SVM基础详解

SULLEY

目录

1 间隔: 分类的直观感受 **2**

2 线性可分支持向量机和硬间隔最大化 **2**

2.1 线性可分支持向量机 2
2.2 函数间隔和几何间隔 3
2.3 间隔最大化 4
2.4 拉格朗日对偶 5
2.4.1 原始问题 6
2.4.2 对偶问题 6
2.4.3 原始问题与对偶问题的关系 7
2.5 用对偶问题进行求解 7

3 线性支持向量机与软间隔最大化 **9**

3.1 线性支持向量机 9
3.2 对偶算法 10
3.3 支持向量 12
3.4 合页损失函数 12

4 非线性支持向量机与核技巧 **13**

4.1 核技巧 13
4.1.1 核函数 13
4.1.2 核技巧在SVM中的应用 14
4.1.3 核函数的充要条件 14
4.1.4 常见的核函数 15
4.2 非线性支持向量机 16

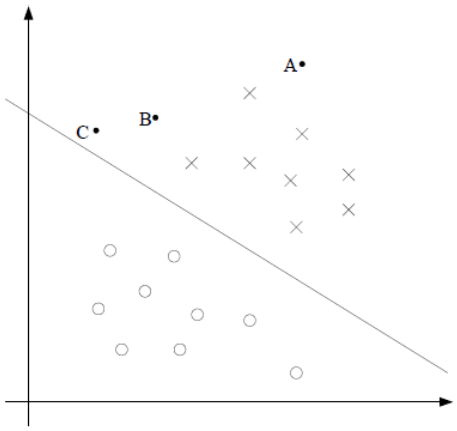
5 序列最小最优化算法SMO **16**

5.1 坐标上升 16
5.2 SMO 17
5.3 变量的选择方法 18
5.3.1 第一个变量的选择 18
5.3.2 第二个变量的选择 19

1 间隔: 分类的直观感受

我们知道在logistic regression中，某样本被分为类1的概率可以用$p(y=1|x;\theta)=h_\theta(x)=g(\theta^T x)$来表示. 如果这个值大于0.5我们就将它分类为1，或者等价地，$\theta^T x \geq 0$. 很显然，如果$\theta^T x$越大，被分类为1的概率也就越大，也就是说，我们就越有“信心”认为它被分类为1. 同样地，若$\theta^T x$越小，我们就越有信心这个样本被分类为0.

我们现在考虑下面这张图，在这里X表示正样本，O表示负样本，直线为决策边界(或**分隔超平面**)，在边界上的点即$\theta^T x=0$.



现在点A离决策边界很远，C很近. 如果我们要去预测A，那么显然我们的预测应该是正的. 如果预测C，我们就不能确保它就是正样本. 也就是说，如果一个点离分隔超平面越远，即间隔越大，我们越有信心对它进行预测.

2 线性可分支持向量机和硬间隔最大化

2.1 线性可分支持向量机

下面我们开始正式讨论SVM. 考虑一个二元分类问题，假设输入空间与特征空间为两个不同的空间，输入空间为欧式空间或离散几何，特征空间为欧式空间或希尔伯特空间，线性(可分)支持向量机假设这两个空间的元素一一对应，非线性支持向量机利用一个从输入空间到特征空间的非线性映射将输入映射为特征向量. 所以说，输入都由输入空间转换到特征空间，**支持向量机的学习是在特征空间中进行的**.

给定一个特征空间上的训练集

$$T=\{(x^1,y^1),\cdots,(x^N,y^N)\},$$

其中$x^i \in \mathcal{X}=\mathbb{R}^n, y^i \in \mathcal{Y}=\{+1,-1\}$. 学习的目标是在特征空间中找到一个分隔超平面，将样本分为两类. 这个分隔超平面对应方程$w\cdot x+b=0$，它由法向量w和截距b决定.

显然，当数据集线性可分时，存在无穷多个这样的分隔超平面可以将两类数据完全正确地分开. 感知机利用误分类数最小的策略，而支持向量机利用**间隔最大化**的原则来学习分割超平面. 如果我们找到这样的参数w^*, b^*使其间隔最大化，那么我们的分类决策函数就为

$$f(x)=\text{sign}(w^*\cdot x+b^*)$$

我们称之为线性可分支持向量机.

2.2 函数间隔和几何间隔

正如我们在1中讨论的，一个点距离分隔超平面的距离可以表示分类预测的确信程度. 在分隔超平面$w \cdot x + b$确定的情况下，$|w \cdot x + b|$能够相对地表示点x距离超平面的远近. 而$w \cdot x + b$与类标记y的符号是否一致能够表示分类是否正确. 所以我们用$y(w \cdot x + b)$来表示分类的正确性及确信程度，这就是函数间隔(functional margin). 注意，如果这个值是正的，那么它越大，就说明我们对它的预测越有信心；如果是负的，那就说明分类错误.

定义 2.1 (函数间隔) *对训练数据集T和超平面(w, b)，我们定义样本点(x^i, y^i)的函数间隔为*

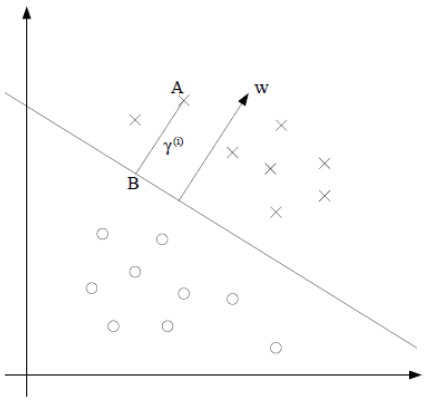
$$\hat{\gamma}^i = y^i(w \cdot x^i + b)$$

再定义超平面关于训练集T的函数间隔为所有这些函数间隔中的最小值

$$\hat{\gamma} = \min_{i=1,\cdots,N} \hat{\gamma}^i$$

虽然函数间隔可以表示分类预测的正确性及确信度，但是选择分离超平面时，只有函数间隔还不够. 因为只要成比例地改变 w, b，例如变为$2w, 2b$，超平面并没有改变，但函数间隔却成为原来的两倍，但我们不能因此说我们有了两倍的确信度. 这启示我们可以对法向量w作一些约束，比如$\|w\| = 1.$，这样就使得间隔是确定的了. 现在，我们就把(w, b)替换为 $(w/\|w\|_2, b/\|w\|_2)$，并使用当前的函数间隔.

现在我们来考虑**几何间隔**. 观察下面的图



注意到w和分隔超平面是垂直的. 我们考虑点A，它表示x^i，计算它到决策边界的距离AB，也就是γ^i的值. 我们知道$w/\|w\|_2$是一个与超平面垂直的单位向量，那么$\gamma^i * w/\|w\|_2$就是其所在的向量，于是垂点B可以表示为 $x^i - \gamma^i * w/\|w\|_2$. 这个点又位于决策边界上，于是满足

$$w^T \left(x^i - \gamma^i \frac{w}{\|w\|} \right) + b = 0$$

从而解得

$$\gamma_i = \left(\frac{w}{\|w\|} \right)^T x^i + \frac{b}{\|w\|}.$$

这个值就表示点A到超平面的真实距离，而不是像函数间隔那样表示的是一个相对的值. 从而我们可以定义几何间隔：

定义 2.2 (几何间隔) *对于给定的训练集T和超平面(w, b)，样本点(x^i, y^i)的几何间隔定义为*

$$\gamma^i = y^i \left(\frac{w}{\|w\|} \cdot x^i + \frac{b}{\|w\|} \right)$$

定义超平面关于训练集的几何间隔为其中的最小值

$$\gamma = \min_{i=1,\cdots,N} \gamma^i.$$

这个几何间隔也是带符号的，当为正时表示正确分类，且值越大，确信度越高；为负就表示分类错误.

注意到函数间隔和几何间隔的关系：

$$\gamma^i = \frac{\hat{\gamma}^i}{\|w\|}, \qquad \gamma = \frac{\hat{\gamma}}{\|w\|}.$$

如果$\|w\| = 1$，即法向量w为单位向量，那么函数间隔等于几何间隔. 如果超平面参数w, b成比例改变(这时候超平面本身没有改变)，函数间隔也成比例改变，但几何间隔不变.

2.3 间隔最大化

支持向量机的基本思想是能够划分训练集并且几何间隔最大. 对线性可分SVM而言，这样的分隔超平面是唯一的. 这里的最大又称为**硬间隔最大化**.其直观解释是：对训练数据集找到几何间隔最大的超平面意味着以充分大的确信度对训练数据进行分类. 不仅要正确分类所有点，还要对未知的新实例有很好的分类预测能力.

于是我们的优化目标就是：

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^i \left(\frac{w}{\|w\|} \cdot x^i + \frac{b}{\|w\|} \right) \geq \gamma, i = 1, \cdots, N \end{aligned}$$

也就是说，我们想要最大化这样的γ，使得对每个样本点分类正确，而且几何间隔至少是γ. 我们再考虑函数间隔和几何间隔的关系，于是改写为

$$\begin{aligned} \max_{\gamma,w,b} \quad & \frac{\hat{\gamma}}{\|w\|} \\ \text{s.t.} \quad & y^i(w^T x^i + b) \geq \hat{\gamma}, i = 1, \cdots, N \end{aligned}$$

在这里，我们去最大化$\hat{\gamma}/\|w\|$，使得所有样本点的函数间隔都至少是$\hat{\gamma}$. 函数间隔的取值并不影响最优化问题的解，这是因为，如果将w, b按比例改变为$\lambda w, \lambda b$，这时函数间隔为$\lambda\hat{\gamma}$，这一改变对上面最优化问题的不等式约束没有影响，对目标函数的优化也没有影响(因为消去了λ)，也就是说，我们可以取这样的λ，它使得$\lambda\hat{\gamma} = 1$，所以我们就可以取$\hat{\gamma} = 1$，得到下面的最优化问题：

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y^i(w^T x^i + b) - 1 \geq 0, i = 1, \cdots, N \end{aligned}$$

这就是一个凸二次规划(convex quadratic programming)问题了. 这个最优化问题的解(w^*, b^*)给出了**最优间隔分类器**，即决策函数$f(x) = \text{sign}(w^* \cdot x + b^*)$. 分隔超平面就是$w^* \cdot x + b^* = 0$.

下面来证明最大间隔分割超平面，即上面的最优化问题的解(w^*, b^*)存在且是唯一的.

(1).存在性

由于训练数据集线性可分，所以最优化问题一定存在可行解. 又由于目标函数有下界，所以必有解，记作(w^*, b^*). 由于训练数据集中既有正类点，又有负类点，所以$w, b = (0, b)$不是最优化的可行解，因为必有$w^* \neq 0$，由此知分隔超平面的存在性.

(2).唯一性

首先证明w^*是唯一的. 假设存在两个最优解$(w_1^*, b_1^*), (w_2^*, b_2^*)$，显然$\|w_1^*\| = \|w_2^*\| = c$为一个常数. 令$w = \dfrac{w_1^* + w_2^*}{2}, b = \dfrac{b_1^* + b_2^*}{2}$，那么$(w, b)$也是一个可行解，从而有

$$c \leq \|w\| \leq \frac{1}{2}\|w_1^*\| + \frac{1}{2}\|w_2^*\| = c$$

从而可以知道$w_1^* = w_2^*$. 因此把这两个最优解写为$(w^*, b_1^*), (w^*, b_2^*)$，再证$b_1^* = b_2^*$. 设$x_1', x_2'$是集合 $\{x^i | y^i = +1\}$中分别对应于$(w^*, b_1^*), (w^*, b_2^*)$使得问题的不等式等号成立的点，$x_1'', x_2''$是集合$\{x^i | y^i = -1\}$ 中分别对应于$(w^*, b_1^*), (w^*, b_2^*)$使得问题的不等式等号成立的点，则由$b_1^* = -\dfrac{1}{2}(w^* \cdot x_1' + w^* \cdot x_1''), b_2^* = -\dfrac{1}{2}(w^* \cdot x_2' + w^* \cdot x_2'')$，得

$$b_1^* - b_2^* = -\frac{1}{2}[w^* \cdot (x_1' - x_2') + w^* \cdot (x_1'' - x_2'')]$$

又因为

$$w^* \cdot x_2' + b_1^* \geq 1 = w^* \cdot x_1' + b_1^*$$
$$w^* \cdot x_1' + b_2^* \geq 1 = w^* \cdot x_2' + b_2^*$$

所以，$w^* \cdot (x_1' - x_2') = 0$. 同理有$w^* \cdot (x_1'' - x_2'') = 0$. 因此

$$b_1^* - b_2^* = 0.$$

我们就证明了唯一性.

在线性可分的情况下，训练数据集的样本点中与分隔超平面距离最近的样本点的实例称为**支持向量**. 支持向量是使不等式约束条件取等号的点，即

$$y^i(w \cdot x^i + b) - 1 = 0.$$

对$y^i = +1$的点，支持向量在超平面

$$H_1 : w \cdot x + b = 1$$

上，对$y^i = -1$的点，支持向量在超平面

$$H_2 : w \cdot x + b = -1$$

上. 注意到H_1, H_2是平行的，它们形成了一条长带，分隔超平面与它们平行且位于正中央. 长带的宽度称为间隔(margin)，容易知道等于$\dfrac{2}{\|w\|}$(这是因为支持向量的函数间隔为1，间隔的一半为几何间隔$1/\|w\|$).

在决定分割超平面时只有支持向量起作用，而其他点不起作用，而支持向量的数量一般来说又是很少的，所以说，SVM由很少的重要的训练样本确定.

2.4 拉格朗日对偶

根据我们的讨论，只要求解上面的最优化问题就行了. 但是，如何进行求解呢？为此，我们引进拉格朗日对偶性(Lagrange duality) 将原始问题转换为对偶问题，通过求解对偶问题来得到原始问题的解.

2.4.1 原始问题

假设$f(x), c_i(x), h_i(x)$是定义在$\mathbb{R}^n$上的连续可微函数，考虑约束最优化问题

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & c_i(x) \le 0, h_j(x) = 0, i = 1, \cdots, k, j = 1, \cdots, l \end{aligned}$$

称该约束最优化问题为**原始问题**.

首先，引进广义拉格朗日函数：

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

其中$\alpha_i \ge 0, x \in \mathbb{R}^n$. 考虑$x$的函数

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

这里下标p表示原始问题.

假设给定某个x，如果x违反约束条件，即存在某个i使得$c_i(x) > 0$或者存在某个j使得$h_j(x) \ne 0$，那么就有

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

若x满足约束条件，则知$\theta_p(x) = f(x)$.

所以如果考虑其极小化问题

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

它是与原始问题等价的，即它们有相同的解. 这就叫广义拉格朗日函数的**极小极大问题**. 这样一来，就把原始问题表示成了广义拉格朗日函数的极小极大问题. 定义原始问题的最优值

$$p^* = \min_x \theta_p(x) = \min_x f(x)$$

称为原始问题的值.

2.4.2 对偶问题

定义

$$\theta_D(\alpha, \beta) = \min_x L(x, \alpha, \beta)$$

再考虑极大化，即

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

问题$\max_{\beta_j, \alpha_i \ge 0} \min_x L(x, \alpha, \beta)$称为广义拉格朗日函数的**极大极小问题**.

可以将极大极小问题表示为约束最优化问题：

$$\begin{aligned} \max_{\beta_j, \alpha_i \ge 0} & \theta_D(\alpha, \beta) = \max_{\beta_j, \alpha_i \ge 0} \min_x L(x, \alpha, \beta) \\ \text{s.t.} \quad & \alpha_i \ge 0, i = 1, 2, \cdots, k \end{aligned}$$

称为原始问题的**对偶问题**. 定义对偶问题的最优值

$$d^* = \max_{\beta_j, \alpha_i \ge 0} \theta_D(\alpha, \beta)$$

称为对偶问题的值.

2.4.3 原始问题与对偶问题的关系

定理 2.1 *若原始问题与对偶问题都有最优值，则*

$$d^* = \max_{\beta_j,\alpha_i\geq 0} \min_x L(x,\alpha,\beta) \leq \min_x \max_{\beta_j,\alpha_i\geq 0} L(x,\alpha,\beta) = p^*$$

证明 由我们已知的关系得

$$\theta_D(\alpha,\beta) = \min_x L(x,\alpha,\beta) \leq L(x,\alpha,\beta) \leq \max_{\beta_j,\alpha_i\geq 0} L(x,\alpha,\beta) = \theta_P(x)$$

所以

$$d^* = \max_{\beta_j,\alpha_i\geq 0} \theta_D(\alpha,\beta) \leq \min_x \theta_P(x) = p^*$$

根据这个定理，我们可以得到，如果x^*和α^*,β^*分别是原始问题和对偶问题的可行解并且$d^* = p^*$，那么x^*和α^*,β^*分别是原始问题和对偶问题的最优解.

定理 2.2 *考虑原始问题和对偶问题，假设函数$f(x),c_i(x)$是凸函数，$h_j(x)$是仿射函数，并且不等式约束是严格可行的，即存在x，对所有的i有$c_i(x) < 0$，则存在x^*,α^*,β^*，使得x^*是原始问题的解，α^*,β^* 是对偶问题的解，并且*

$$p^* = d^* = L(x^*,\alpha^*,\beta^*)$$

定理 2.3 *对原始问题和对偶问题，假设函数$f(x),c_i(x)$是凸函数，$h_j(x)$是仿射函数，并且不等式约束是严格可行的，则 x^*和α^*,β^*分别是原始问题和对偶问题的解的*充要条件*是x^*,α^*,β^*满足下面的* **KKT条件:**

$$\nabla_x L(x^*,\alpha^*,\beta^*) = 0$$
$$\alpha_i^* c_i(x^*) = 0, i = 1,\cdots,k$$
$$c_i(x^*) \leq 0, i = 1,\cdots,k$$
$$\alpha_i^* \geq 0, i = 1,\cdots,k$$
$$h_j(x^*) = 0, j = 1,\cdots,l$$

特别指出，第4个条件为KKT的对偶互补条件，即若$\alpha_i^* > 0$，则$c_i(x^*) = 0$.

2.5 用对偶问题进行求解

我们通过求解对偶问题得到原始问题的最优解，优点是，一是对偶问题求解更容易，二是自然引入核函数，进而推广到非线性分类问题.

首先我们构建拉格朗日函数，对每一个不等式约束引进拉格朗日乘子$\alpha_i \geq 0, i = 1,\cdots,N$，定义拉格朗日函数：

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y^i (w\cdot x^i + b) + \sum_{i=1}^{N} \alpha_i$$

这样原始问题就是

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

其对偶问题为

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

首先我们求极小. 将$L(w,b,\alpha)$分别对w,b求偏导并令为0.

$$\nabla_w L(w,b,\alpha) = w - \sum_{i=1}^{N} \alpha_i y^i x^i = 0$$

$$\nabla_b L(w,b,\alpha) = -\sum_{i=1}^{N} \alpha_i y^i = 0$$

得

$$w = \sum_{i=1}^{N} \alpha_i y^i x^i$$

$$\alpha_i y^i = 0$$

代回得到

$$L(w,b,\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j) + \sum_{i=1}^{N} \alpha_i$$

即

$$\min_{w,b} L(w,b,\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j) + \sum_{i=1}^{N} \alpha_i$$

然后求$\min_{w,b} L(w,b,\alpha)$对α的极大，即是对偶问题

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j) + \sum_{i=1}^{N} \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i y^i = 0 \\ & \alpha_i \geq 0, i = 1, \cdots, N \end{aligned}$$

将求极大转换为求极小，得到等价的对偶最优化问题：

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j) - \sum_{i=1}^{N} \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i y^i = 0 \\ & \alpha_i \geq 0, i = 1, \cdots, N \end{aligned}$$

对线性可分训练数据集，假设对偶最优化问题对α的解为α^*，则可以由α^*求得原始最优化问题对(w,b) 的解(w^*,b^*). 有下面的定理.

定理 2.4 设α^*是对偶最优化问题的解，则存在下标j使得$\alpha_j^* > 0$并且可以按下式求原始最优化问题的解w^*,b^*：

$$w^* = \sum_{i=1}^{N} \alpha_i^* y^i x^i$$

$$b^* = y^j - \sum_{i=1}^{N} \alpha_i^* y^i (x^i \cdot x^j)$$

证明 这时KKT条件2.3成立，则有

$$\nabla_w L(w^*, b^*, \alpha^*) = w^* - \sum_{i=1}^{N} \alpha_i^* y^i x^i = 0$$

$$\nabla_b L(w^*, b^*, \alpha^*) = -\sum_{i=1}^{N} \alpha_i^* y^j = 0$$

$$\alpha_i^*(y^i(w^* \cdot x^i + b^*) - 1) = 0, i = 1, \cdots, N$$

$$y^i(w^* \cdot x^i + b^*) - 1 \geq 0, i = 1, \cdots, N$$

$$\alpha_i^* \geq 0, i = 1, \cdots, N$$

由此得

$$w^* = \sum_{i=1}^{N} \alpha_i^* y^i x^i$$

其中至少有一个$\alpha_j^* > 0$，对这个j有

$$y^j(w^* \cdot x^j + b^*) - 1 = 0$$

将w^*代入就有

$$b^* = y^j - \sum_{i=1}^{N} \alpha_i^* y^i (x^i \cdot x^j)$$

根据这个定理，分隔超平面可以写成

$$\sum_{i=1}^{N} \alpha_i^* y^i (x \cdot x^i) + b^* = 0$$

决策函数可以写成

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i^* y^i (x \cdot x^i) + b^*\right)$$

综上，对于给定的线性可分训练数据集，可以首先求对偶问题的解α^*，再求得w^*, b^*，从而得到分隔超平面.

这时，训练集中对应于$\alpha_i^* > 0$的样本点(x^i, y^i)的实例$x^i \in \mathbb{R}^n$称为**支持向量**. 根据这一定义，支持向量一定在间隔边界上. 由KKT互补条件知，

$$\alpha_i^*(y^i(w^* \cdot x^i + b^*) - 1) = 0, i = 1, \cdots, N$$

如果$\alpha_i^* > 0$则

$$y^i(w^* \cdot x^i + b^*) - 1 = 0$$

或

$$w^* \cdot x^i + b^* = \pm 1$$

这也即x^i一定在间隔边界上，这与之前给出的定义是一致的.

3 线性支持向量机与软间隔最大化

3.1 线性支持向量机

线性可分问题的支持向量机学习算法对线性不可分的训练数据集是不适用的，怎么才能扩展到线性不可分的情况呢？这就需要修改硬间隔最大化为软间隔最大化. 线性不可分意味着某些样本点(x^i, y^i)不能满足函

数间隔大于等于1的约束，为了解决这个问题，可以对每个样本点引入一个松弛变量$\xi_i \geq 0$，使得函数间隔加上松弛变量大于等于1. 这样约束条件就变为

$$y^i(w \cdot x^i + b) \geq 1 - \xi_i$$

同时，对每个松弛变量支付一个代价ξ_i，目标函数变为

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

这里，$C > 0$为惩罚参数，C大时对误分类的惩罚增大，小时对误分类的惩罚减小. 这个目标函数包含两层含义：使 $\frac{1}{2}\|w\|^2$尽量小即间隔尽量大，同时使误分类点的个数尽量少，C是调和二者的系数.

于是，线性不可分的支持向量机的学习问题变成下面的凸二次规划问题(原始问题)：

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y^i(w \cdot x^i + b) \geq 1 - \xi_i, i = 1, \cdots, N \\ & \xi_i \geq 0, i = 1, \cdots, N \end{aligned}$$

在原始问题中，可以证明w是存在且唯一的，但是b的解可能不唯一，而是存在与一个区间. 下面给出线性支持向量机的定义.

定义 3.1 (线性支持向量机) 对于给定的线性不可分的训练集，通过求解凸二次规划问题，即软间隔最大化问题，得到的分隔超平面为

$$w^* \cdot x + b^* = 0$$

以及相应的分类决策函数

$$f(x) = \text{sign}(w^* \cdot x + b^*)$$

称为线性支持向量机.

3.2 对偶算法

仿照线性可分支持向量机，我们容易得到线性支持向量机的对偶问题是

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^i y^j (x^i \cdot x^j) - \sum_{i=1}^{N}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N}\alpha_i y^i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1, \cdots, N \end{aligned}$$

原始最优化问题的拉格朗日函数是

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y^i(w \cdot x^i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i$$

其中$\alpha_i \geq 0, \mu_i \geq 0$.

对偶问题是极大极小问题，首先求L为w,b,ξ的极小：

$$\nabla_w L = w - \sum_{i=1}^{N} \alpha_i y^i x^i = 0$$
$$\nabla_b L = -\sum_{i=1}^{N} \alpha_i y^i = 0$$
$$\nabla_{\xi_i} L = C - \alpha_i - \mu_i = 0$$

得

$$w = \sum_{i=1}^{N} \alpha_i y^i x^i$$
$$\sum_{i=1}^{N} \alpha_i y^i = 0$$
$$C - \alpha_i - \mu_i = 0$$

代入原式，得到

$$\min_{w,b,\xi} L = -\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j) + \sum_{i=1}^{N} \alpha_i$$

再对上式求对α的极大，即得到对偶问题：

$$\begin{aligned} \max_{\alpha} \quad & -\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j) + \sum_{i=1}^{N} \alpha_i \\ \text{s. t.} \quad & \sum_{i=1}^{N} \alpha_i y^i = 0 \\ & C - \alpha_i - \mu_i = 0 \\ & \alpha_i \geq 0 \\ & \mu_i \geq 0, i = 1, \cdots, N \end{aligned}$$

我们消去μ_i只留下α_i，并将约束写为$0 \leq \alpha_i \leq C$. 再将对目标函数求极大转换为求极小，于是得到了我们所说的对偶问题.

定理 3.1 设$\alpha^* = (\alpha_1^*, \cdots, \alpha_N^*)^T$是对偶问题的一个解，若存在一个分量$\alpha_j^*$， $0 < \alpha_j^* < C$，则原始问题的解w^*, b^*可按下式求得：

$$w^* = \sum_{i=1}^{N} \alpha_i^* y^i x^i$$
$$b^* = y^j - \sum_{i=1}^{N} y^i \alpha_i^* (x^i \cdot x^j)$$

证明 原始问题是凸二次规划问题，解满足KKT条件，即得

$$\nabla_w L = w - \sum_{i=1}^{N} \alpha_i y^i x^i = 0$$

$$\nabla_b L = -\sum_{i=1}^{N} \alpha_i y^i = 0$$

$$\nabla_{\xi_i} L = C - \alpha_i - \mu_i = 0$$

$$\alpha_i^*(y^i(w^* \cdot x^i + b^*) - 1 + \xi_i^*) = 0$$

$$\mu_i^* \xi_i^* = 0$$

$$(y^i(w^* \cdot x^i + b^*) - 1 + \xi_i^* \geq 0$$

$$\xi_i^* \geq 0$$

$$\alpha_i^* \geq 0$$

$$\mu_i^* \geq 0, i = 1, \cdots, N$$

由上述式子即可推出结果.

3.3 支持向量

在线性不可分的情况下，将对偶问题的解α^*中对应于$\alpha_j^* > 0$的样本点(x^i, x^j)的实例x^i称为支持向量. 其中，x^i到间隔边界的距离为$\frac{\xi_i}{\|w\|}$.

软间隔的支持向量x^i有下面四种情况：

1. 在间隔边界上，这时$0 < \alpha_j^* < C$，$\xi_j = 0$；
2. 在间隔边界与分隔超平面之间，这时$\alpha_j^* = C$，$0 < \xi_j < 1$，仍然分类正确；
3. 在分割超平面上，这时$\alpha_j^* = C$，$\xi_j = 1$；
4. 在分隔超平面的误分类一侧，这时$\alpha_j^* = C$，$\xi_j > 1$，错误分类.

此外，如果$\alpha_j^* = 0$，那么就位于间隔边界之外.

3.4 合页损失函数

线性支持向量机还有另外一种解释，就是最小化以下的目标函数：

$$\sum_{i=1}^{N} [1 - y^i(w \cdot x^i + b)]_+ + \lambda \|w\|^2$$

第一项是经验损失，函数

$$[1 - y^i(w \cdot x^i + b)]_+$$

称为**合页损失函数**(hinge loss function). 下标+表示取0和自身的最大值. 这就是说，当样本点(x^i, y^i)被正确分类且函数间隔(确信度)$y^i(w \cdot x^i + b)$大于1时损失是0，否则损失是$1 - y^i(w \cdot x^i + b)$. 第二项是正则化项.

定理 3.2 线性支持向量机原始最优化问题：

$$\begin{aligned}\min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y^i(w\cdot x^i + b) \geq 1-\xi_i, i=1,\cdots,N \\ & \xi_i \geq 0, i=1,\cdots,N\end{aligned}$$

等价于最优化问题

$$\min_{w,b} \quad \sum_{i=1}^{N}[1-y^i(w\cdot x^i+b)]_+ + \lambda\|w\|^2$$

证明 我们令$[1-y^i(w\cdot x^i+b)]_+ = \xi_i$，则$\xi_i \geq 0$. 当$1-y^i(w\cdot x^i+b)>0$时，有$y^i(w\cdot x^i+b) = 1-\xi_i$；否则$\xi_i = 0$，有$y^i(w\cdot x^i+b) \geq 1-\xi_i$. 于是满足约束条件. 所以最优化问题可以写成

$$\min_{w,b} \quad \sum_{i=1}^{N}\xi_i + \lambda\|w\|^2$$

取$\lambda = \dfrac{1}{2C}$，则变为

$$\min_{w,b} \quad \frac{1}{C}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i\right)$$

这与原始问题是等价的.

合页损失函数不仅要分类正确，而且确信度足够高的时候损失才是0，也就是说，合页损失函数对学习有更高的要求.

4 非线性支持向量机与核技巧

4.1 核技巧

非线性分类问题是指通过利用非线性模型才能很好地进行分类的问题，无法用线性模型进行直接分类.

用线性分类方法求解非线性分类问题分为两步：首先使用一个变换将原空间的数据映射到新空间；然后在新空间里用线性分类学习方法学习分类模型. 核技巧就是这样一个方法. 其基本想法是通过一个非线性变换将输入空间对应于一个特征空间，使得在输入空间中的超曲面模型对应于特征空间中的超平面模型，这样，分类问题通过在特征空间中求解线性支持向量机就可以完成.

4.1.1 核函数

定义 4.1 设$\mathcal{X}$是输入空间，$\mathcal{H}$为特征空间(希尔伯特空间)，如果存在一个从$\mathcal{X}$到 $\mathcal{H}$的映射

$$\varphi(x) : \mathcal{X} \to \mathcal{H}$$

使得对所有的$z, x \in \mathcal{X}$，函数$K(x,z)$满足

$$K(x,z) = \varphi(x)\cdot\varphi(z)$$

则称$K(x,z)$为核函数，$\varphi(x)$为映射函数.

核技巧的想法是，在学习与预测中只定义核函数，而不显式地定义映射函数. 这是因为，计算$K(x,z)$比较容易，而通过 $\varphi(x),\varphi(z)$并不容易，而且特征空间$\mathcal{H}$一般是高维的，甚至是无穷维的. 可以看到，对于给定的核$K(x,z)$，特征空间$\mathcal{H}$和映射函数$\varphi(x)$的取法并不唯一，可以取不同的特征空间，即便是在同一特征空间里也可以取不同的映射. 比如下面的例子.

假设输入空间是$\mathbb{R}^2$，核函数是$K(x,z)=(x\cdot z)^2$，试求出特征空间$\mathcal{H}$与映射$\varphi:\mathbb{R}^2\to\mathcal{H}$.

取特征空间为$\mathbb{R}^3$，记$x=(x_1,x_2)^T, z=(z_1,z_2)^T$. 由于

$$(x\cdot z)^2=(x_1z_1+x_2z_2)^2=(x_1z_1)^2+2x_1z_1x_2z_2+(x_2z_2)^2$$

所以可以取映射

$$\varphi(x)=(x_1^2,\sqrt{2}x_1x_2,x_2^2)^T$$

容易验证$\varphi(x)\cdot\varphi(z)=K(x,z)$.

还可以取$\mathcal{H}=\mathbb{R}^4$和$\varphi(x)=(x_1^2,x_1x_2,x_1x_2,x_2^2)^T$.

显然，如果$\varphi(x)$和$\varphi(z)$越接近，那么它们的内积，也就是$K(x,z)$就越大；相反，若它们越分开，或者就相互垂直了，那么$K(x,z)$就会很小. 所以，我们可以把$K(x,z)$看作$\varphi(x)$和 $\varphi(z)$有多相似的一种度量.

有了这种观点，我们就可以选择一种我们认为能够作为这种度量的核函数了，比如**高斯核函数**:

$$K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$$

4.1.2 核技巧在SVM中的应用

我们注意到在线性支持向量机的对偶问题中，无论是目标函数还是决策函数都只涉及输入实例与实例之间的内积. 在对偶问题中我们把内积$x^i\cdot x^j$用核函数$K(x,z)=\varphi(x^i)\cdot\varphi(x^j)$来代替，此时目标函数成为

$$W(\alpha)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy^iy^jK(x^i,x^j)-\sum_{i=1}^{N}\alpha_i$$

同样，分类决策函数中的内积也用核函数来代替，

$$f(x)=\text{sign}\left(\sum_{i=1}^{N_s}\alpha_i^*y^iK(x^i,x)+b^*\right)$$

这等价于经过映射函数φ将原来的输入空间变换到特征空间，输入空间中的内积变换为特征空间中的内积，在新的特征空间中学习线性支持向量机. 当映射函数是非线性函数时，学习到的含有核函数的支持向量机也会是非线性分类模型.

在核函数$K(x,z)$给定的情况下，可以利用解线性分类问题的方法求解非线性分类问题的支持向量机. 学习是隐式地在特征空间中进行的，不需要显式地定义特征空间与映射函数. 这样的技巧称为**核技巧**.

4.1.3 核函数的充要条件

现有给定的函数$K(x,z)$，怎么去判断它到底是不是一个核函数呢？我们首先给出一个充要条件.

定义 4.2 (Mercer定理) 设$K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$，则$K(x,z)$为核函数的充要条件是，是任意的 $x^i\in\mathcal{X},i=1,\cdots,m$，$K(x,z)$对应的**Gram**矩阵是**对称半正定矩阵**.

证明　必要性 现在已知$K(x,z)$为核函数，设K为其Gram矩阵，那么对$\forall z \in \mathcal{X}$，我们有

$$\begin{aligned} z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\ &= \sum_i \sum_j z_i \varphi(x^i)^T \varphi(x^j) z_j \\ &= \sum_i \sum_j z_i \sum_k \varphi_k(x^i)\varphi_k(x^j) z_j \\ &= \sum_k \sum_i \sum_j z_i \varphi_k(x^i)\varphi_k(x^j) z_j \\ &= \sum_k \left(\sum_i z_i \varphi_k(x^i) \right)^2 \\ &\geq 0 \end{aligned}$$

充分性 已知对称函数$K(x,z)$的Gram矩阵是半正定的，于是可以构造从$\mathcal{X}$到某个希尔伯特空间$\mathcal{H}$ 的映射：

$$\varphi : x \to K(\cdot, x)$$

又有

$$K(\cdot, x) \cdot f = f(x)$$

并且

$$K(\cdot, x) \cdot K(\cdot, z) = K(x, z)$$

从而

$$K(x,z) = \varphi(x) \cdot \varphi(z)$$

表示$K(x,z)$是$\mathcal{X} \times \mathcal{X}$上的核函数.

注:充分性的证明需要一些预备知识，具体见李航《统计学习方法P118-P121》.

4.1.4　常见的核函数

1. 多项式核函数(polynomial kernal function)

$$K(x,z) = (x \cdot z + 1)^p$$

对应的支持向量机是一个p次多项式分类器，分类决策函数是

$$f(x) = \text{sign}\left(\sum_{i=1}^{N_s} \alpha_i^* y^i (x^i \cdot x + 1)^p + b^* \right)$$

2. 高斯核函数(Gaussian kernal function)

$$K(x,z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2} \right)$$

对应的支持向量机是高斯径向基函数(radial basis function)分类器，分类决策函数是

$$f(x) = \text{sign}\left(\sum_{i=1}^{N_s} \alpha_i^* y^i \exp\left(-\frac{\|x - z\|^2}{2\sigma^2} \right) + b^* \right)$$

3. 字符串核函数(string kernal function)

考虑一个有限字符表Σ，字符串s是从中取出的有限个字符的序列，长度为$|s|$，元素为$s(1)s(2)\cdots s(|s|)$. 两个字符串的连接为st. 所有长度为n的字符串的集合为Σ^n，所有字符串的集合为Σ^*.

考虑字符串s的子串u. 给定一个指标序列$i=(i_1,\cdots,i_{|u|})$，其长度记作$l(i)=i_{|u|}-i_1+1$，如果i是连续的，则$l(i)=|u|$，否则$l(i)>|u|$.

假设$\mathcal{S}$是长度大于等于n的字符串的集合，$s\in\mathcal{S}$. 现在建立$\mathcal{S}$到特征空间$H_n=\mathbb{R}^{\Sigma^n}$的映射$\varphi_n(s)$. $\mathbb{R}^{\Sigma^n}$表示定义在Σ^n上的实数空间，其每一维对应一个字符串$u\in\Sigma^n$，映射$\varphi_n(s)$将字符串s对应于空间$\mathbb{R}_{\Sigma_n}$的一个向量，其在u维上的取值为

$$[\varphi_n(s)]_u=\sum_{i:s(i)=u}\lambda^{l(i)}$$

这里，$0<\lambda\leq 1$是一个衰减参数，$l(i)$表示字符串i的长度，求和在s中所有与u相同的子串上进行.

例如，假设Σ为英文字符集，n为3，$\mathcal{S}$为长度大于等于3的字符串的集合. 考虑将$\mathcal{S}$ 映射到H_3. H_3的一维对应于字符串`asd`. 这时，字符串$[\varphi_3(\texttt{lass das})]_{\texttt{asd}}=2\lambda^5$. 两个字符串$s,t$的字符串核函数是基于映射$\varphi_n$的特征空间的内积：

$$k_n(s,t)=\sum_{u\in\Sigma^n}[\varphi_n(s)]_u[\varphi_n(t)]_u=\sum_{u\in\Sigma^n}\sum_{(i,j):s(i)=t(j)=u}\lambda^{l(i)}\lambda^{l(j)}$$

字符串核函数$k_n(s,t)$给出了字符串s,t中长度等于n的所有子串组成的特征向量的余弦相似度. 直观上，两个字符串相同的子串越多，它们就越相似，字符串核函数的值也越大. 字符串核函数可以由动态规划快速地计算.

4.2 非线性支持向量机

利用核技巧，可以将线性分类的学习方法应用到非线性分类问题中去，将线性支持向量机扩展到非线性支持向量机，只需将线性支持向量机对偶形式中的内积换成核函数.

定义 4.3 (非线性支持向量机) 从非线性分类训练集，通过核函数与软间隔最大化或凸二次规划，学习得到的分类决策函数

$$f(x)=\text{sign}\left(\sum_{i=1}^{N_s}\alpha_i^*y^iK(x^i,x)+b^*\right)$$

称为非线性支持向量机.

5 序列最小最优化算法SMO

凸二次规划问题具有全局最优解，但是当训练集很大的时候，求解过程十分困难，为此，SMO(sequential minimal optimization) 算法将高效地求解.

5.1 坐标上升

考虑求解下面的无限制优化问题

$$\max_{\alpha}W(\alpha_1,\cdots,\alpha_m)$$

我们使用**梯度上升**进行求解

Loop until convergence:{

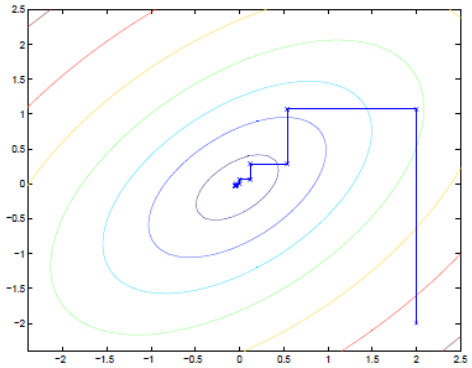
for $i = 1, \cdots, m$:{

$\alpha_i := \arg\max_{\hat{\alpha}_i} W(\alpha_1, \cdots, \alpha_{i-1}, \hat{\alpha}_i, \cdots, \alpha_m)$

}

}

在这个算法的内循环中，我们每次保持一个除了α_i之外的变量固定，然后对α_i进行优化. 优化过程可以用下图表示：



5.2 SMO

现在我们是要求解下面的最优化问题：

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^i y^j K(x^i, x^j) - \sum_{i=1}^{N}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N}\alpha_i y^i = 0 \\ & 0 \le \alpha_i \le C, i = 1, \cdots, N \end{aligned}$$

现在，假如我们固定$\alpha_2, \cdots, \alpha_N$然后对$\alpha_1$使用坐标上升算法. 但是，我们知道

$$\alpha_1 y^1 = -\sum_{i=2}^{N}\alpha_i y^i$$

α_1也被定死了！所有，我们必须每次更新两个来满足限制. 于是我们得到下面的SMO算法：

Repeat till convergence:{

1.选择α_i, α_j.

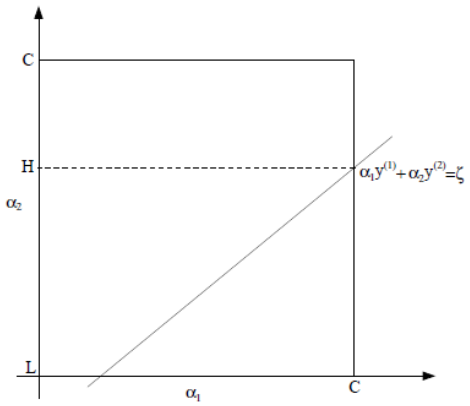
2.根据所选的优化$W(\alpha)$并保持其他的α_k不变.

}

为了检验该算法的收敛性，我们转而检验KKT条件是否满足. 现在，假如我们选择了α_1, α_2并保持其他的不变. 于是

$$\alpha_1 y^1 + \alpha_2 y^2 = -\sum_{i=3}^{N}\alpha_i y^i = \zeta$$

于是我们可以画出关于上式的约束图，如下：



根据约束$0 \leq \alpha_i \leq C$，我们知道α_1, α_2必须位于这个正方形内，又根据约束 $\alpha_1 y^1 + \alpha_2 y^2 = \zeta$，知道$\alpha_1, \alpha_2$位于一条直线上. 从这些约束知道， $L \leq \alpha_2 \leq H$，在这里$L = 0$. 我们可以用α_2去表示α_1

$$\alpha_1 = (\zeta - \alpha_2 y^2) y^1$$

因此，我们的优化目标可以写为

$$W(\alpha) = W((\zeta - \alpha_2 y^2) y^1, \alpha_2, \cdots, \alpha_N)$$

这就是一个关于α_2的二次函数，能被表示为$a\alpha_2^2 + b\alpha_2 + c$，于是我们就能容易地求解这个优化问题，只要令导数为零. 我们令$\alpha_2^{\text{new,unclipped}}$表示新得到的未经剪切的$\alpha_2$，然后根据公式

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

就能得到经过剪切的(也就是满足约束条件的)α_2，由此可以求得

$$\alpha_1^{\text{new}} = \alpha_1^{\text{old}} + y^1 y^2 (\alpha_2^{\text{old}} - \alpha_2^{\text{new}})$$

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，其中至少一个是违反KKT条件的.

5.3.1 第一个变量的选择

SMO选取违反KKT条件最严重的样本点为第一个样本点. 具体地，检验样本点(x^i, y^i)是否满足KKT条件，即

$$\alpha_i = 0 \Leftrightarrow y^i g(x^i) \geq 1$$
$$0 < \alpha_i < C \Leftrightarrow y^i g(x^i) = 1$$
$$\alpha_i = C \Leftrightarrow y^i g(x^i) \leq 1$$

其中$g(x^i) = \sum_{i=j}^{N} \alpha_j y^j K(x^i, x^j) + b$

该检验是在ε范围内进行的. 在检验过程中，首先遍历所有满足$0 < \alpha_i < C$的样本点，即在间隔边界上的支持向量点. 如果都满足，那么遍历整个训练集进行检验.

5.3.2 第二个变量的选择

假设已经找到第一个变量，现在找第二个变量α_2. 第二个变量的选择标准是希望能使α_2有足够大的变化.

实际上，α_2^{new}是依赖于$|E_1 - E_2|$的，其中$E_i = g(x^i) - y^i$. 为了加快计算，一种做法是选择α_2 使得$|E_1 - E_2|$最大. 为了节省计算时间，将所有的E_i保存在一个列表中.

在特殊情况下，采用以下启发式规则选择. 遍历在间隔边界上的支持向量点，依次将其作为对应的α_2试用，直到目标函数有足够的下降. 若找不到合适的，那么遍历训练集. 若仍找不到，就放弃α_1寻求其他的α_1.