第五章 支持向量机

§5.1 背景知识

§5.2 理论推导

§5.3 方法扩展

§5.1 背景知识

一、问题引入

二、间隔与支持向量

问题引入

感知器方法回顾

- 数 据 集：

$$\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$$

- 模　　型：

$$\hat{y} = \mathbf{sgn}(w^T x + w_0)$$

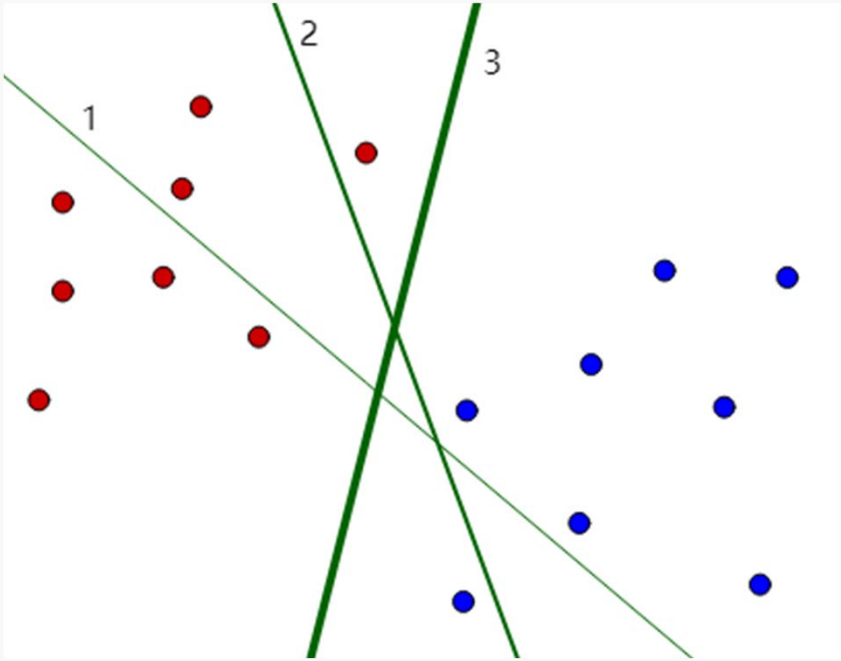
- 惩罚函数：

$$J(w, w_0) = \sum_{y_i(w^T x_i + w_0) \le 0} -y_i(w^T x_i + w_0)$$

- 优化方法：梯度下降法

$$w^{(t+1)} = w^{(t)} - \rho \nabla_w J(w, w_0)$$

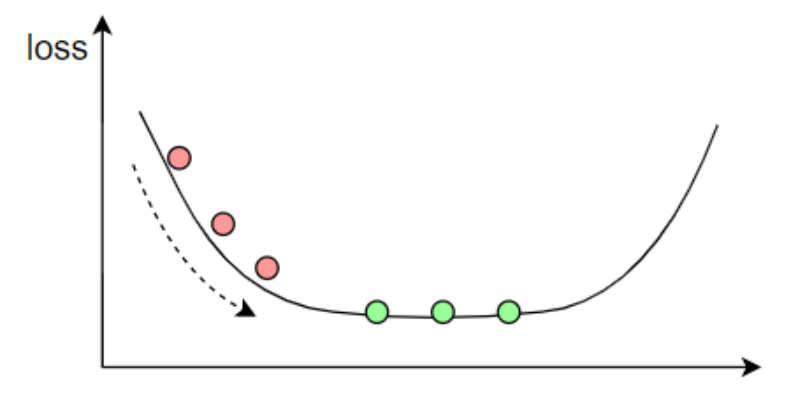
$$w_0^{(t+1)} = w_0^{(t)} - \rho \nabla_{w_0} J(w, w_0)$$

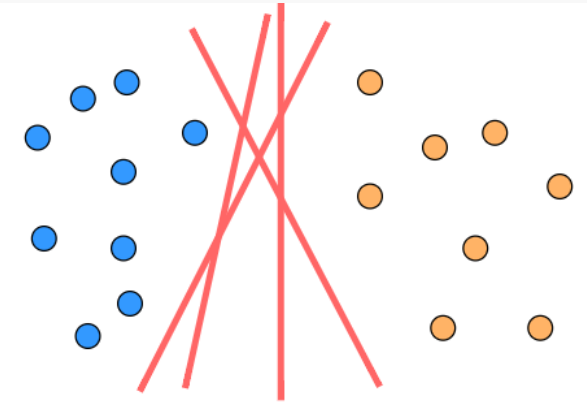


感知器收敛过程示意图

问题引入

- 感知器分类面特点
  - 当训练数据线性可分时，能得到线性分类面
  - **理论上：**满足优化目标的线性分类面通常有无数个
  - **实际上：**受初始化、学习率等因素影响，会收敛到不同解

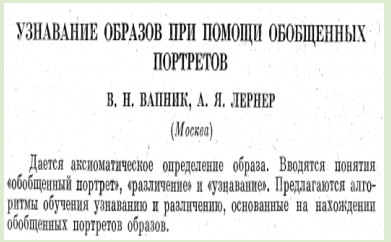




- 感知器是否存在最优分类面？

不一定存在，多数情况下不存在，最优解不能保证

问题引入


УЗНАВАНИЕ ОБРАЗОВ ПРИ ПОМОЩИ ОБОБЩЕННЫХ ПОРТРЕТОВ

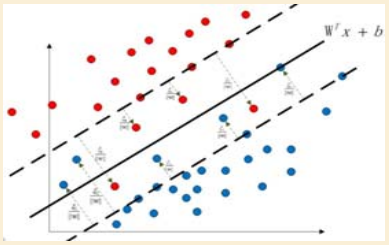
В. Н. ВАПНИК, А. Я. ЛЕРНЕР

(Москва)

Дается аксиоматическое определение образа. Вводятся понятия «обобщенный портрет», «различение» и «узнавание». Предлагаются алгоритмы обучения узнаванию и различению, основанные на нахождении обобщенных портретов образов.

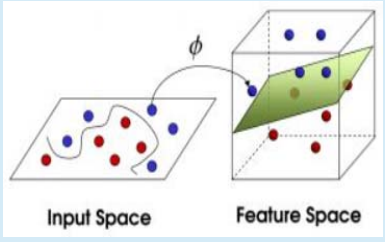
SVM由广义肖像算法发展而来

1963



松弛变量线性规划求解技术

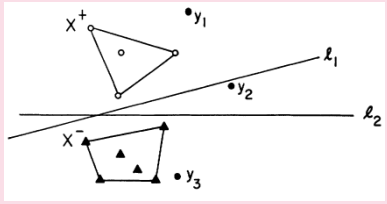
1968



通过核方法得到非线性SVM

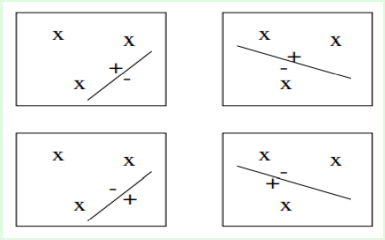
1992

1965



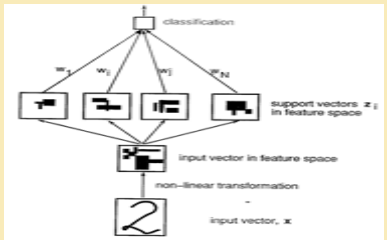
最大间隔决策的理论研究

1971



VC维理论

1995

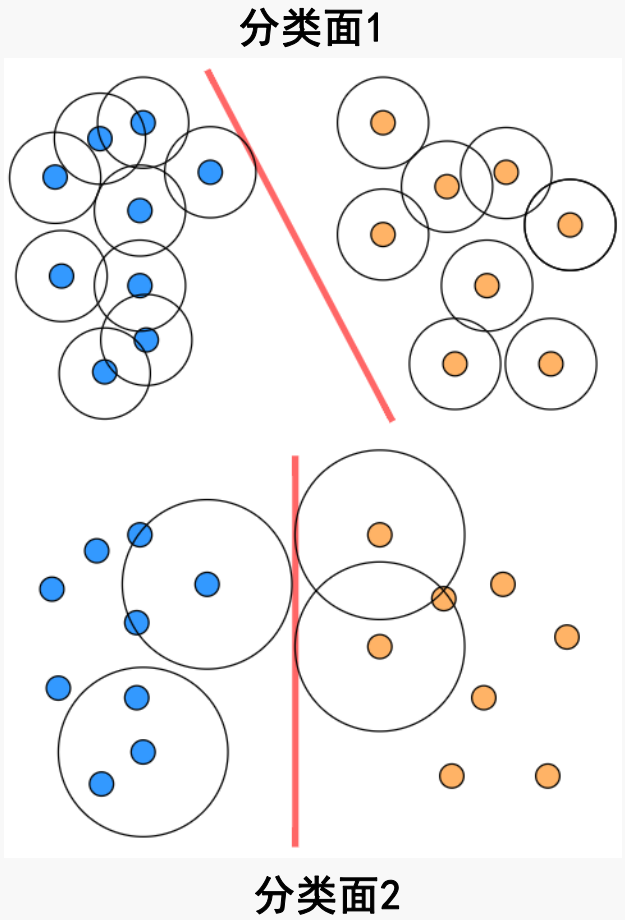


软间隔非线性SVM用于手写字符识别

V. N. Vapnik

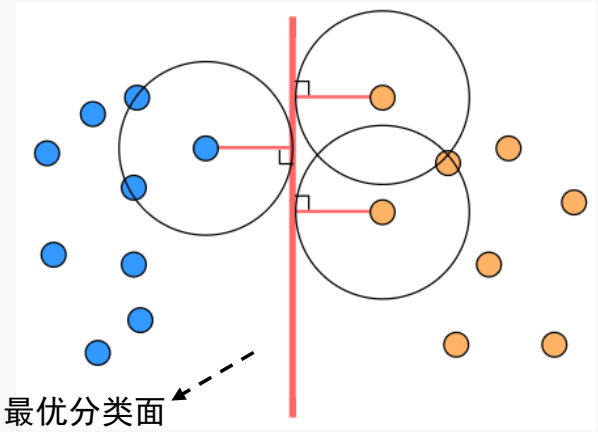
间隔与支持向量

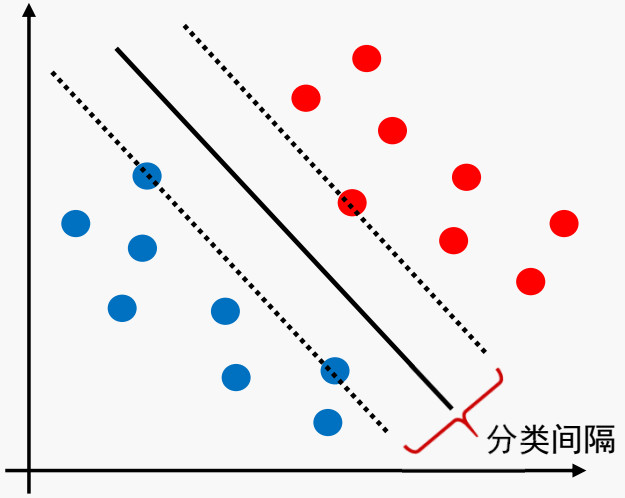
- 线性分类面的优劣
  - 一个好的分类面应该对局部扰动不敏感
  - 右图中，以超球的半径代表扰动的大小。很明显，分类面2比分类面1能“容忍”更大的扰动
  - 在线性可分的情况下，随着超球半径增大，最后可以得到唯一的线性分类面，能“容忍”最大程度的扰动
- 最优超平面
  - 定义：一个超平面，如果它能够将训练样本没有错误地分开，并且两类训练样本中离超平面最近的样本与超平面之间的距离是最大的，则把这个超平面称作为最优分类超平面



间隔与支持向量

- 支持向量：训练样本中**离超平面最近**的样本
- 分类间隔：**两个异类支持向量**到超平面的距离之和
- 最优超平面：一个超平面，如果它能够将训练样本没有错误地分开，并且两类训练样本中离超平面最近的样本与超平面之间的距离是最大，则把这个超平面称作**最优（分类）超平面**，或**最大间隔超平面**





间隔与支持向量

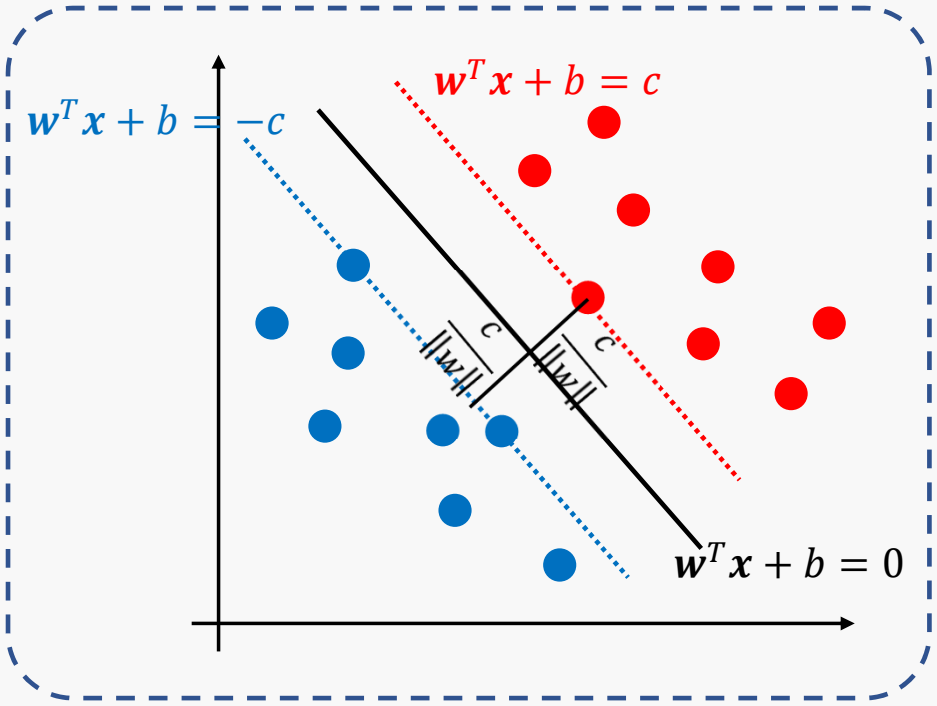
- 如何求解最优超平面呢？
  - 准备工作

1. 最优超平面的必要条件：
   异类支持向量到超平面的距离相等
2. 同一个超平面可对应无数组(w, b, c)
   为确定它们的尺度，不妨固定$c = 1$

  - 数学形式：有约束的优化问题

$$\max_{w,b} \text{Margin}(w, b) = \frac{2c}{\|w\|_2} = \frac{2}{\|w\|_2}$$

$$s.\text{t.}\ \ y_i(w \cdot x_i + b) \geq c = 1, 1 \leq i \leq n$$



分类间隔为$\frac{2c}{\|w\|}$，为什么?

间隔与支持向量

求解最优超平面

$$\max_{w,b} \text{Margin}(w,b) = \frac{2}{\|w\|_2}$$

$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, 1 \leq i \leq n$$

- 上述优化问题等价于下面的形式：

支持向量机的原问题：

$$\min_{w,b} \frac{1}{2}\|w\|_2^2$$

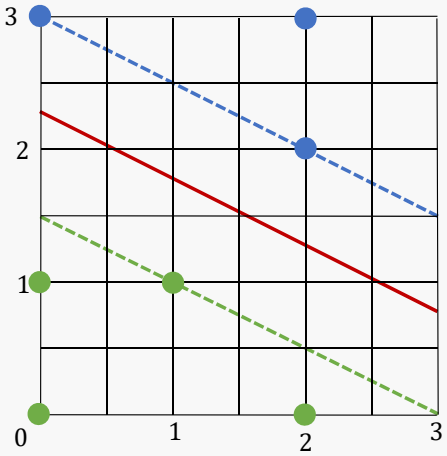
$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, 1 \leq i \leq n$$

注：$\|w\|_2^2$相比$\|w\|_2^{-1}$更容易优化，系数$1/2$仅是为了计算方便。

SVM应用举例

利用线性SVM求解分类问题

$$\omega_1: \{(1,1)^T, (0,1)^T, (2,0)^T, (0,0)^T\}$$
$$\omega_2: \{(2,2)^T, (2,3)^T, (0,3)^T\}$$



解答：通过观察法获取支持向量为

$$\omega_1: (1,1)^T \quad \omega_2: (2,2)^T, (0,3)^T$$

列式计算

$$w^T \begin{pmatrix} 1 \\ 1 \end{pmatrix} + b = 1 \quad w^T \begin{pmatrix} 0 \\ 3 \end{pmatrix} + b = -1 \quad w^T \begin{pmatrix} 2 \\ 2 \end{pmatrix} + b = -1$$

解得

$$w = \begin{pmatrix} -\frac{4}{3} \\ -\frac{2}{3} \end{pmatrix}, \qquad b = 3, \qquad -\frac{4}{3}x_1 - \frac{2}{3}x_2 + 3 = 0$$

§5.2 理论推导

一、对偶问题

二、核函数

三、软间隔与正则化

对偶问题

- 考虑如下有约束优化问题：

$$\min_{\boldsymbol{w}} \ f(\boldsymbol{w})$$
$$\text{s.t.} \ \ g_i(\boldsymbol{w}) \le 0, i = 1, \dots, k$$
$$h_i(\boldsymbol{w}) = 0, i = 1, \dots, l$$

→ 可行域$\mathcal{R}$

- 该问题的拉格朗日函数如下：

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{w}) + \sum_{i=1}^{k} \alpha_i g_i(\boldsymbol{w}) + \sum_{i=1}^{l} \beta_i h_i(\boldsymbol{w})$$
$$\text{s.t.} \quad \alpha_i \ge 0, \qquad i = 1, \dots, k$$

其中$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$为拉格朗日乘子

- 进一步考虑如下函数：

$$\theta(\boldsymbol{w}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \alpha_i \ge 0} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

说明$\theta(\boldsymbol{w})$在$\mathcal{R}$上的取值等于$f(\boldsymbol{w})$，在$\mathcal{R}$外取值为$+\infty$

对偶问题

- 如果$\boldsymbol{w}$的某个取值$\hat{\boldsymbol{w}}$不满足对$g(\boldsymbol{w})$或$h(\boldsymbol{w})$的约束条件，即：

$$\exists\, i, \text{s.t.}\ \ g_i(\hat{\boldsymbol{w}}) > 0 \ \text{ or } \ h_i(\hat{\boldsymbol{w}}) \neq 0$$

则$\theta(\boldsymbol{w})$的取值满足：

$$\theta(\hat{\boldsymbol{w}}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \left[f(\hat{\boldsymbol{w}}) + \sum_{i=1}^{k} \alpha_i g_i(\hat{\boldsymbol{w}}) + \sum_{i=1}^{l} \beta_i h_i(\hat{\boldsymbol{w}}) \right] = +\infty$$

由此易知：

$$\theta(\boldsymbol{w}) = \begin{cases} f(\boldsymbol{w}), & \text{如果}\boldsymbol{w}\text{满足约束} \\ +\infty, & \text{如果}\boldsymbol{w}\text{不满足约束} \end{cases}$$

- 由以上讨论可知，我们仅需求解 $\min_{\boldsymbol{w}} \theta(\boldsymbol{w}) = \min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ 即可。

对偶问题

- 上面两页总结：

$$\min_{w} \ f(w)$$
$$\text{s.t.} \ \ g_i(w) \le 0, i = 1, \dots, k$$
$$h_i(w) = 0, i = 1, \dots, l$$

等价于求解 $\min_{w} \max_{\alpha,\beta:\alpha_i \ge 0} \mathcal{L}(w, \alpha, \beta)$

- 在**一定条件**下，可以进一步得到：

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

且存在解(w^*, α^*, β^*)满足：

1. 使等式两边同时取得最值
2. 满足Karush-Kuhn-Tucker (KKT) 条件

很重要：想要的不是目标函数的最值，而是取最值时的参数值

对偶问题

- 上页的“一定条件”包括：
  - $f(\cdot)$和$g_i(\cdot)$为凸函数
  - $h_i(\cdot)$为线性函数
  - $\forall\, i,\ \exists w,\ \text{s.t.}\ \ g_i(w) < 0$
- KKT条件：
  - $\frac{\partial}{\partial w}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0$
  - $\frac{\partial}{\partial \beta}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0$
  - $\alpha_i^* g_i(w^*) = 0,\ \ i = 1, \dots, k$
  - $g_i(w^*) \le 0,\ \ \ \ i = 1, .., k$
  - $\alpha_i^* \ge 0,\ \ \ \ \ \ \ \ i = 1, \dots, k$

- 在SVM原问题中：
  - $f(w, b) = \frac{1}{2}\|w\|_2^2$为凸函数
  - $g_i(w, b) = 1 - y_i(w \cdot x_i + b)$为凸函数
  - 不存在等式约束条件
  - 对任意样本(x_i, y_i)，显然存在(w, b)满足$y_i(w \cdot x_i + b) > 1$

利用这个结论可进一步推导SVM原问题！

对偶问题

- 支持向量机的原问题：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geq 1, 1\leq i\leq n$$

- 由上页的结论，原问题与下面的问题同解：

$$\max_{\boldsymbol{\alpha}:\alpha_i\geq 0}\ \min_{\boldsymbol{w},b}\ \mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n}\alpha_i\big(1-y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\big)$$

求解上式分为三个步骤：

1. 求解 $\min_{\boldsymbol{w},b}\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})$，得到 $\boldsymbol{w},b$ 关于 $\boldsymbol{\alpha}$ 的表达式，带入 $\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})$
2. 继续求解 $\max_{\boldsymbol{\alpha}:\alpha_i\geq 0}\mathcal{L}(\boldsymbol{w}(\boldsymbol{\alpha}),b(\boldsymbol{\alpha}),\boldsymbol{\alpha})$
3. 利用$\boldsymbol{\alpha}^*$回代求解$(\boldsymbol{w}^*,b^*)$

对偶问题

- 第一步，令 $\nabla_w \mathcal{L}$ 和 $\nabla_b \mathcal{L}$ 等于零，得到：　　这一步是无约束优化，直接求导即可

$$\nabla_w \mathcal{L} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = \mathbf{0} \qquad \nabla_b \mathcal{L} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

- 第二步，将上面两式带入 $\max\limits_{\boldsymbol{\alpha}:\alpha_i \geq 0} \min\limits_{w,b} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha})$ 得到：

支持向量机的对偶问题：

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$
$$\text{s.t.} \quad \alpha_i \geq 0, \qquad i = 1, .., n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

对偶问题

- 第三步，利用α^*回代求解w^*和判别函数（b^*在后面给出求解方法）：

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

$$\hat{y} = \operatorname{sgn}(w^* \cdot x + b^*) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\right)$$

- 回顾(w^*, b^*, α^*)满足的KKT条件：

$$\begin{aligned} \alpha_i^* \left(1 - y_i (w^* \cdot x_i + b^*)\right) &= 0, \quad & i &= 1, \ldots, n \\ 1 - y_i (w \cdot x_i + b) &\le 0, \quad & i &= 1, \ldots, n \\ \alpha_i &\ge 0, \quad & i &= 1, \ldots, n \end{aligned}$$

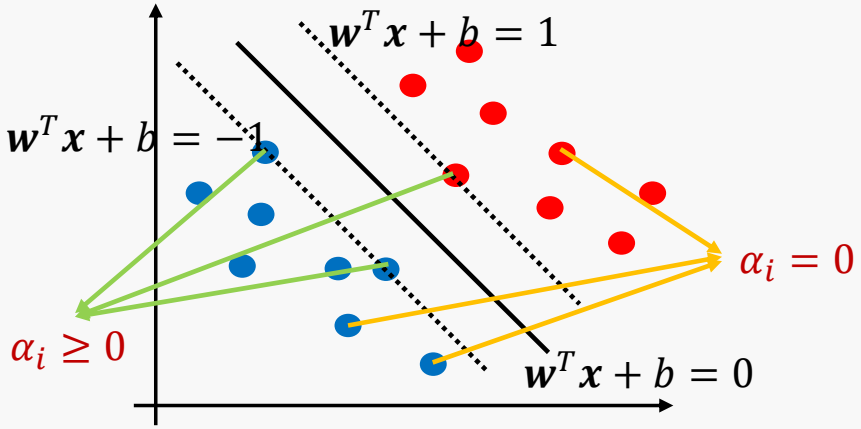
  - 若$\alpha_i > 0$，必有 $y_i(w \cdot x_i + b) = 1$，即(x_i, y_i)为支持向量
  - 若 $y_i(w \cdot x_i + b) > 1$，必有$\alpha_i = 0$，即非支持向量对应的α_i等于0

对偶问题

- 回顾(w^*, b^*, α^*)满足的KKT条件：

$$\begin{aligned} \alpha_i^*\left(1 - y_i(w^* \cdot x_i + b^*)\right) &= 0, \quad i = 1, \ldots, n \\ 1 - y_i(w \cdot x_i + b) &\le 0, \quad i = 1, \ldots, n \\ \alpha_i &\ge 0, \quad i = 1, \ldots, n \end{aligned}$$

  - 若$\alpha_i > 0$，必有 $y_i(w \cdot x_i + b) = 1$，即(x_i, y_i)为支持向量
  - 若 $y_i(w \cdot x_i + b) > 1$，必有$\alpha_i = 0$，即非支持向量对应的α_i等于0



对偶问题

- 从支持向量的角度重新审视(w^*, b^*)以及判别函数，以S表示支持向量的集合
  - w^*的表达式：

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

  - 因为对于支持向量有 $y_i(w \cdot x_i + b) = 1$，所以b^*的表达式：

$$b^* = \frac{1}{|S|} \sum_{i \in S} \left(y_i - \sum_{s \in S} \alpha_s y_s (x_s \cdot x_i) \right)$$

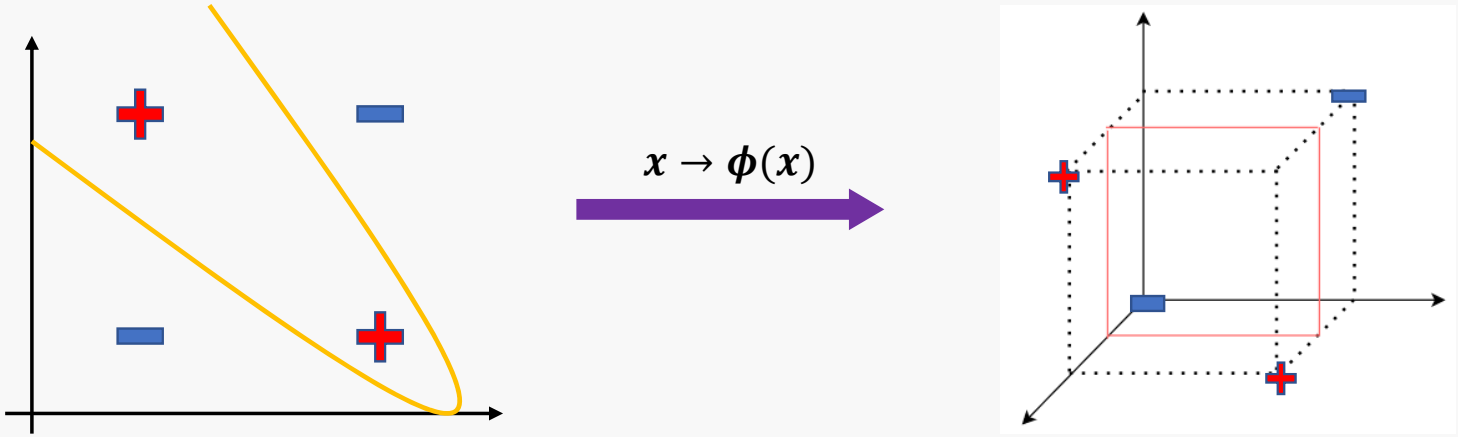
  - 判别函数：

$$\hat{y} = \mathrm{sgn}(w^* \cdot x + b^*) = \mathrm{sgn}\left(\sum_{i \in S} \alpha_i^* y_i (x_i \cdot x) + b^* \right)$$

- 以上三式中，只有支持向量发挥作用！最终模型只与支持向量有关！

核函数

- 前面讨论了训练样本线性可分的情况，如果不满足线性可分条件呢？



- 可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分
- 如果原始空间是有限维，那么一定存在一个高维特征空间使样本可分

核函数

- 令$\boldsymbol{\phi}(x)$表示高维特征向量，则高维特征空间中的分类面可表示为：

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b$$

- 支持向量机的原问题可表示如下：

$$\min_{\boldsymbol{w}, b} \frac{1}{2} \|\boldsymbol{w}\|_2^2$$

$$\text{s.t.} \quad y_i(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1, \qquad i = 1, \ldots, n$$

- 其对偶问题为：

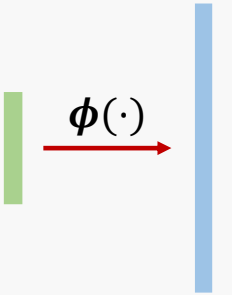
$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \left(\boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j) \right)$$

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

$$\alpha_i \geq 0, \quad i = 1, \ldots, n$$

核函数

- 函数$\boldsymbol{\phi}(\cdot)$又被称为基函数，它将x映射到更高维度的特征空间，这将导致内积运算$\boldsymbol{\phi}(x_i)\cdot\boldsymbol{\phi}(x_j)$的计算量增加
- 多项式基函数是一种常用的基函数
  - (x_1,x_2,x_3)对应的2次基函数为$(x_1,\ x_2,\ x_3,\ x_1x_2,\ x_1x_3,\ x_2x_3,\ x_1^2,\ x_2^2,\ x_3^2)$
  - $\boldsymbol{x}\in\mathbb{R}^d$对应的$p$次基函数的维度为$\sum_{i=1}^{p}\binom{i+d-1}{i}$
  - 当 d 或 p 较大时，多项式基函数将引发维度灾难
- 先计算基函数，然后计算内积不可取：计算量过大！
- 有什么解决方案呢？



可分性变好，
但计算量增大

核函数

- 回顾线性可分情况的对偶问题和判别函数

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$

$$\hat{y} = \operatorname{sgn}\left(\sum_{i \in \mathcal{S}} \alpha_i^* y_i (\boldsymbol{x}_i \cdot \boldsymbol{x}) + b^* \right)$$

对偶问题的求解以及判别函数都只涉及内积$(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$！

可以在原始空间中计算高维空间的内积，以降低计算量！

- 核函数的定义：

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j)$$

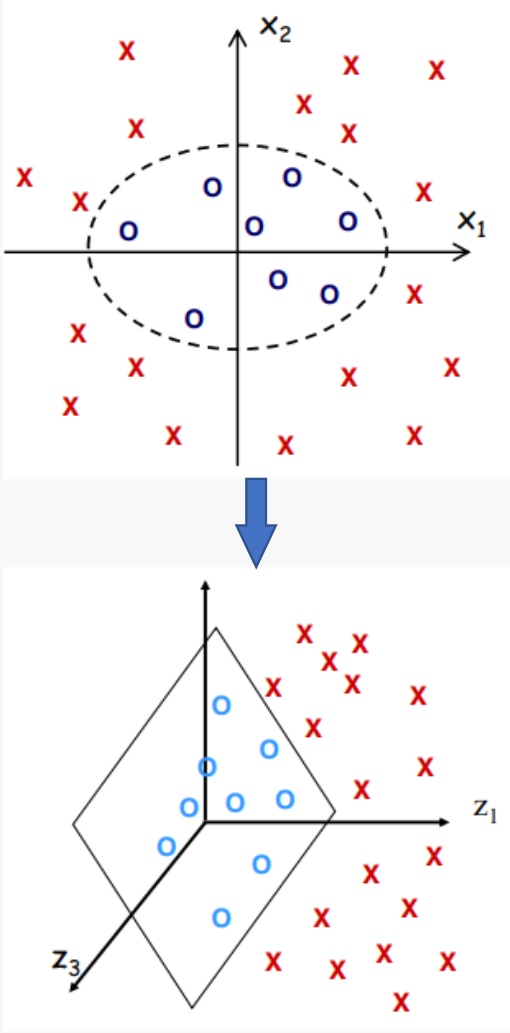
核函数

核函数举例

- 原始特征$x=(x_1,x_2)$，基函数$\phi(x)=\left(x_1^2,\ \sqrt{2}x_1x_2,\ x_2^2\right)$

$$\begin{aligned}\phi(x)\cdot\phi(y)&=\left(x_1^2,\sqrt{2}x_1x_2,x_2^2\right)\cdot\left(y_1^2,\sqrt{2}y_1y_2,y_2^2\right)\\&=x_1^2y_1^2+2x_1x_2y_1y_2+x_2^2y_2^2\\&=(x_1y_1+x_2y_2)^2\\&=\left((x_1,x_2)\cdot(y_1,y_2)\right)^2\\&=(x\cdot y)^2\\&:=K(x,y)\end{aligned}$$

- 通过核函数$K(x,y)=(x\cdot y)^2$，我们能够在原始特征空间（2维）计算高维特征空间（3维）的内积



核函数

- 引入核函数后的支持向量机
  - 对偶问题：

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

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

$$\alpha_i \geq 0, \quad i = 1, \ldots, n$$

  - 求解后可以得到：

$$\begin{aligned} \hat{y} &= \operatorname{sgn}(w \cdot \phi(x) + b) \\ &= \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i (\phi(x_i) \cdot \phi(x)) + b \right) \\ &= \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j) + b \right) \end{aligned}$$

  - 求解过程和判别函数**均不涉及高维基函数**！

核函数

常用核函数列表

- 线性核：

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

- 多项式核：

$$K(x_i, x_j) = (x_i \cdot x_j)^d$$

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

- 径向基函数（RBF）核：

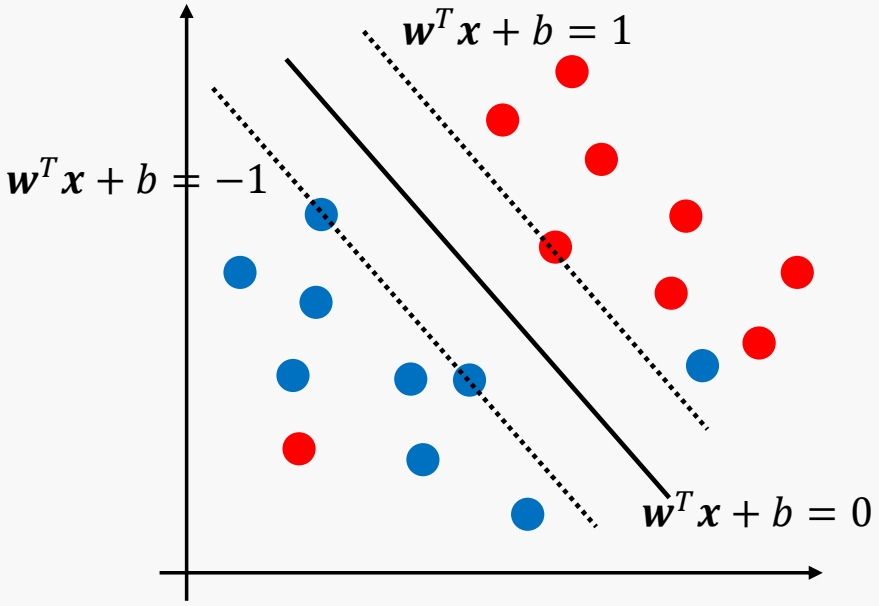
$$K(x_i, x_j) = \exp\left\{-\frac{\|x_i - x_j\|_2^2}{\sigma^2}\right\}$$

- Sigmoid核：

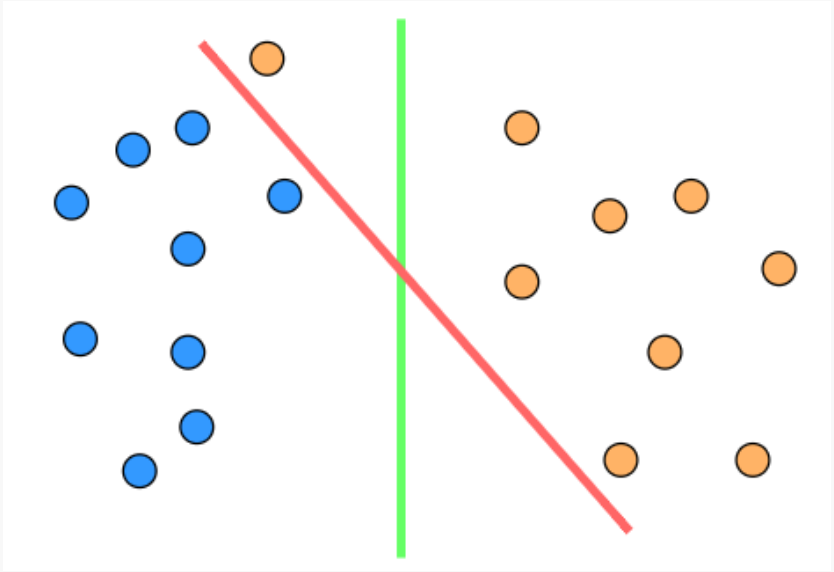
$$K(x_i, x_j) = \tanh(\beta x_i \cdot x_j + \theta)$$

软间隔与正则化

- 在前面的讨论中，我们一直假定训练样本在样本空间或者特征空间中是线性可分的，实际中可能存在一些问题：
  - 很难确定合适的核函数使得训练样本在特征空间线性可分
  - 离群值很可能导致分类面过拟合



非线性可分



对离群值敏感

软间隔与正则化

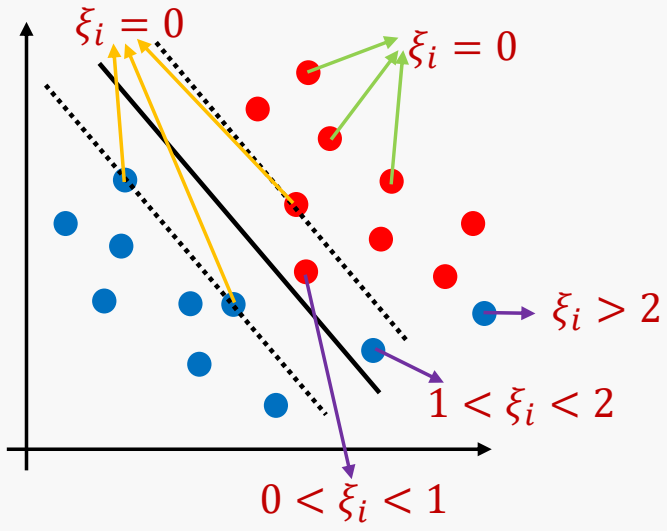
软间隔支持向量机

- 目标：让支持向量机应对非线性可分数据以及缓解其对离群值的敏感性
- 原问题：

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1-\xi_i, \qquad i = 1,\dots,n$$

$$\xi_i \geq 0, \qquad i = 1,\dots,n$$



- 解释：
  - 对于部分样本，允许函数间隔为$1-\xi_i$（$\xi_i > 0$），此时优化目标会增加$C\xi_i$的惩罚项
  - 超参数C用来权衡“大间隔”与“保证训练样本均在间隔带以外”

软间隔与正则化

软间隔支持向量机的对偶问题

$$\max_{\substack{\boldsymbol{\alpha}:\alpha_i\geq 0\\ \boldsymbol{\mu}:\mu_i\geq 0}} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \mathcal{L}(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\mu})$$

$$= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1-\xi_i - y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\right) - \sum_{i=1}^{n}\mu_i\xi_i$$

令$\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\mu})$对$\boldsymbol{w},b,\boldsymbol{\xi}$的偏导数为0，可将上式化简为：

软间隔支持向量机的对偶问题

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \left(\boldsymbol{x}_i\cdot\boldsymbol{x}_j\right)$$

$$\text{s.t.}\quad 0\leq\alpha_i\leq C, \qquad i=1,..,n$$

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

软间隔与正则化

- 与软间隔对应，线性可分条件下的支持向量机也可称为硬间隔支持向量机
- 软间隔与硬间隔支持向量机的相同之处：
  - 两者对偶问题的优化目标函数一致，解法一致
  - 两者的$\boldsymbol{w}$和判别函数的表达式一致

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i$$

$$\hat{y} = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i (\boldsymbol{x}_i \cdot \boldsymbol{x}) + b^* \right)$$

- 软间隔与硬间隔支持向量机的不同之处：
  - 硬间隔版本要求$\alpha_i \geq 0$
  - 软间隔版本要求$0 \leq \alpha_i \leq C$

软间隔与正则化

- 从KKT条件的角度看 $0 \le \alpha_i \le C$：

$$\begin{cases} \alpha_i \ge 0, \quad \mu_i \ge 0, \\ y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \ge 1 - \xi_i \\ \alpha_i(y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1 + \xi_i) = 0, \\ \xi_i \ge 0, \quad \mu_i \xi_i = 0 \end{cases}$$

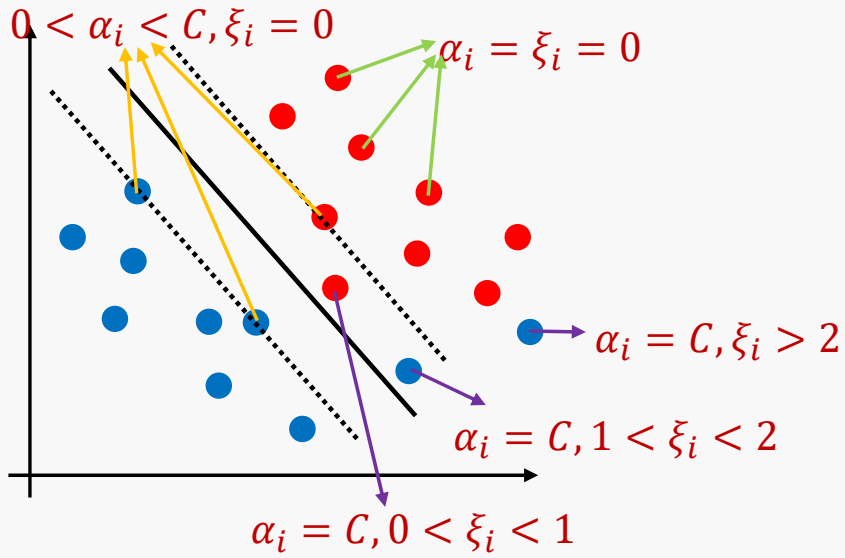
$$\alpha_i + \mu_i = C \quad \longleftarrow \quad \nabla_{\xi_i} \mathcal{L} = 0 \text{要求}$$

  - 当$\boldsymbol{\alpha_i = 0}$时，$\boldsymbol{\mu_i = C}$，$\boldsymbol{\xi_i = 0}$，$\boldsymbol{y_i(w \cdot x_i + b) \ge 1}$，非支持向量
  - 当$\boldsymbol{0 < \alpha_i < C}$时，$\boldsymbol{0 < \mu_i < C}$，$\boldsymbol{\xi_i = 0}$，$\boldsymbol{y_i(w \cdot x_i + b) = 1}$，支持向量，且在最大间隔边界上
  - 当$\boldsymbol{\alpha_i = C}$时，$\boldsymbol{\mu_i = 0}$，$\boldsymbol{\xi_i \ge 0}$，$\boldsymbol{y_i(w \cdot x_i + b) = 1 - \xi_i \le 1}$，支持向量，且越过最大间隔边界

- 软间隔支持向量机的最终模型也仅与支持向量有关！

软间隔与正则化

- 上页结论：
  - 当$\alpha_i = 0$时，$\mu_i = C,\ \xi_i = 0,\ y_i(w \cdot x_i + b) \geq 1$，非支持向量
  - 当$0 < \alpha_i < C$时，$0 < \mu_i < C,\ \xi_i = 0,\ y_i(w \cdot x_i + b) = 1$，支持向量，且在最大间隔边界上
  - 当$\alpha_i = C$时，$\mu_i = 0,\ \xi_i \geq 0,\ y_i(w \cdot x_i + b) = 1 - \xi_i \leq 1$，支持向量，且越过最大间隔边界



§5.3 方法扩展

一、支持向量回归

二、核方法

三、多分类问题

支持向量回归

- 回顾：线性回归

在于如何衡量$f(x_i)$和y_i之间的差异，均方误差是回归任务中最常用的性能度量，因此我们可试图让均方误差最小化，即

$$(w^*, b^*) = \underset{(w,b)}{\operatorname{argmin}} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

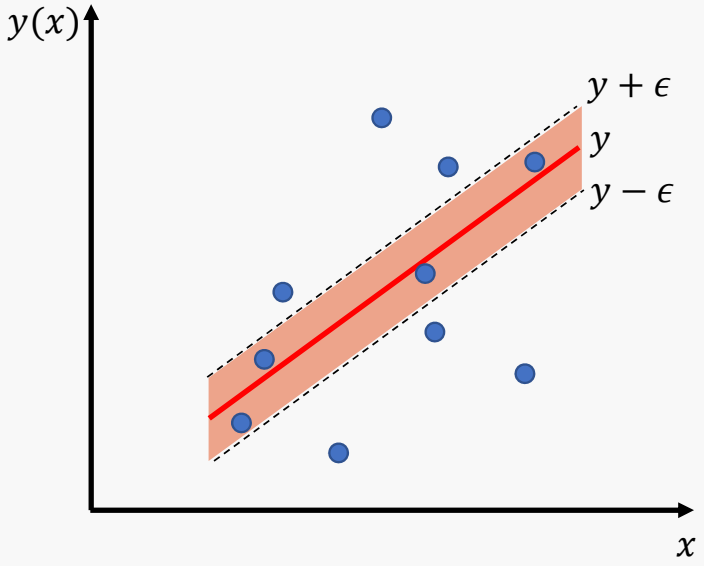
$$= \underset{(w,b)}{\operatorname{argmin}} \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

- 均方误差有非常好的几何意义，它对应了常用的欧几里得距离或简称“欧氏距离”（Euclidean distance）

支持向量回归

- 目标：学习回归模型 $f(x) = w\phi(x) + b$
  - **传统回归模型**：当且仅当$f(x)$与y完全相同时，损失才为零
  - **支持向量回归**：对$f(x)$与y的偏差大于容忍度ϵ才计算损失

- 支持向量的对比
  - 分类（SVM）：当数据点远离分类平面时损失为零，支持向量为离分类平面最近的点
  - 回归（SVR）：当数据点与回归平面足够近时损失为0，支持向量为$|f(x) - y| > \epsilon$的点

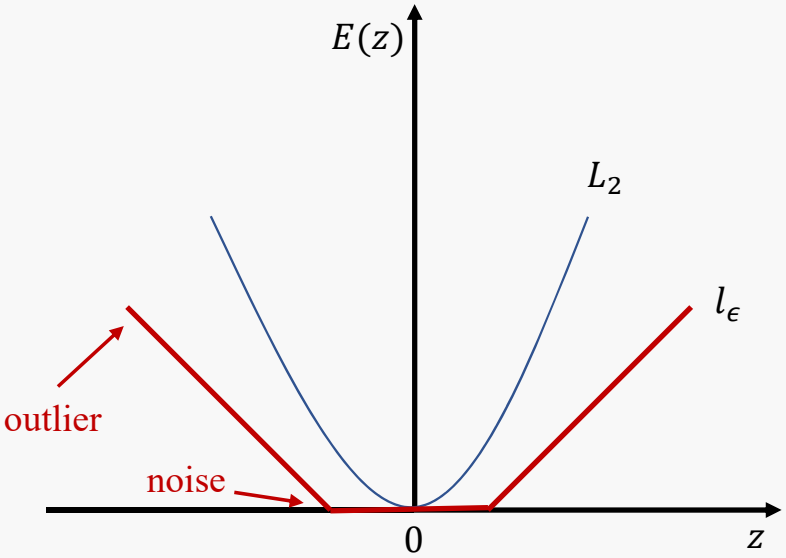


支持向量回归

- 优化目标
  - SVR问题可形式化为：

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

  - 不敏感损失函数$l_\epsilon(z) = \max(0, |z| - \epsilon)$
  - 对噪声（较小偏差）、异常（较大偏差）的鲁棒性均优于L_2损失函数



支持向量回归

- 优化目标
  - 间隔带两侧损失函数可有所不同
  - 引入松弛变量ξ_i和$\hat{\xi}_i$，重写为：

$$\min_{\boldsymbol{w},b,\xi_i,\hat{\xi}_i} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}(\xi_i + \hat{\xi}_i)$$
$$\text{s.t. } f(\boldsymbol{x}_i) - y_i \le \epsilon + \xi_i,$$
$$y_i - f(\boldsymbol{x}_i) \le \epsilon + \hat{\xi}_i,$$
$$\xi_i \ge 0, \hat{\xi}_i \ge 0, i = 1,2,\dots,m.$$

  - 拉格朗日函数：

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$$
$$= \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}(\xi_i + \hat{\xi}_i) - \sum_{i=1}^{m}\mu_i\xi_i - \sum_{i=1}^{m}\hat{\mu}_i\hat{\xi}_i$$
$$+ \sum_{i=1}^{m}\alpha_i(f(\boldsymbol{x}_i) - y_i - \epsilon - \xi_i) + \sum_{i=1}^{m}\hat{\alpha}_i(y_i - f(\boldsymbol{x}_i) - \epsilon - \hat{\xi}_i)$$

支持向量回归

- 优化目标
  - 令L对变量的偏导为零可得：

$$w = \sum_{i=1}^{m} (\widehat{\alpha_i} - \alpha_i) x_i$$

$$0 = \sum_{i=1}^{m} (\widehat{\alpha_i} - \alpha_i)$$

$$C = \alpha_i + \mu_i$$

$$C = \widehat{\alpha_i} + \widehat{\mu_i}$$

  - 得到对偶问题：

$$\min_{\alpha, \widehat{\alpha}} \sum_{i=1}^{m} y_i(\widehat{\alpha_i} - \alpha_i) - \epsilon(\widehat{\alpha_i} + \alpha_i) - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\widehat{\alpha_i} - \alpha_i)\left(\widehat{\alpha_j} - \alpha_j\right) x_i^T x_j$$

$$\text{s.t.} \sum_{i=1}^{m} (\widehat{\alpha_i} - \alpha_i) = 0$$

$$0 \le \alpha_i, \widehat{\alpha_i} \le C$$

支持向量回归

- 优化目标
  - 上述过程需满足KKT条件：

$$\begin{cases} \alpha_i(f(x_i) - y_i - \epsilon - \xi_i) = 0 \\ \widehat{\alpha_i}(y_i - f(x_i) - \epsilon - \widehat{\xi_i}) = 0 \\ \alpha_i\widehat{\alpha_i} = 0, \xi_i\widehat{\xi_i} = 0 \\ (C - \alpha_i)\xi_i = 0, (C - \widehat{\alpha_i})\widehat{\xi_i} = 0 \end{cases}$$

  - 样本落在$\epsilon-$间隔带之外时，α_i和$\widehat{\alpha_i}$才能取非零值，此时样本为支持向量
  - 将w代入回归模型有：

$$f(x) = \sum_{i=1}^{m}(\widehat{\alpha_i} - \alpha_i)x_i^T x + b$$

  - SVR的支持向量满足$(\alpha_i - \widehat{\alpha_i}) \neq 0$
  - 求解b：先求解对偶问题得到α_i，由$\alpha_i(f(x_i) - y_i - \epsilon - \xi_i) = 0$和$(C - \alpha_i)\xi_i = 0$，选取$0 < \alpha_i < C$，则可以得到：

$$b = y_i + \epsilon - \sum_{j=1}^{m}(\widehat{\alpha_j} - \alpha_j)x_i^T x_i$$

核方法

- 考虑特征映射
  - SVM：

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

  - SVR：

$$f(x) = \sum_{i=1}^{m} (\widehat{\alpha_i} - \alpha_i) \kappa(x, x_i) + b$$

  - 若不考虑偏移项，则SVM和SVR学到的模型总是能表示成核函数$\kappa(x, x_i)$的线性组合
  - 更一般的结论（表示定理）：

定理 6.2（表示定理） 令 $\mathbb{H}$ 为核函数 κ 对应的再生核希尔伯特空间，$\|h\|_{\mathbb{H}}$ 表示 $\mathbb{H}$ 空间中关于 h 的范数，对于任意单调递增函数 $\Omega : [0, \infty] \mapsto \mathbb{R}$ 和任意非负损失函数 $\ell : \mathbb{R}^m \mapsto [0, \infty]$，优化问题

$$\min_{h \in \mathbb{H}} \; F(h) = \Omega(\|h\|_{\mathbb{H}}) + \ell\big(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \ldots, h(\boldsymbol{x}_m)\big) \tag{6.57}$$

的解总可写为

$$h^*(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i \kappa(\boldsymbol{x}, \boldsymbol{x}_i) \; . \tag{6.58}$$

核方法

- **核方法**：一系列基于核函数的学习方法
- 核线性判别分析（KLDA）
  - 假设通过映射ϕ，将样本映射到特征空间$\mathbb{F}$执行线性判别分析：

$$h(x) = w^T \phi(x)$$

  - KLDA的学习目标为：

$$\min_w J(w) = \frac{w^T S_b^\phi w}{w^T S_w^\phi w}$$

  - 记两类样本集为X_0, X_1，总样本数为$m = m_0 + m_1$，则第i类样本在特征空间$\mathbb{F}$中的均值为：

$$\mu_i^\phi = \frac{1}{m_i} \sum_{x \in X_i} \phi(x)$$

  - 两个散度矩阵分别为：

$$S_b^\phi = \left(\mu_1^\phi - \mu_0^\phi\right)\left(\mu_1^\phi - \mu_0^\phi\right)^T, S_w^\phi = \sum_{i=0}^{1} \sum_{x \in X_i} \left(\phi(x) - \mu_i^\phi\right)\left(\phi(x) - \mu_i^\phi\right)^T$$

核方法

核线性判别分析（KLDA）

- 将$J(w)$作为损失函数l，令$\Omega \equiv 0$，则由表示定理，$h(x)$可以表示为：

$$h(x) = \sum_{i=1}^{m} \alpha_i \kappa(x, x_i)$$

- 得到：

$$w = \sum_{i=1}^{m} \alpha_i \phi(x_i)$$

- 通常难以直接写出映射ϕ的具体形式，可通过核函数对该问题进行求解。令$\mathbf{K} \in \mathbb{R}^{m \times m}$为核矩阵，令$\mathbf{1}_i \in \{1, 0\}^{m \times 1}$为第i类样本的指示向量，再令：

$$\hat{\mu}_0 = \frac{1}{m_0}\mathbf{K}\mathbf{1}_0, \hat{\mu}_1 = \frac{1}{m_1}\mathbf{K}\mathbf{1}_1, \mathbf{M} = (\hat{\mu}_0 - \hat{\mu}_1)(\hat{\mu}_0 - \hat{\mu}_1)^T, \mathbf{N} = \mathbf{K}\mathbf{K}^T - \sum_{i=0}^{1} m_i \hat{\mu}_i \hat{\mu}_i^T$$

- 原问题等价为：

$$\max_{\alpha} J(\alpha) = \frac{\alpha^T \mathbf{M} \alpha}{\alpha^T \mathbf{N} \alpha} \longrightarrow \text{同LDA解法得到}\alpha \longrightarrow \text{即可写出}h(x)\text{的形式}$$

多分类问题

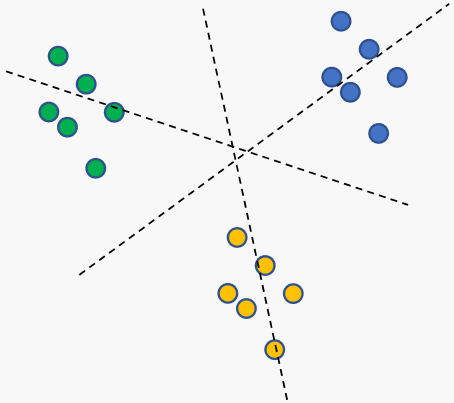
- 成对分类方法（one-against-one）
  - 每两个类之间都构造一个SVM进行判别：

$$\min_{w^{ij},b^{ij},\xi^{ij}} \frac{1}{2}||w^{ij}||^2 + C\sum_{t=1}^{m}\xi_t^{ij}$$

$$s.t.\ \left(w^{ij}\right)^T x_t + b^{ij} \geq 1 - \xi_t^{ij}, \text{if } y_t = i$$

$$\left(w^{ij}\right)^T x_t + b^{ij} \leq -1 + \xi_t^{ij}, \text{if } y_t = j$$

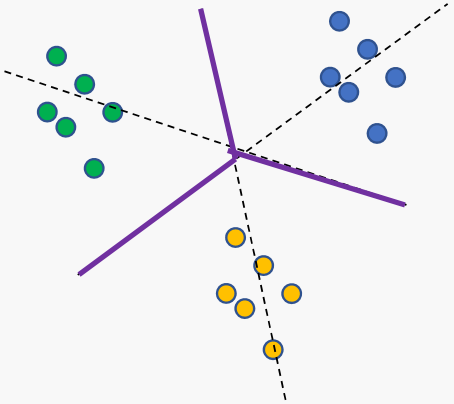
$$\xi_t^{ij} \geq 0, t = 1, 2, \ldots, m.$$



  - 若一共有M个类，则需要训练$\frac{1}{2}M(M-1)$个SVM分类器
  - 训练完成后，对于新数据采用投票策略进行分类：
    - 每个分类器对判定的类别投一票，决策函数为：

$$y_{new}^{ij} = \text{sign}\left[\left(w^{ij}\right)^T x_t + b^{ij}\right]$$

    - 票数最多的类别为最终预测
    - 票数相同时简单选择索引最小的类别



多分类问题

一对多分类方法（one-against-all）

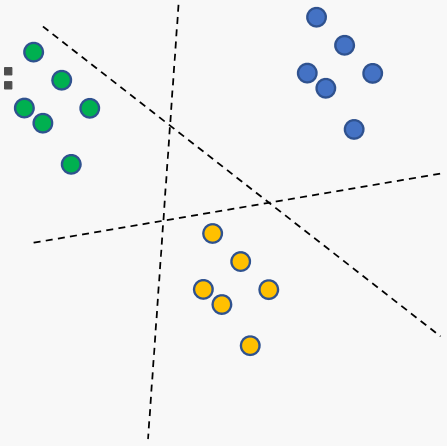
- 对每个类都构造一个SVM进行判别，区分第i类和其余M-1类：

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

$$s.t.\ \left(w^i\right)^T x_t + b^i \geq 1 - \xi_t^i, \text{if } y_t = i$$

$$\left(w^i\right)^T x_t + b^i \leq -1 + \xi_t^i, \text{if } y_t \neq j$$

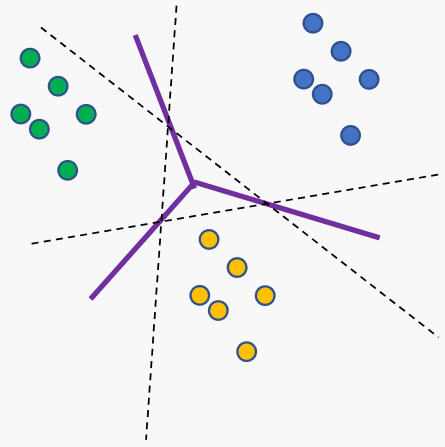
$$\xi_t^i \geq 0, t = 1, 2, \ldots, m.$$



- 训练完成后，对于新数据根据预测值的大小进行分类：
  - 修改决策函数，去掉符号函数：

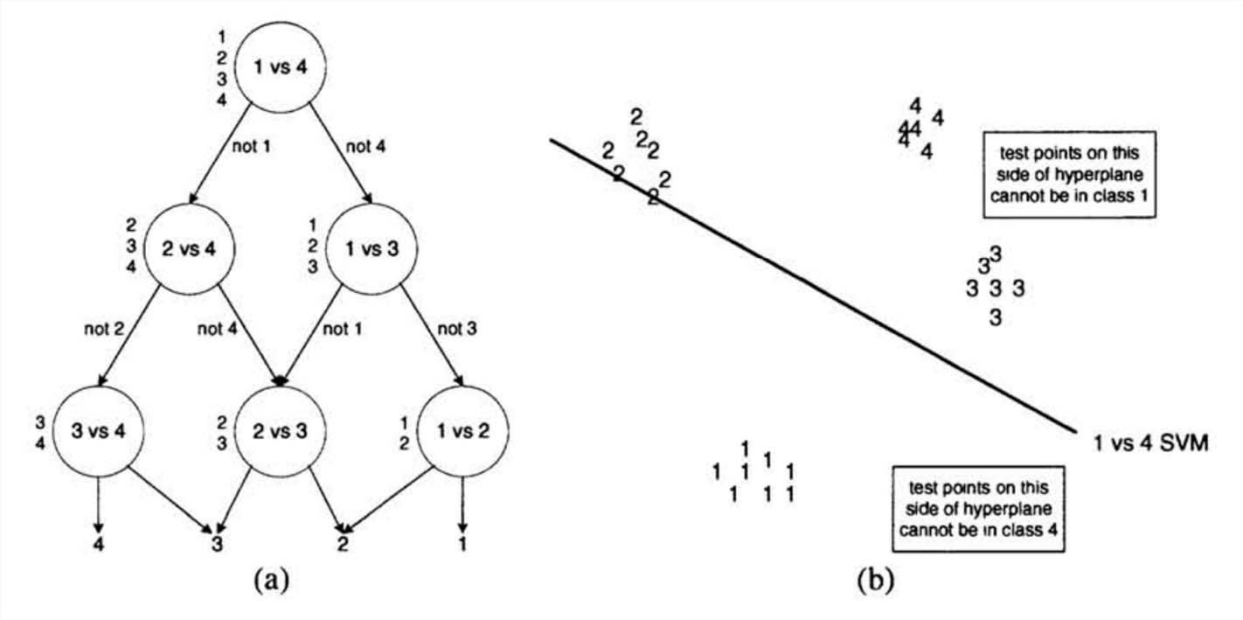
$$d_{new}^i = \left(w^i\right)^T x_t + b^i$$

  - 选取预测值最大的一类作为最终预测
  - **问题：**不同的分类器并非同时训练，且不同的分类器训练时正负样本的比例各有不同，因此去掉符号函数以后预测值可能不具有很好的可比较性



多分类问题

- 有向无环图支持向量机（DAGSVM）
  - 训练阶段与成对分类方法相同
  - 测试时通过有根节点（共有$\frac{1}{2}M(M-1)$个节点）的有向无环图进行判断

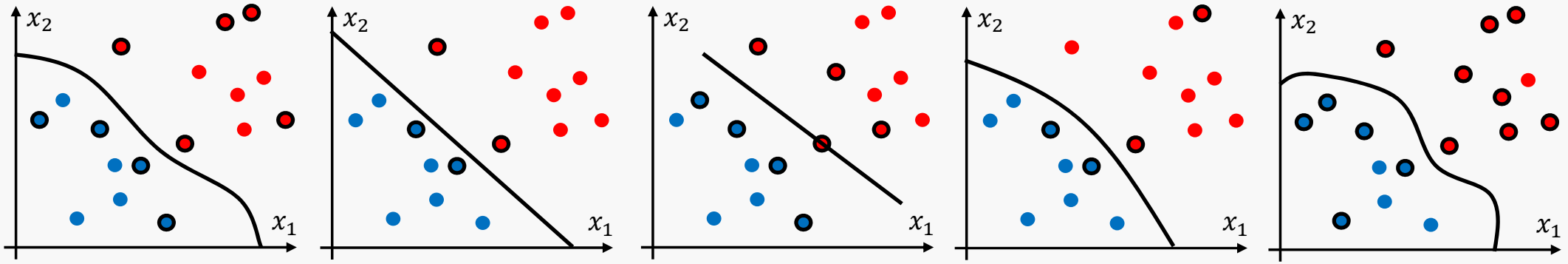


思考题

- 为下面的每个SVM模型标出正确的图形

$$\min \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.}\ \xi_i \geq 0, y_i(\boldsymbol{w}^T\boldsymbol{x} + w_0) \geq 1 - \xi_i, C = 0.1$$

$$\min \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.}\ \xi_i \geq 0, y_i(\boldsymbol{w}^T\boldsymbol{x} + w_0) \geq 1 - \xi_i, C = 1$$

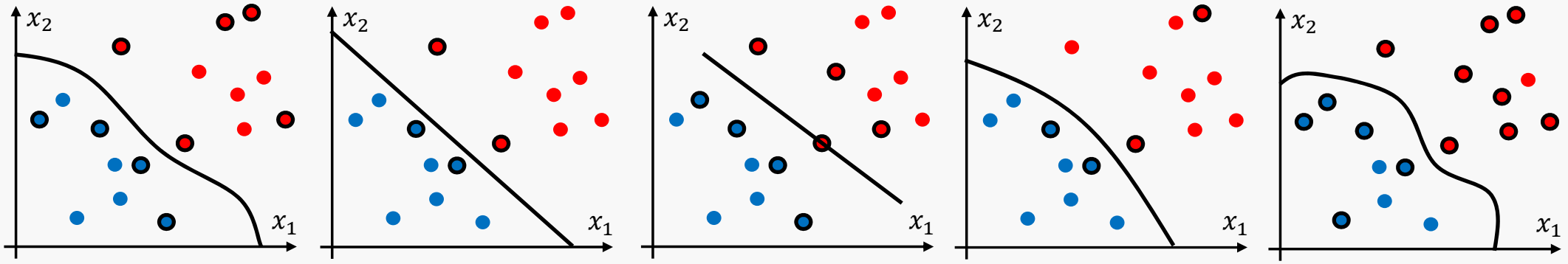


思考题

- 为下面的每个SVM模型标出正确的图形

$$\max\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)$$

$$\text{s.t.}\ \alpha_i\geq 0,\sum_{i=1}^{N}\alpha_i y_i=0,k(\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'+\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'\right)^2$$



思考题

- 为下面的每个SVM模型标出正确的图形

$$\max\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)$$

$$\text{s.t.}\ \alpha_i\geq 0,\sum_{i=1}^{N}\alpha_i y_i=0, k(\boldsymbol{x},\boldsymbol{x}')=\exp\left(-\frac{1}{2}||\boldsymbol{x}-\boldsymbol{x}'||^2\right)$$

$$\max\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)$$

$$\text{s.t.}\ \alpha_i\geq 0,\sum_{i=1}^{N}\alpha_i y_i=0, k(\boldsymbol{x},\boldsymbol{x}')=\exp\left(-||\boldsymbol{x}-\boldsymbol{x}'||^2\right)$$