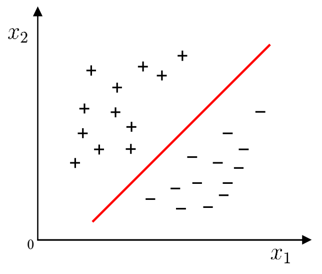
第六章：支持向量机

大纲

- 间隔与支持向量
- 对偶问题
- 核函数
- 软间隔与正则化
- 支持向量回归
- 核方法

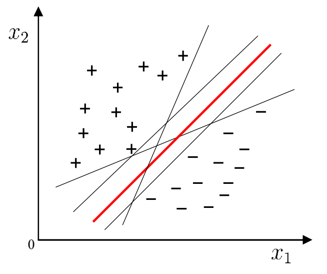
引子

线性模型：在样本空间中寻找一个超平面，将不同类别的样本分开.



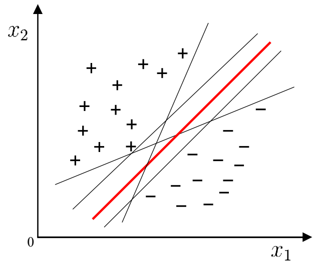
引子

-Q:将训练样本分开的超平面可能有很多，哪一个好呢?



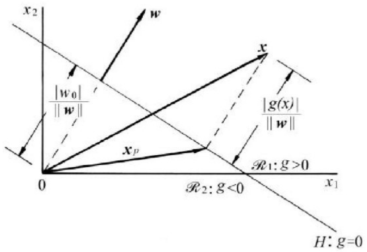
引子

-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"，容忍性好，鲁棒性高，泛化能力最强.

间隔与支持向量



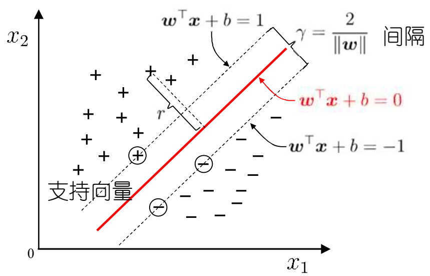
- H:g(x)=w^{T}x+b=0
- 样本空间中任意点x到超平面(w,b) 的距离可写为:

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

超平面(w,b) 关于样本点x的几何间隔

间隔与支持向量

超平面方程: $\boldsymbol{w}^{\top}\boldsymbol{x} + b = 0$



推广性的界

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{h(\ln(2n/h)+1) - \ln(\eta/4)}{n}}$$

h是函数集的VC维,n是样本数。

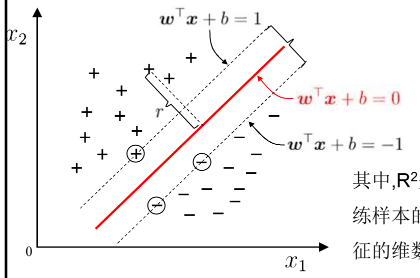
- 学习机器的实际风险由两部分组成:
  - 训练样本的经验风险
  - 置信范围(同置信水平 $1-\eta$ 有关,而且同学习机器的VC维和训练样本数有关。

VC维 confidence

$$R(\alpha) \le R_{emp}(\alpha) + \phi(h/n)$$

间隔与支持向量

超平面方程：$\boldsymbol{w}^\top \boldsymbol{x} + b = 0$



对于规范化的分类超平面，如果权值满足 ||w||≤A,那么这种分类超平面集合的VC维有下面的上界：

h ≤min([R^2A^2], d)+ 1

其中,R^2是样本特征空间中能包含所有训练样本的最小超球体的半径,d是样本特征的维数。

在求最大间隔分类超平面时,最大化分类间隔也就等价于最小化A ,实际上是使VC维上界最小。

支持向量机基本型

- 最大间隔：寻找参数 $\boldsymbol{w}$ 和 b，使得 γ 最大.

$$\arg\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|} \quad \text{s.t. } y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ i = 1,2,\ldots,m.$$

$$\Downarrow$$

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t. } y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ i = 1,2,\ldots,m.$$

带有约束的优化问题
$$\begin{cases} \min_{x\in D} f(x) \\ \text{s.t. } g_i(x) \leq 0, i = 1,2,\ldots,q \\ h_j(x) = 0, j = q+1,\ldots,m \end{cases}$$

其中f(x)是目标函数，g(x)为不等式约束，h(x)为等式约束。

若f(x)，h(x)，g(x)三个函数都是线性函数，则该优化问题称为线性规划。若任意一个是非线性函数，则称为非线性规划。

若目标函数为二次函数，约束全为线性函数，称为二次规划。

若f(x)为凸函数，g(x)为凸函数，h(x)为线性函数，则该问题称为**凸优化**。注意这里不等式约束g(x)<=0则要求g(x)为凸函数，若g(x)>=0则要求g(x)为凹函数。

凸优化的任一局部极值点也是全局极值点，局部最优也是全局最优。

- 已知一个如图所示的训练数据集，其正例点是$x_1=(3,3)^T$，$x_2=(4,3)^T$，负例点是$x_3 =(1,1)$" ,试求最大间隔分离超平面.

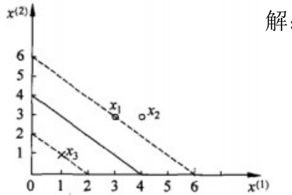


图 7.4 间隔最大分离超平面示例

解：

$$\min_{w,b} \frac{1}{2}(w_1^2 + w_2^2)$$
$$\text{s.t.} \quad 3w_1 + 3w_2 + b \geqslant 1$$
$$4w_1 + 3w_2 + b \geqslant 1$$
$$-w_1 - w_2 - b \geqslant 1$$

求得此最优化问题的解 $w_1 = w_2 = \frac{1}{2}$，$b = -2$．于是最大间隔分离超平面为

$$\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$$

大纲

- 间隔与支持向量
- 对偶问题
- 核函数
- 软间隔与正则化
- 支持向量回归
- 核方法

对偶问题

- 原问题

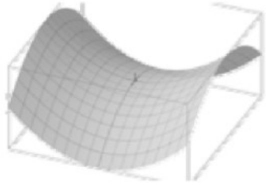
$$\min_{x \in R^n} \quad f(x)$$
$$s.t. \quad c_i(x) \le 0 \quad i = 1, 2..k$$
$$h_j(x) = 0 \quad j =$$

引入拉格朗日函数，由于约束条件有k+l个，

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

其中 α_i, β_j 是拉格朗日乘子， $\alpha_i \ge 0$ ，我们再定

$$\theta_p(x) = \max_{\alpha, \beta, \alpha_i \ge 0} L(x$$



- 可以把原始问题中的s.t.条件去掉，得到原始问题的等价问题：

$$min_x f(x) = min_x max_{\alpha, \beta, \alpha \ge 0} \mathrm{L}（x, \alpha, \beta）$$

对偶问题

$$\min_{x \in R^n} f(x) \quad s.t. \quad c_i(x) \le 0 (i = 1, 2..k), h_j(x) = 0 (j = 1, 2..l)$$

$$\Updownarrow$$

$$\min_{x} \max_{\alpha, \beta, \alpha_i \ge 0} (f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x)) = \min_x \theta_p(x)$$

将 $\max_{\alpha, \beta, \alpha_i \ge 0} (f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x))$ 记做 $P(x)$

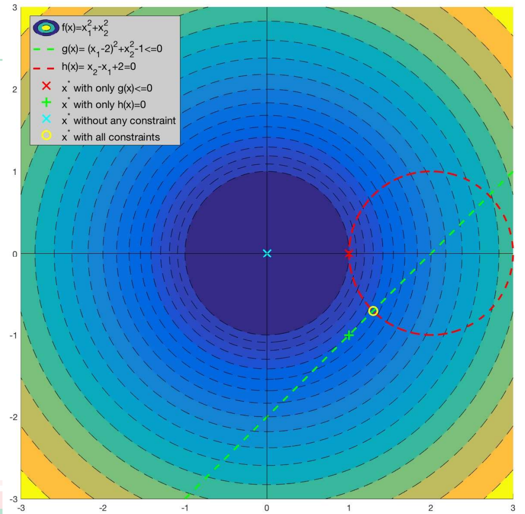
如果x满足 $c_i(x) > 0$或者$h_j(x) \ne 0$ ，则 $P(x)$ 值无限大，则第一步就无解

如果x满足 $c_i(x) \le 0$或者$h_j(x) = 0$ ，又有max下标的条件限制，则一定是

$$\sum_{i=1}^{k} \alpha_i c_i(x) = 0 \text{，} \quad \sum_{j=1}^{l} \beta_j h_j(x) = 0 \text{，则有:}$$

$$\max_{\alpha, \beta, \alpha_i \ge 0} (f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x)) = \max_{\alpha, \beta, \alpha_i \ge 0} (f(x)) = f(x)$$

- 上式两边加 $\min$ ，即证明得等价



KKT条件

针对这一问题，我们可以设计拉格朗日函数如下：

$$L(\boldsymbol{x},\alpha,\beta)=(x_1^2+x_2^2)+\alpha\left[(x_1-2)^2+x_2^2-1\right]+\beta(x_1-x_2-2)$$

根据公式(03)可知：

$$\theta_p(x)=\max_{\alpha,\beta:\alpha\geq 0}\mathcal{L}(x,\alpha,\beta)$$

此时，我们依然可以得到，如果x不满足上面的两个约束条件，即：

1. 若$g(x)>0$；则可以任取α使得$\theta_p(x)$趋于无穷；
2. 若$h(x)\neq 0$；则只有任取β,且$\beta,h(x)$同号，那么$\theta_p(x)$依旧可能趋于无穷；
3. 而只有两个约束条件同时满足，$\theta_p(x)$才可能取到最大值；

对偶问题

$$\frac{1}{2}w^T w+\sum_{i=1}^m\alpha_i-w^T\sum_{i=1}^m\alpha_i y_i x_i-\cancel{\left(\sum_{i=1}^m\alpha_i y_i\right)}b$$

拉格朗日乘子法

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

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

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

$$\boxed{\boldsymbol{w}=\sum_{i=1}^m\alpha_i y_i\boldsymbol{x}_i},\quad 0=\sum_{i=1}^m\alpha_i y_i \longrightarrow -\frac{1}{2}w^T w+\sum_{i=1}^m\alpha_i$$

- 第三步：回代可得

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

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

解的稀疏性

- 最终模型：$f(x)=\operatorname{sgn}((w*,x)+b)=\operatorname{sgn}(\sum_{i=1}^N\alpha_i^* y_i(x_i,x)+b*)$

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

  所有α_i非零的样本公式求解b*后再取平均值

- KKT条件：

$$\begin{cases}\alpha_i\geq 0,\\ y_i f(\boldsymbol{x}_i)\geq 1,\\ \alpha_i(y_i f(\boldsymbol{x}_i)-1)=0.\end{cases}\Rightarrow \text{必有 } \alpha_i=0 \text{ 或 } y_i f(\boldsymbol{x}_i)=1$$

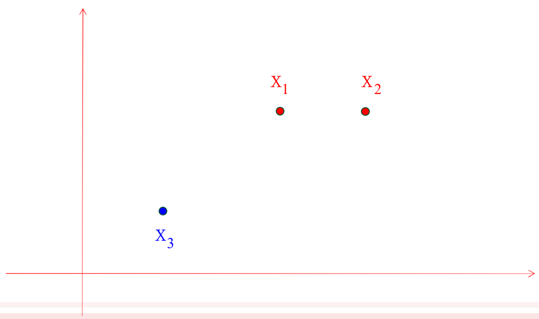
$$y_i f(\boldsymbol{x}_i)>1 \Rightarrow \alpha_i=0$$

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.

https://zhuanlan.zhihu.com/p/38163970

对偶方法重新求解前面的问题

如图所示的训练数据集，其正实例点是$x_1=(3,3)$,$x_2=(4,3)$,负实例$x_3=(1,1)$,试求其线性可分的支持向量机。



对偶方法重新求解前面的问题

例：$\min\limits_{w_1,w_2,b} \frac{1}{2}\left(w_1^2+w_2^2\right)$

$$s.t.\begin{cases}3w_1+3w_2+b\ge 1\\4w_1+3w_2+b\ge 1\\w_1+w_2+b\le -1\end{cases}$$

$$\Rightarrow\quad w_1=w_2=\frac{1}{2},\quad b=-2$$

第一步：转化为对偶问题

$$\min_{\alpha_1,\alpha_2,\alpha_3} -\sum_{i=1}^{3}\alpha_i+\frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)$$

$$s.t.\begin{cases}\alpha_1+\alpha_2-\alpha_3=0\\\alpha_1\ge 0,\alpha_2\ge 0,\alpha_3\ge 0\end{cases}$$

$$\min_{\alpha_1,\alpha_2,\alpha_3}\left\{\frac{1}{2}\left(18\alpha_1^2+25\alpha_2^2+2\alpha_3^2+42\alpha_1\alpha_2-12\alpha_1\alpha_3-14\alpha_2\alpha_3\right)-\alpha_1-\alpha_2-\alpha_3\right\}$$

$$s.t.\begin{cases}\alpha_1+\alpha_2-\alpha_3=0\\\alpha_1\ge 0,\alpha_2\ge 0,\alpha_3\ge 0\end{cases}$$

第二步：代入约束条件 $\alpha_3=\alpha_1+\alpha_2$

目标函数变形为：

$$s(\alpha_1,\alpha_2)=4\alpha_1^2+\frac{13}{2}\alpha_2^2+10\alpha_1\alpha_2-2\alpha_1-2\alpha_2$$

$$\Rightarrow\begin{cases}s_{\alpha_1}(\alpha_1,\alpha_2)=8\alpha_1+10\alpha_2-2\\s_{\alpha_2}(\alpha_1,\alpha_2)=13\alpha_2+10\alpha_1-2\end{cases}$$

$$\Rightarrow\begin{cases}\alpha_1=\frac{3}{2}\\\alpha_2=-1\end{cases}$$ 不符合要求，从而最小值在边界达到。

第三步：利用KKT条件，计算向量w

当 α_1=0 时，

$$s(0,\alpha_2)=\frac{13}{2}\alpha_2^2-2\alpha_2,\ \text{且}\ s(0,\alpha_2)_{\min}\Big|_{\alpha_2=\frac{2}{13}}=-\frac{2}{13}$$

当 α_2=0 时，

$$s(\alpha_1,0)=4\alpha_1^2-2\alpha_1,\ \text{且}\ s(\alpha_1,0)_{\min}\Big|_{\alpha_1=\frac{1}{4}}=-\frac{1}{4}$$

从而：当 $\alpha_1=\frac{1}{4}$，$\alpha_2=0$，时，$s_{\min}=-\frac{1}{4}$. 此时 $\alpha_3=\frac{1}{4}$.

故：$w=\alpha_1y_1x_1+\alpha_3y_3x_3=\left(\frac{1}{2},\frac{1}{2}\right)$

第四步：利用KKT条件，计算b

注意到 $\alpha_1 > 0$, 从而 x_1 为支持向量。

从而 $y_1 f(x_1) = 1 \overset{y_1^2=1}{\Longrightarrow} y_1^2 f(x_1) = y_1 \Rightarrow b = y_1 - w' x_1 = -2$

这样我们就得到了支持向量机(分离超平面)

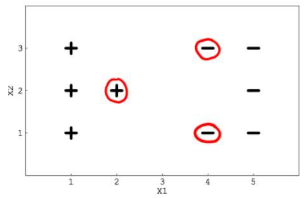
$$\frac{1}{2}x_1 + \frac{1}{2}x_2 - 2 = 0$$

对于新的样本点，我们使用的决策函数为

$$f(x) = \text{sign}\left(\frac{1}{2}x_1 + \frac{1}{2}x_2 - 2\right)$$

如果样本变多，人工计算不现实，需要一种高效的计算算法

- 假设有一个线性SVM分类器用来处理二分类问题，下图显示给定的数据集，其中被红色圈出来的代表支持向量。



  - 若移动其中任意一个红色圈出的点，决策边界是否会变化？
    A.会　B.不会
  - 若移动其中任意一个没有被圈出的点，决策边界会发生变化？
    A.会　B.不会

- 假定现在有一个四分类问题，你要用One-vs-all策略训练一个SVM的模型。请看下面的问题：
- 20. 由题设可知，你需要训练几个SVM模型？
  A.1 B.2 C.3　D.4

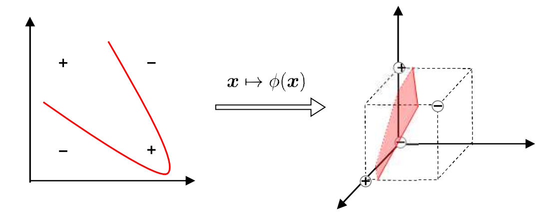
大纲

- 间隔与支持向量
- 对偶问题
- 核函数
- 软间隔与正则化
- 支持向量回归
- 核方法

线性不可分

-Q:若不存在一个能正确划分两类样本的超平面，怎么办?

-A: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.



线性不可分

- 线性判别函数是形式最为简单的判别函数，但是它不能用于复杂情况。

  例：设计一个一维分类器，使其功能为：

$$\text{如果}\begin{cases} x<b \text{ 或 } x>a & \text{则决策}x\in\omega_1 \\ b\le x\le a & \text{则决策}x\in\omega_2 \end{cases}$$

$$\text{判别函数：}\ g(x)=(x-a)(x-b)$$

- 二次函数的一般形式：

$$g(x)=c_0+c_1x+c_2x^2$$

线性不可分

- 二次函数的一般形式：

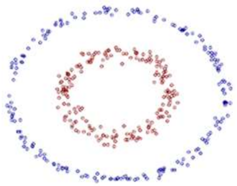
映射X→Y

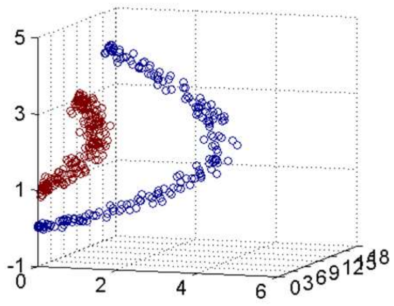
$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}=\begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix},\quad \mathbf{a}=\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}=\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}$$

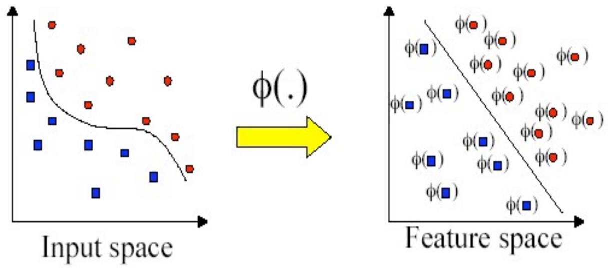
- g(x)又可表示成：

$$g(x)=\mathbf{a}^T\mathbf{y}=\sum_{i=1}^{3}a_iy_i$$

线性不可分







核函数 $K(x_i,x_j) \overset{\Delta}{=} (\phi(x_i)\cdot\phi(x_j))$

33

核支持向量机

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

原始问题

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b) \geq 1,\ i=1,2,\ldots,m.$$

对偶问题

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boxed{\phi(\boldsymbol{x}_i)^\top\phi(\boldsymbol{x}_j)} - \sum_{i=1}^{m}\alpha_i$$
$$\text{s.t. } \sum_{i=1}^{m}\alpha_i y_i = 0,\ \alpha_i \geq 0,\ i=1,2,\ldots,m.$$

预测

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

只以内积的形式出现

核函数

- 基本想法：不显式地设计核映射，而是设计核函数.

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top\phi(\boldsymbol{x}_j)$$

- Mercer定理(充分非必要)：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.

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

核函数

- 基本想法：不显式地设计核映射，而是设计核函数.

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top\phi(\boldsymbol{x}_j)$$

- Mercer定理(充分非必要)：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.

- 常用核函数：

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \boldsymbol{x}_i^\top\boldsymbol{x}_j$	
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	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\delta^2}\right)$	$\delta > 0$为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|}{\delta}\right)$	$\delta > 0$
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, $\beta > 0$, $\theta < 0$

核函数的注意事项：

- 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：
- 1 核函数的线性组合仍为核函数
- 2 核函数的直积仍为核函数

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

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

$$g(x_1)\kappa(x_1, x_2) g(x_2) \text{仍为核函数。}$$

大纲

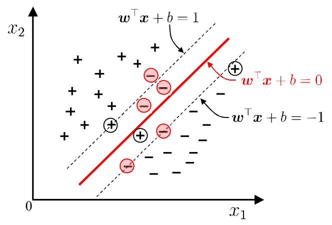
- 间隔与支持向量
- 对偶问题
- 核函数
- 软间隔与正则化
- 支持向量回归
- 核方法

软间隔

-Q:现实中，很难确定合适的核函数使得训练样本在特征空间中线性可分；同时一个线性可分的结果也很难断定是否是有过拟合造成的.

-A:引入"软间隔"的概念，允许支持向量机在一些样本上不满足约束.

不满足约束的样本



0/1损失函数

- 基本想法：最大化间隔的同时，让不满足约束的样本应尽可能少.

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

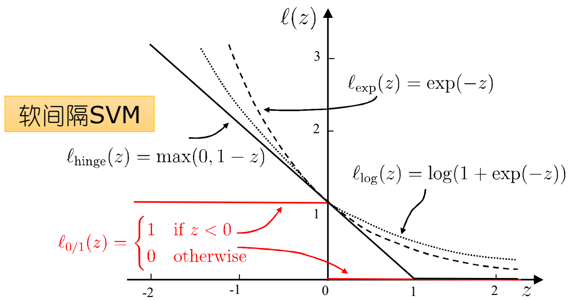
其中$l_{0/1}$是"0/1损失函数"

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

正则化常数 $C > 0$, 如果$C \to \infty$，则等价于要求所有的样本点都分类正确，否则就允许一部分极少的样本分类错误

- 存在的问题：0/1损失函数非凸、非连续，不易优化！

替代损失



替代损失函数数学性质较好，一般是0/1损失函数的上界

软间隔支持向量机

原始问题 $\min\limits_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\max\left(0, 1-y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)\right)$

引入“松弛变量(Slack Variable)” $\xi \geq 0$，

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

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

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

注：每一个样本都对应一个松弛变量 ξ_i,用以表征该样本不满足约束 $y_i f(x_i)\geq 1$ 的程度。

求解软间隔问题：

- 构造Lagrange 函数

$$L(\mathbf{w},b,\alpha,\xi,\mu)=\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{m}\xi_i+\sum_{i=1}^{m}\alpha_i\left(1-\xi_i-y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b\right)\right)-\sum_{i=1}^{m}\mu_i\xi_i$$

$s.t.\ \alpha_i \geq 0, \mu_i \geq 0$

分别对变量求导，并令其为0，得到

$$\mathbf{w}=\sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i$$

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

$$C=\alpha_i+\mu_i$$

软间隔支持向量机

原始问题 $\min\limits_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\max\left(0, 1-y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)\right)$

对偶问题

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

$$\text{s.t.} \sum_{i=1}^{m}\alpha_i y_i = 0,\ 0\leq\alpha_i\leq C,\ i=1,2,\ldots,m.$$

软间隔支持向量机KKT条件

$$
\begin{cases}
\alpha_i \ge 0, \quad \mu_i \ge 0 \\
y_i f(\mathbf{x}_i) - 1 + \xi_i \ge 0 \\
\alpha_i \left(y_i f(\mathbf{x}_i) - 1 + \xi_i \right) = 0 \\
\xi_i \ge 0, \mu_i \xi_i = 0
\end{cases}
$$

根据KKT条件可推得最终模型仅与支持向量有关，也即hinge损失函数依然保持了支持向量机解的稀疏性.

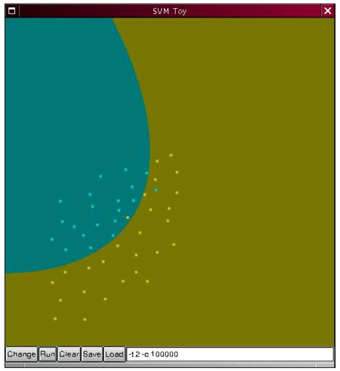
$$\alpha_i^* [y_i (w^T x_i + b) - 1 + \xi_i^*] = 0$$

- 根据这个KKT条件的对偶互补条件，我们有：

$$\alpha_i^* = 0 \Rightarrow y_i (w^* \cdot \phi(x_i) + b) \ge 1$$
$$0 < \alpha_i^* < C \Rightarrow y_i (w^* \cdot \phi(x_i) + b) = 1$$
$$\alpha_i^* = C \Rightarrow y_i (w^* \cdot \phi(x_i) + b) \le 1$$

$$\alpha_i^* = 0 \Rightarrow y_i g(x_i) \ge 1$$
$$0 < \alpha_i^* < C \Rightarrow y_i g(x_i) = 1$$
$$\alpha_i^* = C \Rightarrow y_i g(x_i) \le 1$$

支持向量机实现



正则化

- 支持向量机学习模型的更一般形式

$$\min_f \quad \Omega(f) + C \sum_{i=1}^{m} l(f(\boldsymbol{x}_i), y_i)$$

结构风险，描述模型的某些性质

经验风险，描述模型与训练数据的契合程度

- 通过替换上面两个部分，可以得到许多其他学习模型
  - 对数几率回归(Logistic Regression)
  - 最小绝对收缩选择算子(LASSO)
  - ……

正则化

L_p 范数 (norm) 是常用的正则化项，其中 L_2 范数 $\|\boldsymbol{w}\|_2$ 倾向于 $\boldsymbol{w}$ 的分量取值尽量均衡，即非零分量个数尽量稠密，而 L_0 范数 $\|\boldsymbol{w}\|_0$ 和 L_1 范数 $\|\boldsymbol{w}\|_1$ 则倾向于 $\boldsymbol{w}$ 的分量尽量稀疏，即非零分量个数尽量少.

下图为采用不同核函数或不同的松弛因子得到的 SVM 决策边界。但粗心的实验者忘记记录每个图形对应的模型和参数了。请你帮忙给下面每个模型标出正确的图形。

(1)、$\min\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i\right)$, s.t.

$\xi_i \geq 0,\ y_i\left(\mathbf{w}^T\mathbf{x}+w_0\right) \geq 1-\xi_i,\ i=1,\ldots,N,$

其中 $C=0.1$。

(2)、$\min\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i\right)$, s.t.

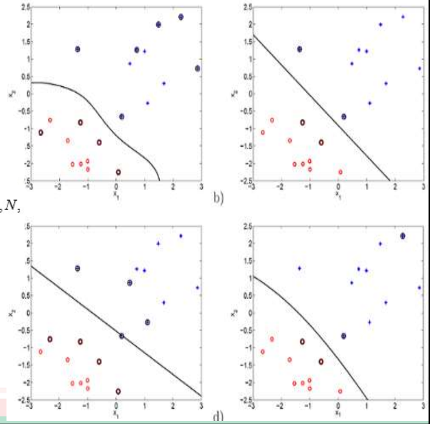
$\xi_i \geq 0,\ y_i\left(\mathbf{w}^T\mathbf{x}+w_0\right) \geq 1-\xi_i,\ i=1,\ldots,N,$

其中 $C=1$。

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

s.t. $0 \leq \alpha_i < C,\ i=1,\ldots,N,\quad \sum_{i=1}^{N}\alpha_i y_i = 0$

其中 $k(\mathbf{x},\mathbf{x}') = \mathbf{x}^T\mathbf{x}' + \left(\mathbf{x}^T\mathbf{x}'\right)^2$。



高效求解方法 – SMO: sequential minimal optimization

- 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.
- 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，回代入对偶问题可得一个单变量的二次规划，该问题具有闭式解.

- 偏移项 b：通过支持向量来确定.

$$\begin{cases} 0 \leq \alpha_1 \leq C \\ 0 \leq \alpha_2 \leq C \\ y_1\alpha_1 + y_2\alpha_2 = -\sum_{i=3}^{n} y_i\alpha_i = \gamma \end{cases} \Rightarrow \alpha_1 = y_1\gamma - y_1 y_2\alpha_2$$

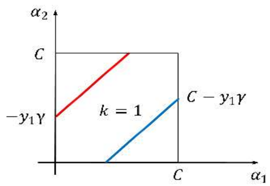


Fig. 1(a)

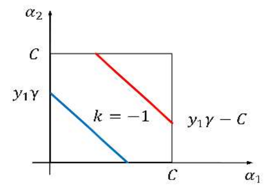


Fig. 1(b)

y_1与y_2异号，此时的直线段斜率为1，如Fig.1(a)所示。分析a2的取值范围：

$$L = \max(0, \alpha_2 - \alpha_1)$$
$$H = \min(C, C + \alpha_2 - \alpha_1)$$

y_1与y_2同号，此时的直线段斜率为-1，如Fig.1(b)所示。分析a2的取值范围：

$$L = \max(0, \alpha_2 + \alpha_1 - C)$$
$$H = \min(C, \alpha_2 + \alpha_1)$$

SMO

- 假如我们通过求导得到的$\alpha^{new,unc}$，则最终的α^{new}应该为：

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

- 如何求导得到的$\alpha_2^{new,unc}$，只需要将目标函数对α_2求偏导数即可，

$$f(\alpha_1,\alpha_2)=\frac{1}{2}K_{11}\alpha_1^2+\frac{1}{2}K_{22}\alpha_2^2+y_1y_2K_{12}\alpha_1\alpha_2+y_1\alpha_1v_1+y_2\alpha_2v_2-\alpha_1-\alpha_2+r$$

$$\alpha_1y_1+\alpha_2y_2=-\sum_{i=3}^{N}\alpha_iy_i=\zeta$$

$$v_1=\sum_{i=3}^{N}\alpha_iy_iK_{i,1}$$

$$v_2=\sum_{i=3}^{N}\alpha_iy_iK_{i,2}$$

两边同时乘上 y_1，由于 $y_iy_i=1$ 得到:

$$\alpha_1=\zeta y_1-\alpha_2y_1y_2$$

$$W(\alpha_2)=\frac{1}{2}K_{11}(\varsigma-\alpha_2y_2)^2+\frac{1}{2}K_{22}\alpha_2^2+y_2K_{12}(\varsigma-\alpha_2y_2)\alpha_2-(\varsigma-\alpha_2y_2)y_1-\alpha_2+v_1(\varsigma-\alpha_2y_2)+y_2v_2\alpha_2$$

$$W(\alpha_2)=\frac{1}{2}K_{11}(\varsigma-\alpha_2y_2)^2+\frac{1}{2}K_{22}\alpha_2^2+y_2K_{12}(\varsigma-\alpha_2y_2)\alpha_2-(\varsigma-\alpha_2y_2)y_1-\alpha_2+v_1(\varsigma-\alpha_2y_2)+y_2v_2\alpha_2$$

对α_2求导

$$\frac{\partial W}{\partial\alpha_2}=K_{11}\alpha_2+K_{22}\alpha_2-2K_{12}\alpha_2-K_{11}\varsigma y_2+K_{12}\varsigma y_2+y_1y_2-1-v_1y_2+y_2v_2$$

$$(K_{11}+K_{22}-2K_{12})\alpha_2=y_2(y_2-y_1+\varsigma K_{11}-\varsigma K_{12}+v_1-v_2)$$
$$=y_2\left[y_2-y_1+\varsigma K_{11}-\varsigma K_{12}+\left(g(x_1)-\sum_{j=1}^{2}y_j\alpha_jK_{1j}-b\right)-\left(g(x_2)-\sum_{j=1}^{2}y_j\alpha_jK_{2j}-b\right)\right]$$

$$(K_{11}+K_{22}-2K_{12})\alpha_2=y_2(y_2-y_1+\varsigma K_{11}-\varsigma K_{12}+v_1-v_2)$$
$$=y_2\left[y_2-y_1+\varsigma K_{11}-\varsigma K_{12}+\left(g(x_1)-\sum_{j=1}^{2}y_j\alpha_jK_{1j}-b\right)-\left(g(x_2)-\sum_{j=1}^{2}y_j\alpha_jK_{2j}-b\right)\right]$$

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

$$E_i=g(x_i)-y_i=\sum_{j=1}^{m}\alpha_j^*y_jK(x_i,x_j)+b-y_i$$

$$(K_{11}+K_{22}-2K_{12})\alpha_2^{new,unc}=y_2((K_{11}+K_{22}-2K_{12})\alpha_2^{old}y_2+y_2-y_1+g(x_1)-g(x_2))$$
$$=(K_{11}+K_{22}-2K_{12})\alpha_2^{old}+y_2(E_1-E_2)$$

$$\alpha_2^{new,unc}=\alpha_2^{old}+\frac{y_2(E_1-E_2)}{K_{11}+K_{22}-2K_{12}}$$

$$W(\alpha_2) = -\frac{1}{2}K_{1,1}(\zeta - \alpha_2 y_2)^2 - \frac{1}{2}K_{2,2}\alpha_2^2 - y_2(\zeta - \alpha_2 y_2)\alpha_2 K_{1,2} - v_1(\zeta - \alpha_2 y_2) - v_2 y_2 \alpha_2 + \alpha_1 + \alpha_2 + C$$

对α_2求导

$$\frac{\partial W}{\partial \alpha_2} = K_{11}\alpha_2 + K_{22}\alpha_2 - 2K_{12}\alpha_2 - K_{11}\varsigma y_2 + K_{12}\varsigma y_2 + y_1 y_2 - 1 - v_1 y_2 + y_2 v_2$$

$$\alpha_2^{new,unc} = \alpha_2^{old} + \frac{y_2(E_1 - E_2)}{K_{11} + K_{22} - 2K_{12}}$$

$$E_i = g(x_i) - y_i = \sum_{j=1}^{m} \alpha_j^* y_j K(x_i, x_j) + b - y_i$$

$$\frac{\partial W}{\partial \alpha_2} = K_{11}\alpha_2 + K_{22}\alpha_2 - 2K_{12}\alpha_2 - K_{11}\varsigma y_2 + K_{12}\varsigma y_2 + y_1 y_2 - 1 - v_1 y_2 + y_2 v_2$$

令其为 0，得到

$$\begin{aligned}(K_{11} + K_{22} - 2K_{12})\alpha_2 &= y_2(y_2 - y_1 + \varsigma K_{11} - \varsigma K_{12} + v_1 - v_2) \\ &= y_2\left[y_2 - y_1 + \varsigma K_{11} - \varsigma K_{12} + \left(g(x_1) - \sum_{j=1}^{2} y_j \alpha_j K_{1j} - b\right) - \left(g(x_2) - \sum_{j=1}^{2} y_j \alpha_j K_{2j} - b\right)\right]\end{aligned}$$

将 $\varsigma = \alpha_1^{old} y_1 + \alpha_2^{old} y_2$ 代入，得到

$$\begin{aligned}(K_{11} + K_{22} - 2K_{12})\alpha_2^{new,unc} &= y_2((K_{11} + K_{22} - 2K_{12})\alpha_2^{old} y_2 + y_2 - y_1 + g(x_1) - g(x_2)) \\ &= (K_{11} + K_{22} - 2K_{12})\alpha_2^{old} + y_2(E_1 - E_2)\end{aligned}$$

SMO 变量选择原则

- 第一个变量是在KKT条件不满足的中间选择，直观来看，KKT条件违背的程度越大，则变量更新后可能会使得目标函数的增幅越大，从而选择违背KKT条件程度越大的变量
- 第二个变量应选择使得目标函数增长最快的变量；常用启发式，也就是两样本的间距最大

- b值的计算

$$b^{new} = \begin{cases} b_1^{new}, & 0 < \alpha_1 < C \\ b_2^{new}, & 0 < \alpha_2 < C \\ (b_1^{new} + b_2^{new})/2, & otherwise \end{cases}$$

SOM算法

- 总体流程
  - 1.初始化,一般情况下令初始的 α_i 全部为0;
  - 2.选取优化变量α_1和α_2,执行相关的优化计算，得到更新后的α_1和α_2
  - 3.开始新的一轮迭代，重复执行上面的第2步，知道全部的 α_i满足KKT条件以及约束条件;

大纲

- 间隔与支持向量
- 对偶问题
- 核函数
- 软间隔与正则化
- 支持向量回归
- 核方法