6.5 SMO算法详解

6.5.1 概论

(一) 概念

支持向量机的学习问题可以形式化为求解凸二次规划问题. 这样的凸二次规划问题具有全局最优解, 并且有许多最优化算法可以用于这一问题的求解. 但是当训练样本容量很大时, 这些算法往往变得非常低效, 以致无法使用.

所以, 如何高效地实现支持向量机学习就成为一个重要的问题. 目前人们已提出许多快速实现算法. 本节讲述其中的序列最小最优化 (sequential minimal optimization, SMO) 算法, 这种算法1998年由Platt提出.

(二) 目标问题--软间隔对偶问题

首先, 软间隔的对偶问题前面已经说过了, 也就是 (6.40)

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j \tag{6.40}$$

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

$0 \leqslant \alpha_i \leqslant C, \quad i = 1, 2, \ldots, m$

同时, 对软间隔支持向量机, KKT条件要求

$$\left\{\begin{array}{l} \alpha_i \geqslant 0, \quad \mu_i \geqslant 0 \\ y_i f\left(\boldsymbol{x}_i\right) - 1 + \xi_i \geqslant 0 \\ \alpha_i \left(y_i f\left(\boldsymbol{x}_i\right) - 1 + \xi_i\right) = 0 \\ \xi_i \geqslant 0, \quad \mu_i \xi_i = 0 \end{array}\right. \tag{6.41}$$

接下来, 我们变换一下, 先把max变换为min, 然后把 $\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$ 用核函数表示为 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, 关于核函数可参考 6.3.2 小节知识.

那么, 就可以变换为:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) - \sum_{i=1}^{N} \alpha_i \tag{7.98}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{7.99}$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, 2, \cdots, N \tag{7.100}$$

在这个问题中, 变量时拉格朗日乘子, 一个变量 α_i 对应一个样本点 $(\boldsymbol{x}_i, y_i)$; 变量的总数等于训练样本容量.

(三) SMO算法的思路

SMO算法是一种启发式算法, 其基本思路是: **如果所有变量的解都满足此最优化问题的KKT条件 (Karush-Kuhn-Tucker conditions) , 那么这个最优化问题的解就得到了. 因为KKT条件是该最优化问题的充分必要条件**. 否则, 选择**两个变量, 固定其他变量, 针对这两个变量构建一个二次规划问题**. 这个二次规划问题关于这两个变量的解应该更接近原始二次规划问题的解, 因为这会使得原始二次规划问题的目标函数值变得更小. 重要的是, 这时子问题可以通过解析方法求解, 这样就可以大大提高整个算法的计算速度. 子问题有两个变量, **一个是违反KKT条件最严重的那一个, 另一个由约束条件自动确定**. 如此, SMO算法将原问题不断分解为子问题并对子问题求解, 进而达到求解原问题的目的.

注意, **子问题的两个变量中只有一个是自由变量**, 假设 α_1, α_2 为两个变量, $\alpha_3, \alpha_4, \cdots, \alpha_N$ 固定, 那么由等式约束 (7.99) 可知:

$$\alpha_1 = -y_1 \sum_{i=2}^{N} \alpha_i y_i$$

如果 α_2 确定, 那么 α_1 也随之确定. 所以子问题中同时更新两个变量.

整个SMO算法包括两个部分: **求解两个变量二次规划的解析方法**和**选择变量的启发式方法**

6.5.2 两个变量二次规划的求解方法

(一) 优化问题的改写

不失一般性, 假设选择的两个变量是 α_1, α_2 , 其他变量 $\alpha_i (i = 3, 4, \cdots, N)$ 是固定的. 于是SMO的最优化问题(7.98) ~ (7.100)的子问题可以写成:

$$\begin{aligned} \min_{\alpha_1, \alpha_2} \quad W(\alpha_1, \alpha_2) = & \frac{1}{2} K_{11} \alpha_1^2 + \frac{1}{2} K_{22} \alpha_2^2 + y_1 y_2 K_{12} \alpha_1 \alpha_2 \\ & - (\alpha_1 + \alpha_2) + y_1 \alpha_1 \sum_{i=3}^{N} y_i \alpha_i K_{i1} + y_2 \alpha_2 \sum_{i=3}^{N} y_i \alpha_i K_{i2} \end{aligned} \tag{7.101}$$

$$\text{s.t.} \quad \alpha_1 y_1 + \alpha_2 y_2 = -\sum_{i=3}^{N} y_i \alpha_i = \zeta \tag{7.102}$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, 2, \cdots, N \tag{7.103}$$

其中, $K_{ij} = K(x_i, x_j), i, j = 1, 2, \cdots, N$, ζ 是常数, 目标函数式 (7.101) 中省略了不含 α_1, α_2 的常数项.

> 注: (7.101)的推导过程
>
> 直接带入即可化简, 可得到结果, 但是需要注意以下两个计算过程:
>
> - 但是要注意一个是 $K_{12} = K_{21}$,
> - $y_1 \alpha_1 \sum_{i=3}^{N} y_i \alpha_i K_{i1}$ =$y_1 \alpha_1 \sum_{j=3}^{N} y_j \alpha_i K_{1j}$, 所以可以合并在一起.

(二) 约束条件

为了求解两个变量的二次规划问题(7.101) ~ (7.103), 首先分析约束条件, 然后在此约束条件下求极小.

由于只有两个变量 α_1, α_2 , 约束可以用二维空间中的图形表示（如图7.8所示）

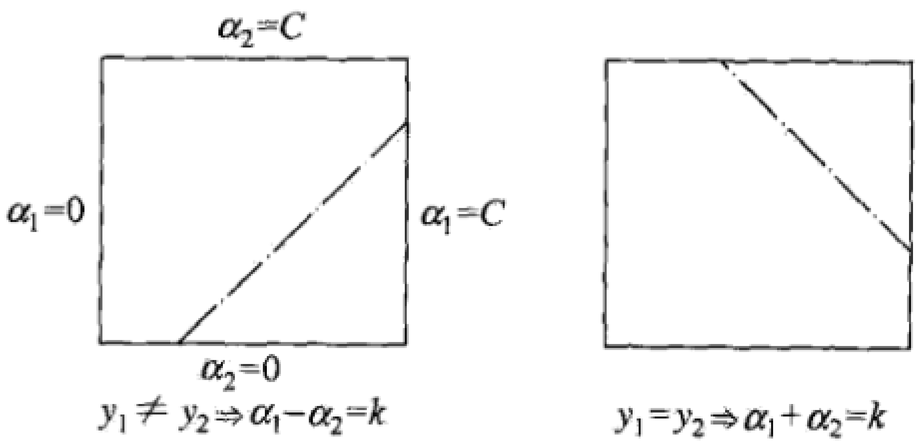


图 7.8 二变量优化问题图示

> 注: 观察约束条件 (7.102) 和 (7.103) , 因为 y_1 和 y_2 的取值只有{-1,1}, 所以当 $y_1 \neq y_2$ 时, 也就变为图 7.8 左边图, $y_1 = y_2$ 时, 也就时图 7.8 右边图

不等式约束 (7.103) 使得 (α_1, α_2) 在盒子 $[0, C] \times [0, C]$ 内, 等式约束 (7.102) 使 (α_1, α_2) 在平行于盒子 $[0, C] \times [0, C]$ 的对角线的直线上. **因此要求的是目标函数在一条平行于对角线的线段上的最优值**. 这使得两个变量的最优化问题成为**实质上的单变量的最优化问题**, 不妨考虑为变量 α_2 的最优化问题.

假设问题 (7.101) ~ (7.103) 的初始可行解为 α_1^{old} , α_2^{old} , 最优解为 α_1^{new} , α_2^{new} , 并且假设在沿着约束方向未经剪辑时 α_2 的最优解为 $\alpha_2^{\text{new,unc}}$.

由于 α_2^{new} 需满足不等式约束 (7.103) , 所以最优值 α_2^{new} 的取值范围必须满足条件

$$L \leqslant \alpha_2^{\text{new}} \leqslant H$$

其中, L 和 H 是 α_2^{new} 所在的对角线段端点的界. 如果 $y_1 \neq y_2$, 即图 7.8 左边图, 则有:

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

如果 $y_1 = y_2$, 即图 7.8 右边图, 则

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

> 注: **L 和 H 的推导过程**
>
> 首先根据原问题的约束条件和初始解,最优解有:
>
> $$\begin{gathered} \alpha_1^{new} y_1 + \alpha_2^{new} y_2 = \alpha_1^{old} y_1 + \alpha_2^{old} y_2 = \zeta \\ 0 \leqslant \alpha_i \leqslant C, \quad i = 1, 2, \cdots, N \end{gathered}$$
>
> **第一种情况, 当 $y_1 \neq y_2$, 即图 7.8 左边图**, 那么有:
>
> $$\alpha_1^{old} - \alpha_2^{old} = \alpha_1^{new} - \alpha_2^{new} = \zeta$$
>
> 进行如下推导:
>
> $$\alpha_2^{new} = \alpha_1^{new} - (\alpha_1^{old} - \alpha_2^{old}) \tag{I}$$
>
> 这里需要注意的一点是, α_2^{new} 是待求解的, α_1^{new} 是变化的.

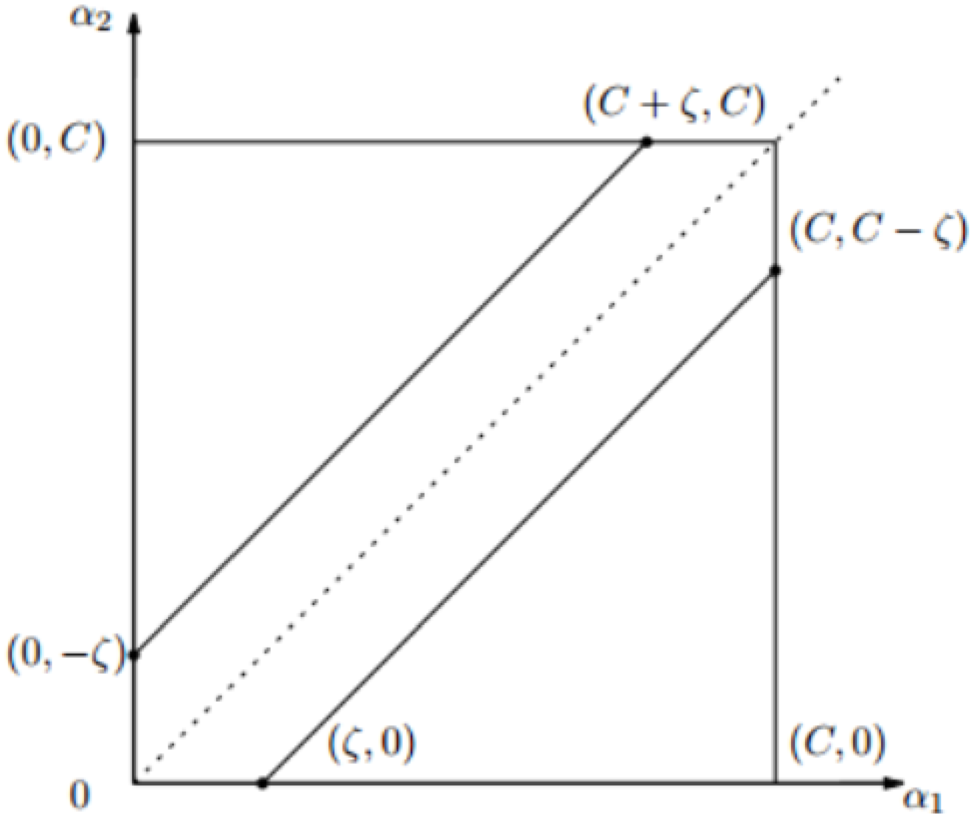
又 $\quad 0 \leqslant \alpha_2^{new} \leqslant C, \quad 0 \leqslant \alpha_1^{new} \leqslant C$, 那么 α_2^{new} 的最小值最小只能到0 , 什么时候取 0 呢, 就是 $(\alpha_1^{old} - \alpha_2^{old}) < 0$ 时, 当 $(\alpha_1^{old} - \alpha_2^{old}) > 0$ 时, 最小值就是 在 $\alpha_1^{new} = 0$ 时, I 式变为:

$\alpha_2^{new} = -(\alpha_1^{old} - \alpha_2^{old})$, 因此, L的取值范围就是 $L = \max\left(0, \alpha_2^{\text{old}} - \alpha_1^{\text{old}}\right)$

同理可以求得 H 的取值范围:

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

具体可以参见下图:

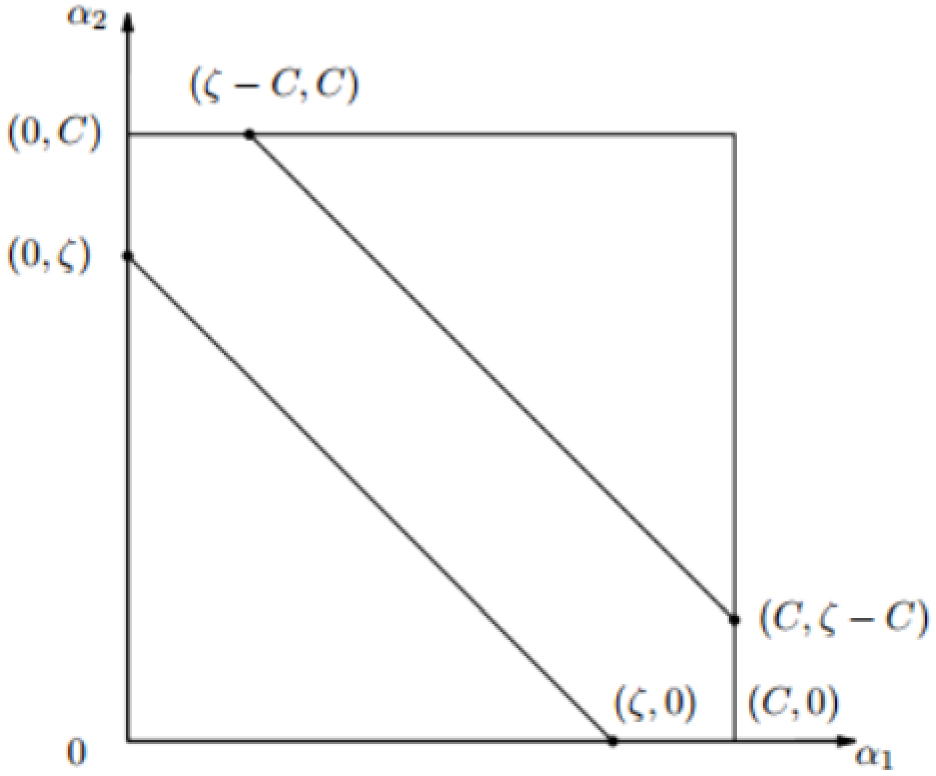


第二种情况, 如果 $y_1 = y_2$, 即图 7.8 右边图,

可以根据同样的方法,推导得到 L 和 H 的取值范围:

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

取值范围如下图:



(三) 两个变量的解

首先求沿着约束方向未经剪辑**即未考虑不等式约束(7.103)时** α_2 的最优解 $\alpha_2^{\text{new,unc}}$, 然后再求剪辑后 α_2 的解 α_2^{new}

为了后面公式的简洁, 记:

$$g(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \tag{7.104}$$

令

$$E_i = g(x_i) - y_i = \left(\sum_{j=1}^{N} \alpha_j y_j K(x_j, x_i) + b \right) - y_i, \quad i = 1, 2 \tag{7.105}$$

当 $i = 1, 2$ 时, $g(x)$ 为 x 的预测值, E_i 为函数 $g(x)$ 对输入 x_i 的预测值与真实输出 y_i 之差.

定理 7.6 两个变量的解

最优化问题 (7.101) ~ (7.103) 沿着约束方向**未经剪辑时的解是**:

$$\alpha_2^{\text{new, unc}} = \alpha_2^{\text{old}} + \frac{y_2 (E_1 - E_2)}{\eta} \tag{7.106}$$

其中,

$$\eta = K_{11} + K_{22} - 2K_{12} = \|\Phi(x_1) - \Phi(x_2)\|^2 \tag{7.107}$$

$\Phi\left(x_1\right)$ 是输入空间到特征空间的映射, $E_i, \quad i=1,2$, 由式 (7.105) 给出.

经剪辑后 α_2 的解是

$$\alpha_2^{\text{new}}=\begin{cases}H, & \alpha_2^{\text{new,unc}}>H \\ \alpha_2^{\text{new,unc}}, & L \leqslant \alpha_2^{\text{new,unc}} \leqslant H \\ L, & \alpha_2^{\text{new,unc}}<L\end{cases} \tag{7.108}$$

由 α_2^{new} 求得 α_1^{new} 是:

$$\alpha_1^{\text{new}}=\alpha_1^{\text{old}}+y_1 y_2\left(\alpha_2^{\text{old}}-\alpha_2^{\text{new}}\right) \tag{7.109}$$

> **注: 关于定理的推导过程.**
>
> 1. 关于**未经剪辑时的解的推导**过程:
>
> 请参考李航<统计学习方法> p127-128的证明
>
> 2. 经剪辑后的解 (7.108) 的解释:
>
> 要使其满足不等式约束必须将其限制在区间 $[L, H]$内, 从而得到 α_2^{new} 的表达式 (7.108)
>
> 3. α_1^{new} 的解 (7.109) 的解释:
>
> 由等式约束 (7.102) , 得到 α_1^{new} 的表达式 (7.109)

6.5.3 变量的选择方法

SMO算法在每个子问题中选择两个变量优化, 其中至少一个变量是违反KKT条件的.

(一) 第 1 个变量的选择

SMO称选择第1个变量的过程为外层循环. 外层循环在训练样本中选取违反 KKT 条件最严重的样本点, 并将其对应的变量作为第1个变量. 具体地, 检验训练样本点 $\left(x_i, y_i\right)$ 是否满足KKT条件, 即

$$\alpha_i=0 \Leftrightarrow y_i g\left(x_i\right) \geqslant 1 \tag{7.111}$$

$$0<\alpha_i<C \Leftrightarrow y_i g\left(x_i\right)=1 \tag{7.112}$$

$$\alpha_i=C \Leftrightarrow y_i g\left(x_i\right) \leqslant 1 \tag{7.113}$$

其中, $g\left(x_i\right)=\sum_{j=1}^N \alpha_j y_j K\left(x_i, x_j\right)+b$.

> 注1 : 关于 (7.111)~(7.113) 的推导:
>
> 1. $\alpha_i=0$
>
> 由(6.39) 知: $C=\alpha_i+\mu_i$, 可得:
>
> $$\mu_i=C$$
>
> 再由对偶问题的 kkt 条件 (6.41) 中的 $\mu_i \xi_i=0$ 可知:
>
> $$\xi_i=0$$
>
> 再由 kkt 条件中的 $y_i f\left(\boldsymbol{x}_i\right)-1+\xi_i \geqslant 0$ (或者原始问题的约束条件, 是一样的), 有:
>
> $$y_i g\left(x_i\right) \geqslant 1$$

> 2. $0 < \alpha_i < C$
>
> 若 $\alpha_i > 0$, 则必有 $y_i f(\boldsymbol{x}_i) = 1 - \xi_i$, 即该样本是支持向量, 由式 (6.39), 即 $C = \alpha_i + \mu_i$ 可知, 若 $\alpha_i < C$, 则 $\mu_i > 0$, 根据 $\mu_i \xi_i = 0$, 进而有 $\xi_i = 0$, 即该样本恰在最大间隔边界上; 所以有, $y_i f(\boldsymbol{x}_i) = 1$, 即 $y_i g(x_i) = 1$
>
> 3. $\alpha_i = C$
>
> 首先, $\xi_i \geqslant 0$, 同时, 由于 $\alpha_i = C$, 那么由 $\alpha_i(y_i f(\boldsymbol{x}_i) - 1 + \xi_i) = 0$ 可得, $(y_i f(\boldsymbol{x}_i) - 1 + \xi_i) = 0$, 所以 $y_i f(\boldsymbol{x}_i) \leqslant 1$

> 注2: 其实 (7.111)~(7.113) 就是 kkt 条件 (6.41) 的充要条件, 两者可以互相推出.

检验是在精度 ε 范围内进行的. 在检验过程中, 外层循环首先遍历所有满足条件 $0 < \alpha_i < C$ 的样本点, 即在间隔边界上的支持向量点, 检验它们是否满足KKT条件. 如果这些样本点都满足KKT条件, 那么遍历整个训练集, 检验它们是否满足KKT条件.

(二) 第 2 个变量的选择

SMO称选择第2个变量的过程为内层循环. 假设在外层循环中已经找到第1个变量 α_1 , 现在要在内层循环中找第2个变量 α_2 . 第2个变量选择的标准是希望能使 α_2 有足够大的变化. 由式 (7.106) 和式(7.108) 可知, 是依赖于$|E_1 - E_2|$ 的, 为了加快计算速度, 一种简单的做法是选择 α_2 , 使其对应的 $|E_1 - E_2|$ 最大. 因为 α_1 已定, E_1 也确定了. 如果 E_1 是正的, 那么选择最小的 E_i 作为E_2; 如果 E_1 是负的, 那么选择最大的 E_i作为E_2. 为了节省计算时间, 将所有 E_i 值保存在一个列表中. 在特殊情况下, 如果内层循环通过以上方法选择的 α_2 不能使目标函数有足够的下降, 那么采用以下启发式规则继续选择 α_2 . 遍历在间隔边界上的支持向量点, 依次将其对应的变量作为 α_2 试用, 直到目标函数有足够的下降. 若找不到合适的 α_2 , 那么遍历训练数据集; 若仍找不到合适的 α_2 , 则放弃第1个 α_1 , 再通过外层循环寻求另外的 α_1

(三) 计算阈值 b 和差值 E_i

在每次完成两个变量的优化后, 都要重新计算阈值b。 当 $0 < \alpha_1^{\text{new}} < \mathrm{C}$ 时, 由 KKT 条件 (7.112) 可知:

$$\sum_{i=1}^{N} \alpha_i y_i K_{i1} + b = y_1$$

> 注: (x_1, y_1) 也满足(7.112), 两边同乘以 y_1 , 有:
>
> $$y_1^2 g(x_i) = y_1$$
>
> 又 $y_1^2 = 1$, 即可得到上述结论

于是, 可得:

$$b_1^{\text{new}} = y_1 - \sum_{i=3}^{N} \alpha_i y_i K_{i1} - \alpha_1^{\text{new}} y_1 K_{11} - \alpha_2^{\text{new}} y_2 K_{21} \tag{7.114}$$

由 E_1 的定义式 (7.105) 有:

$$E_1=\sum_{i=3}^{N}\alpha_i y_i K_{i1}+\alpha_1^{\text{old}}y_1K_{11}+\alpha_2^{\text{old}}y_2K_{21}+b^{\text{old}}-y_1$$

式 (7.114) 的前两项可以通过 E_1 改写为:

$$y_1-\sum_{i=3}^{N}\alpha_i y_i K_{i1}=-E_1+\alpha_1^{\text{old}}y_1K_{11}+\alpha_2^{\text{old}}y_2K_{21}+b^{\text{old}}$$

带入式 (7.114) , 可得:

$$b_1^{\text{new}}=-E_1-y_1K_{11}\left(\alpha_1^{\text{new}}-\alpha_1^{\text{eld}}\right)-y_2K_{21}\left(\alpha_2^{\text{new}}-\alpha_2^{\text{old}}\right)+b^{\text{old}} \tag{7.115}$$

那么, 同样的, 如果 $0<\alpha_2^{\text{new}}<C$, 则有:

$$b_2^{\text{new}}=-E_2-y_1K_{12}\left(\alpha_1^{\text{new}}-\alpha_1^{\text{old}}\right)-y_2K_{22}\left(\alpha_2^{\text{new}}-\alpha_2^{\text{old}}\right)+b^{\text{old}} \tag{7.116}$$

- **如果 α_1^{new} , α_2^{new} 同时满足条件 $0<\alpha_i^{\text{new}}<C, i=1,2$ (也就是 b_1^{new} 和 b_2^{new} 都有效的时候), 他们是相等的, 即 $b^{\text{new}}=b_1^{\text{new}}=b_2^{\text{new}}$**
- **如果 α_1^{new} , α_2^{new} 是 0 或者 C , 那么 b_1^{new} 和 b_2^{new} 以及他们两者之间的数都是符合 KKT 条件的阈值, 这时选择它们的中点作为 $b^{new}=\frac{b_1^{new}+b_2^{new}}{2}$**

6.5.4 SMO算法总结

算法 7.5 (SMO算法)

输入: 训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$, 其中, $x_i\in\mathcal{X}=\mathbf{R}^n$, $y_i\in\mathcal{Y}=\{-1,+1\},\quad i=1,2,\cdots,N$, 精度 $\mathcal{E}$

输出: 近似解 $\hat{\alpha}$

- (1) 取初值 $\alpha^{(0)}=0$, 令 $k=0$;
- (2) 按照 6.5.3 变量的选择方法中第一个变量选择, 选择第一个变量 $\alpha_1^{(k)}$, 按照第二个变量选择方法选择第二个变量 $\alpha_2^{(k)}$, 根据式 (7.106) , 求出新的 $\alpha_2^{\text{new, unc}}$,

$$\alpha_2^{\text{new, unc}}=\alpha_2^{(k)}+\frac{y_2\left(E_1-E_2\right)}{\eta}$$

- (3) 按照下式 (即式 (7.108)) 求出 $\alpha_2^{(k+1)}$

$$\alpha_2^{(k+1)}=\begin{cases}H, & \alpha_2^{\text{new,unc}}>H\\ \alpha_2^{\text{new,unc}}, & L\leqslant\alpha_2^{\text{new,unc}}\leqslant H\\ L, & \alpha_2^{\text{new,unc}}<L\end{cases}$$

- (4) 利用 $\alpha_2^{(k+1)}$ 和 $\alpha_1^{(k+1)}$ 的关系(即式 (7.109)) , 求出 $\alpha_1^{(k+1)}$.

$$\alpha_1^{(k+1)}=\alpha_1^{(k)}+y_1y_2\left(\alpha_2^{(k)}-\alpha_2^{(k+1)}\right)$$

- (5) 按照 6.5.3 变量的选择方法中的 (三) 计算阈值 b 和差值 E_i , 计算 b^{k+1} 和 E_i
- (6) 在精度 $\mathcal{E}$ 范围内检查是否满足如下的终止条件:

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

$$0\leqslant\alpha_i\leqslant C,\quad i=1,2,\cdots,N$$

$$\alpha_i^{k+1} = 0 \Rightarrow y_i g(x_i) \geq 1$$
$$0 < \alpha_i^{k+1} < C \Rightarrow y_i g(x_i) = 1$$
$$\alpha_i^{k+1} = C \Rightarrow y_i g(x_i) \leq 1$$

- (7) 如果满足则结束, 返回 α_i^{k+1}, 否则转到步骤 (2)