第4章

支持向量机

Support Vector Machine

模型 策略 算法
卷积是二维的点积

赫 然
rhe@nlpr.ia.ac.cn
https://rhe-web.github.io/

智能感知与计算研究中心（CRIPAC）
中科院自动化研究所　模式识别国家重点实验室

内容提要

- 支持向量机
  - **4.1 结构风险、经验风险与VC维**
  - 4.2 间隔与支持向量
  - 4.3 对偶问题
  - 4.4 软间隔与正则化
  - 4.5 支持向量回归
  - *4.6 核函数 线性转非线性
  - *4.7 核方法
  - *4.8 核支持向量机

Vladimir Vapnik
美国国家工程院院士

4.1 结构风险、经验风险与VC维

- 几个概念
  - 假设空间
    - 假设空间$H=\{f|y=f(x)\}$：决策函数的集合
    - H通常是由一个参数向量决定的函数簇：$H=\{f|y=f_\theta(x)\}$，参数θ取决于m维欧式空间R^m，称为参数空间。
    - 也可定义为条件概率的集合：$H=\{p|p(y|x)\}$。对应地，有条件概率分布簇：$H=\{p|p_\theta(y|x,\theta\in R^m)\}$
  - 损失函数和风险函数
    - 损失函数度量模型一次预测的好坏
    - 风险函数度量平均意义下模型预测的好坏

4.1 结构风险、经验风险与VC维

- 损失函数
  - 在假设空间H中选取模型 f 作为决策函数
  - 对于给定的输入x，由$f(x)$给出相应的预测，这个值与真实值y可能不一致，因此通常用损失函数来度量错误程度
  - 损失函数是$f(x)$和y的非负实值函数，记作 $L(y, f(x))$

4.1 结构风险、经验风险与VC维

- 风险损失（期望风险）
  - 由于模型的输入输出x,y是变量，遵循联合概率分布$p(x,y)$，所以函数的期望是：

$$R_{exp}(f)=E_p\left[L\left(y,f(x)\right)\right]=\int_{X\times Y}L\left(y,f(x)\right)p(x,y)dxdy$$

  - 这是理论意义上的模型$f(x)$关于联合分布$p(x,y)$的平均意义下的损失，称为风险函数或期望函数
  - 学习的目标：选择风险最小模型。由于联合分布$p(x,y)$是未知的，$R_{exp}(f)$不能直接计算，所以才需要学习。

4.1.1 经验风险

- **定义：**
  - 给定一个训练集$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\}$，模型$f(x)$关于训练数据的平均损失称为经验风险：
    - $R_{emp}(f)=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(x_i))$
  - 期望风险$R_{exp}(f)$是模型关于联合分布的期望损失。
  - 经验风险$R_{emp}(f)$是模型关于训练数据集的平均损失。

4.1.1 经验风险

- 经验风险最小化策略：
  - 该策略认为，经验风险最小的模型就是最优模型：
    - $\min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$
  - 当样本容量很大时，经验风险最小化能够保证有较好的学习效果。
  - 当样本容量很小时，经验风险最小化学习效果未必好，可能产生过拟合。

4.1.2 结构风险

- **结构风险最小化策略：** 最大化后验
  - 防止过拟合
  - 在经验风险上加上表示模型复杂度的正则化项或罚项。可定义：

线性判别分析+单位阵保证可逆

$$R_{srm}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda J(f)$$

极大似然估计

$$\min_{f \in H} R_{srm}(f)$$

函数的函数

  - $J(f)$：模型复杂度，定义在假设空间H上的泛函数。
  - 模型 f 越复杂，复杂度$J(f)$就越大；模型 f 越简单，复杂度$J(f)$就越小。

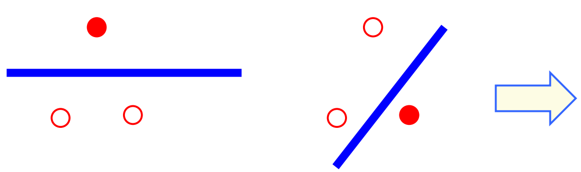
4.1.3 VC维（Vapnik Chervonenkis dimension）

- 定性理解：
  - VC维是衡量假设空间函数复杂度的一种方式。
  - The VC dimension for the set 一组函数 of functions {f(α)} is defined as the maximum number of training points 能分开数据点最大的个数 that can be shattered 打散 by {f(α)}.

正确分类

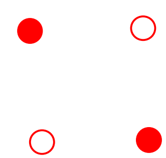
4.1.3 VC维（Vapnik Chervonenkis dimension）

- 定性理解：
  - VC维是衡量假设空间函数复杂度的一种方式。
  - 举个例子，在二分类问题中：
    - 假设空间$H=\{f: y=f(x)\}$是二维平面的线性划分
    - 平面上存在3个点该函数可以将其区分开，一侧取0一侧取1，而4个点却不行，如图：



共8种标记分配方式

存在一种示例集含有3个点，其任意标记分配方式，都可以被假设空间H区分开



任意4个数据点的标记分配方式中，至少有一种不能被线性划分

H的VC维为3

4.1.3 VC维

- 在引入VC维的定义前，先给出几个概念
  - 增长函数$\Pi_{H(m)} = \max_{\{x_1,\ldots,x_m\}} |\{f(x_1), f(x_2), \ldots, f(x_m)\}|$
    - $\Pi_{H(m)}$：假设空间H对m个示例所能赋予标记的最大可能结果数。（例如，在上述的二分类任务中，假设$m = 2$，即T中有2个元素，则其增长函数值为4）
    - $\Pi_{H(m)}$越大$\rightarrow$ H的表示能力越强$\rightarrow$ 对任务的适应能力越强

4.1.3 VC维

- 尽管H包含可能有无穷多个假设，但对于T中示例赋予标记的可能结果数是有限的：
  - 对于m个示例，最多有2^m个可能的结果。
- 对二分类任务来说，H中的假设对T中示例赋予标记的每种可能结果称为对T的一种“对分”。
- 若假设空间H能实现训练集T上的所有对分，即：

$$\Pi_{H(m)} = 2^m$$

则称T能被假设空间H“打散”

4.1.3 VC维

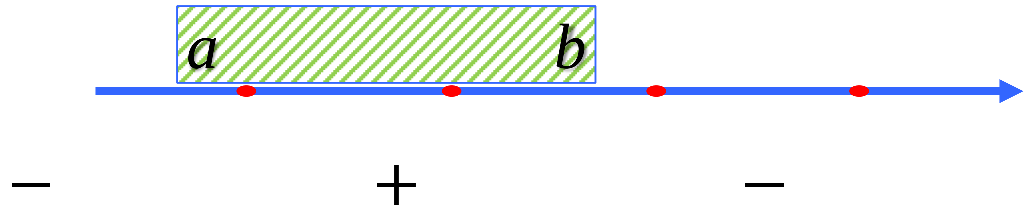
- 现在，我们正式定义VC维：
  - 假设空间H的VC维是指能被H打散的最大示例集的大小，即：

$$VC(H) = \max\{m: \Pi_{H(m)} = 2^m\}$$

  - 泛函
  - $VC(H) = d$表明存在大小为d的示例集能被假设空间H打散。VC维的定义与具体的数据分布无关。
  - 通常这样来计算：
    - 若存在大小为d的示例集能被H打散，但不存在任何$d+1$的示例集能被H打散，则H的VC维是d。

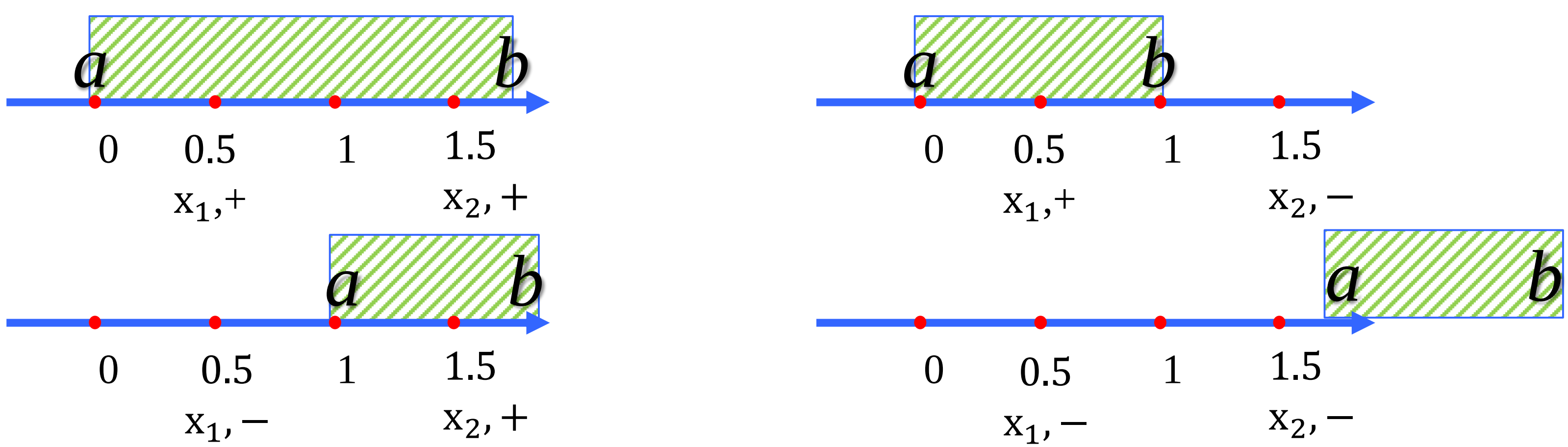
4.1.3 VC维

- 举例：
  - 实数域中的区间$[a,b]$：
    - 令H表示实数域中所有闭区间构成的集合$\{h_{[a,b]}: a,b \in$



4.1.3 VC维

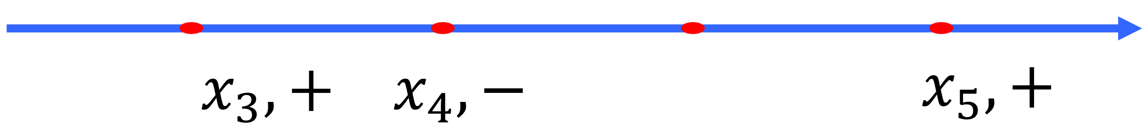
- 举例：
  - 实数域中的区间$[a, b]$：
    - 令$x_1 = 0.5$，$x_2 = 1.5$，共有4种对分，假设空间中存在假设$\{h_{[0,1]}, h_{[0,2]}, h_{[1,2]}, h_{[2,3]}\}$将其打散。所以$H$的VC维至少为2



4.1.3 VC维

– 对任意大小为3的示例集$\{x_3, x_4, x_5\}$，不妨设$x_3 < x_4 < x_5$，则H中不存在任何假设$h_{[a,b]}$能实现对分结果$\{(x_3, +), (x_4, -), (x_5, +)\}$。于是，$H$的VC维为2。

期望风险近似经验风险+vc



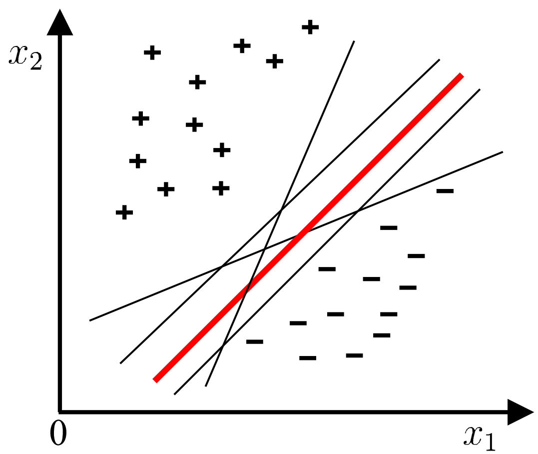
内容提要

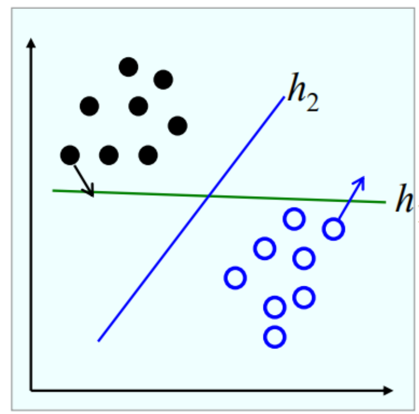
- 支持向量机
  - 4.1 结构风险、经验风险与VC维
  - **4.2 间隔与支持向量**
  - 4.3 对偶问题
  - 4.4 软间隔与正则化
  - 4.5 支持向量回归
  - *4.6 核函数
  - *4.7 核方法
  - *4.8 核支持向量机

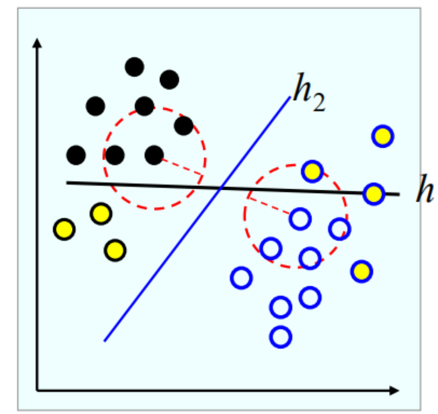
类似损失函数

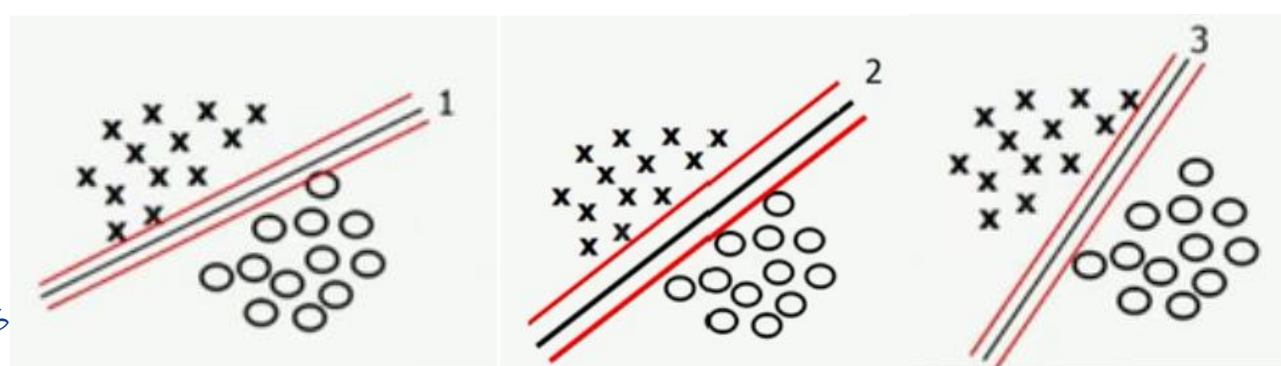
4.2 间隔与支持向量

- 将训练样本分开的超平面可能有很多, 哪条最好呢?
  - 应选“正中间”：
    - 对样本的容忍性好，鲁棒性强，泛化能力强





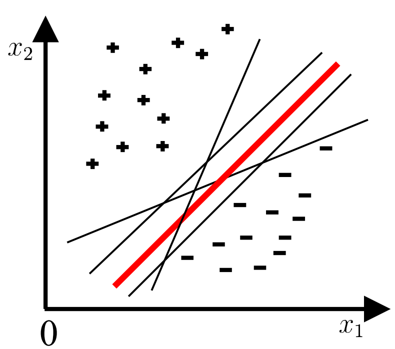




4.2 间隔与支持向量

- 在样本空间中，划分超平面可以通过如下方程来描述：
  - $w^T x + b = 0$
    - 法向量$w = (w_1; w_2; \dots; w_d)$：决定超平面的方向；
    - 位移项b：决定超平面与原点之间的距离
  - 将超平面记为(w, b)，则样本空间中任一点到超平面(w, b)的距离为：

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



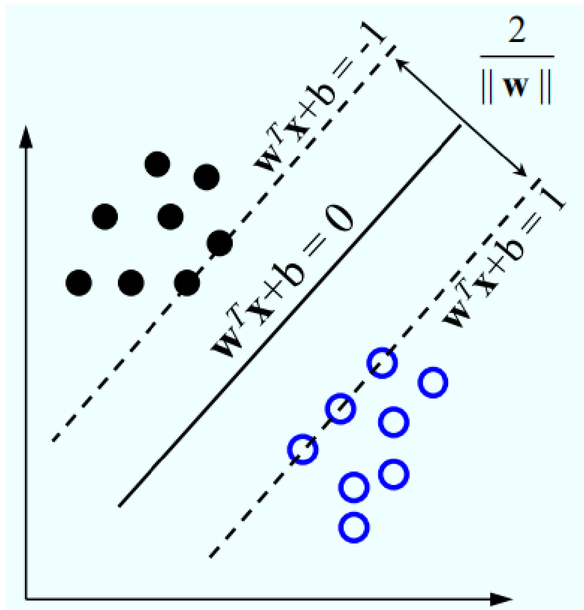
4.2 间隔与支持向量

- 假设超平面(w,b)能将训练样本正确分类，即：
- 对于训练集:

  $T=\{(x_1,y_1),\ldots,(x_n,y_n)\},$

  $y_i\in(+1,-1)$

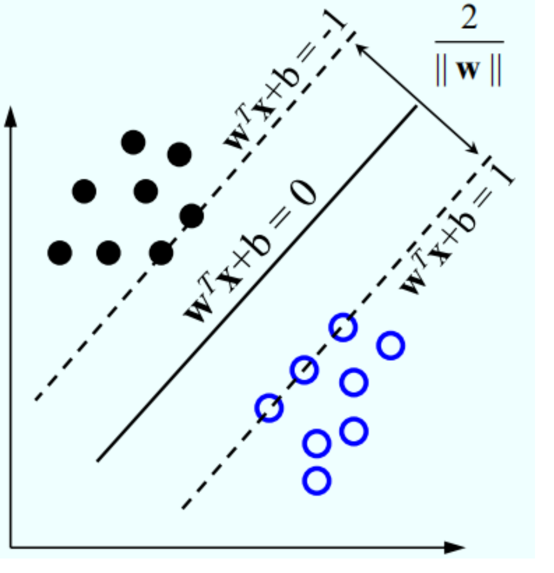
  - 若$y_i=+1$,则有$w^Tx+b>0$;
  - 若$y_i=-1$,则有$w^Tx+b<0$;
  - 令$\begin{cases}w^Tx_i+b\geq+1, y_i=+1\\ w^Tx_i+b\leq-1, y_i=-1\end{cases}$



4.2 间隔与支持向量

- 如图，与超平面最近的几个点使得上式等号成立，他们被称为“支持向量”。
- 两个不同类的支持向量到超平面的距离之和为：

$$\gamma = \frac{2}{||w||}$$ 斜率



  - 它被称为“间隔”。

4.2 间隔与支持向量

- 支持向量机(Support Vector Machine，SVM)
  - 给定训练集：

  $T = \{(x_1, y_1), \ldots, (x_n, y_n)\}, y_i \in (+1, -1)$

    - 任务：
      - 求解最大间隔分类超平面（“正中间”）

最大倒数=最小

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

$$s.t. y_i(w^T x_i + b) - 1 \geq 0, i = 1,2,\ldots,n$$

4.2 间隔与支持向量

- 支持向量机（Support Vector Machine，SVM）
  - 从而估计出：
    - 分类超平面：

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

    - 分类决策函数:

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

$$s.t. f(x) = sign(w^T x + b)$$

内容提要

- 支持向量机
  - 4.1 结构风险、经验风险与VC维
  - 4.2 间隔与支持向量
  - **4.3 对偶问题**
  - 4.4 软间隔与正则化
  - 4.5 支持向量回归
  - *4.6 核函数
  - *4.7 核方法
  - *4.8 核支持向量机

4.3 对偶问题

- 我们希望求解

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

$$s.t.\quad y_i(w^Tx_i+b)-1\geq 0, i=1,2,\dots,n$$

来得到大间隔划分超平面所对应的模型:

$$f(x)=w^Tx+b$$

拉格朗日乘子法

- 在不等式约束最优化问题中，常利用拉格朗日对偶性将原始问题转化为对偶问题进行求解

4.3 对偶问题

*数学知识点——广义拉格朗日函数

$$\min_{x \in R^d} f(x)$$

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

$$h_j(x) = 0, j = 1,2,\dots,l$$

广义拉格朗日函数：

$$L(x,\alpha,\beta) = f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x)$$

4.3 对偶问题

- 具体来说，为每条约束添加拉格朗日乘子$\alpha_i \geq 0$，则该函数的拉格朗日函数可写作：

$$L(w,b,\alpha) = \frac{1}{2}||w||^2 + \sum_{i=1}^{\overset{\text{样本数}}{m}} \alpha_i(1 - y_i(\overset{\text{纬度}}{w}^T x_i + b))$$

其中$\alpha = (\alpha_1; \alpha_2; \ldots; \alpha_m)$，原问题转换为：

$$\max_{\alpha} \min_{w,b} L(w,b,\alpha)$$

$$s.t.\ \ \alpha_i \geq 0, i = 1,2,\ldots,n$$

University of Chinese Academy of Sciences

4.3 对偶问题

- 求解：

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

$$\nabla_b L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i y_i \cancel{x_i} = 0 \quad \Rightarrow \quad 0 = \sum_{i=1}^{m} \alpha_i y_i$$

- 代回：

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

4.3 对偶问题

- 求对偶问题，即求$\min\limits_{w,b} L(w,b,\alpha)$对$\alpha$的极大：

SMO (Sequential Minimal Optimization)算法，略

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

$$s.t. \qquad 0 = \sum_{i=1}^{m}\alpha_i y_i \,, i = 1,2,\ldots,n$$

- 解出α后，求出w和b即可得到模型

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

4.3 对偶问题

*数学知识点——不等式约束优化问题

$$\min f(x)$$
$$s.t. \quad g(x) \le 0$$

- KKT （Karush-Kuhn-Tucker）条件：

$$\begin{cases} \nabla_x L(x,\lambda) = 0 \\ g(x) \le 0 \\ \lambda \ge 0 \\ \lambda g(x) = 0 \end{cases}$$

4.3 对偶问题

- 注意原问题：

$$min_{w,b}\frac{1}{2}||w||^2$$
$$s.t.\ y_i(w^Tx_i+b)-1\geq 0, i=1,2,...,n$$

有不等式约束，因此上述过程需要满足KKT条件，即：

$$\begin{cases} \alpha_i \geq 0 \\ y_i(w^Tx_i+b)-1\geq 0 \\ \alpha_i(\,y_i(w^Tx_i+b)-1) = 0 \end{cases}$$

$$\alpha_i = 0, y_i(w^Tx_i+b)-1\geq 0$$

$$\alpha_i > 0, y_i(w^Tx_i+b)=1$$ 支持向量

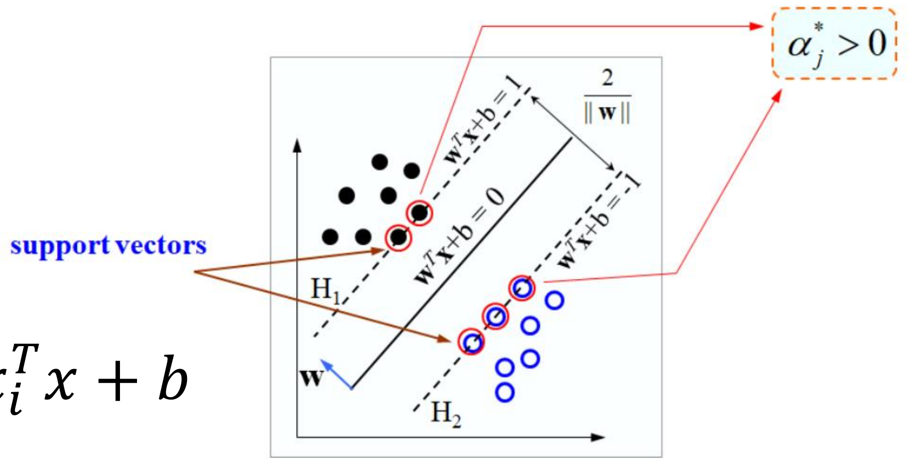
4.3 对偶问题

- 使 KKT 条件中等式成立的点为支持向量

  $y_i(w^T x_i + b) = 1$

  – 最终求得的模型：

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



  – 若$\alpha_i = 0$：则该样本不会对最终的式子有影响；

  – 若$\alpha_i > 0$：此时该点为支持向量；

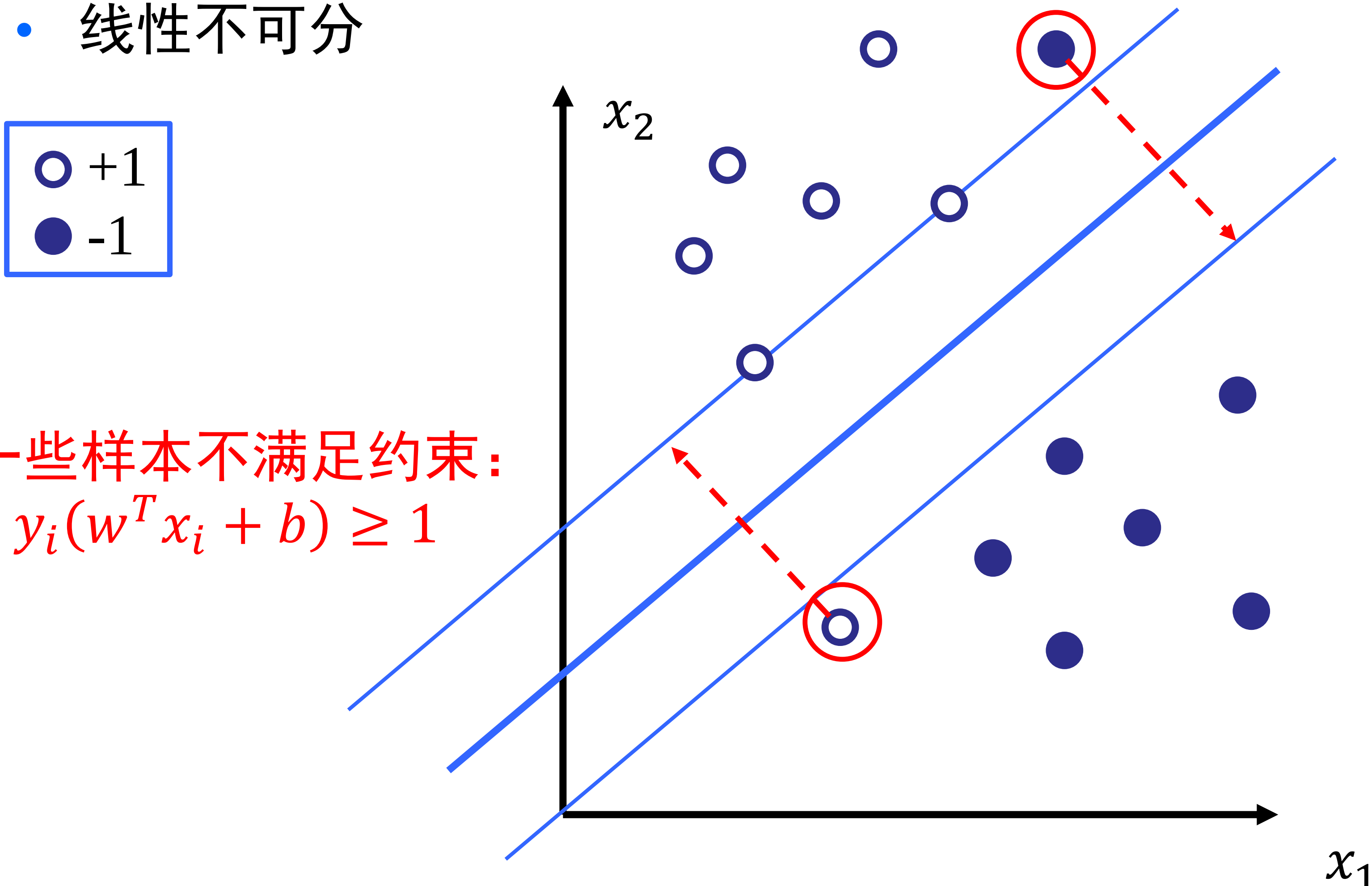
  – SVM的重要性质：最终模型仅与支持向量有关

内容提要

- 支持向量机
  - 4.1 结构风险、经验风险与VC维
  - 4.2 间隔与支持向量
  - 4.3 对偶问题
  - **4.4 软间隔与正则化**
  - 4.5 支持向量回归
  - *4.6 核函数
  - *4.7 核方法
  - *4.8 核支持向量机

4.4 软间隔与正则化

- 线性不可分



有一些样本不满足约束：
$y_i(w^T x_i + b) \geq 1$

4.4 软间隔与正则化

- 当然，在最大化间隔的同时，不满足约束条件的样本应该尽可能少，于是，优化目标可以写为：

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

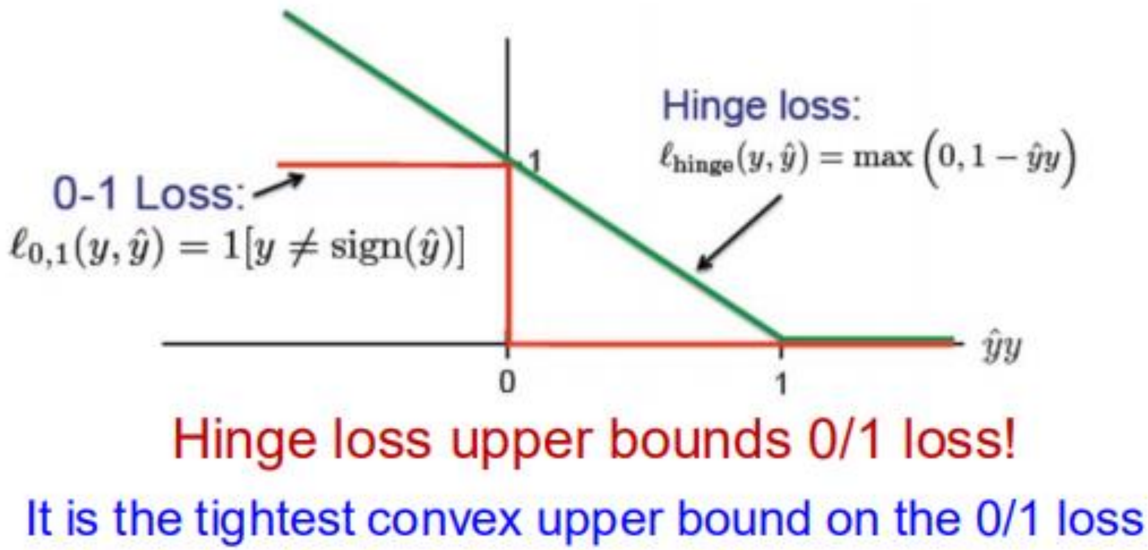
其中C是一个常数，$l_{0/1}$是“0/1损失函数”：

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

4.4 软间隔与正则化

- 然而， $l_{0/1}$非凸、不连续，性质不好，常用hinge损失函数作为替代损失函数：

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



4.4 软间隔与正则化

- 若采用hinge损失，则优化目标变成：

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

- 引入“松弛变量” $\xi_i \geq 0$，可重写为：

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

简化竖直距离

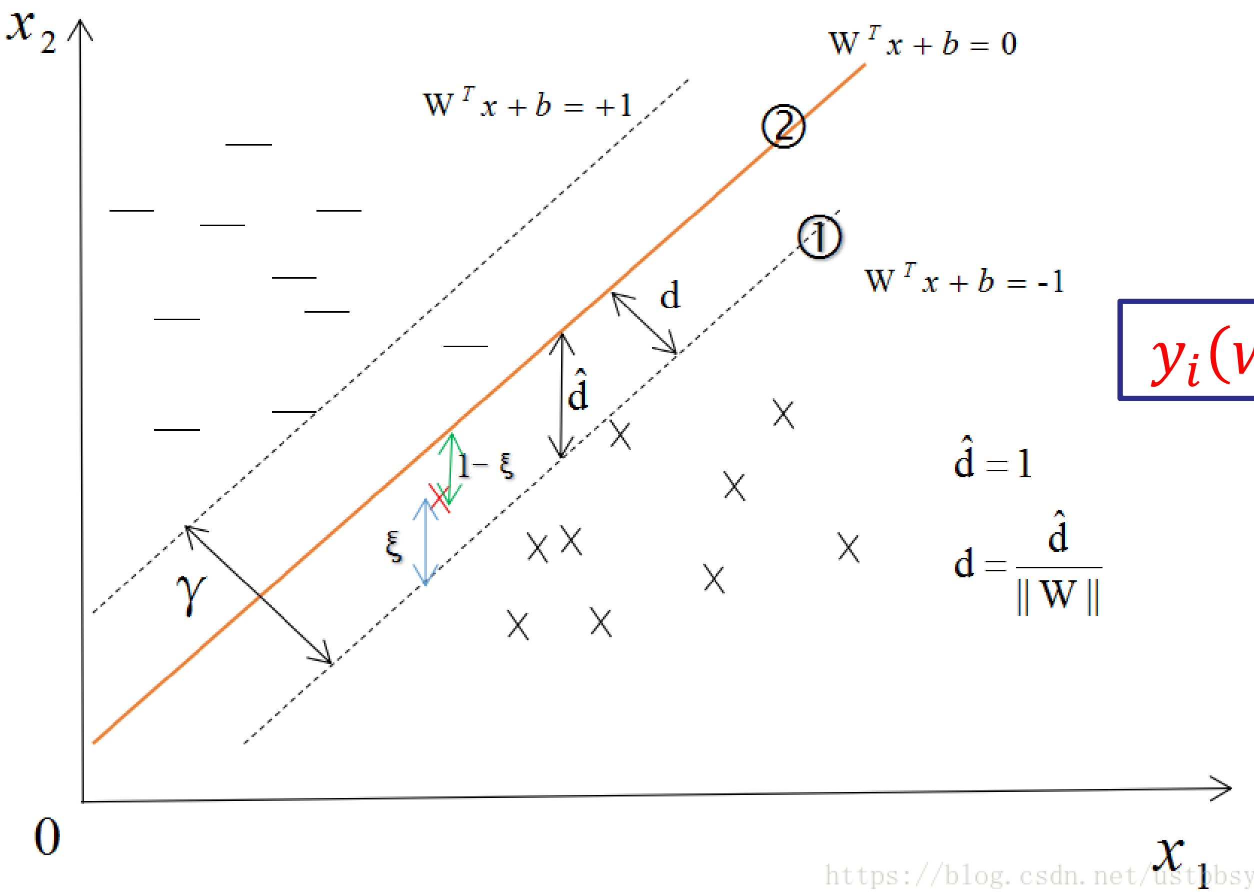
$$s.t.\;\; y_i(w^T x_i + b) \geq 1 - \xi_i \,,$$

$$\xi_i \geq 0,$$

$$i = 1,2,\ldots,n$$

4.4 软间隔与正则化

- 线性不可分



$$y_i(w^T x_i + b) + \xi_i \geq +1$$

4.4 软间隔与正则化

- 硬间隔与软间隔
  - 线性可分情况：
    - $y_i(w^T x_i + b) \geq +1$ 硬间隔
  - 对于线性不可分情况，对约束条件引入松弛变量，允许有少量样本落在两类分类间隔中间：

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, min \overbrace{\sum_{i=1}^{n} \xi_i}^{\text{结构风险}}$$

软间隔

4.4 软间隔与正则化

- 线性不可分——学习模型

体现了表达能力　　　体现了经验风险

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

$$s.t.\quad y_i(w^T x_i + b) \ge 1 - \xi_i, \xi_i \ge 0,\quad i = 1,2,\ldots,n$$

目标函数第一项表示使间隔尽可能大，第二项使得误差分类点的个数尽可能小

4.4 软间隔与正则化

- 软间隔最大化（原始问题）：

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

$$s.t.\quad y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,\ldots,n$$

  - 拉格朗日函数：

$L(w, b, \xi, \alpha, \mu)$

$$= \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i y_i(w^T x_i + b + \xi_i) + \sum_{i}^{n}\alpha_i + \sum_{i}^{n}\mu_i\xi_i$$

  - 拉格朗日对偶

$$\max_{\alpha>0,\mu>0}\ \min_{w,b,\xi} L(w, b, \xi, \alpha, \mu)$$

4.4 软间隔与正则化

- 令$L(w,b,\xi,\alpha,\mu)$对w,b,ξ_i的偏导为0可得：

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i\,，\ 0=\sum_{i=1}^{m}\alpha_i y_i，\ C=\alpha_i+\mu_i$$

- 代回得对偶问题：

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

$$s.t.\ \sum_{i=1}^{n}\alpha_i y_j=0;\ \ 0\le\alpha_i\le C, i=1,2,\ldots,n$$

4.4 软间隔与正则化

- 类似，还需满足KKT条件：

$$\begin{cases} \alpha_i \geq 0, \qquad \mu_i \geq 0 \\ y_i f(x_i) - 1 + \xi_i \geq 0 \\ \alpha_i(y_i f(x_i) - 1 + \xi_i) = 0 \\ \xi_i \geq 0, \qquad \mu_i \xi_i = 0 \end{cases}$$

- 共需满足7条约束：

$$\begin{cases} \alpha_i \geq 0, \qquad \mu_i \geq 0 & (1) \\ y_i f(x_i) - 1 + \xi_i \geq 0 & (2) \\ \alpha_i(y_i f(x_i) - 1 + \xi_i) = 0 & (3) \\ \xi_i \geq 0, \qquad \mu_i \xi_i = 0 & (4) \end{cases}$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \quad (5)$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \quad (6)$$

$$C = \alpha_i + \mu_i \quad (7)$$

4.4 软间隔与正则化

- 由(3),对任意样本(x_i, y_i)，总有:

  $\alpha_i = 0$ 或 $y_i f(x_i) = 1 - \xi_i$

  – 若$\alpha_i = 0$,则该样本不会对$f(x)$有任何影响

$$\begin{cases} \alpha_i \geq 0, \qquad \mu_i \geq 0 & (1) \\ y_i f(x_i) - 1 + \xi_i \geq 0 & (2) \\ \alpha_i(y_i f(x_i) - 1 + \xi_i) = 0 & (3) \\ \xi_i \geq 0, \qquad \mu_i \xi_i = 0 & (4) \end{cases}$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \quad (5)$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \quad (6)$$

$$C = \alpha_i + \mu_i \quad (7)$$

中国科学院大学
University of Chinese Academy of Sciences

4.4 软间隔与正则化

- 若 $\alpha_i > 0$，则 $y_i f(x_i) = 1 - \xi_i$，该样本是支持向量
  - 由(7)，若 $\alpha_i < C$，则 $\mu_i > 0$，有 $\xi_i = 0$，该样本在最大分隔边界上
  - 由(7)，若 $\alpha_i = C$，则 $\mu_i = 0$，
    - 若 $\xi_i \leq 1$，该样本在最大间隔内部
    - 若 $\xi_i > 1$，该样本错分

$$
\begin{cases}
\alpha_i \geq 0, \quad \mu_i \geq 0 & (1) \\
y_i f(x_i) - 1 + \xi_i \geq 0 & (2) \\
\alpha_i (y_i f(x_i) - 1 + \xi_i) = 0 & (3) \\
\xi_i \geq 0, \quad \mu_i \xi_i = 0 & (4)
\end{cases}
$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \quad (5)$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \quad (6)$$

$$C = \alpha_i + \mu_i \quad (7)$$

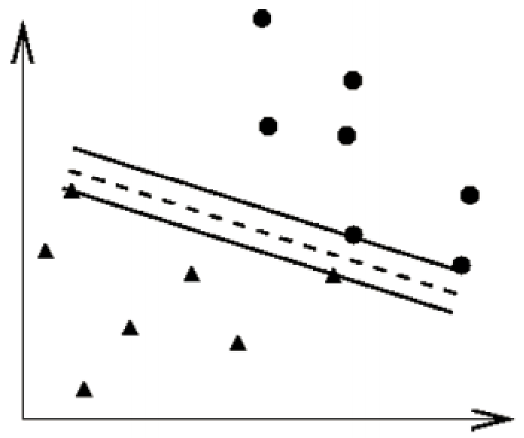
软间隔支持向量机的最终模型仅与支持向量有关！

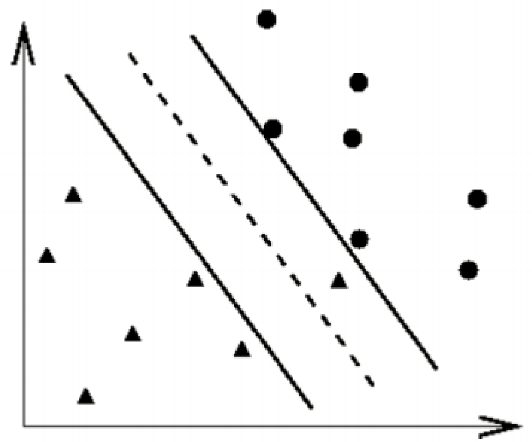
4.4 软间隔与正则化

- **软间隔支持向量**
  - 支撑面（两个类边界）以外的样本点，均有$\alpha_i = 0$
  - 支持向量：$\alpha_i > 0$：包含位于边界上的点、两个类边界以内的、以及错分点（边界以外）
  - 位于类边界上的点，其对应的拉格朗日乘子可能有如下三种情形：
    - $\alpha_i = 0$（正好不是支持向量）、$0 < \alpha_i < C$、$\alpha_i = C$

4.4 软间隔与正则化

- **C的选择对分界面的影响**





C值较大，更加关心错分样本，倾向于产生没有错分样本的分界面

C值较小，更加关心分类间隔，倾向于产生大间隔的分界面

通过选择合适的C值，适当的注意分类间隔，能减少过拟合。

4.4 软间隔与正则化

- **如何确定C?**
  - 交叉验证（cross-validation）：将训练集分成p等份，依次进行p次分类器学习-分类器测试过程。
    - 每次选择$p-1$份数据训练分类器（SVM模型），在剩下的1份数据集上进行测试。
    - 交叉验证的正确率为p次测试的平均结果。
  - 模型参数值设置的技术路线：通过对不同的C值进行交叉验证，取正确率最高的C值，在所有训练数据上重新学习SVM模型。

内容提要

- 支持向量机
  - 4.1 结构风险、经验风险与VC维
  - 4.2 间隔与支持向量
  - 4.3 对偶问题
  - 4.4 软间隔与正则化
  - **4.5 支持向量回归**
  - *4.6 核函数
  - *4.7 核方法
  - *4.8 核支持向量机

4.5 支持向量回归

- 考虑回归问题：
  - 给定训练样本$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}, y_i \in R$，希望学得一个如下一个回归模型，使得$f(x)$与y尽可能接近：

$$f(x) = w^T + b, x \in R^d$$

  - 传统线性最小二乘法（正则化）：

$$\min_{w,b} \sum_{i=1}^{n} (w^T x_i + b - y_i)^2 + \lambda ||w||_2^2$$

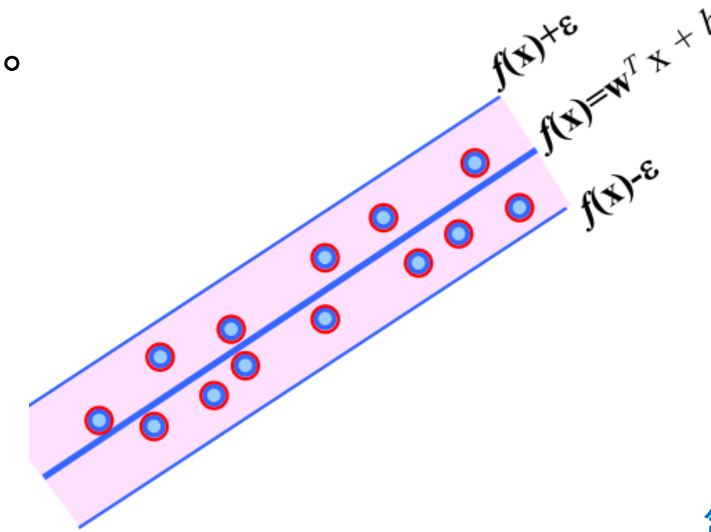
*理论上，$f(x)$可以是任意函数，本节以线性函数为例

中国科学院大学 University of Chinese Academy of Sciences

4.5 支持向量回归

- 支持向量回归(Support Vector Regression, SVR)
  - 对于样本(x, y)，传统回归样本直接基于模型输出$f(x)$与真实输出y之间的差别来计算损失，当且仅当二者完全相同时才为0；
  - 支持向量回归假设$f(x)$与y之间可以有ε的容忍偏差，在这个偏差内，不计算损失。

如果所有的样本点都在一个宽度为2ε的管道内，我们得到了一个很好的回归！



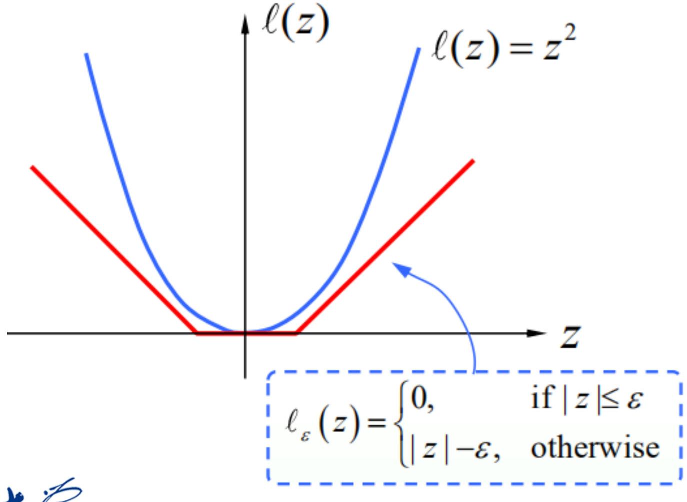
4.5 支持向量回归

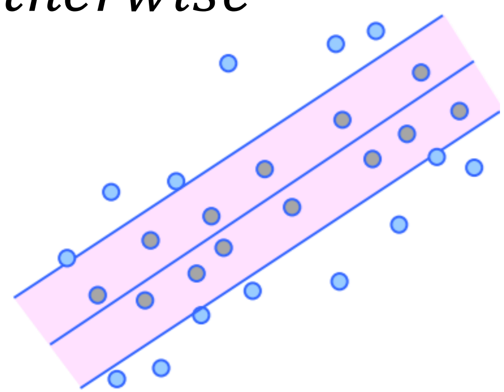
- 则SVR问题可形式化为：

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

其中C是正则化常数，l_ε是ε − 不敏感损失函数：

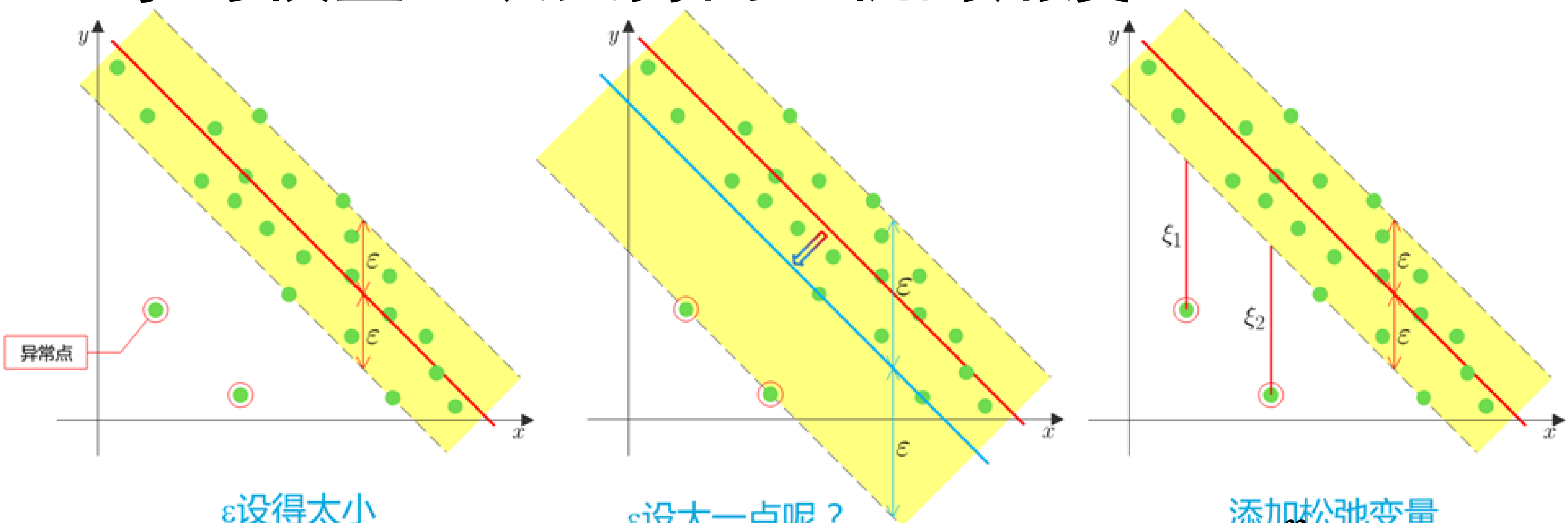
$$l_\varepsilon(z) = \begin{cases} 0, if\ |z| \le \varepsilon \\ |z| - \varepsilon, otherwise \end{cases}$$





4.5 支持向量回归

- 学习模型（从支持向量机的角度）



ε设得太小　　ε设大一点呢？　　添加松弛变量

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

$$s.t.\ \ |f(x_i) - y_i| \le \varepsilon,$$

$$i = 1,2,\dots,n$$

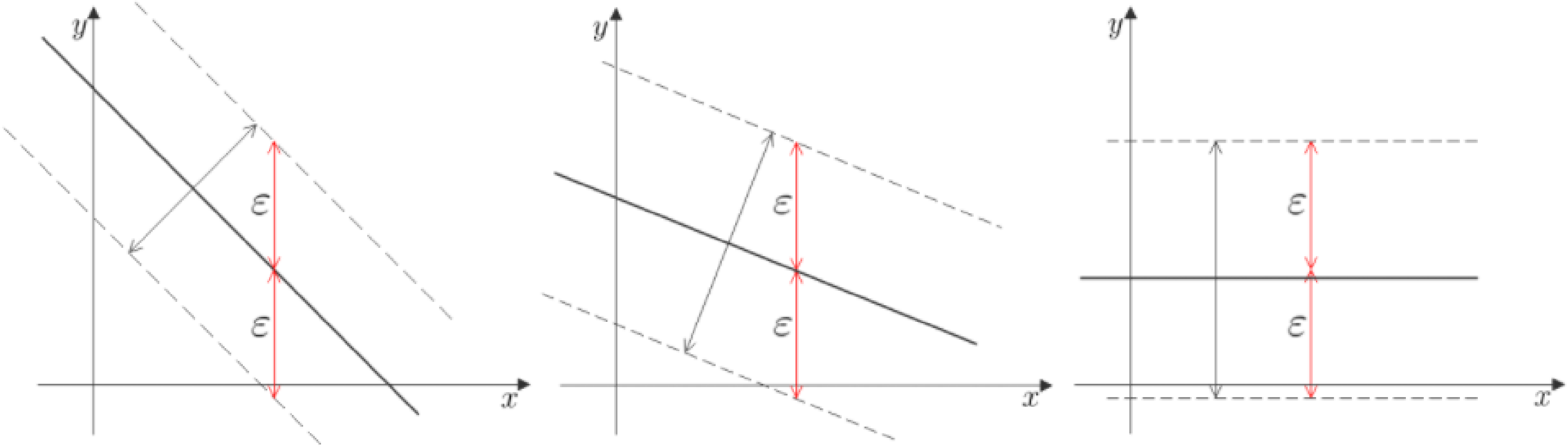
$$\min_{w,b} \frac{1}{2} ||w||_2^2 + C\sum_{i=1}^{n} \xi_i$$

$$s.t.\ \ |f(x_i) - y_i| - \xi_i \le \varepsilon, \xi_i \ge 0,$$

$$i = 1,2,\dots,n$$

4.5 支持向量回归

- 学习模型（从支持向量机的角度）



ε不变的前提下，哪张图中的间隔最大？

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

w越小 越接近e

$$s.t.\ \ |f(x_i) - y_i| \le \varepsilon,$$

$$i = 1,2, \dots, n$$

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

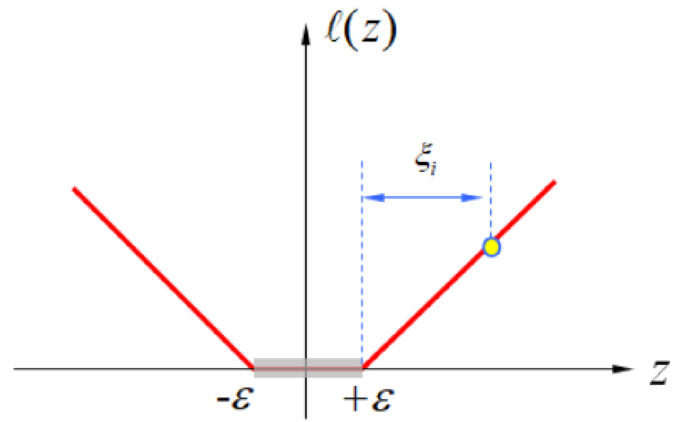
$$s.t.\ \ |f(x_i) - y_i| - \xi_i \le \varepsilon, \xi_i \ge 0,$$

$$i = 1,2, \dots, n$$

4.5 支持向量回归

$$\min_{w,b} \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n}\xi_i$$
$$s.t.\ \ |f(x_i) - y_i| - \xi_i \le \varepsilon,$$
$$\xi_i \ge 0,$$
$$i = 1,2,\dots,n$$

$$\min_{w,b} \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n}\xi_i$$
$$s.t.\ \ -\varepsilon - \xi_i \le f(x_i) - y_i \le \varepsilon + \xi_i,$$
$$\xi_i \ge 0,$$
$$i = 1,2,\dots,n$$



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

4.5 支持向量回归

- 松弛模型
  - 引入松弛变量ξ_i和 $\hat{\xi}_i$，可重写SVR问题的形式：

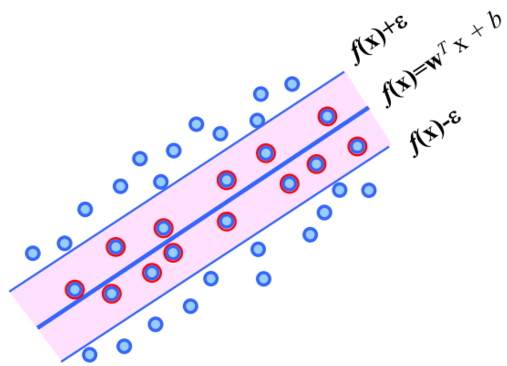
$$\min_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i)$$

$$s.t.\ f(x_i) - y_i \le \varepsilon + \xi_i,$$

$$y_i - f(x_i) \le \varepsilon + \hat{\xi}_i,$$

$$\xi_i \ge 0, \hat{\xi}_i \ge 0, i = 1,2,\ldots,n$$

*间隔带两侧的松弛程度可以不同



在所有样本点中，只有分布在“管壁”上的那一部分样本点决定管道的位置。这一部分训练样本称为“支持向量”。

4.5 支持向量回归

- 松弛模型的广义拉格朗日函数

$$L(w, b, \alpha, \hat{\alpha}, \xi_i, \hat{\xi}_i, \mu, \hat{\mu})$$

$$= \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) + \sum_{i=1}^{n}\mu_i\xi_i - \sum_{i=1}^{n}\hat{\mu}_i\hat{\xi}_i$$

$$+ \sum_{i=1}^{n}\alpha_i(f(x_i) - y_i - \varepsilon - \xi_i) + \sum_{i=1}^{n}\hat{\alpha}_i(y_i - f(x_i) - \varepsilon - \hat{\xi}_i)$$

绝对值

4.5 支持向量回归

- *松弛模型求解
  - 将$f(x)=w^T+b$代入，再令$L(w,b,\alpha,\hat{\alpha},\xi_i,\hat{\xi}_i,\mu,\hat{\mu})$对$w,b,\xi_i,\hat{\xi}_i$的偏导为0可得：

只有一个非0

$$
\begin{cases}
w=\sum_{i=1}^{n}(\alpha_i-\hat{\alpha}_i)x_i \\
0=\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i) \\
C=\alpha_i+\mu_i \\
C=\hat{\alpha}_i+\hat{\mu}_i
\end{cases}
$$

4.5 支持向量回归

- *松弛模型求解
  - 将 (2)-(5)代回，即可得到SVR的对偶问题：

$$\max_{\alpha,\hat{\alpha}} \sum_{i=1}^{n}\left(y_i(\hat{\alpha}_i - \alpha_i) - \varepsilon(\hat{\alpha}_i + \alpha_i)\right)$$

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)x_i^T x_j$$

$$s.t. \sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i) = 0, 0 \le \alpha_i, \hat{\alpha}_i \le C$$

$$f(x) = w^T + b \quad (1)$$

$$w = \sum_{i=1}^{n}(\alpha_i - \hat{\alpha}_i)x_i \quad (2)$$

$$0 = \sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i) \quad (3)$$

$$C = \alpha_i + \mu_i \quad (4)$$

$$C = \hat{\alpha}_i + \hat{\mu}_i \quad (5)$$

4.5 支持向量回归

- *松弛模型求解
  - 同样需要满足KKT条件：

$$\begin{cases} \alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)=0 \\ \hat{\alpha}_i\left(y_i-f(x_i)-\varepsilon-\hat{\xi}_i\right)=0 \\ \alpha_i\hat{\alpha}_i=0 \\ \xi_i\hat{\xi}_i=0 \\ (C-\alpha_i)\xi_i=0 \\ (C-\hat{\alpha}_i)\hat{\xi}_i=0 \end{cases}$$

- 共需满足11条约束：

$$f(x)=w^T+b\ (1)$$

$$w=\sum_{i=1}^{n}(\alpha_i-\hat{\alpha}_i)x_i\ (2)$$

$$0=\sum_{i=1}^{n}(\alpha_i-\hat{\alpha}_i)\ (3)$$

$$C=\alpha_i+\mu_i\ (4)$$

$$C=\hat{\alpha}_i+\hat{\mu}_i\ (5)$$

$$\begin{cases} \alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)=0\ (6) \\ \hat{\alpha}_i\left(y_i-f(x_i)-\varepsilon-\hat{\xi}_i\right)=0\ (7) \\ \alpha_i\hat{\alpha}_i=0\ (8) \\ \xi_i\hat{\xi}_i=0\ (9) \\ (C-\alpha_i)\xi_i=0\ (10) \\ (C-\hat{\alpha}_i)\hat{\xi}_i=0\ (11) \end{cases}$$

4.5 支持向量回归

- *松弛模型求解
  - 由 (6) 可以看出，当且仅当 $f(x_i)-y_i-\varepsilon-\xi_i=0$，$\alpha_i$ 可以取非零值。
    - 即仅当样本(x_i,y_i)不落入ε-间隔带中，相应的α_i和$\hat{\alpha}_i$才能取非零值。
  - 此外，约束 $f(x_i)-y_i-\varepsilon-\xi_i=0$和$y_i-f(x_i)-\varepsilon-\hat{\xi}_i=0$不能同时成立，因此$\alpha_i$和$\hat{\alpha}_i$至少有一个为0

$$f(x)=w^T+b \quad (1)$$

$$w=\sum_{i=1}^{n}(\alpha_i-\hat{\alpha}_i)x_i \quad (2)$$

$$0=\sum_{i=1}^{n}(\alpha_i-\hat{\alpha}_i) \quad (3)$$

$$C=\alpha_i+\mu_i \quad (4)$$

$$C=\hat{\alpha}_i+\hat{\mu}_i \quad (5)$$

$$\begin{cases}\alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)=0 & (6)\\ \hat{\alpha}_i\left(y_i-f(x_i)-\varepsilon-\hat{\xi}_i\right)=0 & (7)\\ \alpha_i\hat{\alpha}_i=0 & (8)\\ \xi_i\hat{\xi}_i=0 & (9)\\ (C-\alpha_i)\xi_i=0 & (10)\\ (C-\hat{\alpha}_i)\hat{\xi}_i=0 & (11)\end{cases}$$

4.5 支持向量回归

- *松弛模型求解
  - 将(2)代入(1)，则SVR的解形如：

$$f(x)=\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)x_i^T x+b$$

  - 由(6)和(10)可知，在获得α_i后，如果$0<\alpha_i<C$，必有$\xi_i=0$，则：

$$b=y_i+\varepsilon-\sum_{j=1}^{n}(\hat{\alpha}_i-\alpha_i)x_j^T x_i$$

*b可以取多个点的平均

$$f(x)=w^T+b \quad (1)$$

$$w=\sum_{i=1}^{n}(\alpha_i-\hat{\alpha}_i)x_i \quad (2)$$

$$0=\sum_{i=1}^{n}(\alpha_i-\hat{\alpha}_i) \quad (3)$$

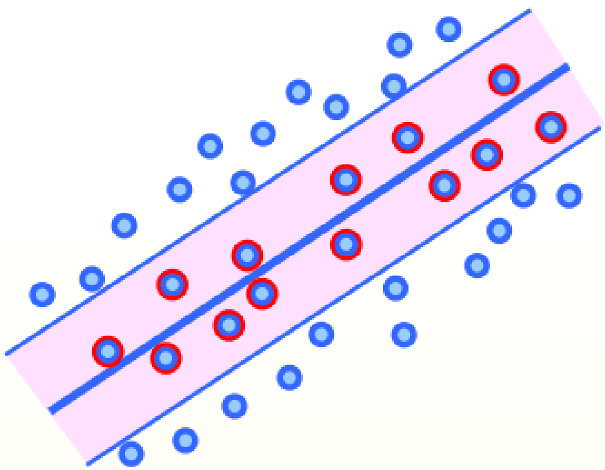
$$C=\alpha_i+\mu_i \quad (4)$$

$$C=\hat{\alpha}_i+\hat{\mu}_i \quad (5)$$

$$\begin{cases} \alpha_i(f(x_i)-y_i-\varepsilon-\xi_i)=0 & (6)\\ \hat{\alpha}_i\left(y_i-f(x_i)-\varepsilon-\hat{\xi}_i\right)=0 & (7)\\ \alpha_i\hat{\alpha}_i=0 & (8)\\ \xi_i\hat{\xi}_i=0 & (9)\\ (C-\alpha_i)\xi_i=0 & (10)\\ (C-\hat{\alpha}_i)\hat{\xi}_i=0 & (11) \end{cases}$$

4.5 支持向量回归

- 支持向量



$$f(x_i) - y_i - \varepsilon - \xi_i = 0 \quad \rightarrow \quad \alpha_i > 0$$
$$y_i - f(x_i) - \varepsilon - \hat{\xi}_i = 0 \quad \rightarrow \quad \hat{\alpha}_i > 0$$

上述条件有一个成立即表示

该点必定落在ε –间隔带之外

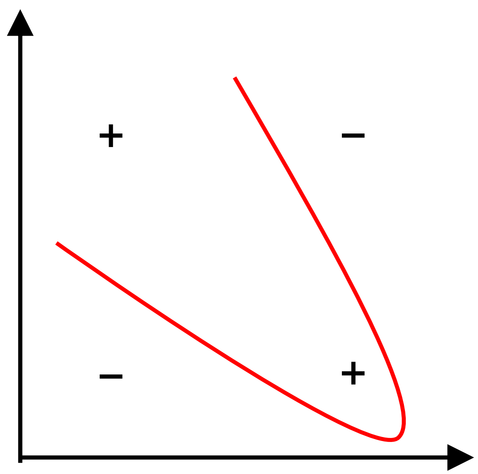
当$\hat{\alpha}_i - \alpha_i \neq 0$时，所对应的点为支持向量，它们必定落在$\varepsilon$–间隔带之外

内容提要

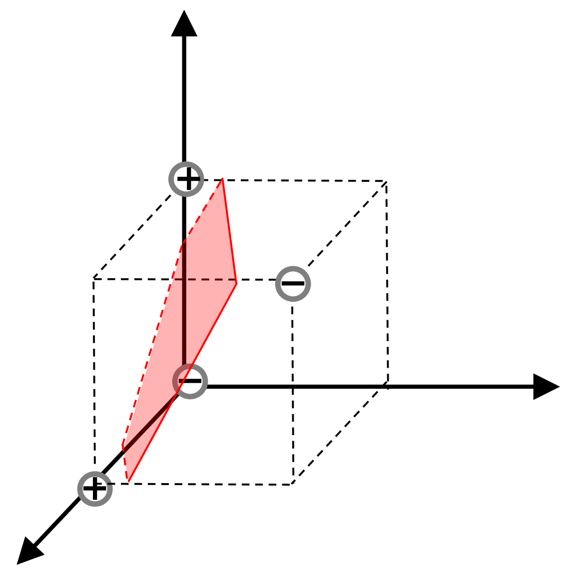
- 支持向量机
  - 4.1 结构风险、经验风险与VC维
  - 4.2 间隔与支持向量
  - 4.3 对偶问题
  - 4.4 软间隔与正则化
  - 4.5 支持向量回归
  - ***4.6 核函数***
  - *4.7 核方法
  - *4.8 核支持向量机

*4.6 核函数

- 线性不可分
  - 不存在一个能正确划分两类样本的超平面，除了软间隔方法，还可以将样本从原始空间映射到一个更高维的特征空间，使得样本在这个空间内线性可分



$x \mapsto \phi(x)$



*4.6 核函数

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

原始问题

$$\min_{w,b} \frac{1}{2}||w||^2$$
$$s.t.\ y_i(w^T\phi(x) + b) \geq 1, i = 1,2,..,n$$

对偶问题

$$\max_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \phi(x_i)^T\phi(x_j) - \sum_{i=1}^{n}\alpha_i$$
$$s.t. \sum_{i=1}^{n}\alpha_i\, y_i = 0, \alpha_i \geq 0, i = 1,2,..,n$$

预测 $f(x) = w^T\phi(x) + b = \sum_{i=1}^{n}\alpha_i\, y_i \boxed{\phi(x_i)^T\phi(x_j)}$+b

只以内积形式出现

中国科学院大学 University of Chinese Academy of Sciences

*4.6 核函数

- 基本想法：
  - 不显示的设计核映射，而设计核函数

$$\kappa\left(x_i, x_j\right) = \phi(x_i)^T \phi(x_j)$$

- Mercer定理（充分不必要）：

  只要一个对称函数所对应的核矩阵半正定，则它就能作为核函数来使用

$$\text{核矩阵K} = \begin{bmatrix} \kappa(x_1,x_1) & \kappa(x_1,x_2) & \kappa(x_1,x_3) & \cdots & \kappa(x_1,x_m) \\ \kappa(x_2,x_1) & \kappa(x_2,x_2) & \kappa(x_2,x_3) & \cdots & \kappa(x_2,x_m) \\ \kappa(x_3,x_1) & \kappa(x_3,x_2) & \kappa(x_3,x_3) & \cdots & \kappa(x_3,x_m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \kappa(x_m,x_1) & \kappa(x_m,x_2) & \kappa(x_m,x_3) & \cdots & \kappa(x_m,x_m) \end{bmatrix}$$

*4.6 核函数

- 基本想法：
  - 不显示的设计核映射，而设计核函数

$$\kappa(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

- Mercer定理（充分不必要）：

只要一个对称函数所对应的核矩阵半正定，则它就能作为核函数来使用

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，$\beta > 0$, $\theta < 0$

*4.6 核函数

- 注意：
  - 核函数选择成为svm的最大变数
  - 经验：文本数据使用线性核，情况不明使用高斯核
  - 核函数的性质：
    - 核函数的线性组合仍为核函数
    - 核函数的直积仍为核函数：

$$\kappa_1 \otimes \kappa_2 (x_i, x_j) = \kappa_1 (x_1, x_2) \otimes \kappa_2 (x_1, x_2)$$

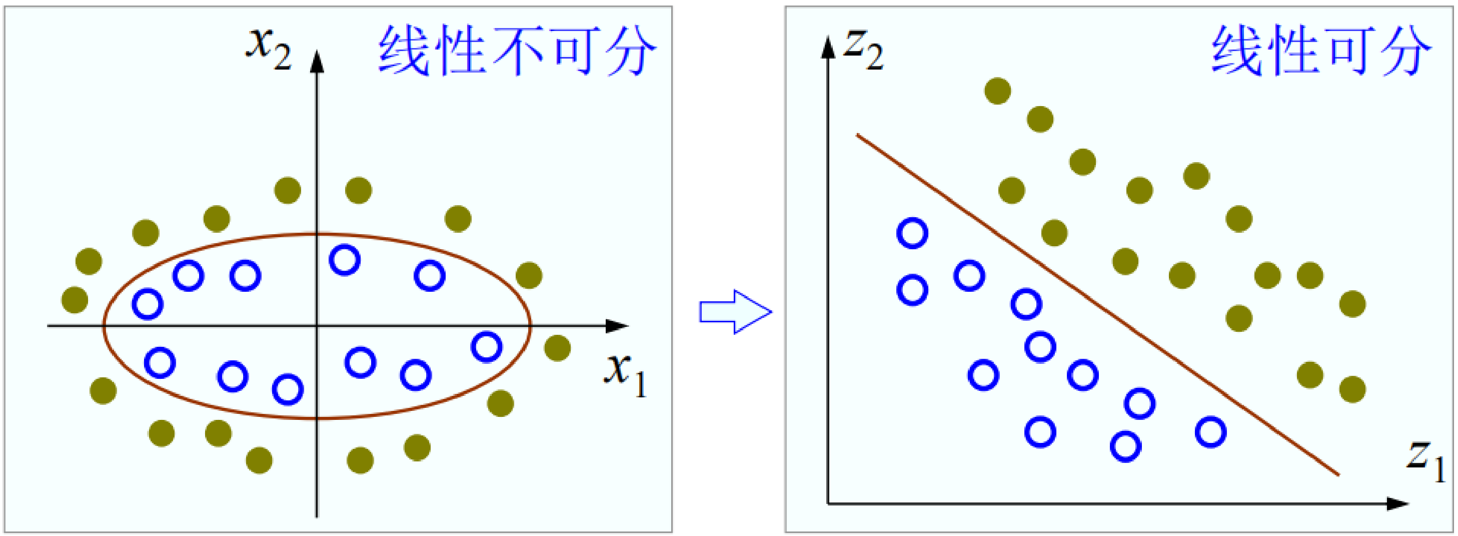
    - 设 $\kappa(x_i, x_j)$ 为核函数，则对于任意函数 g，$g(x_1)\kappa(x_1, x_2)g(x_2)$仍为核函数

内容提要

- 支持向量机
  - 4.1 结构风险、经验风险与VC维
  - 4.2 间隔与支持向量
  - 4.3 对偶问题
  - 4.4 软间隔与正则化
  - 4.5 支持向量回归
  - *4.6 核函数
  - ***4.7 核方法**
  - *4.8 核支持向量机

*4.7 核方法

- 非线性分类问题



椭圆：$w_1x_1^2 + w_2x_2^2 + b = 0$　　直线：$w_1z_1 + w_2z_2 + b = 0$

变换：$z = \phi(x) = (x_1^2, x_2^2)^T$

*4.7 核方法

小样本 核方法 不用nn

- 用线性方法解决非线性问题
  - 第一步，使用一个变换将原空间中的数据映射到新空间
  - 第二步，在新空间里用线性分类学习方法从训练中学习一个分类模型

*4.7 核方法

- 表示定理
  - 令H为核函数K对应的再生核希尔伯特空间，$||h||_H$表示H空间中关于h的范数，对任意单调递增函数$\Omega:[0,\infty]\rightarrow R$和任意非负损失函数$R_m\rightarrow[0,\infty]$，优化问题

$$\min_{h\in H} F(h)=\Omega(||h||_H)+loss(h(x_1),h(x_2),\ldots,h(x_n))$$ 的

解总可以写为 $h^*(x)=\sum_{i=1}^{n}\alpha_i K(x,x_i)$

*4.7 核方法

- 正定核
  - 之前提到过，低维空间到高维空间的映射只以内积形式出现，故可以不构造映射，直接给定一个核函数。
  - $\boldsymbol{K}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ 满足什么条件才能成为核函数？
    - 假定$K(x_1, x_2)$是$X \times X$上的对称函数，并且对于任意的 $x_1, x_2, \ldots, x_m \in X$，$K(x_1, x_2)$ 关于$x_1, x_2, \ldots, x_m$的Gram矩阵是半正定的，则可以依据$K(x_1, x_2)$构造一个希尔伯特空间。

phi的空间

*4.7 核方法

- 正定核
  - 充要条件：设 $K: X \times X \rightarrow R$对称函数(定义在$X \times X$上)，则$K(x_1, x_2)$为正定核的充要条件是对任意 $x_i \in X, i = 1,2, \dots, n, K(x_1, x_2)$ 对应的 Gram 矩阵：$K = \left[K\left(x_i, x_j\right)\right] \in R^{n \times n}$是半正定矩阵。

*4.7 核方法

- Mercer核
  - 设 $K: X \times X \to R$ 是对称函数，$K(x_1, x_2)$为某个特征空间的内积运算的充要条件是，对任意的非零函数 $\phi(x)$, 且$\phi(x)$,平方可积，有：

$$\iint K(x_1, x_2)\phi(x_1)\phi(x_2)dx_1 dx_2 > 0$$

此时， $K(x_1, x_2)$为Mercer核

正定核比Mercer核更具一般性！

内容提要

- 支持向量机
  - 4.1 结构风险、经验风险与VC维
  - 4.2 间隔与支持向量
  - 4.3 对偶问题
  - 4.4 软间隔与正则化
  - 4.5 支持向量回归
  - *4.6 核函数
  - *4.7 核方法
  - ***4.8 核支持向量机**

*4.8 核支持向量机

- KSVM：从对偶问题直接实现SVM核化
  - 训练

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

$$s.t.\ \sum_{i=1}^{n}\alpha_i y_j = 0;\ \ 0 \le \alpha_i \le C, i = 1,2,\dots,n$$

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

$$s.t.\ \sum_{i=1}^{n}\alpha_i y_j = 0;\ \ 0 \le \alpha_i \le C, i = 1,2,\dots,n$$

*4.8 核支持向量机

- KSVM：从对偶问题直接实现SVM核化
  - 预测（对新数据）

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i \left(x_i^T \cdot x\right) + b^*), b^* = y_j - \sum_{i=1}^{n} \alpha_i^* y_i \left(x_i^T \cdot x_j\right)$$

w

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i \, K(x_i, x) + b^*), b^* = y_j - \sum_{i=1}^{n} \alpha_i^* y_i K\left(x_i, x_j\right)$$

测试的时候 和部分支持向量最内积

参考文献

- 周志华，《机器学习》

致谢

- 感谢向世明老师的20版PPT作为原始材料
- 感谢丁雨禾与段俊贤对本PPT的制作与修改

Thank All of You!

(Questions?)

赫 然

rhe@nlpr.ia.ac.cn

https://rhe-web.github.io/

智能感知与计算研究中心(CRIPAC)

中科院自动化研究所· 模式识别国家重点实验室