1、线性可分：

一个训练样本集$\{(X_i, y_i)\}$，在 i=1…N 线性可分，是指存在(ω, b)，使得对 i=1…N，有：

$$y_i(w^T X_i + b) > 0$$

2、支持向量机寻找的最优分类直线应满足：

（1） 该直线分开了两类；

（2） 该直线最大化间隔（margin）；

（3） 该直线处于间隔的中间，到所有支持向量距离相等。

3、间隔 margin 的推导：

法一：设两侧支持向量所在超平面分别为：$H_1: w^T x_+ + b = +1$和$H_2: w^T x_- + b = -1$，而$d_{margin} = |x_+ - x_-| \cos\theta$。因为

$$\begin{aligned} H_1 - H_2 &= 1 - (-1) \\ &= w^T(x_+ - x_-) \\ &= |w||x_+ - x_-| \cos\theta \\ &= |w| d_{margin} \end{aligned}$$

所以$d_{margin} = \frac{2}{\|w\|}$。

法二：点 X$(x_1, x_2, \ldots, x_i)$到超平面$w^T x + b = 0$的距离公式为$d = \frac{|w^T X + b|}{\sqrt{w^T w}}$，而对于支持向量上的点 X，有$|w^T X + b| = \pm 1$，代入得$d = \frac{1}{\|w\|}$。由定义知：$d_{margin} = \frac{2}{\|w\|}$。

4、支持向量机起始形式：最大化 margin

最小化：$\frac{1}{2}\|w\|^2$

限制条件：$y_i(w^T x_i + b) \geq 1, (i = 1 \sim N)$

已知：训练样本集$\{(X_i, y_i)\}$，i=1…N

待求：(w, b)

5、支持向量机的第一改进形式：对于线性不可分情况，需适当放松限制条件

最小化：$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \delta_i$或$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \delta_i^2$

限制条件：$\begin{cases} \delta_i \geq 0, (i = 1 \sim N) \\ y_i(w^T X_i + b) \geq 1 - \delta_i, (i = 1 \sim N) \end{cases}$

6、从低维到高维的映射：

在一个 M 维空间上随机取 N 个训练样本，随机对每个训练样本赋予标签+1 或-1，设这些训练样本线性可分的概率为 P(M)，则当 M 趋于无穷大时，有 P(M)=1。

7、支持向量机的第二改进形式：对于第一改进形式不能解决的非线性可分情况，映射到高维把数据化为线性可分

最小化：$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \delta_i$或$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \delta_i^2$

限制条件：$\begin{cases} \delta_i \geq 0, (i = 1 \sim N) \\ y_i(w^T \varphi(X_i) + b) \geq 1 - \delta_i, (i = 1 \sim N) \end{cases}$

其中w的维度与$\varphi(X_i)$相同。

8、核函数：$K(X_1, X_2) = \varphi(X_1)^T \varphi(X_2)$

注：

（1） 核函数K与映射φ是一一对应的关系；

（2） 核函数$K(X_1,X_2)$不能任意取，而必须满足以下充要条件才能写成$\varphi(X_1)^T\varphi(X_2)$的形式：

①交换性：$K(X_1,X_2)=K(X_2,X_1)$

②半正定性：$\forall C_i(i=1\sim N),\forall N$ 有 $\sum_{i=1}^{N}\sum_{j=1}^{N}C_iC_jK(X_iX_j)\geq 0$

9、支持向量机的第三改进形式：转化为对偶问题

第一步：把原问题标准化：

最小化：$\frac{1}{2}\|w\|^2-C\sum_{i=1}^{N}\delta_i$或$\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\delta_i^2$

限制条件：$\begin{cases}\delta_i\leq 0,(i=1\sim N)\\ 1+\delta_{\mathrm{i}}-\mathrm{y}_iw^T\varphi(X_i)-\mathrm{y}_ib\leq 0,(\mathrm{i}=1\sim N)\end{cases}$

目标函数是凸的，限制条件是线性的，满足强对偶定理。

第二步：把原问题转化为对偶问题：

最大化：$\theta(\alpha,\beta)=\inf_{w,\delta_i,b}\{\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}(\beta_i-C)\delta_i+\sum_{i=1}^{N}\alpha_i[1+\delta_i-y_iw^T\varphi(X_i)-y_ib]\}$

限制条件：$\begin{cases}\alpha_i\geq 0\\ \beta_i\geq 0\end{cases}$

第三步：由于要遍历所有$(\boldsymbol{w},\boldsymbol{\delta_i},\boldsymbol{b})$求最小值，对这三个变量求偏导，并令其等于0：

$$\frac{\partial\theta}{\partial w}=w-\sum_{i=1}^{N}\alpha_i\varphi(X_i)y_i=0\longrightarrow w=\sum_{i=1}^{N}\alpha_iy_i\varphi(X_i)$$

$$\frac{\partial\theta}{\partial\delta_i}=-C+\alpha_i+\beta_i=0\longrightarrow\alpha_i+\beta_i=C$$

$$\frac{\partial\theta}{\partial\mathrm{b}}=-\sum_{i=1}^{N}\alpha_iy_i=0\longrightarrow\sum_{i=1}^{N}\alpha_iy_i=0$$

第四步：将第三步的结果代入第二步的对偶问题中：

$$\theta(\alpha,\beta)=\inf_{w,\delta_i,b}\left\{\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}(\beta_i-C)\delta_i+\sum_{i=1}^{N}\alpha_i[1+\delta_i-y_iw^T\varphi(X_i)-y_ib]\right\}$$

$$=\inf_{w,\delta_i,b}\left\{\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}\beta_i\delta_i-\sum_{i=1}^{N}C\delta_i+\sum_{i=1}^{N}\alpha_i+\sum_{i=1}^{N}\alpha_i\delta_i-\sum_{i=1}^{N}\alpha_iy_iw^T\varphi(X_i)-\sum_{i=1}^{N}\alpha_iy_ib\right\}(展开)$$

$$=\inf_{w,\delta_i,b}\left\{\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}\alpha_i-\sum_{i=1}^{N}\alpha_iy_iw^T\varphi(X_i)-\sum_{i=1}^{N}\alpha_iy_ib\right\}(代入\alpha_i+\beta_i=C，消C)$$

$$=\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}\alpha_i-\sum_{i=1}^{N}\alpha_iy_iw^T\varphi(X_i)\ （代入\sum_{i=1}^{N}\alpha_iy_i=0）$$

$$=\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}\alpha_i-\sum_{i=1}^{N}w^Tw\ （代入w=\sum_{i=1}^{N}\alpha_iy_i\varphi(X_i)）$$

$$= -\frac{1}{2}\| w \|^2 + \sum_{i=1}^{N} \alpha_i \ （合并同类项）$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \varphi(X_i)^{\mathrm{T}} \varphi(X_j) \ （展开）$$

第五步：转化限制条件：

$$\because \beta_i = C - \alpha_i \geq 0$$
$$\therefore \alpha_i \leq C$$

$$又 \because \alpha_i \geq 0$$

$$\therefore 0 \leq \alpha_i \leq C, (i = 1 \dots N)$$

另有未消除的等式要加入到约束条件中：

$$\sum_{i=1}^{N} \alpha_i y_i = 0, (i = 1 \dots N)$$

10、 支持向量机算法的统一流程：

训练过程：

第一步：输入训练数据$\{(\boldsymbol{X_i}, \boldsymbol{y_i})\}, (\boldsymbol{i = 1 \dots N})$，其中$\boldsymbol{y_i = \pm 1}$。求解如下的二次规划问题，解出所有$\boldsymbol{\alpha_i (i = 1 \dots N)}$：

最大化：$\theta(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \varphi(X_i)^T \varphi(X_j)$

约束条件：$\begin{cases} 0 \leq \alpha_i \leq C, (i = 1 \dots N) \\ \sum_{i=1}^{N} \alpha_i y_i = 0, (i = 1 \dots N) \end{cases}$

第二步：根据 KKT 条件，计算 b（推导过程如下）：

由强对偶定理的推论——对偶差距为 0，有：

$$\begin{cases} \alpha_i [1 + \delta_i - y_i w^T \varphi(X_i) - y_i b] = 0 \\ \beta_i \delta_i = 0 \longrightarrow (\mathrm{c} - \alpha_i)\delta_i = 0 \end{cases}$$

对于某个i，若$\alpha_i \neq 0$且$\alpha_i \neq C$（此时该样本位于支持向量上），由 KKT 条件，必有：

$$\begin{cases} \delta_i = 0 \\ 1 + \delta_i - y_i w^T \varphi(X_i) - y_i b = 0 \end{cases}$$

由于$\omega = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)$以及$\varphi(X_i)^T \varphi(X_j) = \mathrm{K}(X_i, X_j)$，代入到$1 + \delta_i - y_i \omega^T \varphi(X_i) - y_i b = 0$，有：

$$1 + \delta_i - \sum_{j=1}^{N} \alpha_i y_i y_j K(X_j, X_i) - y_i b = 0$$

任选一个满足$0 < \alpha_i < C$的α_i，则b可由$b = \frac{1 - \sum_{j=1}^{N} \alpha_j y_i y_j K(X_j, X_i)}{y_i}$算出。

测试过程：

由于：

$$w^{\mathrm{T}} \varphi(\mathrm{X}) + \mathrm{b} = \sum_{i=1}^{N} \alpha_i y_i \varphi(X_i)^T \varphi(X) + b$$

$$= \sum_{i=1}^{N} \alpha_i y_i \mathrm{K}(X_i, X) + b$$

考察测试数据X，预测其类别y:

如果$\sum_{i=1}^{N} \alpha_i y_i K(X_i, X) + b \geq 0$，则$y = +1$;

如果$\sum_{i=1}^{N} \alpha_i y_i K(X_i, X) + b < 0$，则$y = -1$。

11、 求解$\boldsymbol{\alpha_i (i = 1 \ldots N)}$的一般算法：梯度下降算法

$$\frac{\partial \theta(\alpha)}{\partial \alpha} = \begin{bmatrix} \frac{\partial \theta(\alpha)}{\partial (\alpha_1)} \\ \frac{\partial \theta(\alpha)}{\partial (\alpha_2)} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial \theta(\alpha)}{\partial (\alpha_N)} \end{bmatrix}$$

其中:

$$\frac{\partial \theta(\alpha)}{\partial (\alpha_t)} = 1 - \frac{1}{2}\left[y_t \varphi(X_t) \sum_{i=1}^{N} y_i \alpha_i \varphi(X_i)^T - y_t^2 \alpha_t \varphi(X_t)^T \varphi(X_t) + y_t \varphi(X_t) \sum_{j=1}^{N} y_j \alpha_j \varphi\left(X_j\right)^T - y_t^2 \alpha_t \varphi(X_t)^T \varphi(X_t) + 2 y_t^2 \alpha_t \varphi(X_t)^T \varphi(X_t) \right]$$

$$= 1 - y_t \varphi(X_t) \sum_{i=1}^{N} y_i \alpha_i \varphi(X_i)^T$$

$$= 1 - y_t \sum_{i=1}^{N} y_i \alpha_i K(X_i, X_t)$$

该算法的规模正比于训练样本数（α有N个分量），会在实际任务中造成很大的开销。

12、 求解$\boldsymbol{\alpha_i (i = 1 \ldots N)}$的高效算法：序列最小优化算法（Sequential Minimal Optimization，SMO）

第一步：视为一个二元函数

SMO 算法选择同时优化两个参数，固定其他 N-2 个参数。假设选择优化的变量为α_1, α_2，固定其他参数$\alpha_3, \ldots, \alpha_N$，可以简化目标函数为:

$$min \Psi(\alpha_1, \alpha_2) = \frac{1}{2} K_{11} \alpha_1^2 + \frac{1}{2} K_{22} \alpha_2^2 + y_1 y_2 K_{12} \alpha_1 \alpha_2 - (\alpha_1 + \alpha_2) + y_1 v_1 \alpha_1 + y_2 v_2 \alpha_2 + \text{ Constant}$$

其中$v_i = \sum_{j=3}^{N} \alpha_j y_j K\left(x_i, x_j\right), i = 1,2$，Constant 表示常数项（不包含变量$\alpha 1, \alpha 2$的项）。

第二步：转化为一元函数

由等式约束，有:

$$\alpha_1 y_1 + \alpha_2 y_2 = -\sum_{i=3}^{N} \alpha_i y_i = \zeta$$

两边同乘以y_1，解α_1：

$$\alpha_1 = (\zeta - y_2\alpha_2)y_1$$

把上式代入目标函数，消除α_1，并略去常数项：

$$min\Psi(\alpha_2) = \frac{1}{2}K_{11}(\zeta - \alpha_2 y_2)^2 + \frac{1}{2}K_{22}\alpha_2^2 + y_2K_{12}(\zeta - \alpha_2 y_2)\alpha_2 - (\zeta - \alpha_2 y_2)y_1 - \alpha_2 + v_1(\zeta - \alpha_2 y_2) + y_2 v_2\alpha_2$$

第三步：对一元函数求极值点：

对α_2求偏导并令其等于 0：

$$\frac{\partial\Psi(\alpha_2)}{\partial\alpha_2} = (K_{11} + K_{22} - 2K_{12})\alpha_2 - K_{11}\zeta y_2 + K_{12}\zeta y_2 + y_1y_2 - 1 - v_1y_2 + v_2y_2 = 0$$

由于

$$\begin{aligned} f(X_t) &= w^TX_t + b \\ &= \sum_{i=1}^{m}\alpha_i y_i X_i^T X_t + b \\ &= \alpha_1y_1K_{1t} + \alpha_2y_2K_{2t} + \sum_{i=3}^{m}\alpha_i y_i k_{it} + b \end{aligned}$$

当$t=1$时，有$f(X_1) - \alpha_1y_1K_{11} - \alpha_2y_2K_{12} - b = \sum_{i=3}^{m}\alpha_i y_i K_{1i} = v_1$；当$t=2$时，有$f(X_2) - \alpha_1y_1K_{12} - \alpha_2y_2K_{22} - b = \sum_{i=3}^{m}\alpha_i y_i K_{2i} = v_2$，代入上式，消除$v_1$和$v_2$，另外，用$\alpha_1y_1 + \alpha_2y_2 = \zeta$消去$\zeta$，得到：

$$(K_{11} + K_{22} - 2K_{12})\alpha_2^{\text{new, unclipped}} = (K_{11} + K_{22} - 2K_{12})\alpha_2^{\text{old}} + y_2[y_2 - y_1 + f(X_1) - f(X_2)]$$

令$\eta = (K_{11} + K_{22} - 2K_{12})$，$f(X_1) - y_1 = E_1$，$f(X_2) - y_2 = E_2$，解$\alpha_2^{\text{new, unclipped}}$得：

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

第四步：对原始解修剪

考虑约束条件：

$$\begin{cases} 0 \le \alpha_1 \le C \\ 0 \le \alpha_2 \le C \\ \alpha_1y_1 + \alpha_2y_2 = \zeta \end{cases}$$

当$y_1y_2 = -1$，有$\alpha_1 - \alpha_2 = \zeta$，此时

$$\begin{cases} L = \max(0, -\zeta) = \max(0, \alpha_2^{old} - \alpha_1^{old}) \\ H = \min(C, C - \zeta) = \min(C, C + \alpha_2^{old} - \alpha_1^{old}) \end{cases}$$

当$y_1y_2 = 1$，有$\alpha_1 + \alpha_2 = \zeta$，此时

$$\begin{cases} L = \max(0, \zeta - C) = \max(0, \alpha_1^{old} + \alpha_2^{old} - C) \\ H = \min(C, \zeta) = \min(C, \alpha_1^{old} + \alpha_2^{old}) \end{cases}$$

最优解必须在方框且在直线上取得，因此$L \le \alpha_2^{new} \le H$，$a_2^{new}$表示成如下分段函数形式：

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

第五步：求解α_1^{new}

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

第六步：临界情况

若$\eta = K_{11} + K_{22} - 2K_{12} \leq 0$，则$a_2^{new}$需要取临界值$min(L, H)$。

理解一：

$$\begin{cases} \eta < 0，\text{核函数 K 不满足 Mercer 定理，矩阵 K 非正定} \\ \eta = 0，\text{样本}X_1\text{和}X_2\text{输入特征相同} \end{cases}$$

理解二：对$\psi(\alpha_2)$求二阶导，$\frac{\partial^2 \psi(\alpha_2)}{\partial \alpha_2^2} = K_{11} + K_{22} - 2K_{12} = \eta$。

$$\begin{cases} \text{当}\eta < 0\text{ 时，目标函数为凸函数，没有极小值，极值在定义域边界处取得} \\ \text{当}\eta = 0\text{ 时，目标函数为单调函数，同样在边界处取极值} \end{cases}$$

计算方法：

将$\alpha_2^{new} = L$和$\alpha_2^{new} = H$分别代入式$\alpha_1^{\text{new}} = \alpha_1^{\text{old}} + y_1 y_2 \left(\alpha_2^{\text{old}} - \alpha_2^{\text{new}}\right)$中，并记$y_1 y_2 = s$，算出的$\alpha_1^{\text{new}}$分别记为$L_1$和$H_1$，则：

$$L_1 = \alpha_1 + s(\alpha_2 - L)$$
$$H_1 = \alpha_1 + s(\alpha_2 - H)$$

由于

$$v_t = \sum_{i=3}^{N} \alpha_i y_i K_{it} = f(X_t) - \alpha_1 y_1 K_{1t} - \alpha_2 y_2 K_{2t} - b$$

代入到目标函数$\psi(\alpha_1, \alpha_2)$，消掉v_1和v_2，有：

$$\begin{aligned} \psi_L &= \psi(\alpha_1 = L_1, \alpha_2 = L) \\ &= \frac{1}{2} k_{11} L_1^2 + \frac{1}{2} k_{22} L^2 + s k_{12} L_1 L - (L_1 + L) + y_1 v_1 L_1 + y_2 v_2 L \\ &= \frac{1}{2} k_{11} L_1^2 + \frac{1}{2} k_{22} L^2 + s k_{12} L_1 L + L_1(y_1 v_1 - 1) + L(y_2 v_2 - 1) \\ &= \frac{1}{2} k_{11} L_1^2 + \frac{1}{2} k_{22} L^2 + s k_{12} L_1 L + L_1 y_1 (v_1 - y_1) + L y_2 (v_2 - y_2) \\ &= \frac{1}{2} k_{11} L_1^2 + \frac{1}{2} k_{22} L^2 + s k_{12} L_1 L + L_1 y_1 [f(X_1) - y_1 - \alpha_1 y_1 K_{11} - \alpha_2 y_2 K_{12} - b] + L y_2 [f(X_2) - y_2 \\ &\quad - \alpha_1 y_1 K_{12} - \alpha_2 y_2 K_{22} - b] \end{aligned}$$

因为$f(X_1) - y_1 = E_1$，$f(X_2) - y_2 = E_2$，$y_i^2 = 1$，$y_1 y_2 = s$，代入上式：

$$\begin{aligned} \psi_L &= \frac{1}{2} k_{11} L_1^2 + \frac{1}{2} k_{22} L^2 + s k_{12} L_1 L + L_1 [y_1 E_1 - \alpha_1 K_{11} - \alpha_2 s K_{12} - y_1 b] + L [y_2 E_2 - \alpha_1 s K_{12} - \alpha_2 K_{22} - y_2 b] \\ &= \frac{1}{2} k_{11} L_1^2 + \frac{1}{2} k_{22} L^2 + s k_{12} L_1 L + L_1 [y_1 (E_1 - b) - \alpha_1 K_{11} - \alpha_2 s K_{12}] + L [y_2 (E_2 - b) - \alpha_1 s K_{12} - \alpha_2 K_{22}] \end{aligned}$$

记$f_1 = y_1(E_1 - b) - \alpha_1 K_{11} - \alpha_2 s K_{12}$，$f_2 = y_2(E_2 - b) - \alpha_1 s K_{12} - \alpha_2 K_{22}$，则上式可以简化为：

$$\psi_L = \frac{1}{2}k_{11}L_1^2 + \frac{1}{2}k_{22}L^2 + sk_{12}L_1L + L_1f_1 + Lf_2$$

同理，可算出

$$\psi_H = \frac{1}{2}k_{11}H_1^2 + \frac{1}{2}k_{22}H^2 + sk_{12}H_1H + H_1f_1 + Hf_2$$

比较ψ_L和ψ_H的大小，α_2取较小的函数值对应的边界点。

第七步：更新阈值 b

每完成对两个变量的优化后，要对b的值进行更新，因为b的值关系到$f(X)$的计算，即关系到下次优化时E_i的计算。

（1） 如果$0 < \alpha_1^{new} < C$，由 KKT 条件，有

$$y_1(w^TX_1 + b) = 1 = y_1^2$$

即

$$w^TX_1 + b = \sum_{i=1}^{N}\alpha_i y_i K_{i1} + b = y_1$$

因此

$$b_1^{new} = y_1 - \sum_{i=3}^{N}\alpha_i y_i K_{i1} - \alpha_1^{new}y_1K_{11} - \alpha_2^{new}y_2K_{21}$$

又由

$$E_1 = f(X_1) - y_1 = \sum_{i=1}^{N}\alpha_i y_i K_{i1} + b - y_1$$

有

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

代入上式，消掉$y_1 - \sum_{i=3}^{N}\alpha_i y_i K_{i1}$，有：

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

（2） 如果$0 < \alpha_2^{new} < C$，则

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

（3） 如果同时满足$0 < \alpha_t^{new} < C, t = 1,2$，则$b_1^{new} = b_2^{new}$

（4） 如果同时不满足$0 < \alpha_t^{new} < C, t = 1,2$，则$b_1^{new}$和$b_2^{new}$以及它们之间的数都满足 KKT 阈值条件，这时选择它们的中点（更鲁棒）

注：启发式选择变量

上述分析是在从 N 个变量中已经选出两个变量进行优化的方法，下面分析如何高效地选择两个变量进行优化，使得目标函数下降的最快。

（1） 第一个变量的选择（外循环）

Ⅰ. 首先遍历整个样本集，选择违反 KKT 条件的α_i作为第一个变量，接着依据相关规则选择第二个变量，对这两个变量采用上述方法进行优化。

Ⅱ. 当遍历完整个样本集后，遍历非边界样本集$(0 < \alpha_i < C)$中违反 KKT 的α_i作为第一个变量，同样依据相关规则选择第二个变量，对此两个变量进行优化。

III． 当遍历完非边界样本集后，再次回到遍历整个样本集中寻找。即在整个样本集与非边界样本集上来回切换，寻找违反 KKT 条件的α_i作为第一个变量。直到遍历整个样本集后，没有违反 KKT 条件α_i，然后退出。

注：边界上的样本对应的$\alpha_i = 0$或者$\alpha_i = C$，在优化过程中很难变化。然而非边界样本$0 < \alpha_i < C$会随着对其他变量的优化有很大的变化。

KKT 条件如下：

$$\alpha_i = 0 \Leftrightarrow y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1$$
$$\alpha_i = C \Leftrightarrow y^{(i)}\left(w^T x^{(i)} + b\right) \leq 1$$
$$0 < \alpha_i < C \Leftrightarrow y^{(i)}\left(w^T x^{(i)} + b\right) = 1$$

（2） 第二个变量的选择（内循环）

假设在外循环中找个第一个变量记为α_1，第二个变量的选择希望能使α_2有较大的变化。由于α_2依赖于$|E_1 - E_2|$，当E_1为正时，那么选择最小的E_i作为E_2；如果E_1为负，选择最大E_i作为E_2。通常为每个样本的E_i保存在一个列表中，选择最大的$|E_1 - E_2|$来近似最大化步长。

注：有时按照上述的启发式选择第二个变量，不能够使得函数值有足够的下降，这时按下述步骤：

I. 首先在非边界集上选择能够使函数值足够下降的样本作为第二个变量。

II. 如果非边界集上没有，则在整个样本集上选择第二个变量。

III. 如果整个样本集依然不存在，则重新选择第一个变量。

13、 常用的核函数

$$\begin{cases} K(x,y) = x^T y \ （线性核） \\ K(x,y) = (x^T y + 1)^d, d \geq 1 \ （多项式核） \\ K(x,y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}, \sigma > 0 \ （高斯核，又称Rbf核） \\ K(x,y) = e^{-\frac{\|x-y\|}{\sigma}}, \sigma > 0 \ （拉普拉斯核） \\ K(x,y) = \tanh(\beta x^T y + b), \beta > 0, \theta < 0 \ （Sigmoid核） \end{cases}$$

14、 SVR（支持向量回归）

SVM 回归需要定义一个常量$\epsilon > 0$，对于某个样本点(X_i, y_i)，如果$|y_i - W^T \varphi(X_i) - b| \leq \epsilon$，则完全没有损失；如果$|y_i - W^T \varphi(X_i) - b| > \epsilon$，则对应的损失为$|y_i - \omega \cdot \emptyset(x_i) - b| - \epsilon$，即：

$$\text{err}(X_i, y_i) = \begin{cases} 0, |y_i - W^T \varphi(X_i) - b| \leq \epsilon \\ |y_i - \omega \cdot \emptyset(x_i) - b| - \epsilon, |y_i - W^T \varphi(X_i) - b| > \epsilon \end{cases}$$

根据如上定义的损失函数，目标函数如下：

$$min \frac{1}{2} \| W \|_2^2 \ \text{s.t.} \ |y_i - W^T \varphi(X_i) - b| \leq \epsilon (i = 1,2,\dots,N)$$

和 SVM 分类模型类似，SVM 回归模型也可以对每个样本点(X_i, y_i)加入松弛变量$\varepsilon_i \geq 0$，但是这里使用的是绝对值，实际上是两个不等式，也就是说两边都需要松弛变量，我们定义为$\varepsilon_i^V, \varepsilon_i^\Lambda$。则 SVM 回归模型的损失函数度量在加入松弛变量之后变为：

最小化：$\frac{1}{2} \| \omega \|_2^2 + C \sum_{i=1}^N \left(\varepsilon_i^V + \varepsilon_i^\wedge\right)$

约束条件：$\begin{cases} -\epsilon - \varepsilon_i^\vee \leq y_i - W^T \varphi(X_i) - b \leq \epsilon + \varepsilon_i^\wedge \ (i = 1,2,\dots,N) \\ \varepsilon_i^\vee \geq 0, \varepsilon_i^\wedge \geq 0 \ (i = 1,2,\dots,N) \end{cases}$

和 SVM 分类模型一样，我们也可以用拉格朗日函数将目标优化函数变成无约束的形式，即：

$$L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})=\frac{1}{2}$$

$$\| \omega \|_2^2+C\sum_{i=1}^{N}\left(\varepsilon_i^{V}+\varepsilon_i^{\wedge}\right)+\sum_{i=1}^{N}\alpha_i^{V}\left(-\epsilon-\varepsilon_i^{V}-y_i+W^T\varphi(X_i)+b\right)$$

$$+\sum_{i=1}^{N}\alpha_i^{\wedge}(y_i-W^T\varphi(X_i)-b-\epsilon-\varepsilon_i^{\wedge})-\sum_{i=1}^{N}\mu^{V}\varepsilon_i^{V}-\sum_{i=1}^{N}\mu^{\wedge}\varepsilon_i^{\wedge}$$

其中$\alpha_i^{V}\geq 0,\alpha_i^{\wedge}\geq 0,\mu^{V}\geq 0,\mu^{\wedge}\geq 0$是拉格朗日系数。
根据 SVM 回归模型的目标函数的原始形式，我们的目标是：

$$arg\ \min_{W,b,\varepsilon^{V},\varepsilon^{\wedge}}\left\{\max_{\mu^{V}\geq 0,\mu^{\wedge}\geq 0,\alpha^{V}\geq 0,\alpha^{\wedge}\geq 0}L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})\right\}$$

和 SVM 分类模型一样，这个优化目标也满足 KKT 条件（根据强对偶定理可得）。也就是说，我们可以通过拉格朗日对偶将优化问题转化为等价的对偶问题来求解。对偶问题如下：

$$arg\ \max_{\mu^{V}\geq 0,\mu^{\wedge}\geq 0,\alpha^{V}\geq 0,\alpha^{\wedge}\geq 0}\left\{\min_{W,b,\varepsilon^{V},\varepsilon^{\wedge}}L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})\right\}$$

我们首先求优化函数对于$W,b,\varepsilon^{V},\varepsilon^{\wedge}$的极小值，可以通过求偏导数得到：

$$\frac{\partial}{\partial W}L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})=W-\sum_{i=1}^{N}\left(\alpha_i^{\wedge}-\alpha_i^{V}\right)\varphi(X_i)=0\Rightarrow W=\sum_{i=1}^{N}\left(\alpha_i^{\wedge}-\alpha_i^{V}\right)\varphi(X_i)$$

$$\frac{\partial}{\partial b}L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})=\sum_{i=1}^{N}\left(\alpha_i^{V}-\alpha_i^{\wedge}\right)=0\Rightarrow\sum_{i=1}^{N}\left(\alpha_i^{V}-\alpha_i^{\wedge}\right)=0$$

$$\frac{\partial}{\partial\varepsilon^{\vee}}L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})=C-\alpha^{\vee}-\mu^{V}=0$$

$$\frac{\partial}{\partial\varepsilon^{\wedge}}L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})=C-\alpha^{\wedge}-\mu^{\wedge}=0$$

将上面 4 个式子代入$\min L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})$，消去$W,b,\varepsilon^{\vee},\varepsilon^{\wedge}$，得到：

$$\min L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})$$

$$=\sum_{i=1}^{N}(y_i-\epsilon)\alpha_i^{\wedge}-\sum_{i=1}^{N}(y_i+\epsilon)\alpha_i^{V}-\frac{1}{2}\sum_{i=1,j=1}^{N}\left(\alpha_i^{\wedge}-\alpha_i^{V}\right)\left(\alpha_j^{\wedge}-\alpha_j^{V}\right)\varphi(X_i)\cdot\varphi\left(X_j\right)$$

接着再求拉格朗日乘子$\mu^{V},\mu^{\wedge},\alpha^{V},\alpha^{\wedge}$的极大值，得到：

$$\max\min L(W,b,\alpha^{\vee},\alpha^{\wedge},\varepsilon^{\vee},\varepsilon^{\wedge},\mu^{V},\mu^{\wedge})$$

$$=\max\left\{\sum_{i=1}^{N}(y_i-\epsilon)\alpha_i^{\wedge}-\sum_{i=1}^{N}(y_i+\epsilon)\alpha_i^{V}-\frac{1}{2}\sum_{i=1,j=1}^{N}\left(\alpha_i^{\wedge}-\alpha_i^{V}\right)\left(\alpha_j^{\wedge}-\alpha_j^{V}\right)\varphi(X_i)\cdot\varphi\left(X_j\right)\right\}$$

$$=\min\left\{\frac{1}{2}\sum_{i=1,j=1}^{N}\left(\alpha_i^{\wedge}-\alpha_i^{V}\right)\left(\alpha_j^{\wedge}-\alpha_j^{V}\right)\varphi(X_i)\cdot\varphi\left(X_j\right)-\sum_{i=1}^{N}(y_i-\epsilon)\alpha_i^{\wedge}+\sum_{i=1}^{N}(y_i+\epsilon)\alpha_i^{V}\right\}$$

对于这个目标函数，我们依然可以用 SMO 算法来求出对应的$\alpha^{\vee},\alpha^{\wedge}$，进而求出我们的回归模型系数$W$和$b$。
注：支持向量回归模型的系数同样具有稀疏性，说明如下：

在 SVM 分类模型中，KKT 条件的对偶互补条件为：$\alpha_i(y_i(W^T\varphi(X_i)+b)-1)=0$，而在回归模型中，我们的对偶互补条件变成了：

$$\begin{cases}\alpha_i^V\left(\epsilon+\varepsilon_i^V+y_i-W^T\varphi(X_i)-b\right)=0\\ \alpha_i^\wedge(\epsilon+\varepsilon_i^\wedge-y_i+W^T\varphi(X_i)+b)=0\end{cases}$$

根据松弛变量定义条件，如果$|y_i-W^T\varphi(X_i)-b|\le\epsilon$，则有$\varepsilon_i^V=0,\varepsilon_i^\wedge=0$，此时$\epsilon+\varepsilon_i^V+y_i-W^T\varphi(X_i)-b\neq 0,\epsilon+\varepsilon_i^\wedge-y_i+W^T\varphi(X_i)+b\neq 0$，那么要满足对偶互补条件，只有$\alpha_i^V=0,\alpha_i^\wedge=0$。

我们定义样本系数$\beta=\alpha_i^\wedge-\alpha_i^V$，根据上面$W$的计算式：$W=\sum_{i=1}^N\left(a_i^\wedge-\alpha_i^V\right)\varphi(X_i)$，我们发现此时$\beta_i=0$，

也就是说W不受这些在误差范围内的点的影响。对于在边界上或边界外的点，$\alpha_i^\wedge\neq 0,\alpha_i^V\neq 0$，此时$\beta_i\neq 0$。

15、 SVM 的优缺点

主要优点：

I． **高维**：解决高维特征的分类问题和回归问题很有效，在特征维度大于样本数时依然有很好的效果。

II． **铰链损失（Hinge loss）**：仅仅使用一部分支持向量来做超平面的决策，无需依赖全部数据。

III． **非线性**：有大量的核函数可以使用，从而可以很灵活的来解决各种非线性的分类回归问题。

IV． **拟合度**：样本量不是海量数据的时候，分类准确率高，泛化能力强。

主要缺点：

I． **超高维**：如果特征维度远远大于样本数，则 SVM 表现一般。

II． **计算量**：SVM 在样本量非常大，核函数映射维度非常高时，计算量过大，不太适合使用。

III． **选择标准**：非线性问题的核函数的选择没有通用标准，难以选择一个合适的核函数。

IV． **缺失数据**：SVM 对缺失数据敏感。