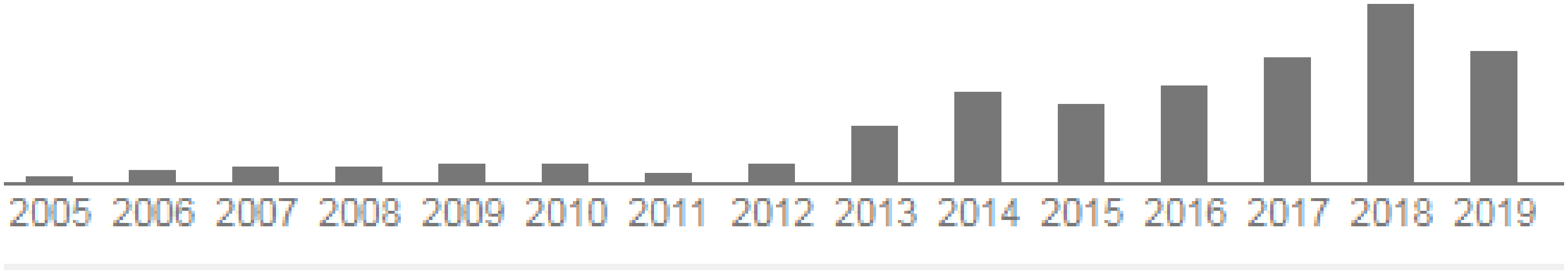
SVM-支持向量机

Support Vector Machine

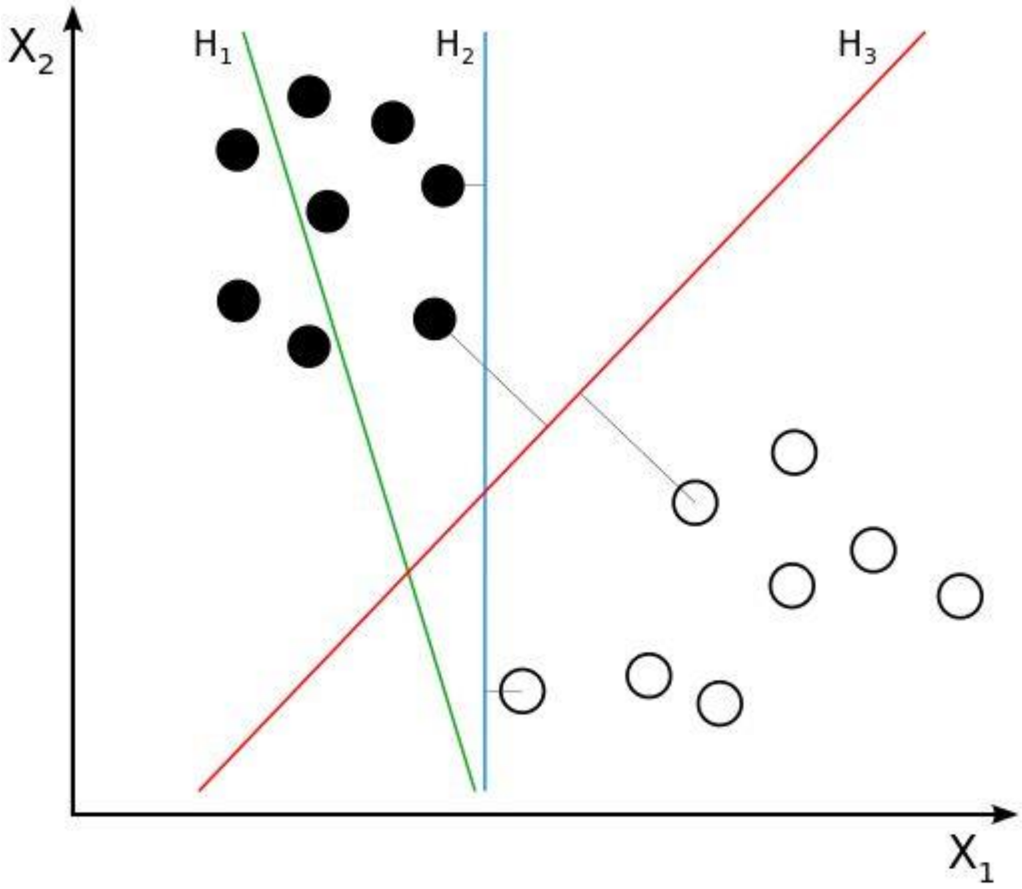
- 提出时间:1964年
- 应用领域：人像识别，文本分类等模式识别问题中均有应用
- 当前标准的前身（软间隔）由Corinna Cortes和Vapnik于1993年提出，并于1995年发表

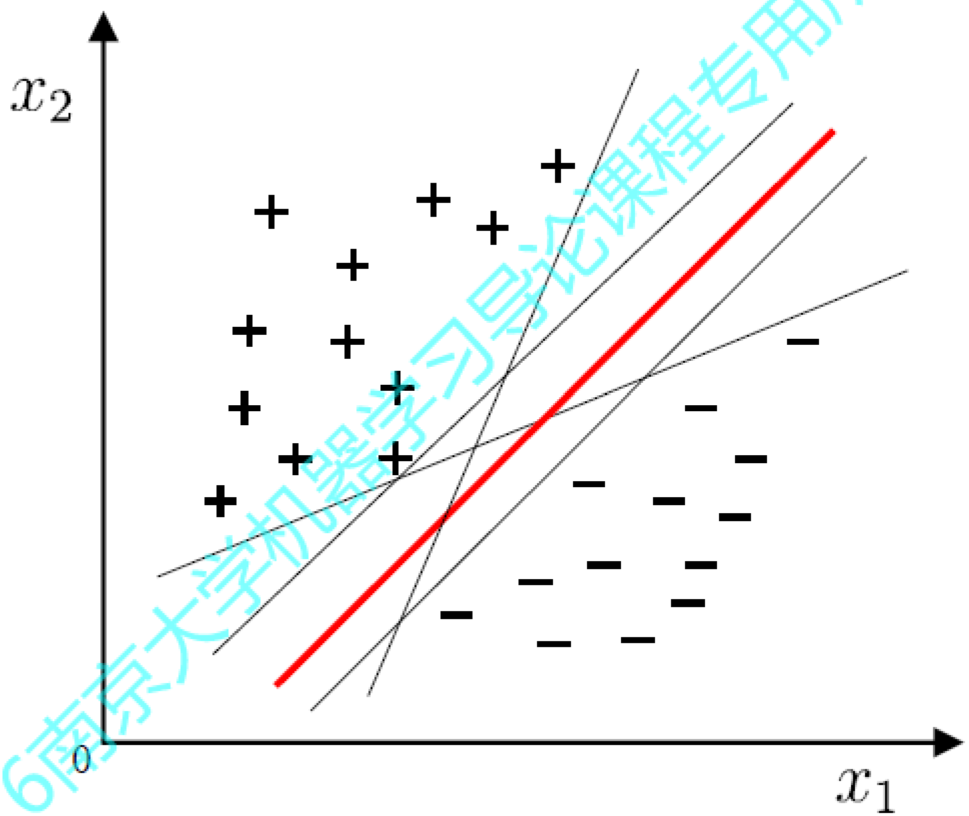


SVM定义

- 是在分类与回归分析中分析数据的监督式学习模型与相关的学习算法。给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM训练算法创建一个将新的实例分配给两个类别之一的模型，使其成为非概率二元线性分类器。SVM模型是将实例表示为空间中的点，这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开。然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。

如何更好地分类?





x_2
x_1
0

超平面

- 在样本空间中，划分超平面可通过如下线性方程来描述：

$$\omega^T x + b = 0 \quad （6.1）$$

$\omega = (\omega_1; \omega_2; \dots; \omega_d)$为法向量，决定超平面的方向， b为位移项，决定了超平面与原点的距离。

- 知道了法向量ω和位移b。
- 样本空间中任意一点x到超平面(ω, b)的距离可表示为：

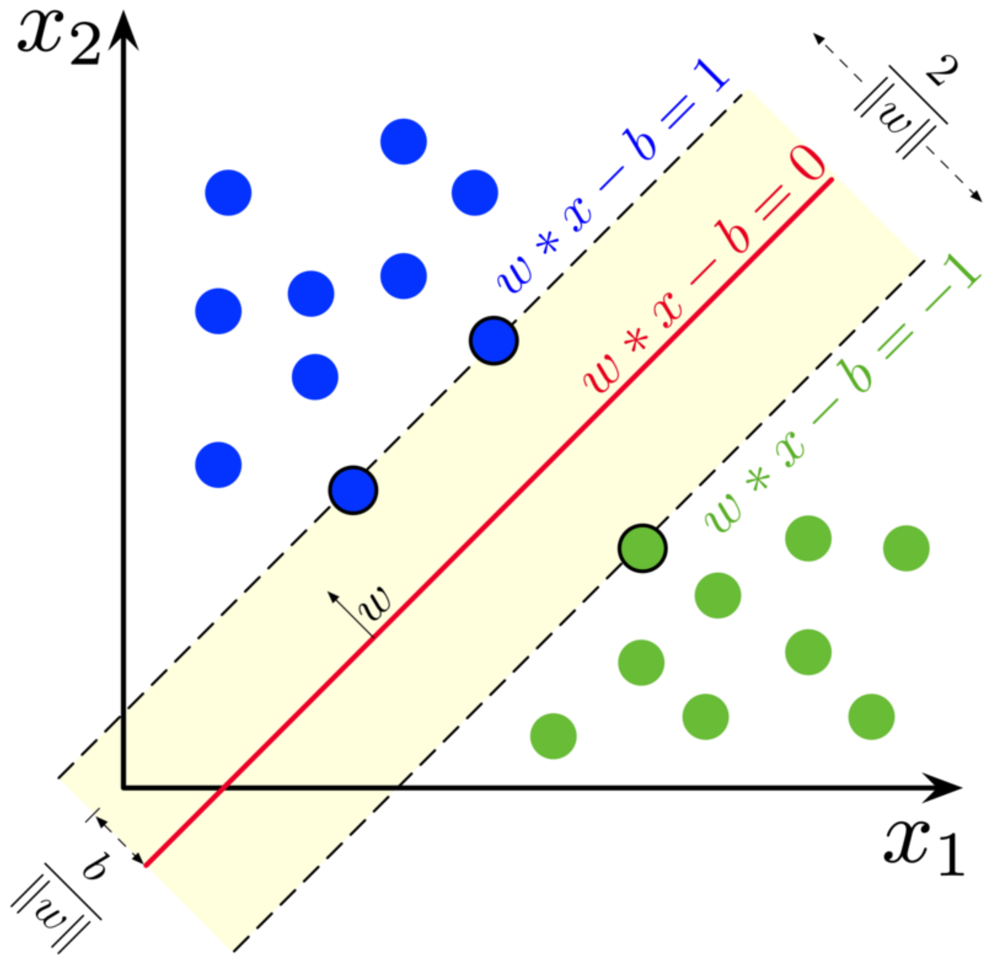
$$\gamma = \frac{|\omega^T x + b|}{||\omega||} \quad （6.2）$$

$||\omega||$:二阶范数

$=\sqrt{\omega^T \omega}$

$$d = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}}$$

支持向量与间隔



$$\begin{cases} \omega^T x_i + b > 0, & y_i = +1; \\ \omega^T x_i + b < 0, & y_i = -1; \end{cases}$$

- 支持向量与间隔：
- 超平面（ω, b）能将训练样本正确分类,距离超平面最近的这几个训练样本点使式(6.3)成立：

$$\begin{cases} \omega^T x_i + b \geq +1, & y_i = +1; \\ \omega^T x_i + b \leq -1, & y_i = -1; \end{cases} \quad (6.3)$$

那么这几个样本点被称为“支持向量（support vector），两个异类支持向量到超平面的距离之和称为间隔：

$$r = \frac{2}{||\omega||} \quad (6.4)$$

$$d = \frac{|c_1 - c_2|}{\sqrt{a^2 + b^2}}$$

$$\begin{cases}\omega^T x_i + b \geq +1, & y_i = +1; \\ \omega^T x_i + b \leq -1, & y_i = -1;\end{cases} \quad (6.3)$$

- 我们要找“最大间隔”的划分超平面，也就是要找能满足式(6.3)中约束的ω, b,使得间隔r最大，即

$$\max_{\omega,b} \frac{2}{||\omega||} \quad (6.5)$$

$$s.t. \quad y_i(\omega^T x_i + b) \geq 1, \qquad i = 1,2,\ldots,m$$

$$r = \frac{2}{||\omega||}$$

- 为求得最大间隔r，显然需要$||\omega||$最小化，这等价于$||\omega||^2$最小化。可以写成：

$$\min_{\omega,b} \frac{1}{2}||\omega||^2 \quad (6.6)$$

$$s.t. \quad y_i(\omega^T x_i + b) \geq 1, \qquad i = 1,2,\ldots,m$$

这就是支持向量机的基本型。

对偶算法

- 为了求解线性可分支持向量机的最优化问题，将式（6.6）作为原始最优化问题，应用拉格朗日对偶性，通过求解对偶问题，得到原始问题的最优解，这就是线性可分支持向量机的对偶算法。对偶问题往往更容易求解，同时自然引入核函数，进而推广到非线性分类问题。

$$\min_{\omega,b} \frac{1}{2}||\omega||^2 \quad （6.6）$$

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

对偶问题

- 对式(6.6)使用拉格朗日乘子法可得到所求问题的“对偶问题”
- 拉格朗日乘子法构造规则：
  - 以约束方程乘以非负的拉格朗日乘子，然后从目标函数中减去

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

即:

$$\mathrm{L}(\omega, b, \alpha) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\omega^T x_i + b)) \quad （6.8）$$

$\alpha = (\alpha_1; \alpha_2; \ldots; \alpha_m)$.

$$\min_{\omega,b} \frac{1}{2}||\omega||^2 \quad (6.6)$$
$$s.t. \quad y_i(\omega^T x_i + b) \geq 1, \qquad i = 1,2,\dots,m$$

对偶问题

- 根据拉格朗日对偶性，原始问题（6.6）的对偶问题是极大极小问题：
  - $\max_{\alpha} \min_{\omega,b} \mathrm{L}(\omega, b, \alpha)$

    要求得对偶问题的解，需要先求$\mathrm{L}(\omega, b, \alpha)$对$\omega, b$的极小，再求对$\alpha$的极大。

求$\min\limits_{\omega,b} L(\omega, b, \alpha)$

- 计算$L(\omega, b, \alpha)$分别对ω和b求偏导为零可得：
- $\omega = \sum_{i=1}^{m} \alpha_i y_i x_i$ （6.9）
- $0 = \sum_{i=1}^{m} \alpha_i y_i$ （6.10）
- 将上面式(6.9)带入该问题的拉格朗日函数,利用(6.10)可得到式:
- $\min\limits_{\omega,b} L(\omega, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j {x_i}^T x_j$ (6.11)

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

  $a_i \geq 0,\ i = 1,2, \ldots, m.$

$$L(\omega, b, \alpha) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\omega^T x_i + b)) \quad (6.8)$$

$\alpha = (\alpha_1; \alpha_2; \ldots; \alpha_m).$

求$\underset{\alpha}{\mathrm{MAX}}\ L(\omega, b, \alpha)$

- 求$\underset{\alpha}{\min}\ L(\omega, b, \alpha)$对α的极大，即得到对偶问题

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j {x_i}^T x_j \quad (6.11)$$

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

$$a_i \geq 0,\ \ i = 1,2,\dots,m.$$

将式(6.11)的目标函数由极大转换成极小，就得到了与之等价的对偶最优化问题，即：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j {x_i}^T x_j - \sum_{i=1}^{m} \alpha_i$$

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

$$a_i \geq 0,\ \ i = 1,2,\dots,m.$$

$$\boxed{\omega = \sum_{i=1}^{m} \alpha_i y_i x_i}$$

$$\boxed{b = y_j - \sum_{i=1}^{m} \alpha_i y_i {x_i}^T x_j}$$

- 求解出α后，求出ω与b即可得到模型

$$f(x) = \omega^T x + b$$
$$= \sum_{i=1}^{m} \alpha_i y_i {x_i}^T x + b \quad （6.12）$$

- 求解α需满足KKT条件，即要求：

$$\begin{cases} \alpha_i \geq 0; \\ y_i f(x_i) - 1 \geq 0; \\ \alpha_i (y_i f(x_i) - 1) = 0. \end{cases} \quad （6.13）$$

对于任意样本(x_i, y_i),总有$\alpha_i = 0$或$y_i f(x_i) = 1$,若α_i=0,则该样本不会在式(6.12)的求和中出现，也就不会对$f(x)$有任何影响；若α_i>0,则必有$y_i f(x_i)$=1,所对应的样本点位于最大间隔边界上，是一个支持向量。

KKT条件

- 一般情况下，最优化问题会碰到以下三种情况：
- （1）无约束条件
- （2）等式约束条件
- （3）不等式约束条件
- 设目标函数f(x)，等式约束为h(X),不等式约束为g(x)

$$
\begin{aligned}
&\min \quad f(X) \\
&s.t. \quad h_j(X)=0 \qquad j=1,2,\ldots,p \\
&\qquad\;\; g_k(X)\le 0 \qquad k=1,2,\ldots q
\end{aligned}
$$

- 则我们定义不等式约束下的拉格朗日函数L，则L表达式为：

- $L(X,\lambda,u)=f(X)+\sum_{j=1}^{p}\lambda_j h_j(X)+\sum_{k=1}^{q}u_k g_k(X)$

- 此时若要求解上述优化问题，必须满足下述条件（也是我们的求解条件）：

$$\left.\frac{\partial L}{\partial X}\right|_{X=X^*}=0 \qquad (1)$$

$$\lambda_j \neq 0, \qquad (2)$$

$$u_k \geq 0, \qquad (3)$$

$$\mu_k g_k(X^*)=0 \qquad (4)$$

$$h_j(X^*)=0 \quad j=1,2,\ldots,p \qquad (5)$$

$$g_k(X^*) \leq 0 \quad k=1,2,\ldots q \qquad (6)$$

这些求解条件就是KKT条件。(1)是对拉格朗日函数取极值时候带来的一个必要条件，(2)是拉格朗日系数约束（同等式情况），(3)是不等式约束情况，(4)是互补松弛条件，(5)、(6)是原约束条件。

线性可分支持向量机学习算法

（1）构造并求解约束最优问题

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j {x_i}^T x_j - \sum_{i=1}^{m}\alpha_i$$

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

$a_i \geq 0,\ i = 1,2,\ldots,m.$

（2）计算

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

$$b = y_j - \sum_{i=1}^{m}\alpha_i y_i {x_i}^T x_j$$

（3）求得分离超平面

$$\omega^T x + b = 0$$

例题：

- 现在有正例点$x_1 = (3,3)^T$, $x_2 = (4,3)^T$, 负例点$x_3 = (1,1)^T$，应用前面的求解算法求线性可分支持向量机。

例 7.2 训练数据与例 7.1 相同.如图 7.4 所示,正例点是 $x_1=(3,3)^{\mathrm{T}}$, $x_2=(4,3)^{\mathrm{T}}$,负例点是 $x_3=(1,1)^{\mathrm{T}}$,试用算法 7.2 求线性可分支持向量机.

解 根据所给数据,对偶问题是

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

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

$$\text{s.t.} \quad \alpha_1+\alpha_2-\alpha_3=0$$

$$\alpha_i \geqslant 0, \quad i=1,2,3$$

解这一最优化问题. 将 $\alpha_3=\alpha_1+\alpha_2$ 代入目标函数并记为

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

对 α_1,α_2 求偏导数并令其为 0,易知 $s(\alpha_1,\alpha_2)$ 在点 $\left(\frac{3}{2},-1\right)^{\mathrm{T}}$ 取极值,但该点不满足约束条件 $\alpha_2 \geqslant 0$,所以最小值应在边界上达到.

当$\alpha_1=0$时，最小值$s\left(0,\frac{2}{13}\right)=-\frac{2}{13}$；当$\alpha_2=0$时，最小值$s\left(\frac{1}{4},0\right)=-\frac{1}{4}$．于是$s(\alpha_1,\alpha_2)$在$\alpha_1=\frac{1}{4},\alpha_2=0$达到最小，此时$\alpha_3=\alpha_1+\alpha_2=\frac{1}{4}$．

这样，$\alpha_1^*=\alpha_3^*=\frac{1}{4}$对应的实例点$x_1,x_3$是支持向量．根据式(7.25)和式(7.26)计算得

$$w_1^*=w_2^*=\frac{1}{2}$$

$$b^*=-2$$

分离超平面为

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

分类决策函数为

$$f(x)=\mathrm{sign}\left(\frac{1}{2}x^{(1)}+\frac{1}{2}x^{(2)}-2\right)$$

■

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j {x_i}^T x_j \quad (6.11)$$
$$s.t \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$
$$a_i \geq 0,\ i = 1,2,\ldots,m.$$

SMO算法

- SMO的基本思路是先固定α_i之外的所有参数，然后求α_i上的极值。由于存在约束$\sum_{i=1}^{m} \alpha_i y_i = 0$,若固定$\alpha_i$之外的其他变量,则$\alpha_i$可由其他变量导出。于是，SMO每次选择两个变量$\alpha_i$和$\alpha_j$，并固定其参数.这样，在参数初始化后，SMO不断执行如下两个步骤直至收敛:
  - 选取一对需要更新的变量α_i和α_j；
  - 固定α_i和α_j以外的参数，求解式(6.11)获得更新后的α_i和α_j
- 完整的SMO算法包括两部分：
  - 求解两个变量的二次规划问题
  - 选择需优化变量的启发式方法

关于α_i和α_j的二次规划问题

- 仅考虑α_i和α_j时，式(6.11)中的约束可重写为:

$$\alpha_i y_i + \alpha_j y_j = c, \quad \alpha_i \geq 0, \quad \alpha_j \geq 0, \quad (6.14)$$

其中

$$c = -\sum_{k \neq i,j} \alpha_k y_k \quad (6.15)$$

是使$\sum_{i=1}^{m} \alpha_i y_i = 0$成立的常数.用

$$\alpha_i y_i + \alpha_j y_j = c$$

消去式(6.11)中的变量α_j，则得到一个关于单变量二次规划问题，仅有的约束是$\alpha_i \geq 0$，不难发现，这样的二次规划问题具有闭式解，于是不必调用数值优化算法即可高效的计算出更新后的α_i和α_j.

$$\boxed{\begin{aligned} &\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y_i (c - \alpha_i y_i) {x_i}^T x_j \quad (6.11) \\ &\quad s.t \ \sum_{i=1}^{m} \alpha_i y_i = 0 \\ &\qquad a_i \geq 0, \ i = 1,2,\ldots,m. \end{aligned}}$$

选择α_i和α_j的启发式方法

- 选取的两变量所对应样本之间的间隔最大，这样的两个变量有很大的差别，与对两个相似的变量进行更新相比，对他们进行更新会使得目标函数的数值产生更大的变化。

偏移量b的计算

- 对任意支持向量(x_s, y_s)都有$y_s f(x_s) = 1$,即

$$y_s(\sum_{i\in S} \alpha_i y_i {x_i}^T x_s + b) = 1，（6.17）$$

两边同乘y_s,

$$ {y_s}^2(\sum_{i\in S} \alpha_i y_i {x_i}^T x_s + b) = y_s$$

因为${y_s}^2$=1

$$(\sum_{i\in S} \alpha_i y_i {x_i}^T x_s + b) = y_s$$

$$b=y_s-\sum_{i\in S} \alpha_i y_i {x_i}^T x_s$$

其中$S = \{i|\alpha_i > 0, i = 1,2, \ldots, m\}$为所有支持向量的下标集.理论上，可选取任意支持向量并通过求解式（6.17）获取b,但现实任务中常采用一种更鲁棒的做法:使用所有支持向量求解的平均值

$$b = \frac{1}{|S|}\sum_{s\in S}(\frac{1}{y_s} - \sum_{i\in S} \alpha_i y_i {x_i}^T x_s).\quad (6.18)$$

总结

- SVM不像逻辑回归一样去拟合样本点，而是在样本中找分隔线，为了评判哪条分隔线最好，引入了几何间隔的概念。之后所有的推导都是去解决目标函数的最优问题上去了，为了更好地解决计算问题，引入了拉格朗日对偶问题，在引入拉格朗日对偶问题后消去了ω，把问题转换为计算α，使得优化函数变为拉格朗日乘子的单一优化问题。