向量机

胡琛

2016 年 4 月 14 日

目录

1 概述

支持向量机 (support vector machines, SVM) 是一种二类分类模型，基本模型是定义在特征空间上的间隔最大的线性分类器。支持向量机还包括核技巧，使其称为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划 (convex quadratic programming) 的问题，也等价于正则化的合页损失函数的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。

支持向量机学习方法包括构建由简至繁的模型：

1. 线性可分支持向量机 (linear support vector machine in linearly separable case)：训练数据可分时，通过硬间隔最大化 (hard margin maximization)，学习一个线性的分类器。

2. 线性支持向量机 (linear support vector machine)：训练数据近似线性可分时，通过软间隔最大化 (soft margin maximization)，也学习一个线性分类器

3. 非线性支持向量机 (non-linear support vector machine)：当训练数据不可分时，通过使用核技巧 (kernel trick) 以及软间隔最大化，学习非线性支持向量机

当输入空间为欧式空间或离散集合、特征空间为希尔伯特空间时，核函数 (kernel function) 表示将输入从输入空间映射到特征空间，得到的特征向量之间的内积。通过使用核函数，可以学习非线性支持向量机，等价于隐式

地在高维的特征空间中学习线性支持向量机。这样的方法被称为核技巧。核方法 (kernel method) 是比支持向量机更为一般的机器学习方法。

2 线性可分支持向量机与硬间隔最大化

2.1 线性可分向量机

2.1.1 数据集

假设给定一个特征空间上的训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}, \tag{1}$$

其中,

$$x_i \in \mathcal{X} = R^n, y_i \in \mathcal{Y} = \{+1, -1\}, i = 1, 2, \ldots, N, \tag{2}$$

x_i 为第 i 个特征向量，也称为实例，y_i 为 x_i 的类标记，当 $y_i = +1$ 时，称 x_i 为正例；当 $y_i = -1$ 时，称 x_i 为负例。(x_i, y_i) 为样本点，再假设训练数据集是线性可分的。

2.1.2 目标

学习的目标是在特征空间中找到一个分离超平面，能将实例分到不同的类。分离超平面对应于方程 $\omega \cdot x + b = 0$，它由法向量 ω 与截距 b 决定的，可用 (ω, b) 表示。分离超平面将特征空间划分为两部分，一部分是正类，一部分是负类，法向量指向的一侧是正类，另一侧是负类。

2.1.3 说明

一般地，当训练数据集线性可分时，存在无穷个分离超平面可将两类数据正确分开，感知机利用误分类最小的策略，求得分离超平面，不过，此时的解有无穷多个。线性可分支持向量机利用间隔最大化求最优分离超平面，此时，解是唯一的。

2.1.4 定义

给定线性可分[1]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得到的分离超平面为：

$$\omega^* \cdot x + b^* = 0 \tag{3}$$

以及相应的分类决策函数：

$$f(x) = sign(\omega^* \cdot x + b^*) \tag{4}$$

称为线性可分支持向量机。

2.1.5 举例

如图 1 所示，红色点与蓝色点就是线性可分的，我们需要做的是找到一个最优化的分割方案。线性可分支持向量机就是去找到一种方案，使得分割线与两边的间隔最大。

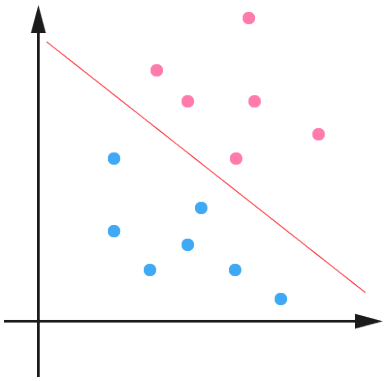


图 1: 二类分类问题

[1]线性可分的定义：给定一个数据集 $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$，其中，$x_i \in \mathcal{X} = R^n, y_i \in \mathcal{Y} = \{+1, -1\}, i = 1, 2, \ldots, N$，如果存在某个超平面 S：$\omega \cdot x + b = 0$ 能够将数据集完全正确地划分到超平面的两侧，即对所有 $y_i = +1$ 的实例 i，有 $\omega \cdot x + b > 0$，对所有 $y_i = -1$ 的实例 i，有 $\omega \cdot x + b < 0$，则称数据集 T 是线性可分数据集 (linear separable data set)；否则，称数据集为线性不可分。

2.2 函数间隔与几何间隔

2.2.1 说明

对于数据集 $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ 上的数据点，它到分隔超平面的距离肯定是关于坐标 $(\vec{x}, \vec{y})$ 的一个函数，考虑到几何上点到面的距离向量必然与超平面的法向量平行，假设超平面的法向量为 $\vec{\omega}$，同时，由于需要考虑平面两侧不同，点到面的距离我们可以人为设置正负，以便与分类标签 $y_i = +1, -1$ 对应，于是，很自然地我们可以写出点到面的距离的函数如下：

$$\bar{\gamma}_i = y_i(\omega \cdot x_i + b) \tag{5}$$

其中，y_i 是数据点的 label，按数据点在分界面的哪一边来定，b 是截距。以上的方式有一个缺点，当我们对上式中 ω 和 b 同时乘以一个因子时，由此确定的超平面是不变的，但是由上式定义出的函数间隔却变为原来的两倍。这意味着，仅靠上式来确定数据点到分离面的距离是不够的。为此，我们可以通过几何上点到面的距离计算公式，来确定数据点到分离面的距离，如图 2 所示的 γ_i，几何上可以由下式给出：

$$\gamma_i = y_i \left(\frac{\omega}{||\omega||} \cdot x_i + \frac{b}{||\omega||} \right) \tag{6}$$

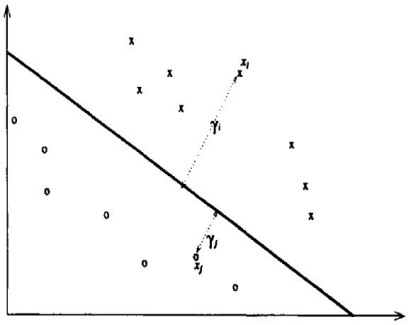


图 2: 几何间隔示意图

然后，我们只需要找到 $\gamma = \min\limits_{i=1,\ldots,N} \gamma_i$ 并使之取极大值，就可以确定我们需要的分隔超平面与数据点到分离面的几何距离了。

2.2.2 定义

- 函数间隔

对于给定的训练数据集 T 和超平面 (ω, b)，定义超平面 (ω, b) 关于样本点 (x_i, y_i) 的函数间隔为：

$$\bar{\gamma_i} = y_i(\omega \cdot x_i + b) \tag{7}$$

定义超平面 (ω, b) 关于训练数据集 T 的函数间隔为超平面 (ω, b) 关于 T 中所有样本点 (x_i, y_i) 的函数间隔最小值，即：

$$\bar{\gamma} = \min_{i=1,\dots,N} \bar{\gamma_i} \tag{8}$$

- 几何间隔

对于给定的训练数据集 T 和超平面 (ω, b)，定义超平面 (ω, b) 关于样本点 (x_i, y_i) 的几何间隔为：

$$\gamma_i = y_i \left(\frac{\omega}{||\omega||} \cdot x_i + \frac{b}{||\omega||} \right) \tag{9}$$

定义超平面 (ω, b) 关于训练数据集 T 的几何间隔为超平面 (ω, b) 关于 T 中所有样本点 (x_i, y_i) 的几何间隔最小值，即：

$$\gamma = \min_{i=1,\dots,N} \gamma_i \tag{10}$$

2.2.3 关系

由定义可知，函数间隔与几何间隔关系如下：

$$\gamma_i = \frac{\bar{\gamma_i}}{||\omega||} \tag{11}$$

$$\gamma = \frac{\bar{\gamma}}{||\omega||} \tag{12}$$

2.3 间隔最大化

对于线性可分训练数据集，分离超平面有无数个，我们的想法是求出分离超平面关于训练数据集的几何间隔，使其取最大值，以此来得到唯一的分离超平面。这里的间隔最大化又被称为硬间隔最大化。对此处理方法的直观解释：对训练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。也就是说，不仅将正负实例点分开，而且对最难分的实例点（离分离超平面最近的点）也有足够大的确定度将它们分开，这样的超平面应该对未知的新实例有很好的分类预测能力。

2.3.1 最大间隔分离超平面

对于我们的想法，用数学语言表达就是：

$$\max_{\omega,b} \quad \gamma \tag{13}$$

$$s.t. \quad y_i\left(\frac{\omega}{||\omega||}\cdot x_i+\frac{b}{||\omega||}\right)\geq\gamma, i=1,2,\ldots,N \tag{14}$$

考虑到函数间隔与几何间隔关系，上式又可以写成，

$$\max_{\omega,b} \quad \bar{\gamma} \tag{15}$$

$$s.t. \quad y_i(\omega\cdot x_i+b)\geq\bar{\gamma}, i=1,2,\ldots,N \tag{16}$$

可以看出，函数间隔的取值 $\bar{\gamma}$ 并不影响最优化问题的解[2]。因此，为了方便计算，我们可以取 $\bar{\gamma}=1$ ，并将其带入上式，同时，考虑到最大化 $\frac{1}{||\omega||}$ 与最小化 $\frac{1}{2}||\omega||^2$ 是等价的[3]，于是，上述最优化问题转换为下面的线性可分向量机学习的最优化问题：

$$\min_{\omega,b} \quad \frac{1}{2}||\omega||^2 \tag{17}$$

$$s.t. \quad y_i(\omega\cdot x_i+b)-1\geq 0, i=1,2,\ldots,N \tag{18}$$

[2]事实上，如果考虑到拉格朗日乘子法的时候，这一点可以更加明显地表现出来。

[3]最大化 $\frac{1}{|x|}$ 等价于最小化 $|x|$ 等价于最小化 $\frac{1}{2}|x|^2$ 。

上式是一个凸二次规划 (convex quadratic programming) 问题。如果求出了约束最优化问题 17 的解 ω^*, b^*，就可以得到最大间隔分离超平面 $\omega^* \cdot x + b^*$ 及分类决策函数 $f(x) = sign(\omega^* \cdot x + b^*$，即线性可分支持向量机模型。

综上，我们可以得到下面的线性可分支持向量机的学习算法--最大间隔法 (maximum margin method):

算法 1：线性可分支持向量机学习算法--最大间隔法

输入：线性可分训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$，其中，$x_i \in \mathcal{X} = R^n, y_i \in \mathcal{Y} = +1, -1, i = 1, 2, \ldots, N$；

输出：最大间隔分离超平面和分类决策函数。

1. 构造并求解约束最优化问题：

$$\begin{aligned} \min_{\omega,b} \quad & \frac{1}{2}||\omega||^2 \\ s.t. \quad & y_i(\omega \cdot x_i + b) - 1 \geq 0, i = 1, 2, \ldots, N \end{aligned} \tag{19}$$

求得最优解 ω^*, b^*

2. 由此得到分离超平面

$$\omega^* \cdot x + b^* = 0 \tag{20}$$

和分类决策函数

$$f(x) = sign(\omega^* \cdot x + b^*) \tag{21}$$

3. 可以证明，最大间隔分离超平面存在且唯一

2.3.2 支持向量与间隔边界

1. 支持向量

在线性可分情况下，训练数据集的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。即使得最优化条件中不等式等号成立的点，

$$y_i(\omega \cdot x_i + b) - 1 = 0 \tag{22}$$

对于 $y_i = +1$ 的正例点，支持向量在超平面 $H1 : \omega \cdot x + b = 1$ ；对于 $y_i = -1$ 的负例点，支持向量在超平面 $H2 : \omega \cdot x + b = -1$ 。

2. 间隔边界

如下图所示，粉色线为 $H1$ ，蓝色线为 $H2$ ，两者之间的距离，称为间隔。分离超平面位于两者中央与两者平行。间隔依赖于分离超平面的法向量 ω ，等于 $\frac{2}{||\omega||}$ ，$H1, H2$ 称为间隔边界。

3. 说明

在决定分离超平面时，只有支持向量起作用，其他实例点并不起作用，移动或添加其他实例点并不影响我们的求解。由于支持向量再确定分离超平面中起着决定性作用，所以将这类分类模型称为支持向量机。支持向量的个数很少，所以支持向量机由很少的“重要的”训练样本决定。

4. 举例

已知训练数据集，其正例点为 $x_1 = (3, 3)^T$ ，$x_2 = (4, 3)^T$ ，负例点为 $x_3 = (1, 1)^T$ ，试求最大间隔分离超平面。

解：

(a) 构造数据集约束最优化问题：

$$\begin{aligned} \min_{\omega,b}: \quad & \frac{1}{2}(\omega_1^2 + \omega_2^2) \\ s.t. \quad & 3\omega_1 + 3\omega_2 + b \geq 1 \\ & 4\omega_1 + 3\omega_2 + b \geq 1 \\ & -\omega_1 - \omega_2 - b \geq 1 \end{aligned} \tag{23}$$

(b) 解最优化问题，

实际上，如果在坐标轴上将三个点画出，可以很容易找到最优解为穿过点 $(0, 4)$ 和 $(4, 0)$ 的直线，利用两个支持向量所在超平面对应的等式 $3\omega_1 + 3\omega_2 + b = 1$ 和 $-\omega_1 - \omega_2 - b = 1$ 可以求出 $b = -2$ ，直线的已知，可以很容易看出 $\omega_1 = \omega_2 = \frac{1}{2}$ 。于是，我

们最终得到最大间隔分离超平面为：

$$\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0 \tag{24}$$

其中，$x_1 = (3,3)^T$ 与 $x_2 = (1,1)^T$ 为支持向量。

2.3.3 学习的对偶算法

1. 拉格朗日对偶性

(a) 原始问题

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

$$\begin{aligned} \min_{x\in \mathrm{R}^n} \quad & f(x) \\ s.t. \quad & c_i(x) \le 0, i = 1,2,\ldots,k \\ & h_j(x) = 0, j = 1,2,\ldots,l \end{aligned} \tag{25}$$

称此问题为原始最优化问题或者原始问题。

引入广义拉格朗日函数 (generalized Lagrange function)

$$L(x,\alpha,\beta) = f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x) \tag{26}$$

这里，$x = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})^T \in \mathrm{R}^n$，$\alpha_i, \beta_j$ 是拉格朗日乘子，$\alpha_i \ge 0$。考虑 x 的函数：

$$\theta_P(x) = \max_{\alpha,\beta:\alpha_i\ge 0} L(x,\alpha,\beta) \tag{27}$$

这里下标 P 表示原始问题。对于某个给定的 x，如果 x 违反原始问题的约束条件，即存在某个 i 使得 $c_i(\omega) > 0$ 或存在某个 j 使得 $h_j(\omega) \ne 0$，那么就有：

$$\theta_P(x) = \max_{\alpha,\beta:\alpha_i\ge 0} \left[f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x) \right] = +\infty \tag{28}$$

因为若某个 i 使约束 $c_i(x) > 0$ ，则可令 $\alpha_i \sim +\infty$ ；若某个 j 使得 $h_j(x) \neq 0$ ，总可令 $\beta_j h_j(x) \sim +infty$ ，而将其他 α_i, β_j 取为 0。因此，如果 x 满足约束条件，那么就有：

$$\theta_p(x) = \begin{cases} f(x) & x\text{满足原始问题约束} \\ +\infty & \text{其他} \end{cases} \tag{29}$$

于是，我们如果考虑极小化问题

$$\min_x \theta_P(x) = \min_x \max_{\alpha,\beta:\alpha\geq 0} L(x,\alpha,\beta) \tag{30}$$

它是与原始问题等价的问题，这样，我们就可以通过求解上式广义拉格朗日的极小极大问题来求解原始约束最优化问题的解。

(b) 对偶问题

定义 $\theta_D(x) = \min_x L(x,\alpha,\beta)$ ，然后考虑极大化 $\theta_D(x)$ ，即：

$$\max_{\alpha,\beta:\alpha_i\geq 0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0} \min_x L(x,\alpha,\beta) \tag{31}$$

问题 $\max_{\alpha,\beta:\alpha_i\geq 0} \min_x L(x,\alpha,\beta)$ 称为广义拉格朗日函数的极大极小问题。

可以将广义拉格朗日函数的极大极小问题表示为约束最优化问题：

$$\begin{aligned} \max_{\alpha,\beta} \theta_D(\alpha,\beta) &= \max_{\alpha,\beta} \min_x L(x,\alpha,\beta) \\ s.t. \quad & \alpha_i \geq 0, i = 1,2,\ldots,k \end{aligned} \tag{32}$$

称为原始问题的对偶问题，定义对偶问题的最优值

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \theta_D(\alpha,\beta) \tag{33}$$

称为对偶问题的值。

(c) 原始问题与对偶问题的关系

定理 1：若原始问题和对偶问题都有最优值，则

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \min_x L(x,\alpha,\beta) \leq \min_x \max_{\alpha,\beta:\alpha_i\geq 0} L(x,\alpha,\beta) = p^* \tag{34}$$

推论 1：设 x^* 和 α^*, β^* 分别是原始问题和对偶问题的可行解，而且有 $d^* = p^*$ ，则 x^* 和 α^*, β^* 分别是原始问题和对偶问题的最优解。

定理 2：考虑原始问题与对偶问题，假设 $f(x)$ 和 $c_i(x)$ 是凸函数[4]，$h_j(x)$ 是仿射函数[5]；并且假设不等式约束 $c_i(x)$ 是严格可行的，即存在 x ，对所有 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，而且

$$p^* = d^* = L(x^*, \alpha^*, \beta^*) \tag{35}$$

定理 3：对原始问题和对偶问题，假设 $f(x)$ 和 $c_i(x)$ 是凸函数，$h_j(x)$ 是仿射函数，并且不等式约束 $c_i(x)$ 是严格可行的，则 x^*, α^*, β^* 分别是原始问题和对偶问题的解的充分必要条件是 x^*, α^*, β^* 必须满足下面的 KKT 条件：

$$\begin{aligned}
\nabla_x L(x^*, \alpha^*, \beta^*) &= 0 \\
\nabla_\alpha L(x^*, \alpha^*, \beta^*) &= 0 \\
\nabla_\beta L(x^*, \alpha^*, \beta^*) &= 0 \\
\alpha^* c_i(x^*) = 0, &\quad i = 1, 2, \ldots, k \\
c_i(x^*) \le 0, &\quad i = 1, 2, \ldots, k \\
\alpha_i^* \ge 0, &\quad i = 1, 2, \ldots, k \\
h_j(x^*) = 0, &\quad i = 1, 2, \ldots, l
\end{aligned} \tag{36}$$

2. 利用对偶问题求解原始问题的最优解

(a) 构建拉格朗日函数

对原始问题 (式 19) 中每个不等式约束引入拉格朗日乘子 (Lagrange multiplier) $\alpha_i \ge 0, i = 1, 2, \ldots, N$ ，定义拉格朗日函数：

[4]简单从理解上讲，$f''(x) \ge 0$ 对应凸函数，$f''(x) < 0$ 对应凹函数。

[5]简单而言，仿射函数是指一阶多项式组成的函数，譬如 $y = ax + b$ 。

$$L(\omega, b, a) = \frac{1}{2}||\omega||^2 - \sum_{i=1}^{N} \alpha_i y_i (\omega \cdot x_i + b) + \sum_{i=1}^{N} \alpha_i \tag{37}$$

其中，$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ 为拉格朗日乘子向量。

(b) 原始问题的对偶问题由之前讨论可知，与原始问题等价的极小极大问题是：

$$\min_{\omega,b} \max_{\alpha} L(\omega, b, \alpha) \tag{38}$$

与之对偶的极大极小问题是：

$$\max_{\alpha} \min_{\omega,b} L(\omega, b, \alpha) \tag{39}$$

考虑到 $\frac{1}{2}||\omega||^2$ 和 $-y_i(\omega \cdot x_i + b) + 1$ 均为凸函数，满足定理 2 的条件，我们知道，原始问题与对偶问题的解是存在的。下面对原始问题的对偶问题进行转换：

i. 求 $\min_{\omega,b} L(\omega, b, \alpha)$

$$\begin{aligned} \nabla_\omega L(\omega, b, \alpha) &= \omega - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \\ \nabla_b L(\omega, b, \alpha) &= \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned}$$

有：

$$\begin{aligned} \omega &= \sum_{i=1}^{N} \alpha_i y_i x_i \\ \sum_{i=1}^{N} \alpha_i y_i &= 0 \end{aligned} \tag{40}$$

将上式代入拉格朗日函数 37 可以得到：

$$\begin{aligned} L(\omega, b, \alpha) &= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i \end{aligned}$$

即：

$$\min_{\omega,b,\alpha} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \tag{41}$$

ii. 求 $\min_{\omega,b} L(\omega, b, \alpha)$ 对 α 的极大，即是对偶问题

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \\ s.t. \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & \alpha_i \geq 0, i = 1, 2, \ldots, N \end{aligned} \tag{42}$$

将上式问题转为取极小，可以得到下面的等价的最优化问题：

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \\ s.t. \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & \alpha_i \geq 0, i = 1, 2, \ldots, N \end{aligned} \tag{43}$$

于是，设 $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*)^T$ 是对偶最优化问题 36 的解，由 α^* 可以求出原始最优化问题 17 的解 ω^* 和 b^*，并且有如下定理：定理 4[6]：设 α^* 是对偶最优化问题的解，则存在下标 j，使得 $\alpha_j^* > 0$，并可按下式求得原始最优化问题的解 ω^*, b^*：

$$\begin{aligned} \omega^* &= \sum_{i=1}^{N}\alpha_i^* y_i x_i \\ b^* &= y_j - \sum_{i=1}^{N}\alpha^* y_i (x_i \cdot x_j) \end{aligned} \tag{44}$$

3. 总结

算法 2：线性可分支持向量机学习算法

[6]利用 KKT 条件，$\nabla_\omega L(\omega^*, b^*, \alpha^*) = \nabla_b L = 0$ 可以求出 ω^*，代入 $\alpha_i^*(y_i(\omega^* \cdot x_i + b^*) - 1 = 0$ 可以求出 b^*。

输入：线性可分训练集 $T=(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$，其中 $x_i\in\mathcal{X}=R^n, y_i\in\mathcal{Y}=+1,-1, i=1,2,\ldots,N$ ；

输出：分离超平面和分类决策函数

(a) 构造并求解约束最优化问题：

$$\begin{aligned}\min_{\alpha}\quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i\\ s.t.\quad & \sum_{i=1}^{N}\alpha_i y_i=0\\ & \alpha_i\geq 0, i=1,2,\ldots,N\end{aligned}$$

求解最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\ldots,\alpha_N^*)^T$

(b) 计算

$$\omega^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i$$

并选择 α^* 的一个正分量 $\alpha_j^*>0$ ，计算

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

(c) 求得分离超平面

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

分离决策函数：

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

(d) 定义 (支持向量)：从上式可以看出，ω^* 和 b^* 只依赖于训练数据中 $\alpha_i^*>0$ 的点，我们将训练数据中对应 $\alpha_i^*>0$ 的实例点 ($\alpha_i^*\in \mathrm{R}^n$) 称为支持向量。

根据这一定义，支持向量一定位于间隔边界上。因为由 KKT 互补条件：

$$\alpha_i^*(y_i(\omega^*\cdot x_i+b^*)-1)=0, i=1,2,\ldots,N$$

可知，若 $\alpha_i^* > 0$ ，则对应实例 x_i 有：

$$y_i(\omega^* \cdot x_i + b) - 1 = 0$$

或

$$\omega^* \cdot x_i + b = \pm 1$$

4. 举例已知数据集正例点为 $x_1 = (3,3)^T$, $x_2 = (4,3)^T$ ，负例点为 $x_3 = (1,1)^T$ ，试求最大间隔分离超平面。

   解：1) 构造数据集约束最优化问题：

$$\begin{aligned}
\min_{\alpha}: \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \\
& = \frac{1}{2}18\alpha_1^2 + 25\alpha_2^2 + 2\alpha_3^2 + 42\alpha_1\alpha_2 - 12\alpha_1\alpha_3 - 14\alpha_2\alpha_3) - \alpha_1 - \alpha_2 - \alpha_3 \\
s.t. \quad & \alpha_1 + \alpha_2 - \alpha_3 = 0 \\
& \alpha_i \geq 0, i = 1,2,3
\end{aligned}$$

   2) 解最优化问题，将 $\alpha_3 = \alpha_1 + \alpha 2$ 带入目标函数并记为

$$s(\alpha_1, \alpha_2) = 4\alpha_1^2 + \frac{13}{2}\alpha_2^2 + 10\alpha_1\alpha_2 - 2\alpha_1 - 2\alpha_2 \tag{45}$$

   对 α_1, α_2 分别求偏导并令其为 0，可以知道 $s(\alpha_1, \alpha_2)$ 在点 $(\frac{3}{2}, -1)^T$ 取极值，但是该点的 α_2 不满足约束 $\alpha_i \geq 0$ ，所以最小值应在边界上达到：

   - 当 $\alpha_1 = 0$ 时，最小值 $s(0, \frac{2}{13}) = -\frac{2}{13}$ ；当 $\alpha_2 = 0$ 时，相应的 $s(\frac{1}{4}, 0) = -\frac{1}{4}$ ，于是，在 $\alpha_1 = \frac{1}{4}, \alpha_2 = 0$ 时，s 达到最小值，此时 $\alpha_3 = \frac{1}{4}$
   - $\alpha_1^* = \alpha_3^* = \frac{1}{4}$ 对应的实例点 x_1, x_3 为支持向量，此时

$$\omega_1^* = \omega_2^* = \frac{1}{2}b^* = -2$$

     于是，分离超平面为：

$$\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$$

分类决策函数为

$$f(x) = \frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2$$

3 线性支持向量机与软间隔最大化

3.1 线性支持向量机

通常情况下，训练数据不是简单的可分数据集，而是有一些特异点 (outlier)，当将这些特异点除去后，剩下的大部分样本点组成的集合是线性可分的。

对类似训练样本的每个样本点 (x_i, y_i) 引入一个松弛变量 $\xi_i \geq 0$，使函数间隔加上松弛变量大于等于 1。这样，约束条件变为：

$$y_i(\omega \cdot x_i + b) \geq 1 - \xi_i \tag{46}$$

同时，对每个松弛变量 ξ_i，支付一个代价 ξ_i，目标函数由原来的 $\frac{1}{2}||\omega||^2$ 变为

$$\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\xi_i \tag{47}$$

这里，$C > 0$ 称为惩罚参数，一般由应用问题决定，C 值大的时候对误分类的惩罚增大，反之则减小。最小化目标函数包含两层意思：使 $\frac{1}{2}||\omega||^2$ 尽可能小以使最大间隔尽可能大，同时使得误分类点的个数尽可能少，C 是调和两者的系数。

线性不可分的线性支持向量机的学习问题现在变成了凸二次规划问题：

$$\begin{aligned} \min_{\omega,b,\xi} \quad & \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\xi_i \\ s.t. \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1, 2, \ldots, N \\ & \xi_i \geq 0, i = 1, 2, \ldots, N \end{aligned} \tag{48}$$

可以证明以上问题的解是存在的，而且 ω 的解是唯一的，b 的解不唯一，而是存在于一个区间内。于是，有：

定义 (线性支持向量机) ：对于给定的线性不可分的训练数据集，通过求解凸二次规划问题，即软间隔最大化问题 48 ，得到的分离超平面为：

$$\omega^* \cdot x + b^* = 0 \tag{49}$$

以及相应的分类决策函数

$$f(x) = sign(\omega^* \cdot x + b^*) \tag{50}$$

称为线性支持向量机。

3.2 学习的对偶算法

原始最优化问题的拉格朗日函数为：

$$\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(w \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i \tag{51}$$

其中， $\alpha_i \geq 0, \mu_i \geq 0$ 。

利用 KKT 条件，即 $\nabla_\omega L(\omega, b, \xi, \alpha, \mu) = \nabla_b L = \nabla_\xi L = 0$ 和不等式 $\alpha_i \geq 0, \mu_i \geq 0$ 可以推导出原始问题的对偶问题：

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \\ s.t. \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, N \end{aligned}$$

定理 5：设 α^* 是对偶最优化问题的解, 若存在下标 j , 使得 $0 < \alpha_j^* < C$ ，并可按下式求得原始最优化问题的解 ω^*, b^* ：

$$\begin{aligned} \omega^* &= \sum_{i=1}^{N}\alpha_i^* y_i x_i \\ b^* &= y_j - \sum_{i=1}^{N}\alpha^* y_i (x_i \cdot x_j) \end{aligned} \tag{52}$$

由此，有算法总结如下：

算法 3：线性支持向量机学习算法

输入：线性训练集 $T=(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$，其中 $x_i\in\mathcal{X}=R^n, y_i\in\mathcal{Y}=+1,-1, i=1,2,\ldots,N$ ；

输出：分离超平面和分类决策函数

1. 选择惩罚参数 $C>0$，构造并求解凸二次规划问题：

$$\begin{aligned}\min_{\alpha}\quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i\\ s.t.\quad & \sum_{i=1}^{N}\alpha_i y_i=0\\ & 0\le\alpha_i\ge C, i=1,2,\ldots,N\end{aligned}$$

求解最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\ldots,\alpha_N^*)^T$

2. 计算 $\omega^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i$，选择 α^* 的一个分量 α_j^* 适合条件 $0<\alpha_j^*<C$，计算：

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

3. 求得分离超平面

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

分离决策函数：

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

4. b^* 取值的说明对于任一适合条件 $0<\alpha_j^*<C$ 的 α_j^*，都可以用来求解 b^*，因此，实际计算的适合，可以取所有符合条件的样本点上的平均。

3.3 支持向量

在线性不可分情况下，解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\ldots,\alpha_N^*)^T$ 中对应于 $\alpha_i^*>0$ 的样本点 (x_i,y_i) 的实例 x_i 称为支持向量 (KKT 条件中的互补松弛条件)。如下图所示，分离超平面由实线表示，间隔边界由虚线表示，正例点由"$\circ$"表

示，负例点由" x " 表示，图中还标出了实例 x_i 到间隔边界的距离 $\frac{1-\xi_i}{||\omega||}$ (图片标的有问题)

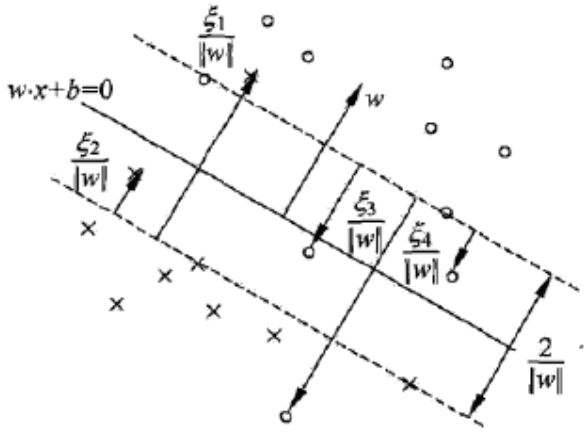


图 3: 软间隔支持向量

软间隔支持向量 x_i 或者在间隔边界上，或者在间隔边界与分离超平面之间，或者在分离超平面误分的一侧，如果 $\alpha^* < C$ (对应 $\mu > 0$ ，考虑到互补松弛条件 $\mu_i^* \xi_i^* = 0$)，那么 $\xi_i = 0$ ；若 $\alpha_i^* = C$ ， $0 < \xi_i < 1$ ， 则分类正确，x_i 在间隔边界与分离超平面之间；若 $\alpha^* = C$ ， $\xi_i > 1$ ， 则 x_i 位于分离超平面误分一侧。

3.4 合页损失函数

线性支持向量机学习还有另外一种解释，就是最小化以下目标函数：

$$\sum_{i=1}^{N}[1 - y_i(\omega \cdot x_i + b)]_+ + \lambda||\omega||^2 \tag{53}$$

目标函数第一项是经验损失或经验风险。函数

$$L(y(\omega \cdot x + b)) = [1 - y_i(\omega \cdot x + b)]_+ \tag{54}$$

称为合页损失函数 (hinge loss function)。下标"+" 表示以下取正值的函数：

$$[z]_+ = \begin{cases} z & z > 0 \\ 0 & z \leq 0 \end{cases} \tag{55}$$

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) $y_i(\omega \cdot x_i + b) > 1$ 时，损失是 0，否则损失为 $1 - y_i(\omega \cdot x_i + b)$ 。

定理 6：线性支持向量机原始最优化问题：

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

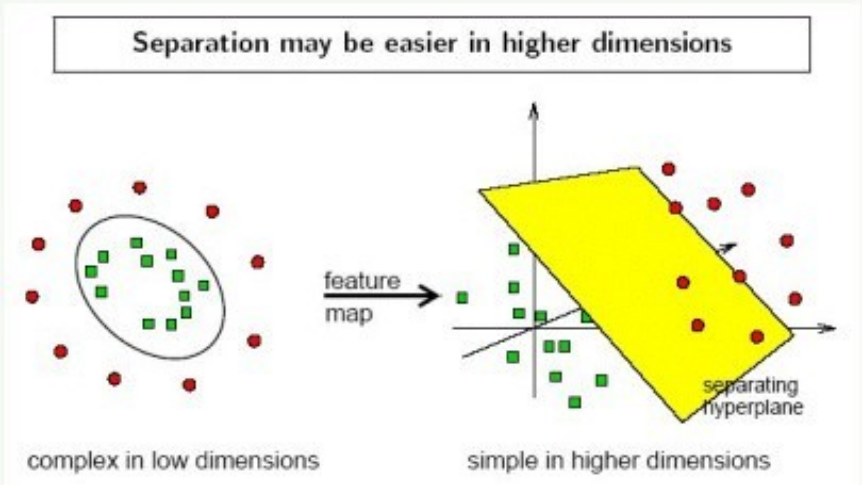
等价于最优化问题

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

4 核函数

4.1 核技巧

如下图所示，通过非线性变换，将非线性问题转变为线性问题，这样的方法属于核技巧。



设原空间为 $\mathcal{X} \subset \mathrm{R}^2, x = (x_1, x_2)^T \in \mathcal{X}$ ，新空间为 $\mathcal{Z} \subset \mathrm{R}^2, z = (z_1, z_2)^T \in \mathcal{Z}$ ，定义从原空间到新空间的变换：

$$z = \phi(x) = ((x^{(1)})^2, (x^{(2)})^2)^T$$

经过变换 $z=\phi(x)$，原空间 $\mathcal{X}\subset \mathrm{R}^2$ 变换为新空间 $\mathcal{Z}\subset \mathrm{R}^2$，原空间中的点相应地变换为新空间中的点，原空间中的椭圆

$$\omega_1(x^{(1)})^2+\omega_2(x^{(2)})^2+b=0$$

变换为新空间中的直线

$$\omega_1 z^{(1)}+\omega_2 z^{(2)}+b=0$$

在新变换空间中，直线 $\omega_1 z^{(1)}+\omega_2 z^{(2)}+b=0$ 将变换后的正负例点正确分开。这样，原空间的非线性可分问题就转变为新空间中的线性可分问题。

总结上例，用线性方法求解非线性问题分为两步 (核技巧)：

- 使用一个变换将原空间中的数据映射到新空间
- 在新空间用线性可分学习方法从训练数据中学习分类模型

定义 (核函数)：设 $\mathcal{X}$ 是输入空间 (欧氏空间 R^n 的子集或离散集合)，又设 $\mathcal{H}$ 是特征空间 (希尔伯特空间)，如果存在一个从 $\mathcal{X}$ 到 $\mathcal{H}$ 的映射

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

使得对所有 $x,z\in\mathcal{X}$，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，式中 $\phi(x)\cdot\phi(z)$ 为 $\phi(x)$ 和 $\phi(z)$ 的内积。

在线性支持向量机中，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。在对偶问题的目标函数中，内积 $x_i\cdot x_j$ 可以用核函数 $K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$ 代表，此时，对偶问题的目标函数变为：

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

同样，分类觉得函数中的内积也可以用核函数代替，于是，分类决策函数变为

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

4.2 常用核函数:not fully understand:

1. 多项式核函数 (polynomial kernel function)

$$K(x,z)=(x\cdot z+1)^p \tag{56}$$

对应的支持向量机是一个 p 次多项式分类器，在此情形下，分类决策函数成为：

$$f(x)=sign(\sum_{i=1}^{N_S}\alpha_i^* y_i(x_i\cdot x)^p+b^*) \tag{57}$$

2. 高斯核函数 (Gaussian kernel function)

$$K(x,z)=exp\left(-\frac{||x-z||^2}{2\sigma^2}\right) \tag{58}$$

对应的支持向量机是高斯径向基函数 (radial basis function) 分类器。在此情形下，分类决策函数成为

$$f(x)=\sum_{i=1}^{N_s}sign\left(\alpha_i^* y_i exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)+b^*\right) \tag{59}$$

3. 字符串核函数

考虑一个有限字符表 Σ 。字符串 s 是从 Σ 中取出的有限个字符的序列，包括空字符串。字符串 s 的长度用 $|s|$ 表示，它的元素记做 $s(1)s(2)\ldots s(|s|)$ 。两个字符串的连接记做 st ，所有长度为 n 的字符串的集合记做 Σ^n ，所有字符串的集合记做 $\Sigma^*=\cup_{n=0}^{\infty}\Sigma^n$ 。

考虑字符串 s 的字串 u , 给定一个指标序列 $i=(i_1,i_2,\ldots,i_{|u}),1\le i_1<i_2<\ldots<i_{|u|}\le|s|$ ， s 的字串定义为 $u=s(i)=s(i_1)s(i_2)\ldots s(i_{|u|})$ ，其长度记做 $l(i)=i_{|u|}-i_1+1$ 。如果 i 是连续的，则 $l(i)=|u|$ ；否则，$l(i)>|u|$ 。

假设 $\mathcal{S}$ 是长度大于或等于 n 的字符串集合，s 是 $\mathcal{S}$ 的元素，现在建立字符串集合 S 到特征空间 $\mathcal{H}=\mathrm{R}^{\Sigma^n}$ 的映射 $\phi_n(s)$ 。R^{Σ^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 $\phi_n(s)$ 将字符串 s 对应于空间 R^{Σ^n} 的一个向量，其在 u 上的取值为：

$$[\phi_n(s)]_u = \sum_{i;s(i)=u} \lambda^{l(i)} \tag{60}$$

这里，$0 < \lambda \leq 1$ 表示一个衰减参数，$l(i)$ 表示字符串 i 的长度，求和在 s 中所有与 u 相同的字串上进行。

譬如，假设 Σ 是应为字符集，n 为 3，$\mathcal{S}$ 为长度大于等于 3 的字符串集合。考虑将字符集 $\mathcal{S}$ 映射到特征空间 $\mathcal{H}_3$ 。$\mathcal{H}_3$ 的一维对应字符串 asd 。这时，字符串“Nasdaq”与“lass das”在这一维上的值分别是 $[\phi_3(Nasdaq)]_{asd} = \lambda^3$ 和 $[\phi_3(lassdas)]_{asd} = 2\lambda^5$ ，在第一个字符串里，asd 是连续的字串，第二个字符串里，asd 是长度为 5 的不连续字串，共出现 2 次。两个字符串 s 和 t 上的字符串核函数是基于映射 ϕ_n 的特征空间中的内积：

$$k_n(s,t) = \sum_{u\in\Sigma^n} [\phi_n(s)]_u [\phi_n(t)]_u = \sum_{u\in\Sigma^n} \sum_{(i,j):\, s(i)=t(j)=u} \lambda^{l(i)}\lambda^{l(j)} \tag{61}$$

字符串核函数 $k_n(s,t)$ 给出了字符串 s 和 t 中长度等于 n 的所有字串组成的特征向量的余弦相似度 (cosine similarity) 。直观上，两个字符串相同的字串越多，他们越相似，字符串核函数的值就越大。字符串核函数可以由动态规划快速计算。