支持向量机

cloud

2016.12.21

1 基本介绍

支持向量机是一种二分类模型，它的最基本模型是定义在特征空间上间隔最大的线性分类器。对于非线性可分的情况，支持向量机也包含核技巧，即把样本投影到更高维的空间使其线性可分。支持向量机的学习策略是间隔最大化，可形式化为一个求解凸二次规划的问题。支持向量机本质就是求解凸二次规划的最优化算法。在这里，支持向量机分为线性可分支持向量机，线性支持向量机和非线性支持向量机。假定给定一个特征空间上的训练数据集如下所示：

$$T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})\}$$

$$x^{(i)} \in \mathcal{X} = R^n, y^{(i)} \in \mathcal{Y} = \{+1, -1\}, i = 1, 2, ..., N$$

$x^{(i)}$ 是第 i 个样本，是特征向量。$y^{(i)}$ 是 $x^{(i)}$ 的类标记，当 $y^{(i)} = +1$ 时 $x^{(i)}$ 为正例，当 $y^{(i)} = -1$ 时 $x^{(i)}$ 为负例，$(x^{(i)}, y^{(i)})$ 为样本点。

2 线性可分支持向量机

2.1 基本定义

线性可分支持向量机是在特征空间中找到一个分离超平面，能将实例分到两个不同的类。分离超平面对应于方程 $w \cdot x + b = 0$，它由法向量 w 和截距 b 决定。那么我们求解的目标是给定线性可分训练数据集，通过间隔最大化或等价地求解相应的凸二次规划问题学习到的分离超平面为：

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

以及相应的分类决策函数：

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

称为线性可分支持向量机，$sign$ 函数是当括号里的值大于 0 就为 1 表示正例，小于 0 为 -1 表示负例。

分离超平面将特征空间划分为正类和负类两部分，那在线性可分的情况下可能有无数个分离超平面将样本进行正确的划分，那支持向量机算法选取哪个分离超平面呢？线性可分支持向量机对应着将两类数据正确划分并且间隔最大的分离超平面，简单来说就是我们可以认为最优分离超平面已知，然后最优分离超平面肯定可以保证距离它最近样本的距离是相比较其它分离超平面最近距离是最大的。

线性可分的情况可以用下图表示，如果在二维空间中，分离超平面就是一条直线，将两类样本正确的分开。在这里上半区是正例，下半区是负例。

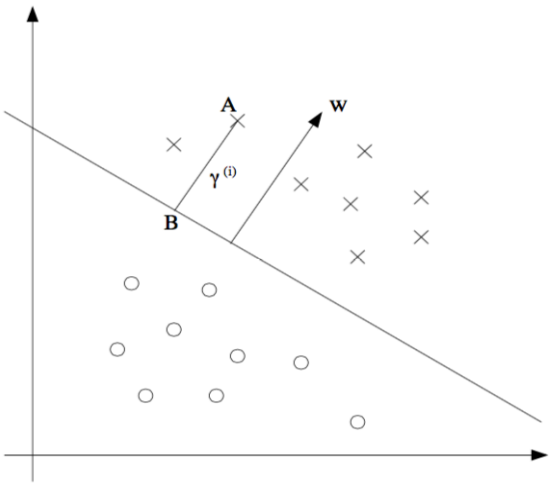

Figure 1: 二分类问题

2.2 目标函数

（1）前面已经说到线性可分支持向量机对应着将两类数据正确划分并且间隔最大的分离超平面，因为线性可分支持向量机不允许样本点误分，所以这里的间隔最大化也称之为硬间隔最大化。那如何定义硬间隔最大呢？对于训练数据集中每一个样本点 $(x^{(i)}, y^{(i)})$，定义样本点距离超平面的函数间隔为：

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

之所以如上式定义函数间隔，那是因为如果 $(w \cdot x^{(i)} + b) < 0$ 那么它就为负例，对应的 $y^{(i)}$ 为 -1，那么它们相乘得到的就是一个正数来表示间隔，正例的原理同上。

那么所有样本点中距离超平面的函数间隔最小值：

$$\hat{\gamma} = \min_{i=1,...,N} \hat{\gamma}_i$$

（2）函数间隔可以表示分类预测的正确性和确信度，但是当成比例的改变 w, b 时，超平面没有改变但是函数距离却会成比例的改变。所以要对超平面的法向量 w 加规范化约束使得 $||w|| = 1$，这时函数间隔就成为几何间隔，相关表示如下。

对于训练数据集中每一个样本点 $(x^{(i)}, y^{(i)})$，定义样本点距离超平面的几何间隔为：

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

所有样本点中距离超平面的几何间隔最小值：

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

（3）下面要考虑如何求得一个几何间隔最大的分离超平面。间隔最大化的直观解释是：超平面不仅将正负实例点分开，而且对最难分的点即距离超平面最近的点也有足够大的确信度将它们分开，这样的超平面有很好的分类预测能力。这个问题可以表示为下面的约束最优化问题：

$$\max_{w,b} \quad \gamma$$

$$s.t. \quad y^{(i)}(\frac{w}{||w||} \cdot x^{(i)} + \frac{b}{||w||}) \geqslant \gamma \quad i = 1, 2, ..., N$$

即我们希望最大化超平面关于训练数据集的几何间隔 γ，约束条件表示的是超平面关于每个训练样本点的几何间隔至少是 γ。

（4）考虑到函数间隔和几何间隔的关系，可以将这个问题改写成：

$$\max_{w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$s.t. \quad y^{(i)}(w \cdot x^{(i)} + b) \geqslant \hat{\gamma} \quad i = 1, 2, ..., N$$

（5）函数间隔 $\hat{\gamma}$ 的取值并不影响最优化问题的求解，那么我们可以取 $\hat{\gamma} = 1$ 代入上式可以得到 $\frac{1}{||w||}$，最大化 $\frac{1}{||w||}$ 等价于最小化 $\frac{1}{2}||w||^2$，于是得到下面的线性可分支持向量学习的最优化问题：

$$\min_{w,b} \quad \frac{1}{2}||w||^2 \tag{1}$$

$$s.t. \quad y^{(i)}(w \cdot x^{(i)} + b) - 1 \geqslant 0 \quad i = 1, 2, ..., N \tag{2}$$

2.3 最优化问题的参数求解

得到上面的最优化问题，我们需要求解最优分离超平面的参数 (w^*, b^*)，在这里由于 (1)(2) 是一个凸二次规划问题，我们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。

首先构建拉格朗日函数，为此对每一个不等式约束 (2) 引入拉格朗日乘子 $\alpha_i \geqslant 0 \quad i = 1, 2, ..., N$，定义拉格朗日函数如下，其中 $\alpha = (\alpha_1, \alpha_2, ..., \alpha_N)^T$。

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i y^{(i)}(w \cdot x^{(i)} + b) + \sum_{i=1}^{N} \alpha_i \tag{3}$$

根据拉格朗日对偶性，原始问题的对偶问题是极大极小问题：

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

所以为了得到对偶问题的解，需要先求 $L(w, b, \alpha)$ 对 w, b 的极小，再求对 α 的极大。

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

首先将拉格朗日函数 $L(w, b, \alpha)$ 分别对 w, b 求偏导并令其为 0，这个时候把 α 看成常数。

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

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

得到 $w = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$ 和 $\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$ 代入拉格朗日函数 (3)，即得 $\min\limits_{w,b} L(w, b, \alpha)$：

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

$\min\limits_{w,b} L(w, b, \alpha)$ 被求了出来，下面要求 $\min\limits_{w,b} L(w, b, \alpha)$ 对 α 的极大。

(2) 求 $\min\limits_{w,b} L(w, b, \alpha)$ 对 α 的极大，即是对偶问题：

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

$$s.t. \quad \sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$

$$\alpha_i \geqslant 0 \quad i = 1, 2, ..., N$$

（3）由求极大转化为求极小，就得到下面与之等价的最优化问题：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)}(x^{(i)}\cdot x^{(j)}) - \sum_{i=1}^{N}\alpha_i \tag{4}$$

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y^{(i)} = 0 \tag{5}$$

$$\alpha_i \geqslant 0 \quad i = 1, 2, ..., N \tag{6}$$

最终求解原始问题 (1)(2) 转换成求解对偶问题 (4) ∼ (6)，那么我们可以通过 (4) ∼ (6) 求解 α。对线性可分训练数据集，假设求解 (4) ∼ (6) 的解为 $\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)$，可由 α^* 求得原始最优化问题 (1)(2) 对 (w, b) 的解 (w^*, b^*)。

2.4 通过 α^* 可求解 (w^*, b^*)

上面我们已经求解出来 α^*，下面要求解 (w^*, b^*)。首先给出定理，设 $\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)$ 是对偶问题 (4) ∼ (6) 的解，则存在下标 j，使得 $\alpha_j^* > 0$，并可按照 $w^* = \sum_{i=1}^{N}\alpha_i^* y^{(i)}x^{(i)}$ 和 $b^* = y^{(j)} - \sum_{i=1}^{N}\alpha_i^* y^{(i)}(x^{(i)}\cdot x^{(j)})$ 求得原始最优化问题 (1)(2) 的解 (w^*, b^*)。

证明：对偶问题的解可以求出原始问题的解，根据 KTT 条件，拉格朗日函数 (3) 满足下列条件：

$$\nabla_w L(w^*, b^*, \alpha^*) = w^* - \sum_{i=1}^{N}\alpha_i^* y^{(i)}x^{(i)} = 0 \tag{7}$$

$$\begin{aligned}
\nabla_b L(w^*, b^*, \alpha^*) &= -\sum_{i=1}^{N}\alpha_i^* y^{(i)} = 0 \\
\alpha_i^*(y^{(i)}(w^*\cdot x^{(i)} + b^*) - 1) &= 0 \\
y^{(i)}(w^*\cdot x^{(i)} + b^*) - 1 &\geqslant 0 \\
\alpha_i^* &\geqslant 0, \quad i = 1, 2, ..., N
\end{aligned} \tag{8}$$

根据公式 (7) 可以求得：

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

那么其中至少有一个 $\alpha_j^* > 0$（反证法：因为如果所有的 $\alpha_j^* = 0$，那么由公式 (7) 可得 $w^* = 0$，而 $w^* = 0$ 不是原始最优化问题的解，所以产生矛盾）。所以一定存在 $\alpha_j^* > 0$，那么这时通过公式 (8) 可得 $y^{(j)}(w^*\cdot x^{(j)} + b) - 1 = 0$，然后把 $w^* = \sum_{i=1}^{N}\alpha_i^* y^{(i)}x^{(i)}$ 代入，并注意 $(y^{(j)})^2 = 1$ 可求得：

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

即可得到分离超平面为：

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

分类决策函数为：

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

2.5 支持向量

我们已经求出来最优分离超平面 $w^* \cdot x + b^* = 0$，在二维空间中如下图实线所示。样本点中与分离超平面距离最近的样本点的实例称为支持向量，即在虚线上的点。支持向量在公式 (8) 中是 $\alpha_i^* > 0$ 的点，根据式 (8) 就有 $y^{(i)}(w^* \cdot x^{(i)} + b) - 1 = 0$，所以对于 $y^{(i)} = +1$ 的正例点，支持向量在超平面 $H_1 : w^* \cdot x + b^* = 1$，对 $y^{(i)} = -1$ 的负例点，支持向量在超平面 $H_2 : w^* \cdot x + b^* = -1$。

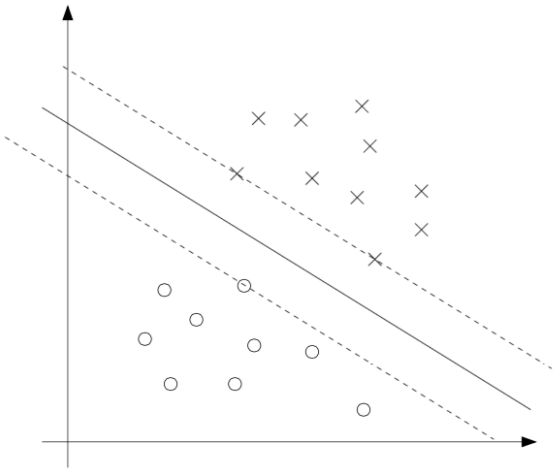

Figure 2: 支持向量

由上图可知，H_1 和 H_2 平行且没有实例点落在它们中间，它们之间的距离称为间隔，间隔依赖于分离超平面的法向量 w 为 $\frac{2}{||w||}$。在决定分离超平面时只有支持向量起作用，其它实例点不起作用，所以支持向量机是由很少的重要训练样本确定的。

3 线性支持向量机

3.1 目标函数

上面我们讨论的是线性可分支持向量机，但是在很多时候按照上面的方法得到的分离超平面不一定是最优的，如下图所示：

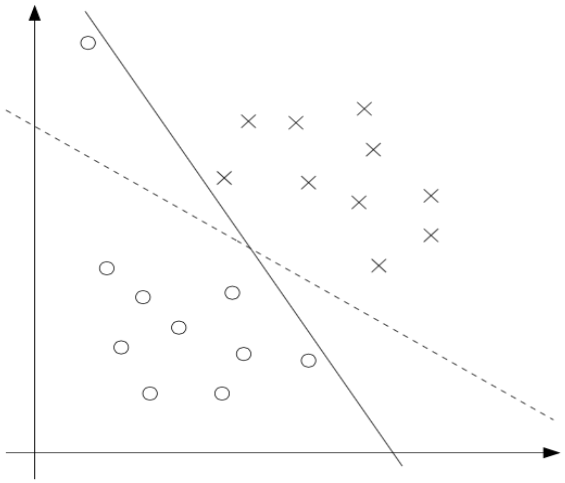

Figure 3: 非最优情况

在上图中实线是利用上节的解法求出来的分离超平面，但是实际上虚线作为分离超平面会更加的合理，但却会出现误分类的情况。像这种允许存在误分类点的情况，我们称之为线性不可分情况，求解的目标称之为软间隔最大化。为了求得更佳合理即软间隔最大化的的分离超平面，可以对每个样本点 $(x^{(i)}, y^{(i)})$ 引入一个松弛变量 $\varepsilon_i \geqslant 0$，使得函数间隔加上松弛变量大于等于 1，那么约束条件由原来的 $y^{(i)}(w \cdot x^{(i)} + b) \geqslant 1$ 变成：

$$y^{(i)}(w \cdot x^{(i)} + b) \geqslant 1 - \varepsilon_i$$

同时对每个松弛变量 ε_i，支付一个代价 ε_i，目标函数变为：

$$\min_{w,b} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\varepsilon_i$$

$C > 0$ 称为惩罚系数由应用问题决定，C 值大则误分类惩罚变大，C 值小则误分类惩罚变小。最小化目标函数的含义是使间隔尽量大，误分类点尽量小，C 是两者的调和系数。

线性支持向量机的学习问题就变成了如下的凸二次规划问题：

$$\min_{w,b} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\varepsilon_i \tag{9}$$

$$s.t. \quad y^{(i)}(w \cdot x^{(i)} + b) \geqslant 1 - \varepsilon_i \quad i = 1, 2, ..., N \tag{10}$$

$$\varepsilon_i \geqslant 0 \quad i = 1, 2, ..., N \tag{11}$$

3.2 最优化问题的参数求解

为不等式约束 (10) 和 (11) 分别引入拉格朗日乘子 $\alpha_i \geqslant 0$ 和 $\mu_i \geqslant 0$，那么原始最优化问题 (9)(10)(11) 的拉格朗日函数是：

$$L(w, b, \varepsilon, \alpha, \mu) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\varepsilon_i - \sum_{i=1}^{N}\alpha_i(y^{(i)}(w \cdot x^{(i)} + b) - 1 + \varepsilon_i) - \sum_{i=1}^{N}\mu_i\varepsilon_i \tag{12}$$

为了得到对偶问题的解，需要先求 $L(w, b, \varepsilon, \alpha, \mu)$ 对 w, b, ε 的极小，再求对 α 的极大。

(1) 求 $\min\limits_{w,b,\varepsilon} L(w, b, \varepsilon, \alpha, \mu)$

首先将拉格朗日函数 $L(w, b, \varepsilon, \alpha, \mu)$ 分别对 w, b, ε 求偏导并令其为 0，这个时候把 α, μ 看成常数。

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

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

$$\nabla_{\varepsilon_i} L(w, b, \varepsilon, \alpha, \mu) = C - \alpha_i - \mu_i = 0$$

得到 $w = \sum_{i=1}^{N}\alpha_i y^{(i)} x^{(i)}$ 和 $\sum_{i=1}^{N}\alpha_i y^{(i)} = 0$ 以及 $C - \alpha_i - \mu_i = 0$ 代入拉格朗日函数 (12)，即得：

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

$\min\limits_{w,b,\varepsilon} L(w, b, \varepsilon, \alpha, \mu)$ 被求了出来，下面要求 $\min\limits_{w,b,\varepsilon} L(w, b, \varepsilon, \alpha, \mu)$ 的极大。

(2) 求 $\min\limits_{w,b,\varepsilon} L(w, b, \varepsilon, \alpha, \mu)$ 对 α 的极大，即是对偶问题：

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

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y^{(i)} = 0 \tag{13}$$

$$C - \alpha_i - \mu_i = 0 \tag{14}$$

$$\alpha_i \geqslant 0 \tag{15}$$

$$\mu_i \geqslant 0 \quad i = 1, 2, ..., N \tag{16}$$

（3）由求极大转化为求极小，然后对 (13) ∼ (16) 进行变换，利用等式约束 (14) 消去 μ_i，从而只留下变量 α_i，这样就可以把 (13) ∼ (16) 的约束写成：

$$0 \leqslant \alpha_i \leqslant C$$

就得到下面与之等价的最优化问题：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)}(x^{(i)}\cdot x^{(j)}) - \sum_{i=1}^{N}\alpha_i \tag{17}$$

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y^{(i)} = 0 \tag{18}$$

$$0 \leqslant \alpha_i \leqslant C \quad i = 1, 2, ..., N \tag{19}$$

最终求解原始问题 (9) ∼ (11) 转换成求解对偶问题 (17) ∼ (19)，那么我们可以通过 (17) ∼ (19) 求解 α。对线性可分训练数据集，假设求解 (17) ∼ (19) 的解为 $\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)$，可由 α^* 求得原始最优化问题 (9) ∼ (11) 对 (w, b) 的解 (w^*, b^*)。

3.3 通过 α^* 可求解 (w^*, b^*)

上面我们已经求解出来 α^*，下面要求解 (w^*, b^*)。首先给出定理，设 $\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)$ 是对偶问题 (17) ∼ (19) 的一个解，则存在下标 j，使得 $0 < \alpha_j^* < C$，并可按照 $w^* = \sum_{i=1}^{N}\alpha_i^* y^{(i)}x^{(i)}$ 和 $b^* = y^{(j)} - \sum_{i=1}^{N}\alpha_i^* y^{(i)}(x^{(i)}\cdot x^{(j)})$ 求出原始最优化问题 (9) ∼ (11)(w^*, b^*)。

证明：对偶问题的解可以求出原始问题的解，根据 KTT 条件，拉格朗日函数 (3) 满足下列条件：

$$\nabla_w L(w^*, b^*, \varepsilon^*, \alpha^*, \mu^*) = w^* - \sum_{i=1}^{N}\alpha_i^* y^{(i)}x^{(i)} = 0 \tag{20}$$

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

$$\nabla_{\varepsilon_i} L(w^*, b^*, \varepsilon^*, \alpha^*, \mu^*) = C - \alpha_i^* - \mu_i^* = 0$$

$$\alpha_i^*(y^{(i)}(w^*\cdot x^{(i)} + b^*) - 1 + \varepsilon_i^*) = 0 \tag{21}$$

$$y^{(i)}(w^*\cdot x^{(i)} + b^*) - 1 + \varepsilon_i^* \geqslant 0$$

$$\mu_i^*\varepsilon_i^* = 0 \tag{22}$$

$$\varepsilon_i^* \geqslant 0$$

$$\alpha_i^* \geqslant 0$$

$$\mu_i^* \geqslant 0, \quad i = 1, 2, ..., N$$

根据公式 (20) 可以求得：

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

若存在 $0 < \alpha_j^* < C$，根据 $C - \alpha_j^* - \mu_j^* = 0$ 知道 $\mu_j^* = 0$，再根据式 (22) 得 $\varepsilon_j^* = 0$，最后由式 (21) 得 $y^{(j)}(w^* \cdot x^{(j)} + b^*) - 1 = 0$，等式两边同时乘以 $y^{(j)}$ 可解得 b^* 为：

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

在这里 b^* 并不是唯一的，对任何一个满足 $0 < \alpha_j^* < C$ 的 α_j^* 都可以求出对应的 b^*，所以实际计算可以取在所有符合条件的样本点上的平均值。

进一步即可得到分离超平面为：

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

分类决策函数为：

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

3.4 支持向量

在线性不可分的情况下，将对偶问题 (17) ~ (19) 的解 $\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)^T$ 中 $\alpha_i^* > 0$ 的样本点 $(x^{(i)}, y^{(i)})$ 的实例 $x^{(i)}$ 称为支持向量，如下图所示。在这里上半区是正例，下半区是负例，因为截图的原因所以导致和前面的图片正负例表示的点相反。

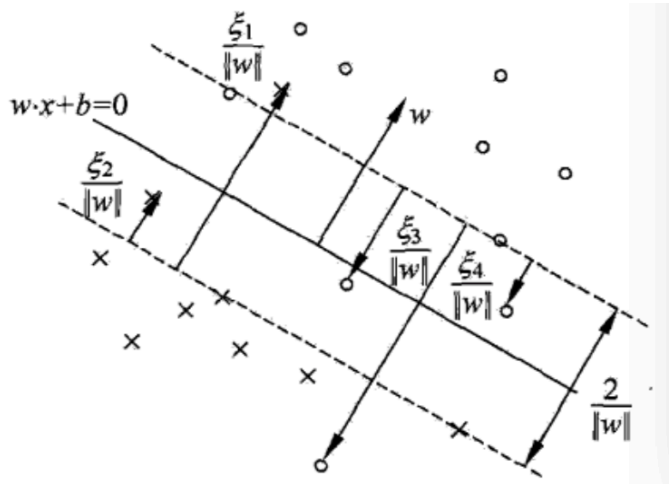

Figure 4: 软间隔的支持向量

在这里，软间隔的支持向量可能在间隔边界上，有可能在间隔边界和分离超平面之间，也有可能在分离超平面误分类一侧。根据上面 KKT 条件所满足的条件可以总结为如下 4 种情况：

（1）当 $\alpha_i^* < C$，则 $\varepsilon_i = 0$：支持向量恰好落在间隔边界上。

（2）当 $\alpha_i^* = C$，则 $0 < \varepsilon_i < 1$：则分类正确，支持向量落在间隔边界和分离超平面之间。

（3）当 $\alpha_i^* = C$，则 $\varepsilon_i = 1$：支持向量落在分离超平面上。

（4）当 $\alpha_i^* = C$，则 $\varepsilon_i > 1$：支持向量落在分离超平面误分类的一侧。

3.5 合页损失函数

下面从损失函数的角度来分析目标函数的合理性，首先合页损失函数如下所示：

$$[z]_+ = \begin{cases} z, & z > 0 \\ 0, & z \leqslant 0 \end{cases}$$

将支持向量机的损失函数写成：

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

也就是说当样本点 $(x^{(i)}, y^{(i)})$ 被正确分类且函数间隔 $y^{(i)}(w \cdot x^{(i)} + b) > 1$ 时，损失为 0, 否则损失为 $1 - y^{(i)}(w \cdot x^{(i)} + b)$。

然后对损失函数加上系数为 λ 的 w 的 L_2 范数，损失目标函数为：

$$\min_{w,b} \sum_{i=1}^{N} [1 - y^{(i)}(w \cdot x^{(i)} + b)]_+ + \lambda ||w||^2 \tag{23}$$

令 $[1-y^{(i)}(w \cdot x^{(i)} + b)]_+ = \varepsilon_i$, 和原始问题 (9) $\sim$ (11) 作比较。因为本式中 $\varepsilon_i \geqslant 0$, 所以约束条件 (11) 成立。又因为本式中当 $1 - y^{(i)}(w \cdot x^{(i)} + b) > 0$ 时有 $y^{(i)}(w \cdot x^{(i)} + b) = 1 - \varepsilon_i$，当 $1 - y^{(i)}(w \cdot x^{(i)} + b) \leqslant 0$ 时，$\varepsilon_i = 0$ 有 $y^{(i)}(w \cdot x^{(i)} + b) \geqslant 1 - \varepsilon_i$，所以约束条件 (10) 成立。所以优化问题 (23) 可以写成：

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

若取 $\lambda = \frac{1}{2C}$，则上式可以写成：

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

与式 (9) 等价。合页损失函数不仅要求分类正确，而且样本点距离分离超平面足够大的时候损失才是 0，所以说合页损失函数对学习有着更高的要求。

4 非线性支持向量机

4.1 基本定义

前面的两种支持向量机都是在线性的情况下进行分类，所谓线性就是在二维空间中直线作为分离超平面，三维空间中平面作为分离超平面的情况。但是在很多情况下通过线性函数无法正确划分正负样本。如下图所示，在原空间中无法通过直线将样本正确分开，需要通过一个椭圆曲线（非线性）才能将样本正确分开，在这里设原空间为 $\mathcal{X} \subset R^2, x = (x_1, x_2)^T \in \mathcal{X}$，$x_1, x_2$ 表示样本的两个特征，在图中分别表示横纵坐标。但是非线性的问题往往不好求解，所以需要将非线性的问题转化为线性问题。

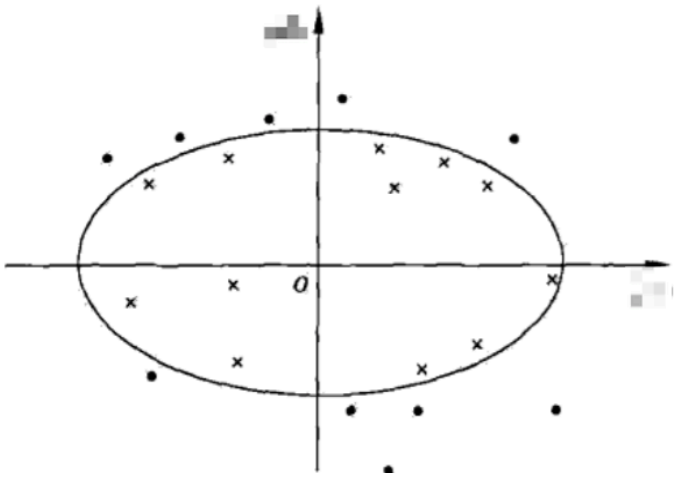

Figure 5: 原空间中的非线性情况

所以可以通过将原空间的数据映射到新空间，也可以理解为从低维空间映射到高维空间，这样就可以把原空间的椭圆变换为新空间的直线，如下所示：

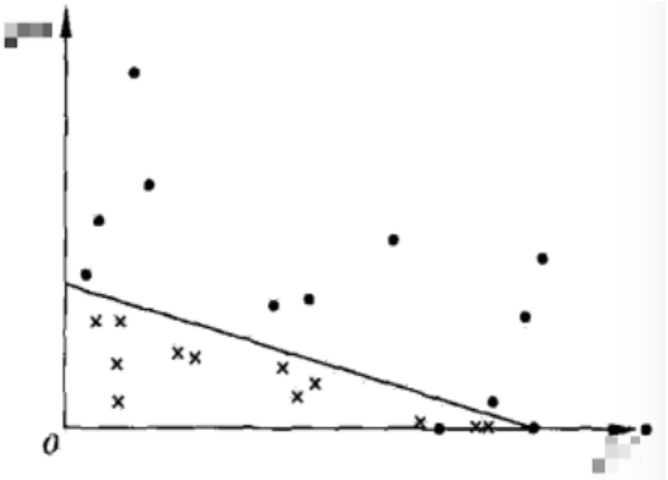

Figure 6: 通过核方法变换之后的空间

新空间为 $\mathcal{Z} \subset R^2, z = (z_1, z_2)^T \in \mathcal{Z}$。定义从原空间到新空间的映射为：

$$z = \phi(x) = ((x_1)^2, (x_2)^2)^T$$

上面的例子说明用线性分类方法求解非线性分类问题问题分为两步：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从训练数据中学习分类模型。核技巧就是属于这样的方法。

4.2 核函数与核技巧

（1）核函数的定义

设 $\mathcal{X}$ 是输入空间，$\mathcal{H}$ 为特征空间，如果存在一个从 $\mathcal{X}$ 到 $\mathcal{H}$ 的映射

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

使得对于所有的样本 $x, z \in \mathcal{X}$，函数 $K(x, z)$ 满足条件

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

则称 $K(x, z)$ 为核函数，$\phi(x)$ 为映射函数。

（2）核技巧

在学习与预测中只定义核函数 $K(x, z)$，而不显示地定义映射函数 ϕ。ϕ 是输入空间到特征空间的映射，特征空间一般是高维甚至无穷维的。对于给定的核函数，特征空间 $\mathcal{H}$ 和映射函数 ϕ 的取法不唯一，可以取不同的特征空间，即使再同一特征空间里也可以取不同的映射。

4.3 目标函数

我们注意到在线性支持向量机的对偶问题上，无论是目标函数还是决策函数都只涉及输入实例与输出实例之间的内积。在对偶问题的目标函数 (17) 中的内积 $x^{(i)} \cdot x^{(j)}$ 可以用核函数 $K(x^{(i)}, x^{(j)}) = \phi(x^{(i)}) \cdot \phi(x^{(j)})$ 来代替，因此在非线性情况下对偶目标函数可以写成：

$$W(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} K(x^{(i)}, x^{(j)}) - \sum_{i=1}^{N} \alpha_i$$

分类决策函数的内积也可以用核函数代替写成：

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i^* y^{(i)} \phi(x^{(i)}) \cdot \phi(x) + b^*) = sign(\sum_{i=1}^{N} \alpha_i^* y^{(i)} K(x^{(i)}, x) + b^*)$$

从上面的目标函数和分类决策函数可知，在核函数给定的条件下，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。那么下面来选择什么样的核函数就是需要讨论的问题。

4.4 正定核

首先要确定对于给定函数 $K(x,z)$，满足什么条件才能成为核函数？通常所说的核函数就是正定核函数，下面给出正定核的充要条件：

设 $K:\mathcal{X}\times\mathcal{X}\to R$ 是对称函数，则 $K(x,z)$ 为正定核的充要条件是对于任意的 $x^{(i)}\in\mathcal{X}, i=1,2,...,m$，$K(x,z)$ 对应的 $Gram$ 矩阵 $K=[K(x^{(i)},x^{(j)})]_{m\times m}$ 为半正定矩阵。

4.5 常用核函数

（1）多项式核函数

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

对应的支持向量机时一个 p 次多项式分类器，在此情形下分类决策函数为：

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

（2）高斯核函数

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

对应的支持向量机是高斯径向基函数分离器，在此情形下分类决策函数为：

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

5 序列最小优化算法

5.1 基本定义

前面的推导都是在 α 的解已知的情况下，那如何求解 α 呢？可以将支持向量的学习问题形式化为求解凸二次规划问题，但是当样本容量很大的时候这种算法会变的很低效。为了快速的求得全局最优解，序列最小优化算法 (SMO) 是一种重要的思路。

SMO 算法要解如下凸二次规划的对偶问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)}K(x^{(i)},x^{(j)})-\sum_{i=1}^{N}\alpha_i \tag{24}$$

$$s.t. \quad \sum_{i=1}^{N} \alpha_i y^{(i)} = 0 \tag{25}$$

$$0 \leqslant \alpha_i \leqslant C \quad i = 1, 2, ..., N \tag{26}$$

其基本思路是：

（1）如果所有变量 α_i 的解都满足此最优化问题的 KKT 条件，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要条件。

（2）否则选择两个变量，固定其它变量，针对这两个变量构造一个二次规划问题。这个问题有两个变量，一个是违反 KKT 条件最严重的那一个，另一个由约束条件 (25) 自动确定。

假设选择两个变量 α_1, α_2，那么其它变量 $\alpha_i (i = 3, 4, ...N)$ 是固定的，于是 SMO 最优化问题 (24) ∼ (26) 的子问题可以写成：

$$\begin{aligned} \min_{\alpha_1, \alpha_2} W(\alpha_1, \alpha_2) =& \frac{1}{2} K_{11} \alpha_1^2 + \frac{1}{2} K_{22} \alpha_2^2 + y^{(1)} y^{(2)} K_{12} \alpha_1 \alpha_2 \\ &+ y^{(1)} \alpha_1 \sum_{i=3}^{N} y^{(i)} \alpha_i K_{i1} + y^{(2)} \alpha_2 \sum_{i=3}^{N} y^{(i)} \alpha_i K_{i2} - (\alpha_1 + \alpha_2) \end{aligned} \tag{27}$$

$$s.t. \quad \alpha_1 y^{(1)} + \alpha_2 y^{(2)} = -\sum_{i=3}^{N} y^{(i)} \alpha_i = \varsigma \tag{28}$$

$$0 \leqslant \alpha_i \leqslant C \quad i = 1, 2 \tag{29}$$

其中 $K_{ij} = K(x_i, x_j), i, j = 1, 2, ..., N$，$\varsigma$ 是常数，目标函数 (27) 省略了不含 α_1, α_2 的常数项。

5.2 两个变量二次规划的求解方法

（1）为了求解 (27) ∼ (29)，首先需要分析约束条件，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的图形如下图表示。不等式约束 (29) 可以表示在下图所示的 $[0, C] \times [0, C]$ 盒子内，等式约束 (28) 使 (α_1, α_2) 在平行于盒子的对角线的直线上。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，在这里考虑为变量 α_2 的最优化问题。

首先假设问题 (27) ∼ (29) 的初始可行解为 $\alpha_1^{old}, \alpha_2^{old}$，最优解为 $\alpha_1^{new}, \alpha_2^{new}$，并假设在沿着约束方向未经剪辑即不考虑不等式约束 (28) 时 α_2 的最优解为 $\alpha_2^{new,unc}$。由于 α_2^{new} 需满足约束 (29)，所以它的取值范围必须满足约束条件 $L \leqslant \alpha_2^{new} \leqslant H$，$L$ 和 H 是 α_2^{new} 所在对角线段端点的界。

如果 $y^{(1)} \neq y^{(2)}$，那么 $y^{(1)}$ 和 $y^{(2)}$ 异号，在这里假定 $y^{(1)} = 1, y^{(2)} = -1$，那么约束 (28) 可以写成 $\alpha_1 - \alpha_2 = \varsigma$ 代表了下面左图的虚线方程。此时虚线与盒子的交点分别是 $(\varsigma, 0)$ 和 $(C, C - \varsigma)$，就有

$0 \leqslant \alpha_2^{new} \leqslant C + \alpha_2 - \alpha_1$，将初始可行解 $\alpha_1^{old}, \alpha_2^{old}$ 代入即得到 $0 \leqslant \alpha_2^{new} \leqslant C + \alpha_2^{old} - \alpha_1^{old}$。同理如果 $y^{(1)} = -1, y^{(2)} = 1$ 可得 $\alpha_2^{old} - \alpha_1^{old} \leqslant \alpha_2^{new} \leqslant C$，就可以用下面的形式表示：

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

如果 $y^{(1)} = y^{(2)}$，就如下面右图所示，推理过程同上。则

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

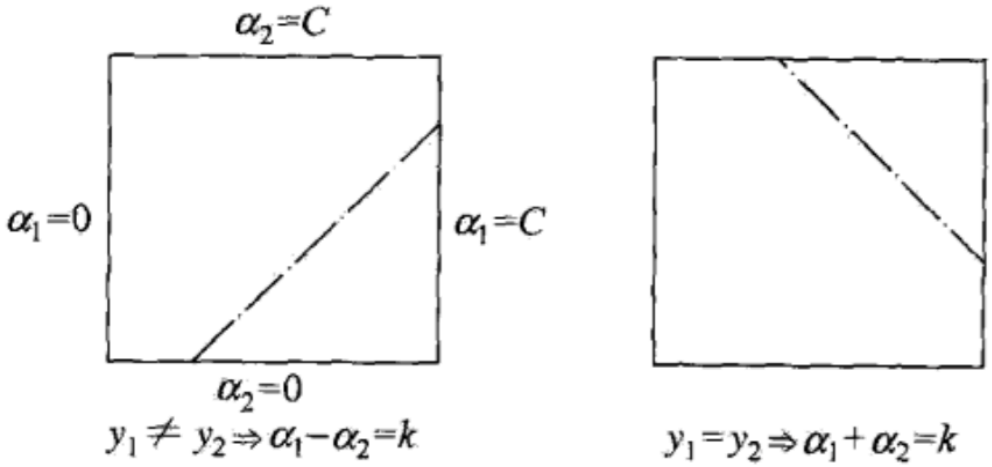

Figure 7: 二变量优化问题图示

（2）下面首先要沿着约束方向未经剪辑时 α_2 的最优解 $\alpha_2^{new,unc}$，然后再求剪辑后的解 α_2^{new}，在这里通过定理直接叙述这个结果。记 $g(x) = \sum_{i=1}^{N} \alpha_i y^{(i)} K(x_i, x) + b$，用 E_i 表示函数 $g(x)$ 对输入 $x^{(i)}$ 的预测值与真实输出 $y^{(i)}$ 的差，记为 $E_i = g(x^{(i)}) - y^{(i)} = \sum_{i=1}^{N} \alpha_i y^{(i)} K(x^{(i)}, x) + b - y^{(i)}, i = 1, 2$。

最优化问题 (27) ~ (29) 沿着约束方向未经剪辑时的解如下，其中 $\phi(x)$ 是输入空间到特征空间的映射。

$$\alpha_2^{new,unc} = \alpha_2^{old} + \frac{y^{(2)}(E_1 - E_2)}{\eta} \tag{30}$$

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

经剪辑后 α_2 的解是：

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

得到 α_2^{new} 后需要求解 α_1^{new}。通过约束 (28) 可知 $\alpha_1^{new}y^{(1)}+\alpha_2^{new}y^{(2)}=\alpha_1^{old}y^{(1)}+\alpha_2^{old}y^{(2)}=\varsigma$，所以可得：

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

最终得到了最优化问题 (27) $\sim$ (29) 的解 $\alpha_1^{new},\alpha_2^{new}$。

5.3 变量选择方法

SMO 算法在每个子问题中选择两个变量优化，其中至少有一个变量是违反 KKT 条件的。

（1）第 1 个变量的选择

将选择第 1 个变量的过程称为外层循环。外层循环再训练样本中选择违反 KKT 条件最严重的样本点，并将其对应的变量作为第一个变量。检验样本点 $(x^{(i)},y^{(i)})$ 是否满足 KKT 条件，即：

$$y^{(i)}g(x^{(i)})=\begin{cases}\geqslant 1 & \{x^{(i)}|\alpha_i=0\}\\ =1 & \{x^{(i)}|0<\alpha_i<C\}\\ \leqslant 1 & \{x^{(i)}|\alpha_i=C\}\end{cases} \tag{31}$$

其中 $g(x^{(i)})=\sum_{j=1}^{N}\alpha_j y^{(j)}K(x^{(i)},x^{(j)})+b$。该检验是在 ε 的范围内进行的，在检验过程中外层循环首先遍历所有满足 $0<\alpha_i<C$ 的样本点即在间隔边界上的支持向量点，如果这些样本点都满足 KKT 条件，那么就遍历全部样本点检验它们是否都满足 KKT 条件。

（2）第 2 个变量的选择

第 2 个变量的选择称之为内层循环，在找到第 1 个变量 α_1 的后，要在内层循环找到有足够大变化的第 2 个变量 α_2。由式 (30) 可知，α_2^{new} 依赖于 $|E_1-E_2|$ 的，所以为了加快计算速度，选择 α_2 时使其 $|E_1-E_2|$ 最大。将所有的 E_i 保存在一个列表中。

在特殊情况下，如果通过上面方法在内层循环选择的 α_2 不能使目标函数有足够的下降，那么可以采用启发式规则继续选择 α_2。即遍历在间隔边界上的支持向量点对应的变量依次作为 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。

（3）计算阈值 b 和差值 E_i

每次完成两个变量的优化之后，都要重新计算阈值 b。由 (31) 可知 $\sum_{i=1}^{N}\alpha_i y^{(i)}K_{i1}+b=y^{(1)}$，于是

$$b_1^{new}=y^{(1)}-\sum_{i=3}^{N}\alpha_i y^{(i)}K_{i1}-\alpha_1^{new}y^{(1)}K_{11}-\alpha_2^{new}y^{(2)}K_{21} \tag{32}$$

由于 $E_1=\sum_{i=3}^{N}\alpha_i y^{(i)}K_{i1}+\alpha_1^{old}y^{(1)}K_{11}+\alpha_2^{old}y^{(2)}K_{21}+b^{old}-y^{(1)}$ 那么式 (32) 的前两项可写成 $y^{(1)}-\sum_{i=3}^{N}\alpha_i y^{(i)}K_{i1}=-E_1+\alpha_1^{old}y^{(1)}K_{11}+\alpha_2^{old}y^{(2)}K_{11}+b^{old}$，代入 (32) 得

$$b_1^{new}=-E_1-y^{(1)}K_{11}(\alpha_1^{new}-\alpha_1^{old})-y^{(2)}K_{21}(\alpha_2^{new}-\alpha_2^{old})+b^{old}$$

同样如果 $0 < \alpha_2^{new} < C$，有

$$b_2^{new} = -E_2 - y^{(1)}K_{12}(\alpha_1^{new} - \alpha_1^{old}) - y^{(2)}K_{22}(\alpha_2^{new} - \alpha_2^{old}) + b^{old}$$

如果 α_1^{new}，α_2^{new} 同时满足大于 0 小于 C 的条件，那么 $b_1^{new} = b_2^{new}$。如果这两个参数的值是 0 或 C，那么 b_1^{new} 和 b_2^{new} 以及它们之间的数都是符合 KTT 条件的阈值，这时选择他们的中点作为 b^{new}。

由于每次完成两个变量的优化之后，还必须更新对应的 E_i 值，并将其保存在列表中。其中 S 是所有支持向量的集合。

$$E_i^{new} = \sum_S y^{(j)}\alpha_j K(x^{(i)}, x^{(j)}) + b^{new} - y^{(i)}$$

5.4 SMO 算法总结

输入：训练数据集 $T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})\}$，其中 $x_i \in \mathcal{X} = R^n, y_i \in \mathcal{Y} = -1, +1, i = 1, 2, ..., N$，精度为 ε。

输出：近似解 $\hat{\alpha}$。

（1）取初值 $\alpha^{(0)} = 0$，另 $k = 0$。

（2）选取最优化变量 $\alpha_1^{(k)}, \alpha_1^{(k)}$，解析求解两个变量的最优化问题 (27) $\sim$ (29)，求得最优解 $\alpha_1^{(k+1)}, \alpha_1^{(k+1)}$，更新 α 为 $\alpha^{(k+1)}$。

（3）若在精度 ε 范围内满足如下的停机条件，其中 $g(x^{(i)}) = \sum_{j=1}^N \alpha_j y^{(j)} K(x^{(i)}, x^{(j)}) + b$。

$$\sum_{i=1}^N \alpha_i y^{(i)} = 0$$

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

$$y^{(i)} g(x^{(i)}) = \begin{cases} \geqslant 1 & \{x^{(i)} | \alpha_i = 0\} \\ = 1 & \{x^{(i)} | 0 < \alpha_i < C\} \\ \leqslant 1 & \{x^{(i)} | \alpha_i = C\} \end{cases}$$

则转（4），否则令 $k = k + 1$ 转（2）。

（4）取 $\hat{\alpha} = \alpha^{(k+1)}$。

6 总结

本文的笔记内容基本都是参考统计学习方法支持向量机的内容，按照自己认为好理解的方式进行排版和注释，时间关系无法将所有形式化的语言口语化，也是感觉这样表述不够准确，在 SMO 算法上理解不够透彻，希望多多交流指正。

7 参考文献

1. 李航博士的统计学习方法
2. 斯坦福大学吴恩达课件