凸优化在支持向量机的应用

控制一班谢硕

2023/1/9

摘要

支持向量机 (SVM) 是一种基于统计学习理论的机器学习方法，由于其优越的学习性能，已经成为当前模式识别、数据挖掘等机器学习领域的研究热点。本文对于支持向量机理论进行了推导，以深刻体会凸优化理论在支持向量机理论中的应用，并加强自己对凸优化的理解。

Abstract

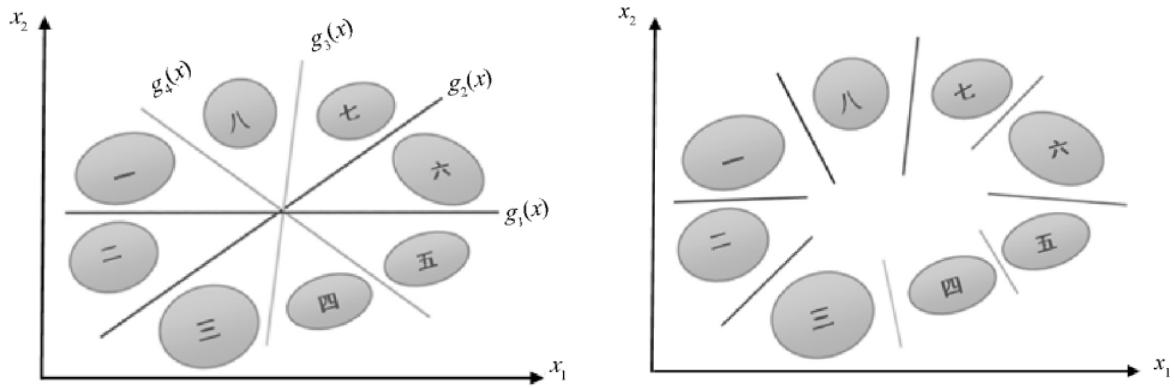
Support vector machine (SVM) is a machine learning method based on statistical learning theory, which has become a research hotspot in the field of machine learning such as pattern recognition and data mining due to its superior learning performance. This paper derives the theory of support vector machines to deeply appreciate the application of convex optimization theory in support vector machine theory and strengthen my own understanding of convex optimization.

目录

1 凸优化理论在线性可分支持向量机的应用

1.1 问题描述

设有数目不同的 8 个学生团体，按单位坐成 8 个学生团体，按单位坐成 8 个圆团，利用四条直线 $g_1(x)$,$g_2(x)$,$g_3(x)$,$g_4(x)$, 把它们按如图 1(a)、(b) 所示的形式分隔开，我们要找出哪几个团队之间的分隔线最合理。通过观察相邻团队到分隔线的间距，易看出图 1 中的团队一与二、六与七、七与八的分隔间距较大相对合理，而图 2 中相邻两个团队之间的分隔间距都较大，因此都比较合理。



(a) 团队间任意分隔　　(b) 团队间最佳分隔

图 1: 团队间分隔方式

1.2 点到直线 l 的距离 d

对于如图 2 所示的直线 $l : x_2 = kx_1 + b$, 可以变形为 $kx_1 - x_2 + b = 0$。为了方便，写成向量形式，令 $x = (x_1, x_2)^T$, 则直线 l 可写成 $g(x) = (k, -1)(x_1), {x_2}^T + b = w^T x + b = w \cdot x + b = 0$。

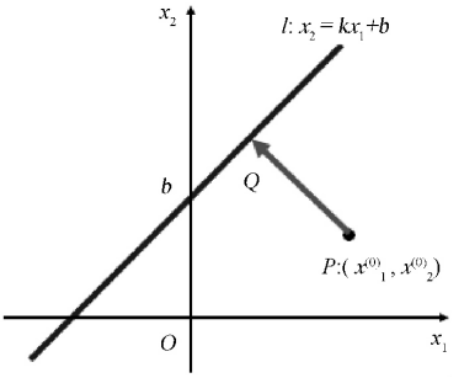


图 2: 点到直线的距离

从上可知，改变 w 和 b 的数值，可分别确定直线的方向和位置，它是确定最佳分类线的基础。由初等数学知，点 $P(x_1^0, y_2^0)$ 到直线 l 的距离为

$$d = \frac{\left|kx_1{}^{(0)} - x_2{}^{(0)} + b\right|}{\sqrt{k^2 + (-1)^2}} = \frac{\left|w^T x^{(0)} + b\right|}{\|w\|} \tag{1}$$

如图 3，当两直线 $l_1 : x_2 = k_1 x_1 + b_1$ 和 $l_2 : x_2 = k_2 x_1 + b_2$ 平行时。有 $k_1 = k_2$, 只有 b_1 和 b_2 不同，所以有平行直线间的距离为

$$d = \frac{|b_1 - b_2|}{\sqrt{k^2 + (-1)^2}} = \frac{|b_1 - b_2|}{\|w\|} \tag{2}$$

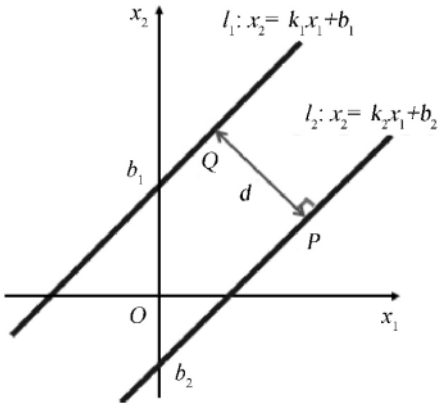
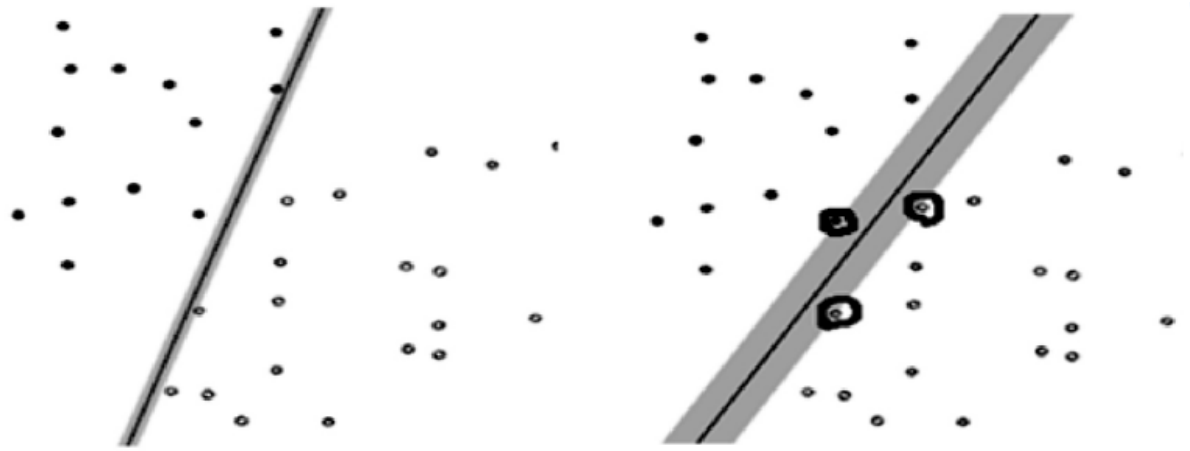


图 3: 平行线间的距离

1.3 将 d 引入支持向量机 (SVM)

所谓 SVM，就是寻找一条分类线，使之到两类样本中最近点的距离最大的一种机器学习算法。对图 4(a)，要用一条直线，将图中的实心点和空心点分开，显然，图上的

这条直线就是一条分类线，而由于图 4(b) 的分隔线显然分类效果更好，因为其分类间隔更宽。



(a) 任意分隔线　　(b) 最佳分隔线

图 4: 寻找一条分类线

而对图 5，已知实心、空心两类共 L 个训练样本 $\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{L}$, 样本特征 $x^i = \left(x_1^{(i)}, x_2^{(i)}\right)$, 样本分类值为 $y^{(i)} \in \{-1, 1\}$。

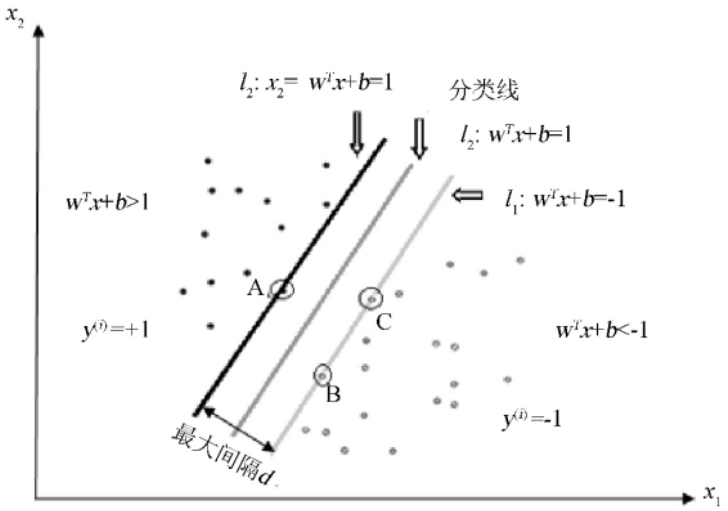


图 5: SVM 分类图

设分类先为 $l : g(x) = w^T x + b = 0$,l 右下侧的线为 $l_1 : g(x) = w^T x + b = -1$，$l$ 左上侧的线为 $l_2 : g(x) = w^T x + b = 1$。则离分类线 l 最近的样本 x 满足 $|g(x) = 1|$(即在面 l_1 或 l_2 上), 这样的 x 称为支持向量。把分类值 $y^{(i)}$ 考虑进去，两类样本点满足的统一表达式为 $(w^T x^{(i)} + b)y^{(i)} \geq 1$，$i = 1, 2, \cdots, L$。这时 l^1 与 l^2 之间的距离为:$d = \frac{2}{||w||}$。

为了能把两类样本分得更开，就希望在满足 $(w^T x^{(i)} + b)y^{(i)} \geq 1$，$i = 1, 2, \cdots, L$ 的条件下，$d = \frac{2}{||w||}$ 最大，这显然是一个条件极值问题。

1.4 条件极值的求法

求解上面的问题，即求 $d = \frac{2}{||w||}$ 在约束条件 $(w^T x^{(i)} + b)y^{(i)} \geq 1$，$i = 1, 2, \cdots, L$ 下的最大值。求法如下：

步骤 1：将约束条件写成 $(w^T x^{(i)} + b)y^{(i)} - 1 \geq 0$，$i = 1, 2, \cdots, L$；

步骤 2：为了便于计算，将求 $d = \frac{2}{||w||}$ 最大，转化为求 $\frac{1}{2}||w||^2$ 最小。

步骤 3：利用拉格朗日函数求极值，设拉格朗日函数为：

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{i=1}^{L} \alpha_i \left[(w^T x^{(i)} + b)y^{(i)} - 1\right] \tag{3}$$

其中，Lagrange 系数 $\alpha_i \geq 0$，$i = 1, 2, \cdots, L, \alpha = (\alpha_1, \alpha_1, \cdots, \alpha_L)^T$。

步骤 4：对 (3) 式分别关于 w 和 b 求偏导数，并令它们等于零，得

$$\begin{cases} \frac{\partial L(w,b,\alpha)}{\partial w_k} = w_k - \sum_{i=1}^{L} \alpha_i y^{(i)} x_k^i = 0, k = 1, 2, \cdots, n \\ \frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^{L} \alpha_i y^{(i)} = 0, k = 1, 2, \cdots, n \end{cases} \tag{4}$$

即

$$\begin{cases} w = \sum_{i=1}^{L} \alpha_i y^{(i)} x^{(i)} \\ \sum_{i=1}^{L} \alpha_i y^{(i)} = 0 \end{cases} \tag{5}$$

步骤 5：把第 (3) 式展开，再将 (5) 式的等量关系代入，就变成

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{L} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1\right] \tag{6}$$

$$= \frac{1}{2}||w||^2 - w^T \sum_{i=1}^{L} \alpha_i y^{(i)} x^{(i)} - b \sum_{i=1}^{L} \alpha_i y^{(i)} + \sum_{i=1}^{L} \alpha^i \tag{7}$$

$$= \sum_{i=1}^{L} \alpha^i - \frac{1}{2} \left(\sum_{i=1}^{L} \alpha_i y^{(i)} x^{(i)} \right)^T \left(\sum_{i=1}^{L} \alpha_i y^{(i)} x^{(i)} \right) = Q(\alpha) \tag{8}$$

要求 $Q(\alpha)$ 在边界约束 $\alpha_i \geq 0$，$i = 1, 2, \cdots, L$ 和等式约束 $\sum_{i=1}^{L} \alpha_i y^{(i)} = 0$ 下的最值点。显然，$Q(\alpha)$ 是关于 α 的二次函数，这是一个在凸集约束下的优化问题，且具有

上面的边界约束和线性等式约束．求解该问题的标准形式为

$$\begin{cases} minf(\alpha) = -Q(\alpha) = \frac{1}{2}\alpha^T H\alpha + C^T\alpha \\ s.t. \quad A_{eq}^T\alpha = 0 \\ \qquad lb \leq \alpha \end{cases} \tag{9}$$

其中，$M = \left[x^{(1)}, x^{(2)}, \cdots, x^{(L)}\right] \times \begin{bmatrix} y^{(1)} & & & \\ & y^{(2)} & & \\ & & \cdot & \\ & & \cdot & y^{(L)} \end{bmatrix}$，$H = M^T M$，$C = [-1, -1, \cdots, -1]^T$，$A_{eq} = [y^{(1)}, y^{(2)}, \cdots, y^{(L)}]^T$，$lb = [0, 0, \cdots, 0]^T$。如图 6 所示，可以直观的理解二次二次凸函数在唯一凹点 α^* 处取到最小值 (最优解)。

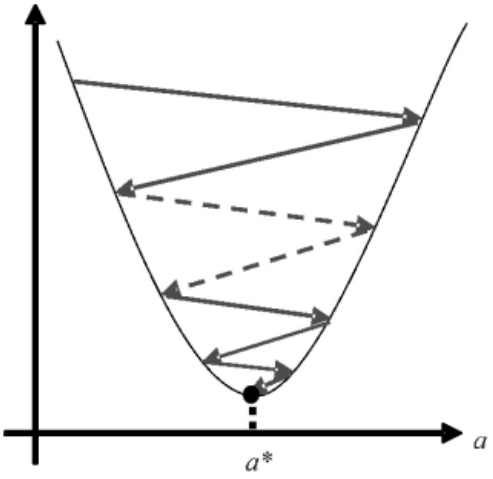


图 6: 二次凸函数的最值

对于该问题，则可利用 Matlab 中专门求解二次规划问题的 quadprog 函数，如下式

$$\begin{cases} minf(x) = -\frac{1}{2}x^T Hx + c^T x \\ s.t. \quad Ax \leq b \\ \qquad A_{eq}x = b_{eq} \\ \qquad lb \leq x \leq ub \end{cases} \tag{10}$$

其中,$x = quadprog(H, c, A, b, A_{eq}, b_{eq}, lb, ub, x_0, options)$。针对本问题，可求出唯一全局最优解 α^*，且这个优化问题的解是在边界条件上取得，即满足

$$\alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1\right] = 0, i = 1, 2, \cdots, L \tag{11}$$

则①当 $y^{(i)}(w^Tx^{(i)}+b)-1>0$ 时，必有 $\alpha_i=0$;

②当 $\alpha_i>0$ 时，必有 $y^{(i)}(w^Tx^{(i)}+b)-1>0$ 时，必有 $\alpha_i=0$。说明 $x^{(i)}$ 是支持向量，它们通常只是占总样本中的很少一部份。

只要求出 $Q(\alpha)$ 的最值点 α^*，则由式（5）就可得

$$w^*=\sum_{i=1}^{L}\alpha_i^*y^{(i)}x^{(i)} \tag{12}$$

再由式（ 1 ）的条件或式（11）选出的支持向量直接求得

$$b^*=-\frac{\max\limits_{y_i=-1}(w^*)^Tx^{(i)}+\max\limits_{y_i=-1}(w^*)^Tx^{(i)}}{2} \tag{13}$$

上式由根据离分类线最近的正的函数间隔等于离分类线最近的负的函数间隔所得。

步骤 6：从而得到最优分类线

$$g(x)=(w^*)^Tx+b^*=\sum_{i=1}^{L}\alpha_i^*y^{(i)}(x^{(i)}\cdot x)+b^*=0 \tag{14}$$

则两类问题的分类函数为：

$$S_{(x)}=sgn\left\{\sum_{i=1}^{L}\alpha_i^*y^{(i)}(x^{(i)}\cdot x)+b^*\right\}=\begin{cases}+1,when\sum_{i=1}^{L}\alpha_i^*y^{(i)}(x^{(i)}\cdot x)+b^*>0\\-1,when\sum_{i=1}^{L}\alpha_i^*y^{(i)}(x^{(i)}\cdot x)+b^*<0\\0,when\sum_{i=1}^{L}\alpha_i^*y^{(i)}(x^{(i)}\cdot x)+b^*=0\end{cases} \tag{15}$$

综合上述方法，可得出如下的 SVM 分类步骤为：

①通过已知样本的特征 $x^{(i)}$ 和类别样本 $y^{(i)}(\pm1)$，求得最佳分类参数 α_i^* 和 b^*；

②将待判别的样本特征 x 代入分类函数式 (15)，就可得到分类结果。

2 凸优化理论在非线性可分支持向量机的应用

2.1 问题引入

通过利用非线性模型才能很好地进行分类的问题称为非线性分类问题，如图 7 所示。两类数据分别分布为两个圆圈的形状，不论是任何高级的分类器，只要它是线性的，就没法处理，SVM 也不行。因为这样的数据本身就是线性不可分的，是一个典型的非线性可分问题。

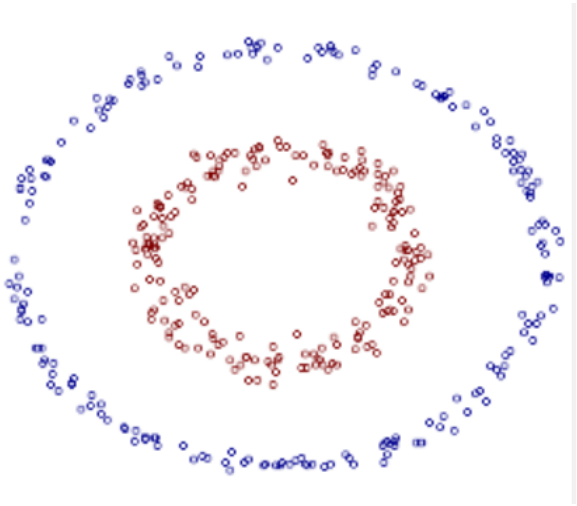


图 7: 典型的非线性可分问题

非线性问题往往不好求解，所以我们希望用线性分类问题的方法来解决这个问题。所采用的方法是进行非线性变换。

2.2 核函数与核技巧

首先，核函数有如下定义。

设 χ 是输入空间，$\aleph$ 为特征空间，如果存在一个从 χ 到 $\aleph$ 的映射：

$$\phi(x) : \chi \to \aleph \tag{16}$$

使得对所有 $x, z \in \chi$，函数 $K(x, z)$ 满足条件：

$$K(x, z) = \phi(x) \cdot \phi(z) \tag{17}$$

则称 $K(x, z)$ 为核函数，$\phi(x)$ 为映射函数，式中 $\phi(x) \cdot \phi(z)$ 为内积。

常用的核函数有：

线性核：

$$K(x_i, x_j) = x_i^T x_j \tag{18}$$

多项式核：

$$K(x_i, x_j) = (x_i^T x_j)^p, \qquad p \geq 1 \tag{19}$$

高斯核：

$$K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2}), \qquad \sigma > 1 \tag{20}$$

拉普拉斯核：

$$K(x_i, x_j) = exp(-\frac{||x_i - x_j||}{\sigma}), \qquad \sigma > 0 \tag{21}$$

线性分类方法求解非线性分类问题一般分为两步：①使用一个变换将原空间的数据映射到新空间。②在新空间里使用线性分类学习方法从训练数据中学习分类模型. 核技巧就是属于上诉介绍的方法，应用到支持向量机的基本思想就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。核技巧的思想为：在学习与预测中只定义核函数 $K(x, z)$，而不显式地定义映射函数 ϕ. 为通常计算 $K(x, z)$ 比较容易，而通过 $\phi(x)$ 和 $\phi(z)$ 的内积来计算 $K(x, z)$ 并不容易。ϕ 是输入空间到特征空间的映射，特征空间 $\aleph$ 往往是高维的，甚至是无穷维。且对于给定的核 $K(x, z)$，特征空间 $\aleph$ 和映射函数 ϕ 的取法并不唯一。

对于线性可支持向量机，无论是目标函数还是决策函数（分离超平面）都只涉及输入实例与实例之间的内积。在对偶问题的目标函数中的内积 x_i，x_j，可以用核函数 $K(x, z) = \phi(x) \cdot \phi(z)$ 来代替。此时对偶问题的目标函数成为：

$$W(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N}\alpha_i \tag{22}$$

分类决策函数变为：

$$S_{(x)} = sgn\left\{\sum_{i=1}^{N_S}\alpha_i^* y^{(i)} K(x, x_i) + b^*\right\} \tag{23}$$

这就等价于：经过映射函数 ϕ 将原来的输入空间变换到一个新的特征空间，将输入空间中的内积 $x_i \cdot x_j$ 变换为特征空间中的内积 $\phi(x_i) \cdot \phi(x_j)$，在新的特征空间里，从训练样本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是非线性模型。在核函数 $K(x, z)$ 给定的条件下，可以利用求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进行，不需要显式地定义特征空间和映射函数，这样的技巧称为核技巧。

2.3 非线性支持向量机学习算法

假设输入训练数据集为：$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，其中，$x_i \in \chi = R^n, y_i \in y = \{-1, +1\}, i = 1, 2, \cdots, N$; 输出为分类决策函数。具体算法步骤如下：

步骤 1：选取适当的核函数 $K(x,z)$ 和适当的参数 C，构造并求解如下优化问题：

$$\begin{cases} \min\limits_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j) - \sum_{i=1}^{N}\alpha_i \\ s.t. \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \\ \qquad 0 \le \alpha_i \le C, i = 1,2,\cdots,N \end{cases} \tag{24}$$

求得最优解 $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*)$

步骤 2：选择 α^* 的一个正分量 $0 < \alpha_j < C$，计算：

$$b^* = y_j - \sum_{i=1}^{N}\alpha_i y_i K(x_i, x_j) \tag{25}$$

步骤 3：构造决策函数：

$$S_{(x)} = sgn\left\{\sum_{i=1}^{N_S}\alpha_i^* y^{(i)} K(x, x_i) + b^*\right\} \tag{26}$$

当 $K(x,z)$ 是正定核函数时，该问题为凸二次规划问题，解是存在的。

2.4 序列最小最优化算法 (SMO)

对于上述的凸二次规划问题具有全局最优解，也有许多最优化算法可以用于求解这一问题，但是当训练数据集容量很大时，这些算法往往变得非常低效。下面介绍序列最小最优化算法 (SMO 算法)，此算法可以快速求解此问题。

SMO 算法是将大优化问题分解为多个小优化问题求解的，这些小优化问题往往很容易求解，并且对它们进行顺序求解的结果与将它们作为整体来说求解的结果是一样的。SMO 算法的目标是求解一系列 α 和 b，一旦求出这些 α，就很容易计算出权重向量 w 并且得到分离超平面。

由于 KKT 条件是该最优化问题的充分必要条件，故如果所有变量 (即拉格朗日乘子 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。而此最优化问题的 KKT 条件为：

$$\begin{cases} \alpha_i = 0 \Leftrightarrow y_i g(x_i) \ge 1 \\ 0 < \alpha_i < C \Leftrightarrow y_i g(x_i) = 1 \\ \alpha_i = C \Leftrightarrow y_i g(x_i) \le 1 \end{cases} \tag{27}$$

其中：

$$g(x_i) = \sum_{j=1}^{N} \alpha_j y_j K(x_i, x_j) + b \tag{28}$$

想要所有变量都满足上述 KKT 条件，可以先固定 α_i 之外的所有参数，然后求 α_i 上的极值。由于约束条件 $\sum_{i=1}^{N} \alpha_i y_i = 0$ 的存在，若固定其它变量，则 α_i 可由其他变量导出。

下面介绍 SMO 算法具体步骤：

输入为训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，其中，$x_i \in \chi = R^n, y_i \in y = \{-1, +1\}, i = 1, 2, \cdots, N;$，精度为 ε，输出为近似解 α。

步骤 1：取初始值 $\alpha^{(0)} = 0$，令 $k = 0$；

步骤 2：选取优化变量 $\alpha_1^{(k)}$，$\alpha_2^{(k)}$，解析求解两个变量的最优化问题，求得最优解 $\alpha_1^{(k+1)}$，$\alpha_2^{(k+1)}$。

步骤 3：若在精度 ε 范围内满足停止条件：

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{29}$$

$$0 \leq \alpha_i \leq C, i = 1, 2, \cdots, N \tag{30}$$

$$y_i \cdot g(x_i) = \begin{cases} \geq 1, & \{x_i | \alpha_i = 0\} \\ = 1, & \{x_i | 0 < \alpha_i < C\} \\ \leq 1, & \{x_i | \alpha_i = C\} \end{cases} \tag{31}$$

其中：

$$g(x_i) = \sum_{j=1}^{N} \alpha_j y_j K(x_i, x_j) + b \tag{32}$$

则转步骤 4，否则令 $k = k + 1$，转步骤 2。

步骤 4：取 $\alpha = \alpha^{(k+1)}$

3 SMO 算法仿真

使用聚类生成器 ($make_blobs$) 和圆环图 ($make_circles$) 生成两类样本共 500 个。对样本 1 和样本 2 使用 SMO 算法进行分类，使用高斯核函数进行空间映射，之后使用 SMO 算法以完成支持向量机进行非线性分类的任务。

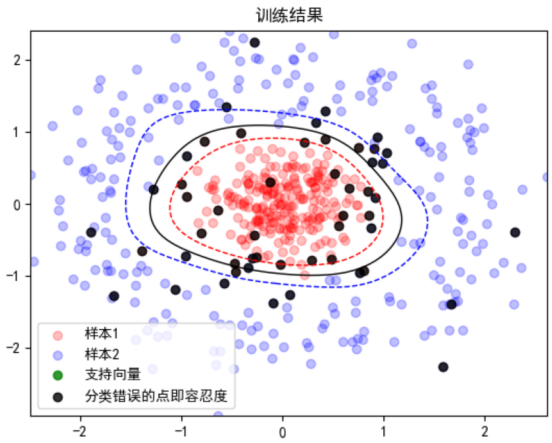


图 8: 仿真结果