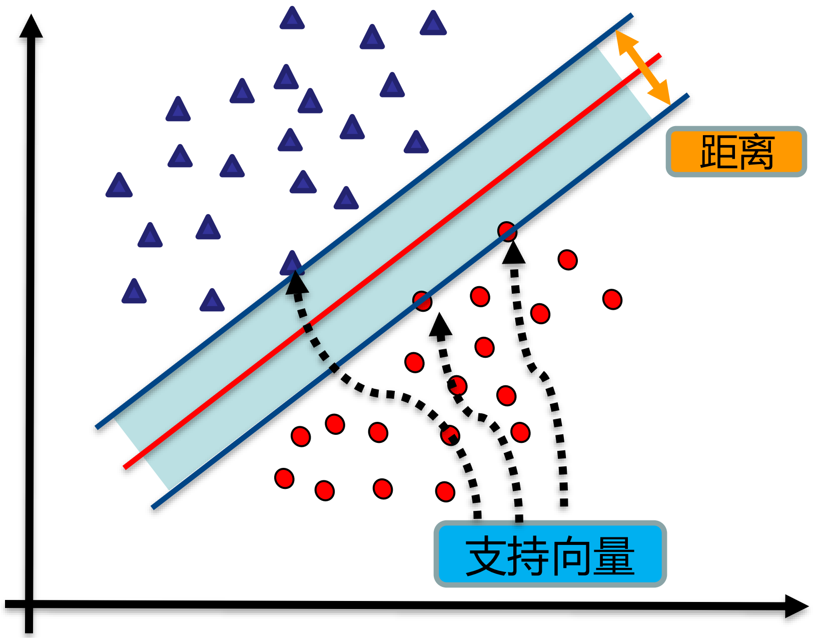
- 红色的决策线对输入扰动更加鲁棒（robust）
- 神经网络也存在鲁棒性问题，典型代表：对抗样本

2.支持向量机概述

支持向量机（Support Vector Machine，SVM）

是一类以监督学习方式对数据进行二分类的分类模型

- SVM核心思想是寻找一个分类超平面，使得样本点与超平面的距离最大化
- SVM也被称为最大间隔分类器（Large Margin Classifier）
- 支持向量（Support Vector）：距离超平面最近的点



2.支持向量机概述

背景知识

任意超平面可以用下面这个线性方程来描述：

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

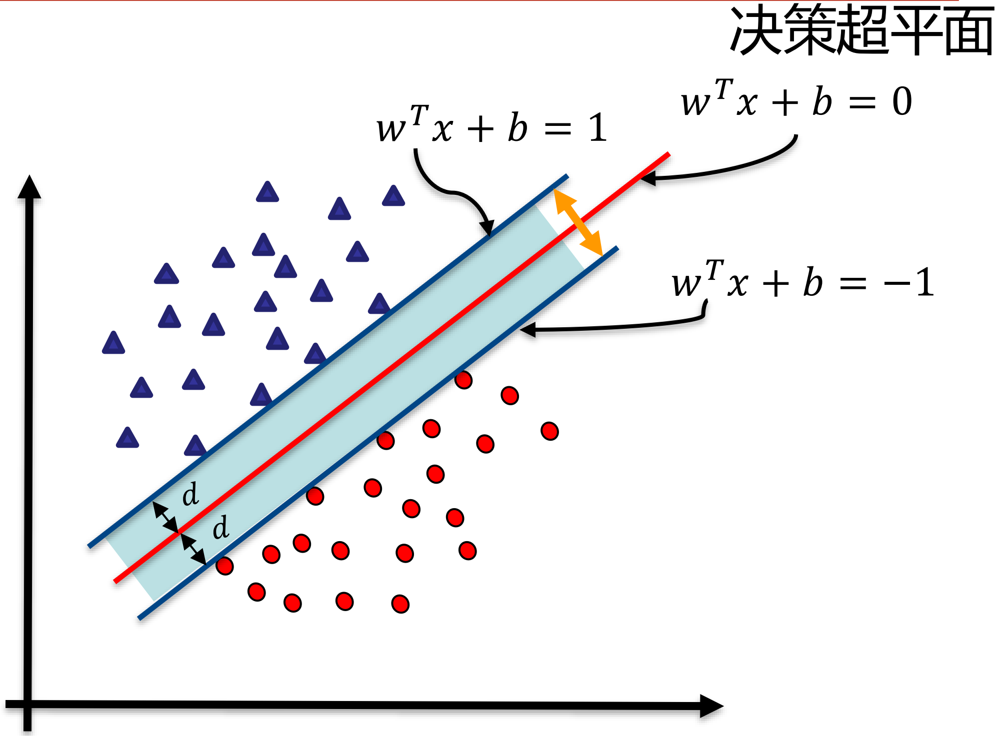
二维空间点 (x, y)到直线 $Ax + By + C = 0$的距离公式是：

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

扩展到 n 维空间后，点 $x = (x^{(1)}, x^{(2)} \ldots x^{(\mathrm{n})})$ 到超平面

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

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



如图所示，根据支持向量的定义，假设支持向量到决策超平面的距离为 d，其他点到决策超平面的距离大于 d

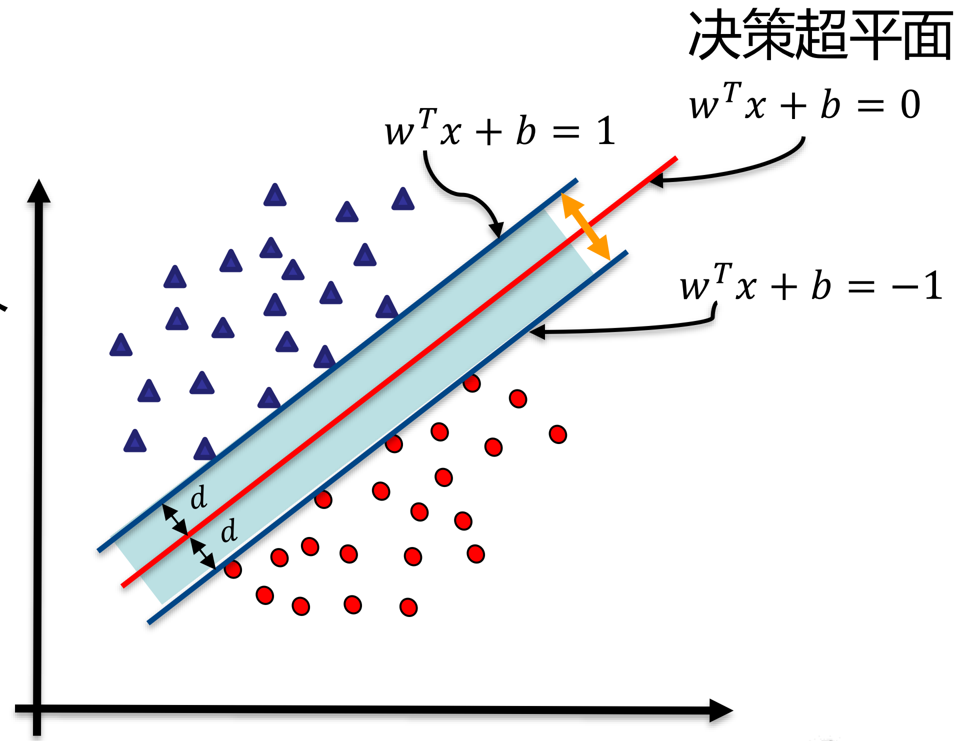
2.支持向量机概述

- 给定训练数据集$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$，其中

  $x_i \in R^n, y_i \in \{+1, -1\}, i = 1,2, \ldots, N$

假设超平面(w, b)将线性可分数据集正确分类，即对于$(x_i, y_i) \in D$，若$y_i = +1$，则$w^T x_i + b > 0$；若$y_i = -1$，则$w^T x_i + b < 0$，令

$$\begin{cases} w^T x_i + b \geq +1, y_i = +1 \\ w^T x_i + b \leq -1, y_i = -1 \end{cases}$$



- 将以上两个方程合并，可得：$y_i(w^T x_i + b) \geq 1$

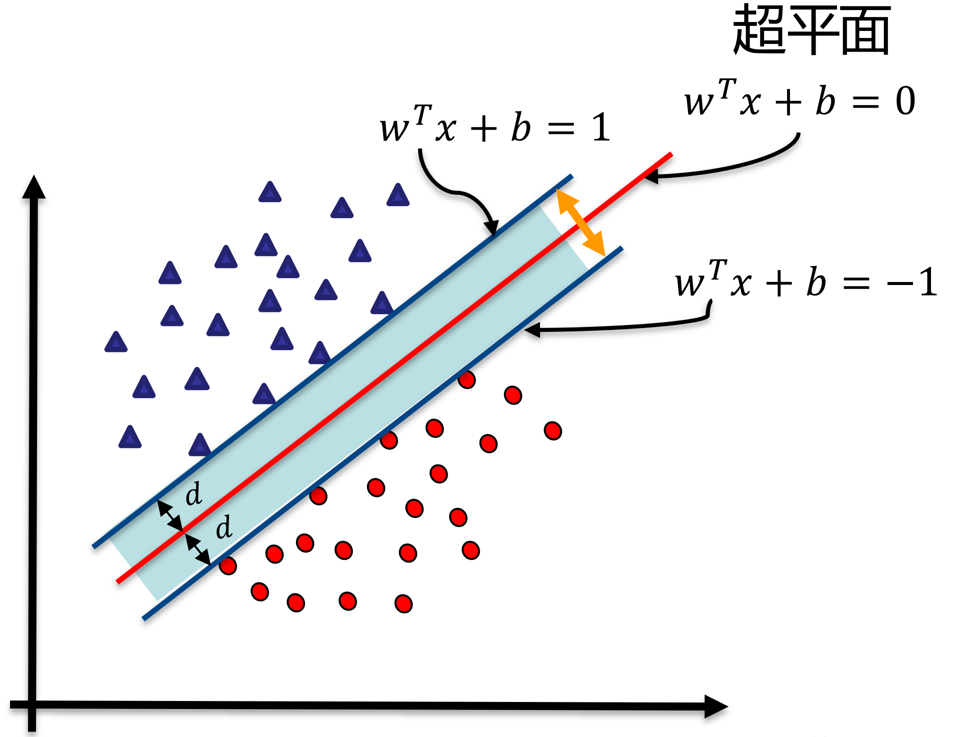
- $y_i(w^T x_i + b) \geq 1$为约束条件，即要求决策超平面(w, b)将所有样本分类正确

2.支持向量机概述

- 支持向量到超平面的距离可以写为：$d=\frac{|w^Tx+b|}{||w||}=\frac{1}{||w||}$
- 两类（正类和负类）支持向量到超平面的距离之和为$\gamma=\frac{2}{||w||}$，也被称为“间隔”，即优化目标

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

$$s.t.\ y_i(w^Tx_i+b)\geq 1, i=1,2,\dots,N$$



2.支持向量机概述

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

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

$$\Updownarrow$$

$$\min_{w,b} \frac{1}{2} ||w||^{\boxed{2}}$$ 方便后续求导

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

- 即支持向量机（SVM）的基本形式

3.支持向量机求解

原优化问题：$$\min_{w,b} \frac{1}{2}\|w\|^2$$

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

拉格朗日函数：$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i (w^T x_i + b) + \sum_{i=1}^{N} \alpha_i$

- 线性可分SVM满足强对偶条件，即$d^* = p^*$，原问题最优值等于对偶问题最优值
- 通过解对偶问题，得到最优解$\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_N^*)$，得到原问题的最优解$(w^*, b^*)$

3.支持向量机求解

求 $\min_{w,b} L(w,b,\alpha)$ $\quad L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i y_i(w^T x_i+b)+\sum_{i=1}^{N}\alpha_i$

$$\frac{\partial}{\partial w}L(w,b,\alpha)=w-\sum_{i=1}^{N}\alpha_i\, y_i x_i=0 \quad\Rightarrow\quad w=\sum_{i=1}^{N}\alpha_i\, y_i x_i$$

$$\frac{\partial}{\partial b}L(w,b,\alpha)=-\sum_{i=1}^{N}\alpha_i\, y_i=0 \quad\Rightarrow\quad \sum_{i=1}^{N}\alpha_i\, y_i=0$$

3.支持向量机求解

$$\boxed{w=\sum_{i=1}^{N}\alpha_i\, y_i x_i \qquad \sum_{i=1}^{N}\alpha_i\, y_i=0}$$ 代入 $$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i y_i(w^T x_i+b)+\sum_{i=1}^{N}\alpha_i$$

$$\min_{w,b} L(w,b,\alpha)=\frac{1}{2}w^T w-\sum_{i=1}^{N}\alpha_i y_i(w^T x_i+b)+\sum_{i=1}^{N}\alpha_i$$

$$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i y_i\left(\left(\sum_{j=1}^{N}\alpha_j\, y_j x_j\right)\cdot x_i+b\right)+\sum_{i=1}^{N}\alpha_i$$

$$=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i$$

3.支持向量机求解

构建拉格朗日对偶问题

求 $\min\limits_{w,b} L(w,b,\alpha)$ 对α的极大

$$\max_{\alpha}\min_{w,b} L(w,b,\alpha) = \max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i$$

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

$$\alpha_i \geq 0, i = 1,2,\ldots,N$$

最小化

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i$$

3.支持向量机求解

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i$$

$$s.t. \sum_{i=1}^{N}\alpha_i y_i = 0 \qquad \alpha_i \geq 0, i = 1,2,\dots,N$$

设$\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_N^*)$为拉格朗日对偶问题最优解，则原始问题最优解$w^*$和$b^*$如下

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

$$\boxed{b^* = ?}$$

■ KKT条件

x^*和(λ^*, v^*)分别是原问题和对偶问题的最优解，且强对偶成立，则

$$\begin{cases} \nabla f_0(x^*) + \sum_{i=1}^{m}\lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(x^*) = 0 \\ f_i(x^*) \leq 0, i = 1, \dots, m \\ h_i(x^*) = 0, i = 1, \dots, p \\ \lambda_i^* \geq 0, i = 1, \dots, m \\ \boxed{\lambda_i^* f_i(x^*) = 0, i = 1, \dots, m} \end{cases}$$

$$\alpha_i^*(y_i(w^T x_i^* + b^*) - 1) = 0$$

3.支持向量机求解

KKT条件

x^*和(λ^*, v^*)分别是原问题和对偶问题的最优解，且强对偶成立，则

$$
\begin{cases}
\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(x^*) = 0 \\
f_i(x^*) \le 0, i = 1, \dots, m \\
h_i(x^*) = 0, i = 1, \dots, p \\
\lambda_i^* \ge 0, i = 1, \dots, m \\
\boxed{\lambda_i^* f_i(x^*) = 0, i = 1, \dots, m}
\end{cases}
$$

$$\alpha_i^*(y_i(w^T x_i^* + b^*) - 1) = 0$$

至少存在一个$\alpha_j^* > 0$，（反证法：若α^*均为0，则$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i = 0$，然而$w^* = 0$不是原始优化问题的解），因此

$$y_j(w^{*T} x_j^* + b^*) - 1 = 0$$

将$w^* = \sum_{i=1}^{N} \alpha_i y_i x_i$代入上式，并利用$y_j^2 = 1$可得

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j)$$

3.支持向量机求解

线性可分支持向量机学习算法

第1步：根据原始优化问题，写出拉格朗日函数

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

第2步：求$\min\limits_{w,b} L(w,b,\alpha)$，并代入$L(w,b,\alpha)$

$$\min_{w,b} L(w,b,\alpha)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i$$

第3步：求解拉格朗日对偶问题，即

$$\max_{\alpha}\min_{w,b} L(w,b,\alpha)=\max_{\alpha}\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i$$

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

$$\alpha_i\geq 0, i=1,2,\dots,N$$

求得最优解$\alpha^*=(\alpha_1^*,\alpha_2^*,\dots,\alpha_N^*)$

第4步：根据KKT条件可得原优化问题最优解w^*和b^*

$$w^*=\sum_{i=1}^{N}\alpha_i y_i x_i \qquad b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x_j)$$

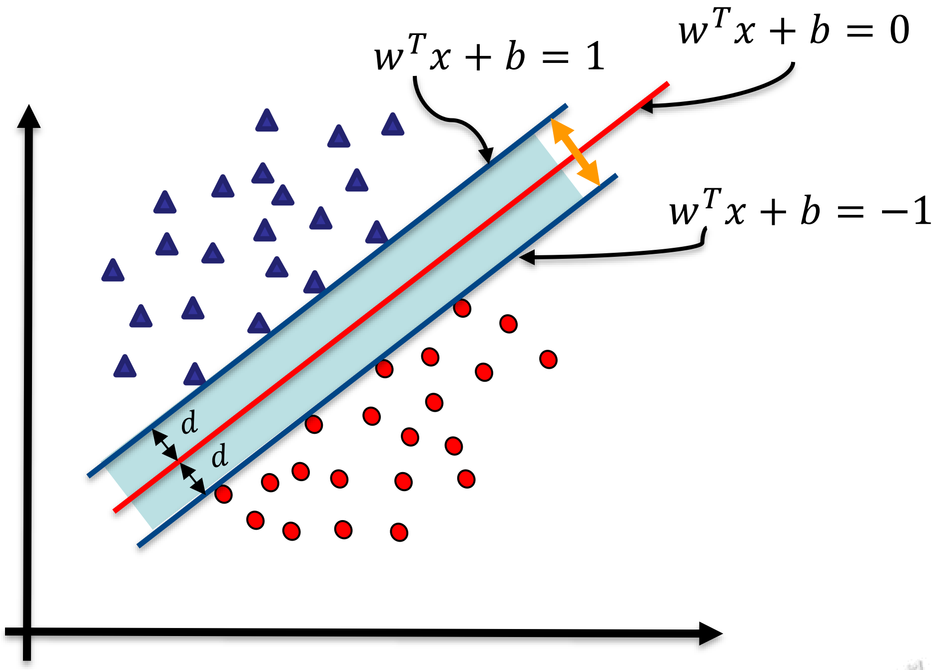
第5步：构建决策超平面以及分类决策函数

决策超平面：$w^{*T}x+b^*=0$

决策函数：$f(x)=sign(w^{*T}x+b^*)$

本章目录

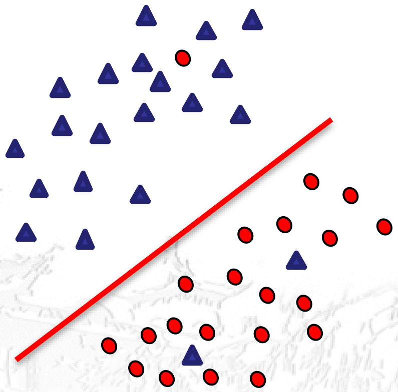
1.线性支持向量机



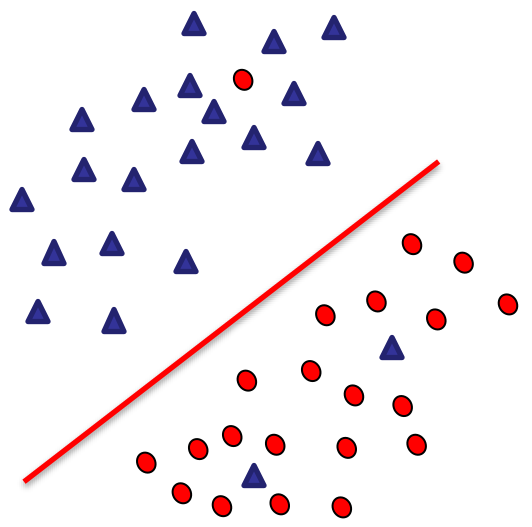
$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t.\, y_i(w^Tx_i + b) \geq 1, i = 1,2, \dots, N$$

- 以上假设训练样本是线性可分的，即存在一个超平面将两类样本完全分开
- 然而现实情况很复杂（如标签错误），应允许支持向量机对部分样本出错



1.线性支持向量机



- 允许一些样本不满足约束 $y_i(w^T x_i + b) \geq 1$

- 在最大化间隔的同时，不满足约束的样本应尽可能少，目标如下

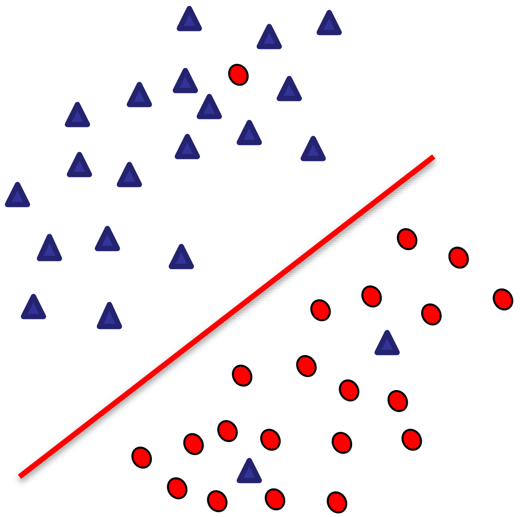
$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} l_{0/1}(y_i(w^T x_i + b) - 1)$$

- $C > 0$为人为指定的惩罚参数，$l_{0/1}$代表0/1损失函数

$$l_{0/1} = \begin{cases} 1, if\ z < 0 \\ 0, otherwise \end{cases}$$

- 常量C允许部分样本不满足约束 $y_i(w^T x_i + b) \geq 1$

1.线性支持向量机

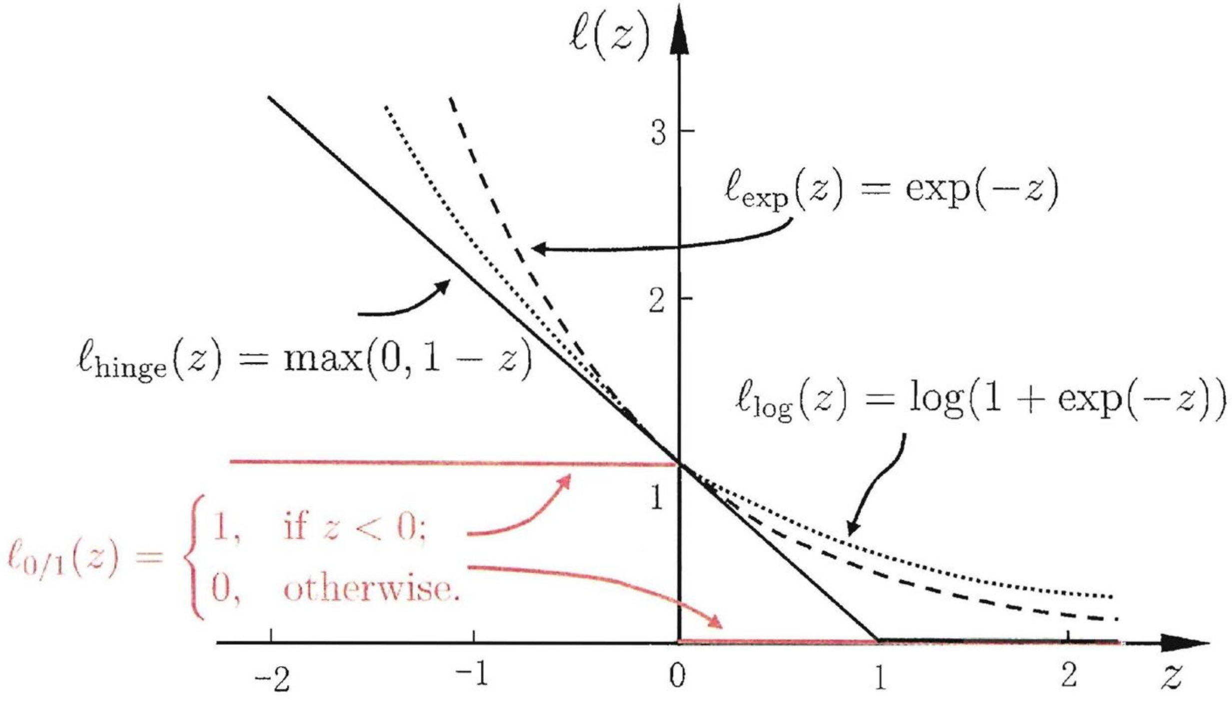


$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} l_{0/1}(y_i(w^T x_i + b) - 1)$$

- $l_{0/1}$非凸、不连续，以上目标函数不易求解。用数学性质较好的“代理损失函数”（surrogate loss function），替代$l_{0/1}$
- 常用代理损失函数:
  - hinge损失函数（hinge loss）：$l_{hinge}(z) = \max(0,1-z)$
  - 指数损失函数（exponential loss）：$l_{exp}(z) = e^{-z}$
  - 对率损失函数（logistic loss）：$l_{log}(z) = \log(1+e^{-z})$

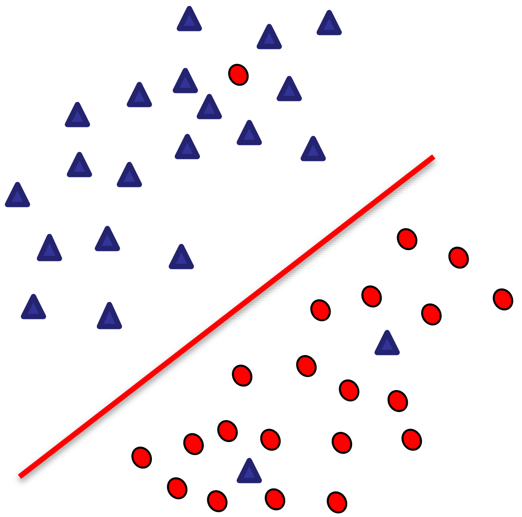
1.线性支持向量机

- 常用代理损失函数：
  - hinge损失函数（hinge loss）： $l_{hinge}(z) = \max(0,1-z)$
  - 指数损失函数（exponential loss）： $l_{exp}(z) = e^{-z}$
  - 对率损失函数（logistic loss）： $l_{log}(z) = \log(1+e^{-z})$



hinge, 铰链

1.线性支持向量机



$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} l_{0/1}(y_i(w^T x_i + b) - 1)$$

$l_{hinge}(z) = \max(0,1 - z)$ 代替$l_{0/1}$

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \max\left(0,1 - y_i(w^T x_i + b)\right)$$

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

(/ksaɪ/)

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

$$s.t.\, y_i(w^T x_i + b) \geq 1 - \xi_i$$

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

优化目标及约束

1.线性支持向量机

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

$$s.t.\, y_i(w^T x_i + b) \geq 1 - \xi_i$$

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

1. 写出拉格朗日函数$L(w,b,\xi,\alpha,\mu)$

2. 求$L(w,b,\xi,\alpha,\mu)$对w,b,ξ的极小，并代入$L(w,b,\xi,\alpha,\mu)$

1.线性支持向量机

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

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

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

1. 写出拉格朗日函数$L(w,b,\xi,\alpha,\mu)$

$$L(w,b,\xi,\alpha,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\big(1-\xi_i-y_i(w^T x_i+b)\big) - \sum_{i=1}^{N}\mu_i\xi_i$$

其中$\alpha_i \geq 0, \mu_i \geq 0$为拉格朗日乘子

2. 求$L(w,b,\xi,\alpha,\mu)$对w,b,ξ的极小，并代入$L(w,b,\xi,\alpha,\mu)$

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

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

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

代入拉格朗日函数，得

$$\min_{w,b,\xi} L(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) + \sum_{i=1}^{N}\alpha_i$$

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

$$C - \alpha_i - \mu_i = 0$$

$$\alpha_i \geq 0$$

$$\mu_i \geq 0, i = 1,2,\ldots,N$$

1.线性支持向量机

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$s.t.\ y_i(w^T x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0, i = 1,2,\dots,N$$

3. 拉格朗日对偶问题

$$\max_{\alpha}\min_{w,b,\xi} L(w,b,\xi,\alpha,\mu) = \max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i$$

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

$$\left.\begin{array}{l} C - \alpha_i - \mu_i = 0 \\ \alpha_i \geq 0 \\ \mu_i \geq 0, i = 1,2,\dots,N \end{array}\right\} \text{消去}\mu_i \quad \left.\begin{array}{l} C - \alpha_i \geq 0 \\ \\ \alpha_i \geq 0 \end{array}\right\} \text{合并} \quad 0 \leq \alpha_i \leq C$$

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i$$
$$s.t.\sum_{i=1}^{N}\alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C, i = 1,2,\dots,N$$

1.线性支持向量机

线性支持向量机的拉格朗日对偶问题

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

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

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

C为为人为指定的
惩罚参数

线性可分支持向量机的拉格朗日对偶问题

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

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

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

- 设$\alpha^*=(\alpha_1^*,\alpha_2^*,\ldots,\alpha_N^*)$为拉格朗日对偶问题最优解，则原问题最优解$w^*$和$b^*$如下

$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i \qquad b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x_j)$$

由KKT中的互补松弛条件可得

1.线性支持向量机

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

超平面：$w^{*T} x + b^* = 0$ $\longrightarrow$ $\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_i) + b^* = 0$

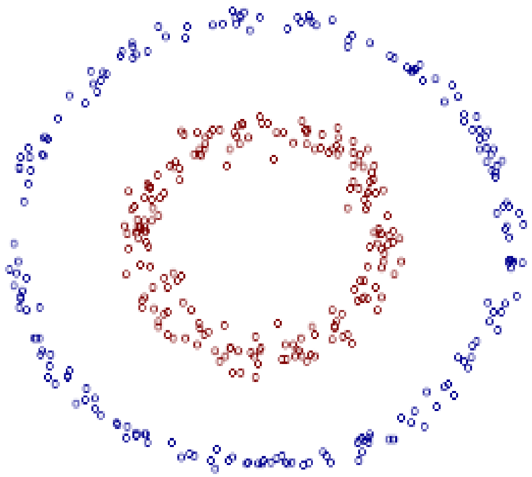
$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j)$$

决策函数：$f(x) = sign(w^{*T} x + b^*)$ $\longrightarrow$ $f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_i) + b^*\right)$

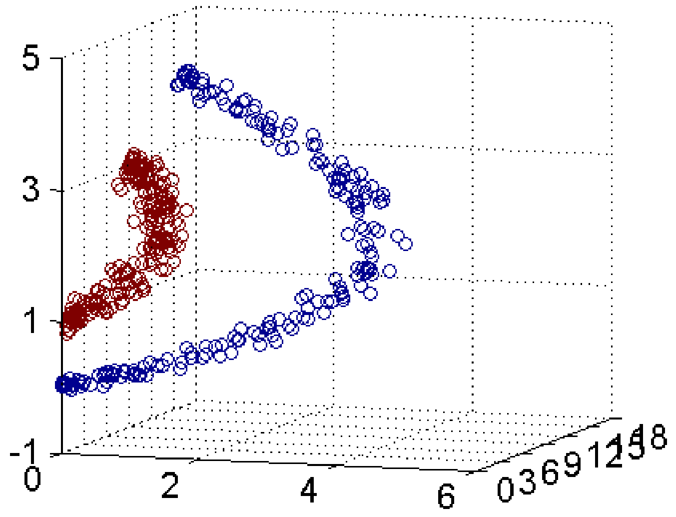
其中 $b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j)$

本章目录

3.非线性支持向量机



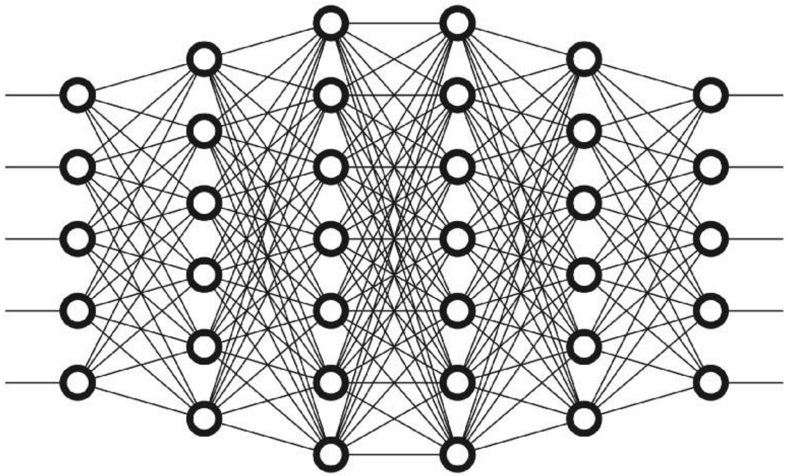
线性不可分



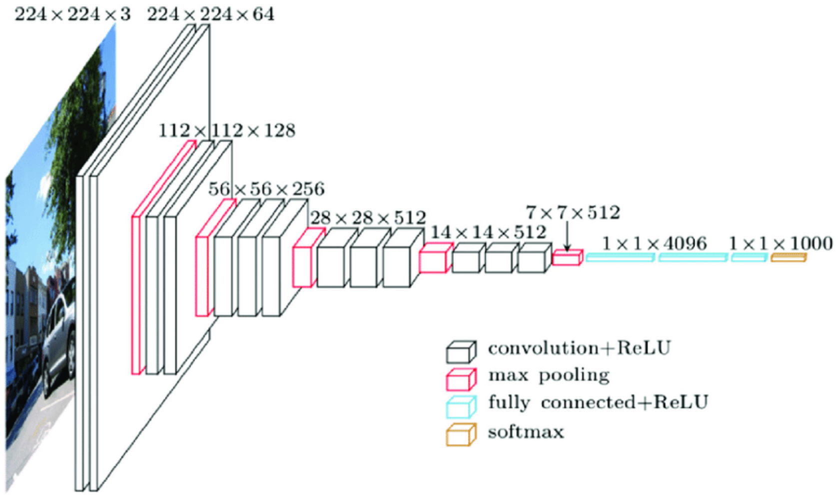
高维下可分

- 无法找到一个超平面将两类样本分开
- 解决思路：对输入x作用非线性变换ϕ，将x 从原始空间映射至高维空间，使得$\phi(x)$可分

3.非线性支持向量机



全连接神经网络



卷积神经网络（CNN）

- 随着层数增加，隐藏层的神经元数量增加
- 随着层数增加，特征的通道数（channel）增加

- 高维空间的特征具有更强的可分能力
- 对于CNN而言，简单地提高特征的通道数即可有效提高性能

ResNet ⟶ Wide ResNet

3.非线性支持向量机

线性可分支持向量机

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

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

拉格朗日对偶问题如下:

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i$$

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

$$\alpha_i \geq 0, i = 1,2,\dots,N$$

非线性支持向量机，优化目标及约束如下:

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

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

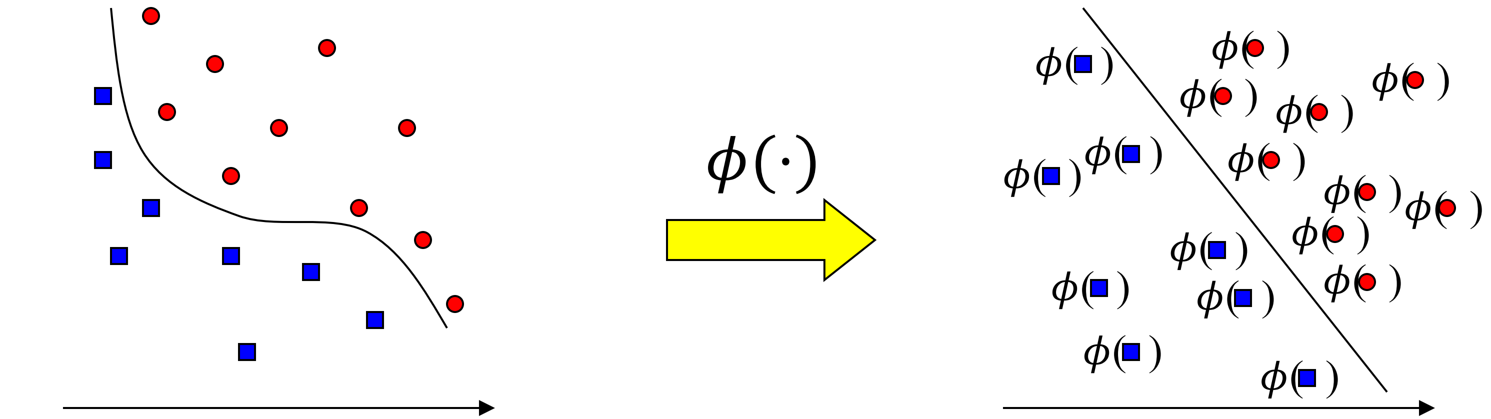
拉格朗日对偶问题如下:

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (\phi(x_i) \cdot \phi(x_j)) + \sum_{i=1}^{N}\alpha_i$$

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

$$0 \leq \alpha_i \leq C, i = 1,2,\dots,N$$

3.非线性支持向量机



输入空间线性不可分　　　　特征空间线性可分

3.非线性支持向量机

非线性支持向量机，优化目标及约束如下:

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

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

拉格朗日对偶问题如下:

$$\max_{\alpha}\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\big(\phi(x_i)\cdot\phi(x_j)\big)+\sum_{i=1}^{N}\alpha_i$$

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

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

- 直接定义非线性映射函数ϕ较困难
- 引入核函数$K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$，$x_i$与$x_j$在特征空间中的内积等于在原始输入空间通过函数$K(x_i,x_j)$计算的结果，从而避开定义$\phi$

3.非线性支持向量机

非线性支持向量机，优化目标及约束如下:

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

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

拉格朗日对偶问题如下:

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\big(\phi(x_i)\cdot\phi(x_j)\big)+\sum_{i=1}^{N}\alpha_i$$

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

$$0 \leq \alpha_i \leq C, i=1,2,\dots,N$$

引入核函数后，拉格朗日对偶问题如下:

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

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

$$0 \leq \alpha_i \leq C, i=1,2,\dots,N$$

3.非线性支持向量机

线性可分支持向量机

$$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i \qquad b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j)$$

超平面： $$\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_i) + b^* = 0$$

决策函数： $$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_i) + b^*\right)$$

非线性支持向量机

$$w^* = \sum_{i=1}^{N} \alpha_i^* y_i \phi(x_i) \qquad b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i \phi(x_i) \cdot \phi(x_j)$$

$$K(x_i, x_j)$$

超平面： $$\sum_{i=1}^{N} \alpha_i^* y_i \phi(x) \cdot \phi(x_i) + b^* = 0$$

$$K(x, x_i)$$

决策函数： $$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i \phi(x) \cdot \phi(x_i) + b^*\right)$$

$$K(x, x_i)$$

3.非线性支持向量机

常用核函数有：

线性核函数

$$K(x_i, x_j) = x_i^T x_j$$

多项式核函数

$$K(x_i, x_j) = (x_i^T x_j)^d$$

$d \geq 1$，为多项式的次数

高斯核函数

$$K(x_i, x_j) = exp(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2})$$

$\sigma > 0$，为高斯核的带宽（width ）

- 这三个常用的核函数中，只有高斯核函数需要调参 σ
- 核函数的选择依赖于经验

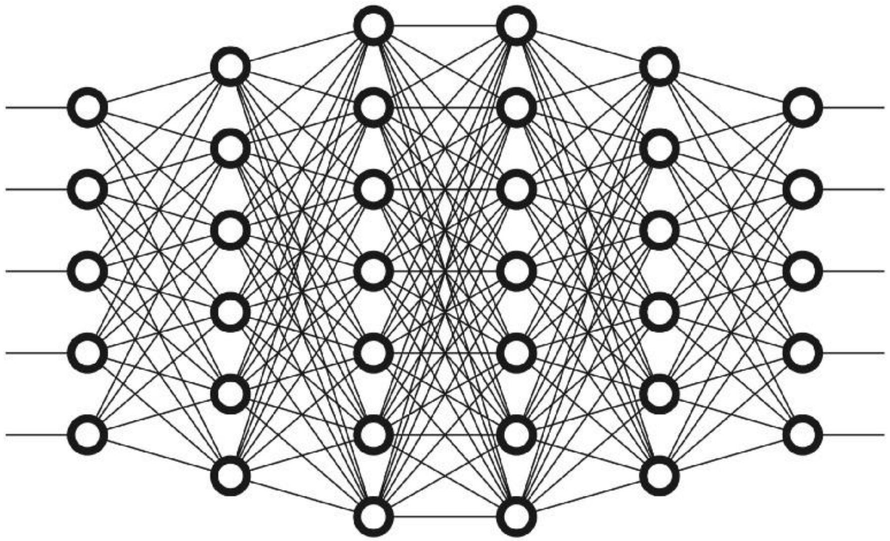
3.非线性支持向量机

$$K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$$

$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i\phi(x_i) \qquad b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i\underbrace{\phi(x_i)\cdot\phi(x_j)}_{K(x_i,x_j)}$$

超平面： $$\sum_{i=1}^{N}\alpha_i^* y_i\underbrace{\phi(x)\cdot\phi(x_i)}_{K(x,x_i)}+b^*=0$$

决策函数： $$f(x)=sign\left(\sum_{i=1}^{N}\alpha_i^* y_i\underbrace{\phi(x)\cdot\phi(x_i)}_{K(x,x_i)}+b^*\right)$$



- 深度神经网络本质是一个复杂的非线性函数F，$F: x\rightarrow y$
- 通用（万能）逼近定理（Universal Approximation Theorem）：神经网络可以近似任意的复杂函数，并且可以达到任意近似精度

本章目录

3.序列最小优化算法

拉格朗日对偶问题:

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

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

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

- 如何求解α^*?

- 序列最小优化算法（Sequential Minimal Optimization）：每次选择α_i和α_j，并固定其他参数，求α_i和α_j的极值，直至收敛。求解思路如下:
  - 选择一对需要更新的变量α_i和α_j
  - 固定α_i和α_j之外的参数，求∇_{α_i}和∇_{α_j}

- SMO算法的两个主要部分:
  - 求解两个变量α_i和α_j的解析方法
  - 选择待优化变量α_i和α_j的方法

3.序列最小优化算法

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

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

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

求解两个变量α_i和α_j

- 记$K_{ij} = K(x_i,x_j)$
- 选择α_1和α_2为待优化变量，其余视为常量

$$\min_{\alpha_1,\alpha_2} W(\alpha_1,\alpha_2) = \frac{1}{2}K_{11}\alpha_1^2 + \frac{1}{2}K_{22}\alpha_2^2 + y_1 y_2 K_{12}\alpha_1\alpha_2 -(\alpha_1+\alpha_2) + y_1\alpha_1\sum_{i=3}^{N} y_i\alpha_i K_{i1} + y_2\alpha_2\sum_{i=3}^{N} y_i\alpha_i K_{i2}$$

$$s.t. \quad \alpha_1 y_1 + \alpha_2 y_2 = -\sum_{i=3}^{N} y_i\alpha_i = \tau$$

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

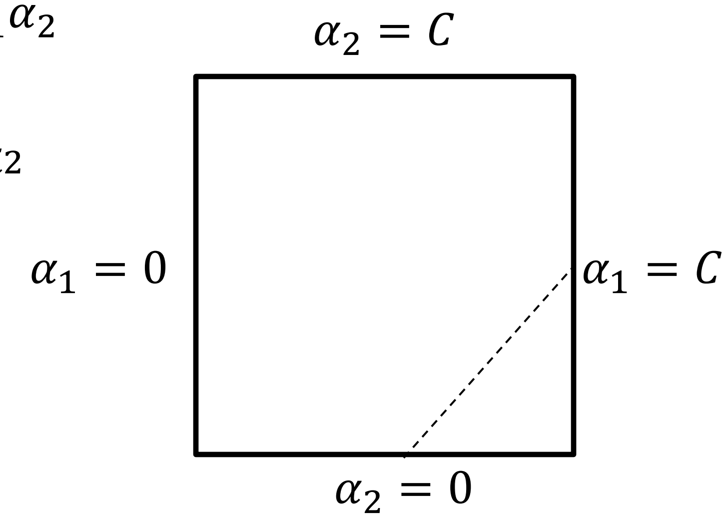
3.序列最小优化算法

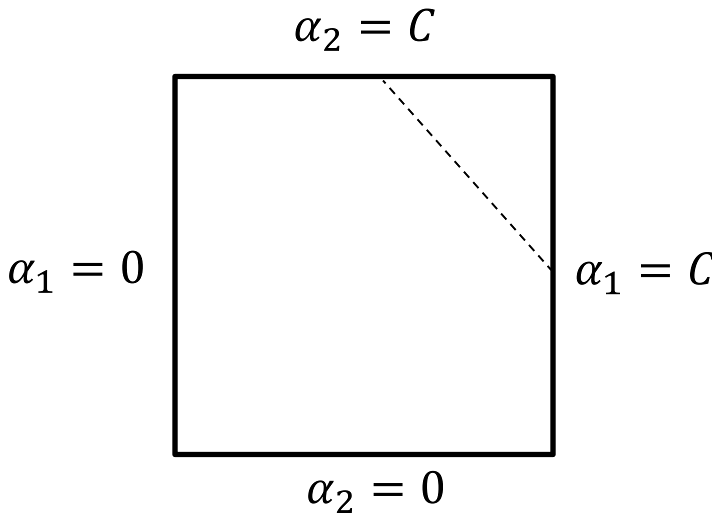
$$\min_{\alpha_1,\alpha_2} W(\alpha_1,\alpha_2)=\frac{1}{2}K_{11}\alpha_1^2+\frac{1}{2}K_{22}\alpha_2^2+y_1y_2K_{12}\alpha_1\alpha_2$$

$$-(\alpha_1+\alpha_2)+y_1\alpha_1\sum_{i=3}^{N}y_i\alpha_iK_{i1}+y_2\alpha_2\sum_{i=3}^{N}y_i\alpha_iK_{i2}$$

$$s.t.\quad \alpha_1y_1+\alpha_2y_2=-\sum_{i=3}^{N}y_i\alpha_i=\tau$$

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

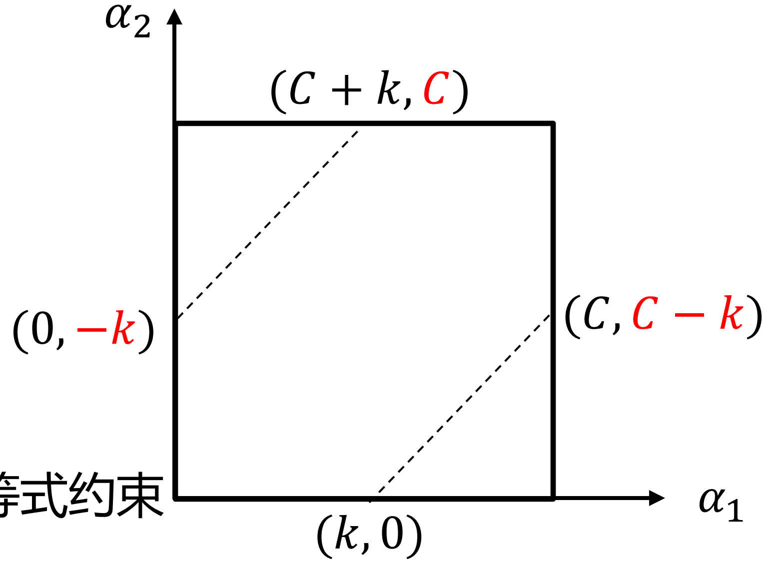


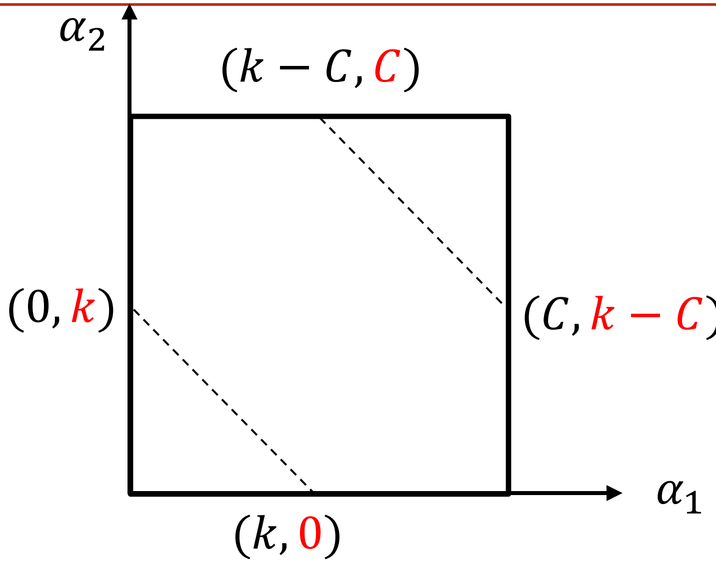


- $y_1\neq y_2, \alpha_2-\alpha_1=k$
- $y_1=y_2, \alpha_1+\alpha_2=k$

- 设初始可行解为$\alpha_1^{old},\alpha_2^{old}$，最优解为$\alpha_1^{new},\alpha_2^{new}$
- 设α_2未被截断（clip）的解为$\alpha_2^{new,unc}$

3.序列最小优化算法





α_i需满足以下等式和不等式约束

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

$y_1\alpha_1 + y_2\alpha_2 = \tau$

$L \le \alpha_2^{new} \le H$

- $y_1 \ne y_2, \alpha_1 - \alpha_2 = k$

$$L = \max(0, -k) = \max\left(0, \alpha_2^{old} - \alpha_1^{old}\right)$$

$$H = \min(C, C-k) = \min\left(C, C + \alpha_2^{old} - \alpha_1^{old}\right)$$

- $y_1 = y_2, \alpha_1 + \alpha_2 = k$

$$L = \max(0, k-C) = \max\left(0, \alpha_1^{old} + \alpha_2^{old} - C\right)$$

$$H = \min(C, k) = \min\left(C, \alpha_1^{old} + \alpha_2^{old}\right)$$

3.序列最小优化算法

- 1. 先求沿着约束方向未截断的解$\alpha_2^{new,unc}$
- 2. 再根据上界（H）和下界（L）得到优化后的解α_2^{new}

预测值 $$g(x) = w^T\phi(x) + b = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b$$

误差 $$E_i = g(x_i) - y_i = \left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right) - y_i$$

3.序列最小优化算法

- 1. 求沿着约束方向未截断的解$\alpha_2^{new,unc}$

$$\alpha_2^{new,unc} = \alpha_2^{old} + \frac{y_2(E_1 - E_2)}{\eta}$$

$$\eta = K_{11} + K_{22} - 2K_{12}$$

具体推导参考教材

- 2. 根据上界（H）和下界（L）得到优化后的解α_2^{new}

$$\alpha_2^{new} = \begin{cases} H, & \alpha_2^{new,unc} > H \\ \alpha_2^{new,unc}, & L \le \alpha_2^{new,unc} \le H \\ L, & \alpha_2^{new,unc} < L \end{cases}$$

$$\begin{cases} \alpha_1^{old} y_1 + \alpha_2^{old} y_2 = \tau \\ \alpha_1^{new} y_1 + \alpha_2^{new} y_2 = \tau \end{cases}$$

$$\alpha_1^{new} y_1 + \alpha_2^{new} y_2 = \alpha_1^{old} y_1 + \alpha_2^{old} y_2$$

$$\Big\downarrow \; y^2 = 1$$

$$\alpha_1^{new} = \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^{new})$$

3.序列最小优化算法

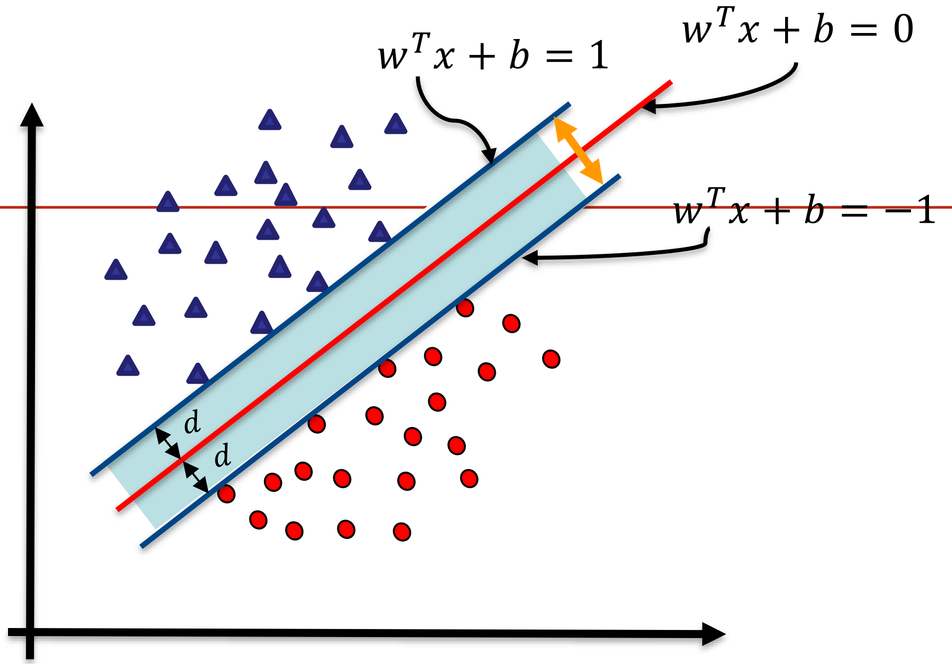
- 选择待优化变量α_i和α_j，包含两层循环
  - 外循环选择第一个变量
  - 内循环选择第二个变量
- SMO算法结束的条件： 所有α_i满足KKT条件
- KKT条件

x^*和(λ^*, v^*)分别是原问题和对偶问题的最优解，且强对偶成立，则

$$\begin{cases} \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(x^*) = 0 \\ f_i(x^*) \le 0, i = 1, \dots, m \\ h_i(x^*) = 0, i = 1, \dots, p \\ \lambda_i^* \ge 0, i = 1, \dots, m \\ \boxed{\lambda_i^* f_i(x^*) = 0, i = 1, \dots, m} \end{cases}$$

互补松弛条件

- 当原问题为凸问题时，且不等式约束为凸函数，等式约束为仿射变换，则KKT条件为充要条件，且对偶间隙为0 。即满足KKT条件的解为最优解，最优解一定满足KKT条件



$$\begin{cases} w^T \phi(x_i) + b \ge +1, y_i = +1 \\ w^T \phi(x_i) + b \le -1, y_i = -1 \end{cases}$$

- 将以上两个方程合并，可得约束条件： $y_i(w^T \phi(x_i) + b) \ge 1$
- 即 $y_i g(x_i) \ge 1$

其中： $g(x_i) = w^T \phi(x_i) + b = \sum_{j=1}^{N} \alpha_j y_j K(x_i, x_j) + b$

3.序列最小优化算法

- KKT条件

x^*和(λ^*, v^*)分别是原问题和对偶问题的最优解，且强对偶成立，则

$$
\begin{cases}
\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(x^*) = 0 \\
f_i(x^*) \le 0, i = 1, \dots, m \\
h_i(x^*) = 0, i = 1, \dots, p \\
\lambda_i^* \ge 0, i = 1, \dots, m \\
\boxed{\lambda_i^* f_i(x^*) = 0, i = 1, \dots, m} \quad \text{互补松弛条件}
\end{cases}
$$

- SVM的约束条件

$$y_i g(x_i) \ge 1$$

- 外循环：确定第1个变量，即选择违反KKT互补松弛条件最严重的样本，即：

$$0 < \alpha_i < C \leftrightarrow y_i g(x_i) = 1$$

$$\alpha_i = 0 \leftrightarrow y_i g(x_i) \ge 1$$

$$\alpha_i = C \leftrightarrow y_i g(x_i) \le 1$$

其中：$g(x_i) = w^T x_i + b = \sum_{j=1}^{N} \alpha_j y_j K(x_i, x_j) + b$

- 内循环：假设外循环已找到第1个变量α_1，选择与α_1差别最大的样本，从而使得目标函数下降最快，可选择最大$|E_1 - E_2|$对应的α_2

谢 谢！