支持向量机

Kuangnan Fang

Email: xmufkn@xmu.edu.cn

支持向量机

支持向量机（support vector machine, SVM）是 Cortes 和 Vapnik 在 1995 年提出来的方法，由于它在分类任务（尤其是文本分类）中表现出卓越的性能，很快成为机器学习的主流方法，并掀起了“统计学习”（statistical learning）的高潮。本章将以最常用的二分类问题为例，先介绍支持向量机的原理。

支持向量机

根据支持向量机的发展历程以及模型繁简程度一般可分为三类：

- 最大间隔分类器（maximal margin classifier，线性可分支持向量机）
- 支持向量分类器（support vector classifier，线性支持向量机）
- 非线性支持向量机（nonlinear support vector machine）

线性可分支持向量机

超平面（*hyperplane*）

指p维线性空间中维度为$p-1$维的线性子空间。例如，二维空间的超平面是一条直线；三维空间的超平面是一个平面。图1给出的直线$1-2X_1-X_2$ 即为二维空间的一个超平面。

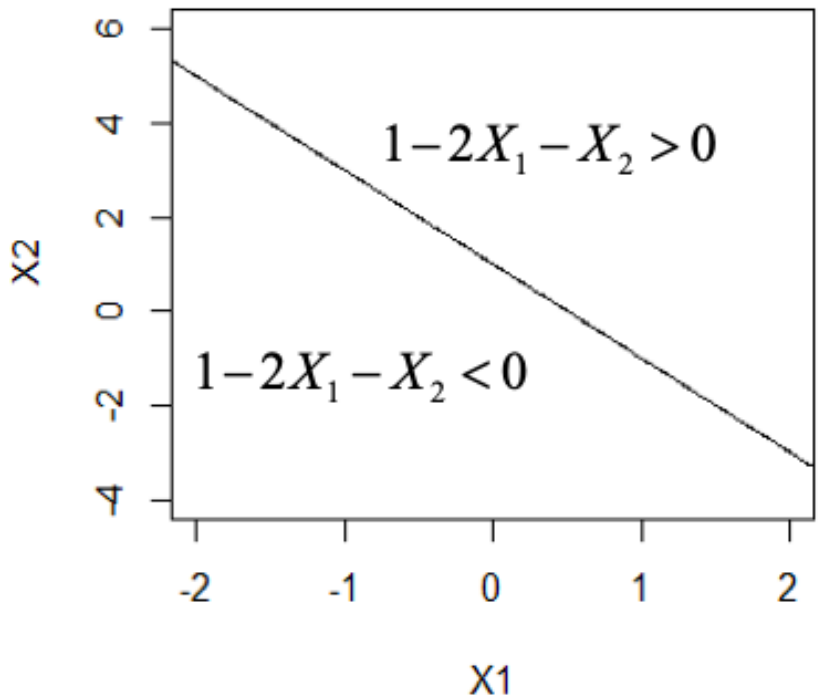


图 1: 超平面$1-2x_1-x_2=0$

线性可分支持向量机

- p维空间的超平面

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0 \tag{1}$$

- 任何满足式（1）的点 X 都会落在超平面上。那么，对于不满足式（1）的点 X，若

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p > 0$$

说明此时的 X 位于超平面的一侧。

- 若

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p < 0$$

则此时的X位于超平面的另一侧。因此，可以说超平面将空间分成了两个部分。

线性可分支持向量机

- 法向量（*normal vector*） $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$
- 超平面：$\beta_1 x_1 + \beta_2 x_2 - 6 = 0$；法向量：$(\beta_1 = 0.8, \beta_2 = 0.6)$
- 法向量的方向与超平面是正交的，沿着法向量方向移动超平面，不改变法向量β的值，只改变截距项β_0 的值。

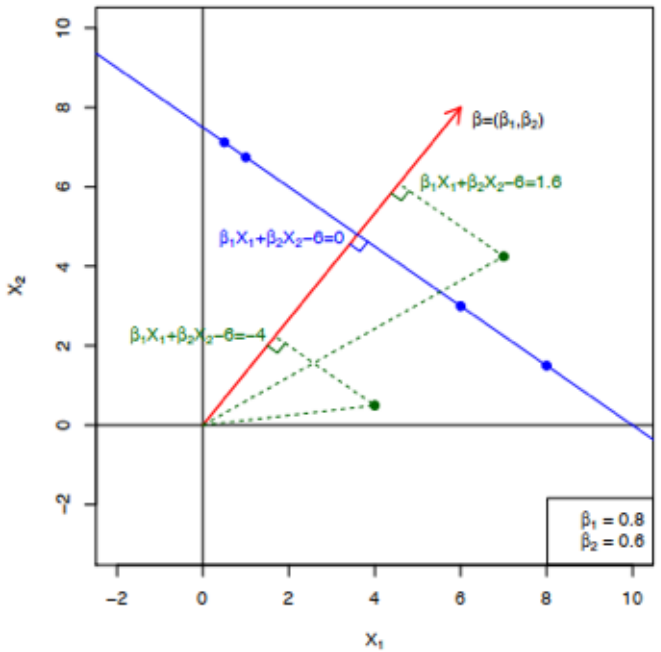


图 2:　法向量与超平面

线性可分支持向量机

- 假设给定一个训练数据集

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$$

其中，$x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\top$，$y_i \in \{-1, 1\}$，其中$-1$和1代表两个不同的类别，即负例和正例。

- 如果存在某个超平面$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$ 能够将数据集D 的正负实例点完全正确划分到超平面的两侧，则称数据集D 是线性可分数据集，反之，则称数据集D 是线性不可分数据集。

线性可分支持向量机

给定线性可分训练数据集，通过间隔最大化方法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

$$\beta_0 + \beta^\top x = 0$$

以及相应的分类决策函数

$$c(x) = sign(\beta_0 + \beta^\top x)$$

称之为 **线性可分支持向量机**。

线性可分支持向量机

对于线性可分的情形，假设我们可以构造一个超平面把两个类别的观测点完全分割开来：

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \beta_0 + \beta^\top x = 0$$

那么这个超平面应该满足：

$$\begin{aligned} f(x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \beta_0 + \beta^T x_i > 0, \text{ 若 } y_i = 1 \\ f(x_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \beta_0 + \beta^T x_i < 0, \text{ 若 } y_i = -1 \end{aligned} \tag{2}$$

线性可分支持向量机

例1 假设有10个观测点，它们分属于两个类别，其中观测点1-5是一类，观测点6-10是另一类。现在想建立一个分类器，使得对给定的任意一个新的观测点，都能将它正确分类。那么，如何利用支持向量机建立这样的分类器呢？

表 1: 两个类别的10个观测点数据

Obs	1	2	3	4	5	6	7	8	9	10
x_1	0.5	1	1.5	1	2.5	2.5	3	3	4	4
x_2	3	2.5	3.5	2	3.8	1	2	1.5	3	1
y	-1	-1	-1	-1	-1	1	1	1	1	1

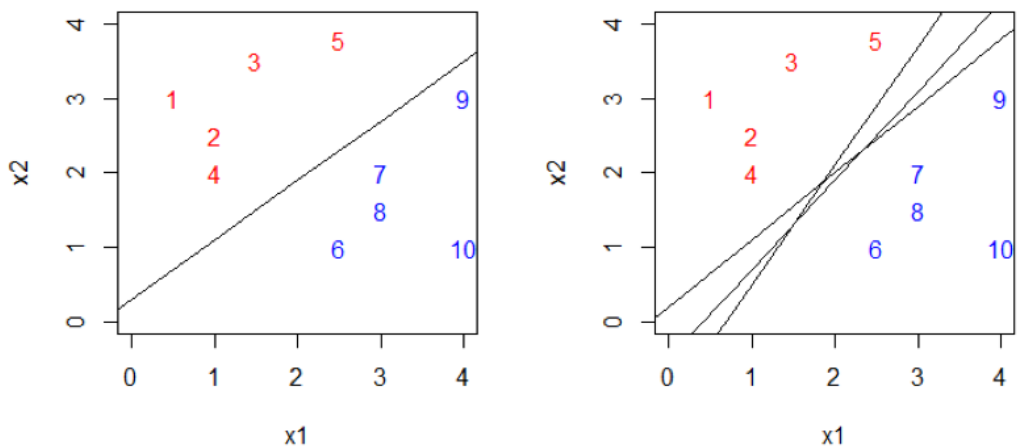


图 3: 左：例1观测点的分割超平面；右：例1观测点可能的多个分割超平面

线性可分支持向量机

函数间隔

对于给定的训练集D和超平面$f(x) = \beta_0 + \beta^\top x = 0$，定义超平面与样本点$(x_i, y_i)$的函数间隔为

$$M_i = y_i(\beta_0 + \beta^\top x_i)$$

定义超平面关于训练集D的函数间隔为

$$M = \min_{i=1,\cdots n} M_i$$

线性可分支持向量机

几何间隔

对于给定的训练集D和超平面$\beta_0 + \beta^\top x = 0$，定义超平面与样本点$(x_i, y_i)$的几何间隔为

$$\tilde{M}_i = \frac{y_i}{\|\beta\|}\left(\beta_0 + \beta^T x_i\right)$$

定义超平面关于训练集D的几何间隔为

$$\tilde{M} = \min_{i=1,\cdots n} \tilde{M}_i$$

几何间隔与函数间隔之间的关系

$$\tilde{M}_i = \frac{M_i}{\|\beta\|} \tag{3}$$

$$\tilde{M} = \frac{M}{\|\beta\|} \tag{4}$$

函数间隔可以衡量分类的正确性和可信度，但是没法根据函数间隔确定分割超平面，如果超平面参数β_0和β都乘以常数c，超平面并未改变，几何间隔不变，但是函数间隔是原来的c倍

线性可分支持向量机

一般来说，若对于给定的训练集，我们可以构造某个超平面将它们分割开来，那么将这个超平面上移或下移或旋转，只要不碰到原有的那些观测点，那么我们就能得到另外的超平面。对于线性可分的训练集，理论上我们可以找到无穷多个超平面。

哪一个超平面才是“最合适”的呢？

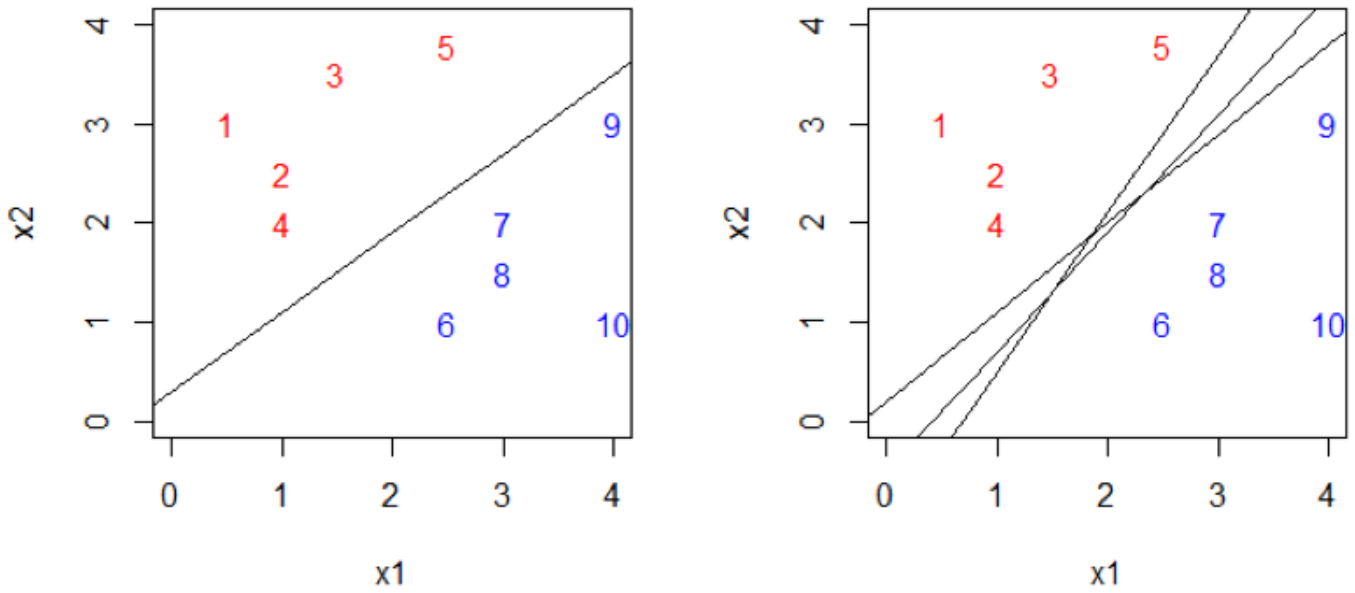


图 4: 左：例1观测点的分割超平面；右：例1观测点可能的多个分割超平面

线性可分支持向量机

- 若观测点离超平面距离越远，则对于该观测点的判断会更加有信心，这也表明了，间隔实际上是代表了误差的上限
- 选择与这些观测点具有最大间隔的超平面作为分类器，称它为**最大间隔分类器（*maximal margin classifier*）**
- 构建一个最大间隔分类器，求解几何间隔最大的超平面

$$\max_{\beta_0,\beta} \quad \tilde{M}$$

$$\text{s.t.} \quad y_i \frac{f(x_i)}{\|\beta\|} = y_i \left(\frac{\beta_0}{\|\beta\|} + \frac{\beta^T}{\|\beta\|} x_i \right) \geq \tilde{M}, i = 1, 2, \cdots, n \quad (5)$$

线性可分支持向量机

- 根据几何间隔与函数间隔的关系式（4），将式（5）改写为：

$$\max_{\beta_0,\beta} \quad \frac{M}{\|\beta\|}$$

$$\text{s.t.} \qquad y_i f(x_i) = y_i(\beta_0 + \beta^T x_i) \geq M, \quad i = 1, 2, \cdots, n \qquad (6)$$

- 或者改写为：

$$\max_{\beta_0,\beta} \quad M$$

$$\text{s.t.} \qquad \|\beta\| = \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i f(x_i) = y_i \left(\beta_0 + \beta^T x_i\right) \geq M, \quad i = 1, 2, \cdots, n \qquad (7)$$

线性可分支持向量机

- 对于式(7)的优化问题，函数间隔实际上是不唯一的，因此第一个约束条件是令$\|\beta\| = 1$，实际上是保证了求解上述最优化问题时能得到参数的唯一解。
- 对于式（6），函数间隔M的取值并不影响最优化问题的解，即不妨令$M = 1$，且最大化$\frac{1}{||\beta||}$等价于最小化$2||\beta||^2$

$$\min_{\beta_0,\beta} \quad \frac{1}{2}\|\beta\|^2$$

$$\text{s.t.} \quad y_i\left(\beta_0 + \beta^T x_i\right) \geq 1, i = 1, 2, \cdots, n \tag{8}$$

- 凸规划（*Convex quadratic programming*）问题，由于所有观测点线性可分，故可行域非空，该问题有解。

线性可分支持向量机

算法*1* 线性可分支持向量机的最大间隔算法

- 输入：线性可分训练集$D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，其中$x_i$是特征向量，类别标签$y_i\in\{-1,1\}$，$i=1,2,\cdots,n$
- 输出：最大间隔分割超平面和分类决策函数
  - （1）求解约束最优化问题：

$$\min_{\beta_0,\beta}\frac{1}{2}\|\beta\|^2$$

$$\text{s.t. } y_i\left(\beta_0+\beta^{\top}x_i\right)\geq 1, i=1,2,\cdots,n$$

解得最优解$\hat{\beta}_0,\hat{\beta}$

  - （2）求得分割超平面

$$f(x)=\hat{\beta}_0+\hat{\beta}^{\top}x=0$$

和分类决策函数

$$c(x)=\text{sign}(f(x))=\text{sign}(\hat{\beta}_0+\hat{\beta}^{\top}x)$$

线性可分支持向量机

- 在线性可分情形下，训练集中与分割超平面距离最近的样本点的观测（observations）称为 **支持向量（*support vector*）**
- 支持向量是使式（8）约束条件等号成立的点，即

$$y_i(\beta_0 + x_i^\top \beta) - 1 = 0$$

- $B_1 : y_i(\beta_0 + x_i^\top \beta) = 1 \qquad B_2 : y_i(\beta_0 + x_i^\top \beta) = -1$

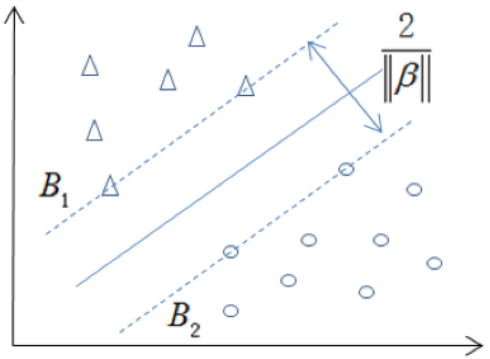


图 5: 支持向量

线性可分支持向量机

- 式（8）是一个凸二次规划问题，对该问题直接求解会比较复杂，往往将该原问题（*primal problem*）转化为对偶问题*(dual problem)*来求解。
- 首先构造原问题的拉格朗日函数（*lagrange function*）

$$L(\beta, \beta_0, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\beta^\top x_i + \beta_0)) + \sum_{i=1}^{n} \alpha_i \quad (9)$$

其中，$\alpha_i \geq 0, i = 1, \cdots, n$ 为 拉格朗日乘子 *(lagrange multiplier)*，$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^T$ 为拉格朗日乘子向量

- 根据朗格朗日对偶性，原问题的对偶问题是极大极小问题：

$$\max_{\alpha} \min_{\beta, \beta_0} L(\beta, \beta_0, \alpha)$$

线性可分支持向量机

解:

(1)求 $\min_{\beta,\beta_0} L(\beta,\beta_0,\alpha)$

将拉格朗日函数 $L(\beta,\beta_0,\alpha)$ 分别对 β 和 β_0 求偏导并令其等于 0:

$$\begin{cases} \frac{\partial L(\beta,\beta_0,\alpha)}{\partial \beta} = \beta - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \\ \frac{\partial L(\beta,\beta_0,\alpha)}{\partial \beta_0} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}$$

可得

$$\begin{cases} \beta = \sum_{i=1}^{N} \alpha_i y_i x_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases} \tag{11}$$

将（11）代入（9）可得

$$\min_{\beta,\beta_0} L(\beta,\beta_0,\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{n} \alpha_i$$

线性可分支持向量机

解:

(2)求 $\max_{\alpha} \min_{\beta,\beta_0} L(\beta, \beta_0, \alpha)$, 即

$$\begin{aligned} &\max -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{n} \alpha_i \\ &s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \\ &\qquad \alpha_i \geq 0, i = 1, 2, \cdots, n \end{aligned} \tag{12}$$

将式（12）的目标函数最大化问题转为最小问题即可得到式（10）。

设式（10）最优化问题的解为 $\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_n)^{\top}$, 可以由$\hat{\alpha}$ 求解出原问题（9)的解 $\hat{\beta}$ 和 $\hat{\beta}_0$

线性可分支持向量机

定理1设 $\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_n)^\top$ 为对偶问题 (10) 的解, 则存在下标 j, 使得 $\hat{\alpha}_j > 0$,按下式求得原问题（9）的解：

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \tag{13}$$

$$\hat{\beta}_0 = y_j - \sum_{i=1}^{n} \hat{\alpha}_i y_i (x_i \cdot x_j) \tag{14}$$

线性可分支持向量机的对偶算法

算法2 线性可分支持向量机的最大间隔算法

- 输入：线性可分训练集$D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，其中$x_i$是特征向量，类别标签$y_i\in\{-1,1\}$，$i=1,2,\cdots,n$
- 输出：最大间隔分割超平面和分类决策函数
  - （1）求解对偶问题的最优化：

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

    解得最优解$\hat{\alpha}=(\hat{\alpha}_1,\cdots,\hat{\alpha}_n)^{\top}$
  - （2）求原问题的解

$$\hat{\beta}=\sum_{i=1}^{N}\hat{\alpha}_i y_i x_i$$
$$\hat{\beta}_0=y_j-\sum_{i=1}^{N}\hat{\alpha}_i y_i(x_i\cdot x_j)$$

  - （3）求得分割超平面

$$f(x)=\hat{\beta}_0+\hat{\beta}^T x=0$$

    和分类决策函数

$$c(x)=\text{sign}(f(x))=\text{sign}(\hat{\beta}_0+\hat{\beta}^T x)$$

支持向量

- 由式（13）和式（14）可以发现$\hat{\beta}$和$\hat{\beta}_0$ 的解只依赖于$\hat{\alpha}_i > 0$所对应的样本点(x_i, y_i)，而与其他样本点无关。我们将训练数据集中对应$\hat{\alpha}_i > 0$的x_i称为支持向量。
- 根据KKT互补条件可知

$$\hat{\alpha}_i \left(y_i \left(\hat{\beta}_0 + \hat{\beta}^T x_i \right) - 1 \right) = 0, i = 1, \ldots, n$$

- 对应$\hat{\alpha}_i > 0$，必有$y_i \left(\hat{\beta}_0 + \hat{\beta}^T x_i \right) - 1$，即$\hat{\beta}_0 + \hat{\beta}^T x_i = \pm 1$。因此支持向量一定在间隔边界上。

SMO算法

那如何求解式（10）？其实这也是一个二次规划问题，可以用二次规划的算法求解。但是该问题的参数个数等于训练样本数，当样本量比较大的时候，同时求解所有 $\alpha_i(i = 1, 2 \ldots, n)$计算很大，为了解决该问题，Platt(1998)提出了SMO（sequential minimal optimization）算法。 SMO 算法基本思想是每次只选取两个参数α_i和α_j进行优化，固定其他参数，又由于有$\sum_{i=1}^{n} \alpha_i y_i = 0$ 约束，也就是α_j可以写成关于 α_i的函数形式，实际求解的问题就变成了关于单变量α_i的二次规划问题，这样的二次规划具有封闭解，不需要调用数值优化算法。不断循环执行以上步骤直至收敛。

线性不可分的情形

最大间隔分类器是针对线性可分情形，但是实际中很多数据是线性不可分的

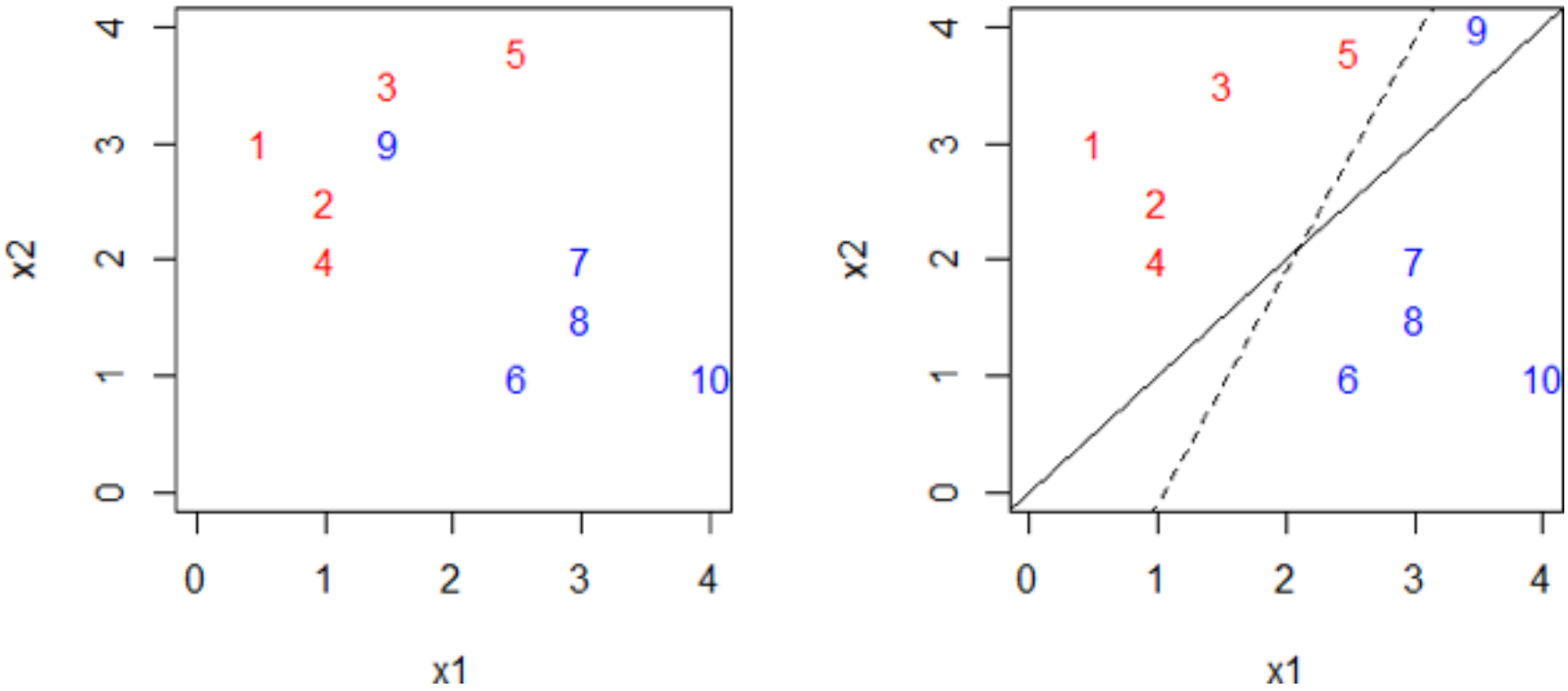


图 6: 左：改变观测点9的位置，此时最大间隔超平面不存在；右：轻微移动观测点9的位置，此时最大间隔超平面发生巨大变化，由黑色实线变为虚线

最大间隔分类器对异常值（outlier）敏感，因为最大间隔超平面其实是由少数几个训练观测点，即支持向量所决定的，所以它对这些观测点的局部扰动的反应是非常灵敏的。

线性不可分的情形

为了提高最大间隔分类器的稳定性以及对测试数据分类的效果，有必要对间隔概念进行扩展，引入软间隔（soft margin）概念，所以原来的间隔也往往被称为硬间隔（hard margin）。软间隔只要求超平面能将大部分不同类别的观测点区别开来就好，并不需要将所有训练集观测点完美区分开来。

软间隔又包含两种情况

- 允许部分观测点穿过边界，但对观测点的分类仍然是正确的
- 允许部分观测点数据分类错误

将由软间隔建立的分类器称为支持向量分类器（support vector classifier）或者线性支持向量机（linear support vector machine）。最大间隔分类器也称为硬间隔分类器（hard margin classifier），是线性支持向量机的特例。由于实际数据往往是线性不可分的，所以线性支持向量机具有更广的适用性。

线性不可分的情形

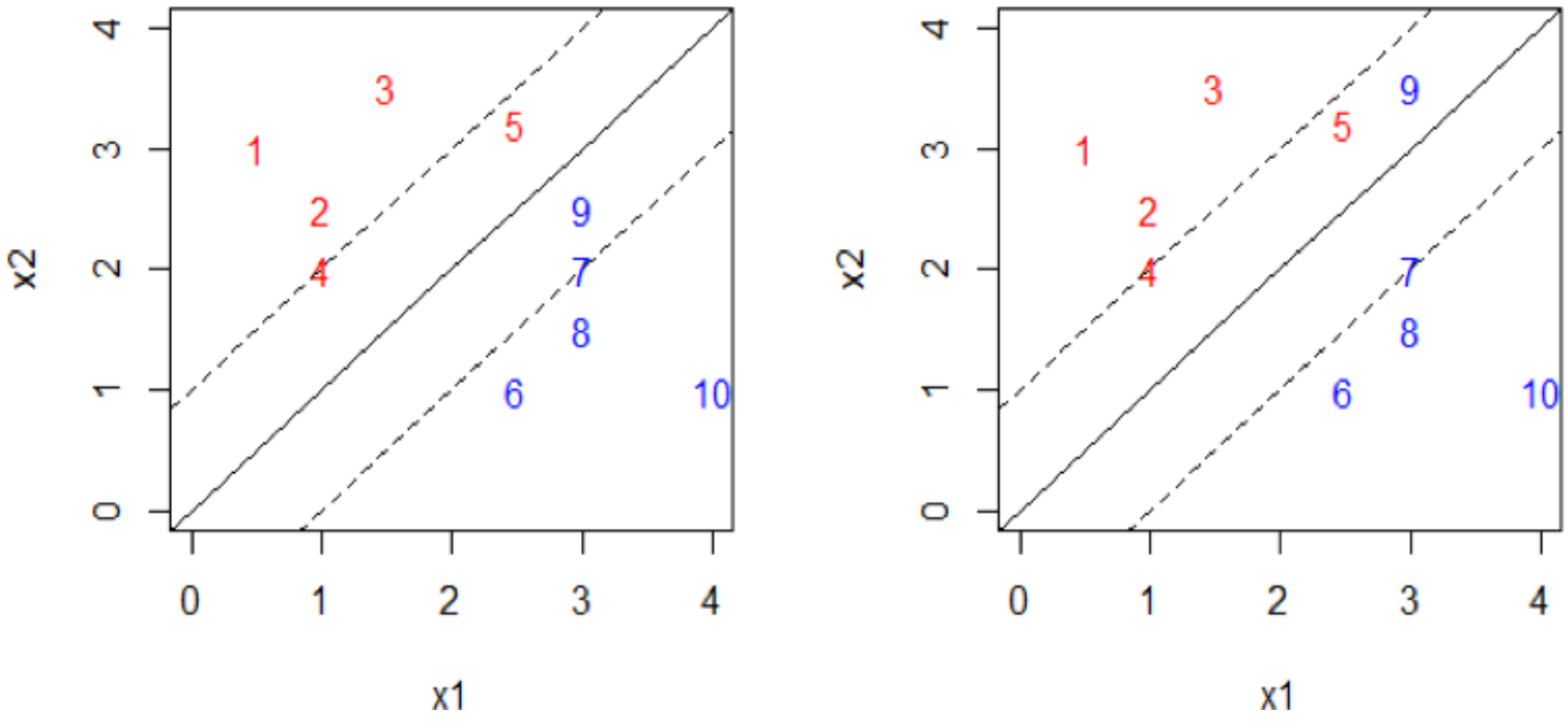
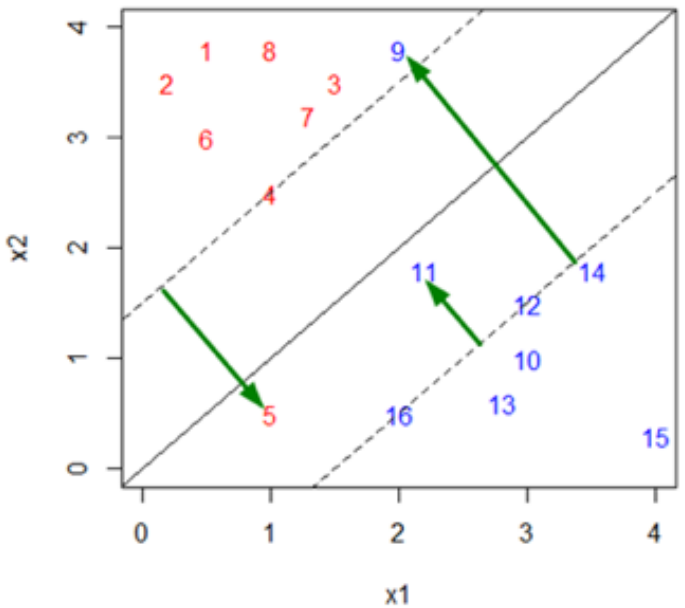


图 7: 例1观测点的软间隔；左：允许部分观测点穿过边界（如观测点5、9），但此时对所有观测点的分类仍然是正确的；右：允许部分观测点（如观测9）分类错误

线性支持向量机

$$
\begin{aligned}
& \max_{\beta_0,\beta,\varepsilon} M \\
\text{s.t.}\ & \|\beta\| = \sum_{j=1}^{p} \beta_j^2 = 1 \\
& y_i f(x_i) = y_i \left(\beta_0 + x_i^T \beta\right) \geq M(1-\varepsilon_i), \forall i \\
& \sum_{i=1}^{n} \varepsilon_i \leq C, \quad \varepsilon_i \geq 0
\end{aligned}
\tag{15}
$$

最优化问题（15）与（7）的最大区别就在于多了松弛变量（slack variable）ε_i



线性支持向量机

- 第一个约束条件保证了求解上述最优化问题时能得到参数的唯一解，以及此时可以用$y_i f(x_i)$表示观测点到超平面的距离。
- 第二个约束条件，现在不等号右边不再是简单的间隔 M，而是多乘了一项$1-\varepsilon_i$。可以根据ε_i 的值判断第i个观测点的位置。
  - 如果$\varepsilon_i = 0$，即$1-\varepsilon_i = 1$，此时不等号右边还是等于M，则观测点i是落在边界正确的一侧（图8中除点 5、 9、 11 以外的点）
  - 如果$0 < \varepsilon_i < 1$，此时不等号右边的值小于M，则说明此时允许观测点i穿过边界，但是分类还是正确的，图8中的点11
  - 如果$\varepsilon_i > 1$，此时不等号右边的值小于0，则此时允许观测点i穿过超平面，即分类是错误的（图8中的点5、9）。

线性支持向量机

- 对于第三个约束， C 是所有松弛变量和的上界，也就是能容忍观测点穿过边界的数量或者说程度。随着 C 增大，能容忍观测点点穿过边界的程度增大，则此时间隔越宽。

那么， C 一般要怎么选择呢？其实这又涉及了偏差-方差的权衡问题。当 C 越大时，我们能容忍观测点穿过边界的程度增大，间隔越宽，则此时能够降低方差，但却可能因拟合不足而产生较大的偏差；相反，C 越小，间隔越窄，这时分类器很有可能会过度拟合数据，即虽然降低了偏差，但可能产生较大的方差。所以在实际问题中，一般也是通过交叉验证的方法来确定 C。

线性支持向量机

最优化问题（10）的对偶问题为：

$$\begin{aligned} &\min_{\beta_0,\beta} \|\beta\| \\ &\text{s.t. } y_i f(x_i) = y_i\left(\beta_0 + x_i^T\beta\right) \geq 1-\varepsilon_i \\ &\sum_{i=1}^{n}\varepsilon_i \leq C, \quad \varepsilon_i \geq 0, i = 1,2,\cdots n \end{aligned} \tag{16}$$

还可以写成下面的形式

$$\begin{aligned} &\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + \lambda\sum_{i=1}^{n}\varepsilon_i \\ &\text{s.t. } y_i f(x_i) = y_i\left(\beta_0 + x_i^T\beta\right) \geq 1-\varepsilon_i \\ &\varepsilon_i \geq 0, i = 1,2,\cdots n \end{aligned} \tag{17}$$

线性支持向量机

原始问题（17）的对偶问题是

$$\begin{aligned} &\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n} a_i \\ &\text{s.t.} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ &0 \le \alpha_i \le C, i = 1, 2, \cdots, n \end{aligned} \tag{18}$$

解：原始问题（17）的拉格朗日函数为

$$L(\beta, \beta_0, \varepsilon, \alpha, \gamma) = \frac{1}{2}\|\beta\|^2 + \lambda\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{N}\alpha_i\left(y_i\left(\beta x_i + \beta_0\right) - 1 + \varepsilon_i\right) - \sum_{i=1}^{N}\gamma_i\varepsilon_i \tag{19}$$

其中，$\alpha_i \ge 0, \quad \gamma_i \ge 0$

线性支持向量机

(1) 求 $L(\beta, \beta_0, \varepsilon, \alpha, \gamma)$ 对 $\beta, \beta_0, \varepsilon$ 的极小，由

$$\begin{cases} \frac{\partial L(\beta,\beta_0,\varepsilon,\alpha,\gamma)}{\partial \beta} = \beta - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \\ \frac{\partial L(\beta,\beta_0,\varepsilon,\alpha,\gamma)}{\partial \beta_0} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \\ \frac{\partial L(\beta,\beta_0,\varepsilon,\alpha,\gamma)}{\partial \varepsilon_i} = \lambda - \alpha_i - \gamma_i = 0 \end{cases}$$

得

$$\begin{cases} \beta = \sum_{i=1}^{N} \alpha_i y_i x_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \lambda - \alpha_i - \gamma_i = 0 \end{cases} \tag{20}$$

将式（20）代入式（19）得

$$\min_{\beta,\beta_0,\varepsilon} L(\beta, \beta_0, \varepsilon, \alpha, \gamma) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i$$

线性支持向量机

(2) 求 $\min_{\beta,\beta_0,\varepsilon} L(\beta,\beta_0,\varepsilon,\alpha,\gamma)$ 对 α 的极大问题，可得对偶问题

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \tag{21}$$

$$\text{s.t.} \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{22}$$

$$\lambda - \alpha_i - \gamma_i = 0 \tag{23}$$

$$\alpha_i \geq 0 \tag{24}$$

$$\gamma_i \geq 0, \quad i = 1, 2, \cdots, N \tag{25}$$

将约束（23）-（25）等价写成

$$0 \leq \alpha_i \leq \lambda,$$

再将目标函数的极大化问题转换为极小化问题即可得对偶问题(18)。

线性支持向量机

定理2 设 $\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_n)^T$是对偶问题（15）-（17）的一个解，若存在一个$\hat{\alpha}_j$ 且满足 $0 < \hat{\alpha}_j < \lambda$则可得原始问题的解：

$$\begin{cases} \hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \\ \hat{\beta}_0 = y_j - \sum_{i=1}^{N} y_i \hat{\alpha}_i \left(x_i \cdot x_j\right) \end{cases}$$

定理2 利用KKT条件即可以得到证明，此处证明省略，作为习题请读者证明。

线性支持向量机对偶学习算法

算法 *10-2* 线性支持向量机对偶学习算法

- 输入：线性可分训练集$D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，其中$x_i$是特征向量，类别标签$y_i\in\{-1,1\}$，$i=1,2,\cdots,n$
- 输出：分割超平面和分类决策函数
  - （1）选择调和参数λ，构造并求解对偶问题：

$$\min\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jy_iy_j(x_i\cdot x_j)-\sum_{i=1}^{n}a_i$$

$$\text{s.t.}\ \sum_{i=1}^{n}\alpha_iy_i=0$$

$$0\le\alpha_i\le\lambda,i=1,2,\cdots,n$$

解得最优解$\hat{\alpha}=(\hat{\alpha}_1,\cdots,\hat{\alpha}_n)^\top$

线性支持向量机对偶学习算法

- （2）求原问题的解：选择一个 $\hat{\alpha}_j$ 且满足 $0 < \hat{\alpha}_j < \lambda$, 得

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i$$
$$\hat{\beta}_0 = y_j - \sum_{i=1}^{N} \hat{\alpha}_i y_i (x_i \cdot x_j)$$

- （3）求得分割超平面 $f(x) = \hat{\beta}_0 + \hat{\beta}^T x = 0$ 和分类决策函数$c(x) = \text{sign}(f(x)) = \text{sign}(\hat{\beta}_0 + \hat{\beta}^T x)$

线性支持向量机

根据 $\hat{\alpha}_i$ 和松他变量 ε_i 的取值，可分为如下情况:

- 当 $\hat{\alpha}_i = 0$, 则 $\varepsilon_i = 0$, 点 x_i 落在间隔外边;
- $0 < \hat{\alpha}_j < \lambda$ 则 $\varepsilon_i = 0$, 点 x_i 落在间隔边界上;
- $\hat{\alpha}_i = \lambda$ 且 $0 < \varepsilon_i < 1$, 点 x_i 落在间隔边界与超平面之间;
- $\hat{\alpha}_i = \lambda$ 且 $\varepsilon_i = 1$, 点 x_i 落在超平面上
- $\hat{\alpha}_i = \lambda$ 且 $\varepsilon_i > 1$, 点 x_i 落在超平面的另一侧

线性支持向量机

- 解可以描述为内积的形式：

$$f(x) = \beta_0 + \sum_{i=1}^{n} \hat{\alpha}_i y_i (x \cdot x_i) \tag{26}$$

- 有且仅有支持向量对应的α_i是非零的，若
  用$S = \{i : \hat{\alpha}_i > 0\}$表示支持向量观测点的指标集合，则有

$$f(x) = \beta_0 + \sum_{i \in S} \hat{\alpha}_i y_i (x \cdot x_i) \tag{27}$$

线性支持向量机

- 另一种求解方法：最小化如下目标函数：

$$\sum_{i=1}^{n}\left[1-y_i\left(\beta_0+\beta^T x_i\right)\right]_+ + \tau\|\beta\|^2 \tag{28}$$

- “损失函数+惩罚函数”的形式：
  - 第1项为合页损失函数（*hinge loss function*）

$$\left[1-y_i(\beta_0+\beta^T x_i)\right]_+ = \begin{cases} 1-y_i\left(\beta_0+\beta^T x_i\right), & 1-y_i\left(\beta_0+\beta^T x_i\right)>0 \\ 0, & 1-y_i\left(\beta_0+\beta^T x_i\right)\le 0 \end{cases}$$

  - 第2项为惩罚函数
- 软间隔最大化方法和合页损失函数法是等价的

线性支持向量机

定理3 线性支持向量机软间隔最大化问题：

$$\begin{aligned} &\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{n} \varepsilon_i \\ &\text{s.t. } y_i f(x_i) = y_i\left(\beta_0 + x_i^T \beta\right) \geq 1 - \varepsilon_i \\ &\varepsilon_i \geq 0, i = 1, 2, \cdots n \end{aligned} \tag{29}$$

等价于最小化

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left[1 - y_i\left(\beta_0 + \beta^T x_i\right)\right]_+ + \tau\|\beta\|^2 \tag{30}$$

线性支持向量机

- 合页损失函数由于在点（1,0）上不可导，直接优化合页损失函数也是比较困难的。
- Rosset 和Zhu (2007)：Huberized 合页损失函数

$$l\left(y_i f_i\right)=\begin{cases}0, & y_i f_i>1\\ \frac{\left(1-y_i f_i\right)^2}{2\delta}, & 1-\delta<y_i f_i\leq 1\\ \frac{1-t-\delta}{2\delta}, & y_i f_i\leq 1-\delta\end{cases}$$

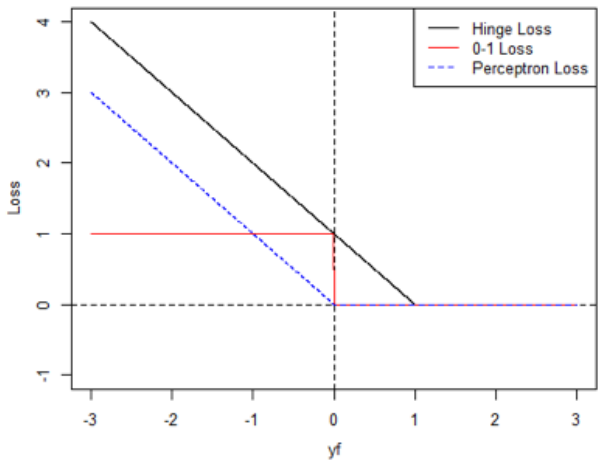


图 9: 合页损失函数图

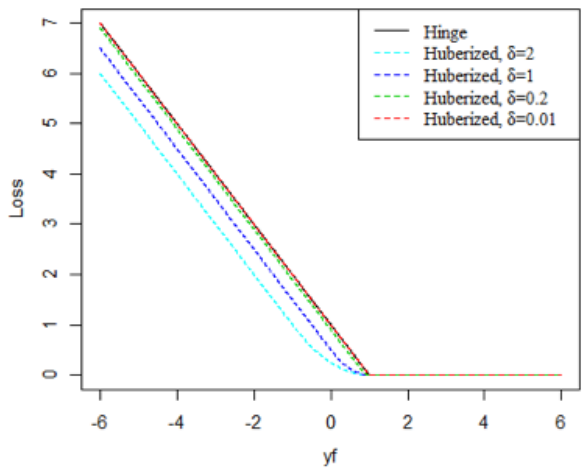


图 10: Huberized 合页损失函数

非线性支持向量机

但是在有些情况下，决策边界是非线性的，这就需要建立非线性支持向量机。

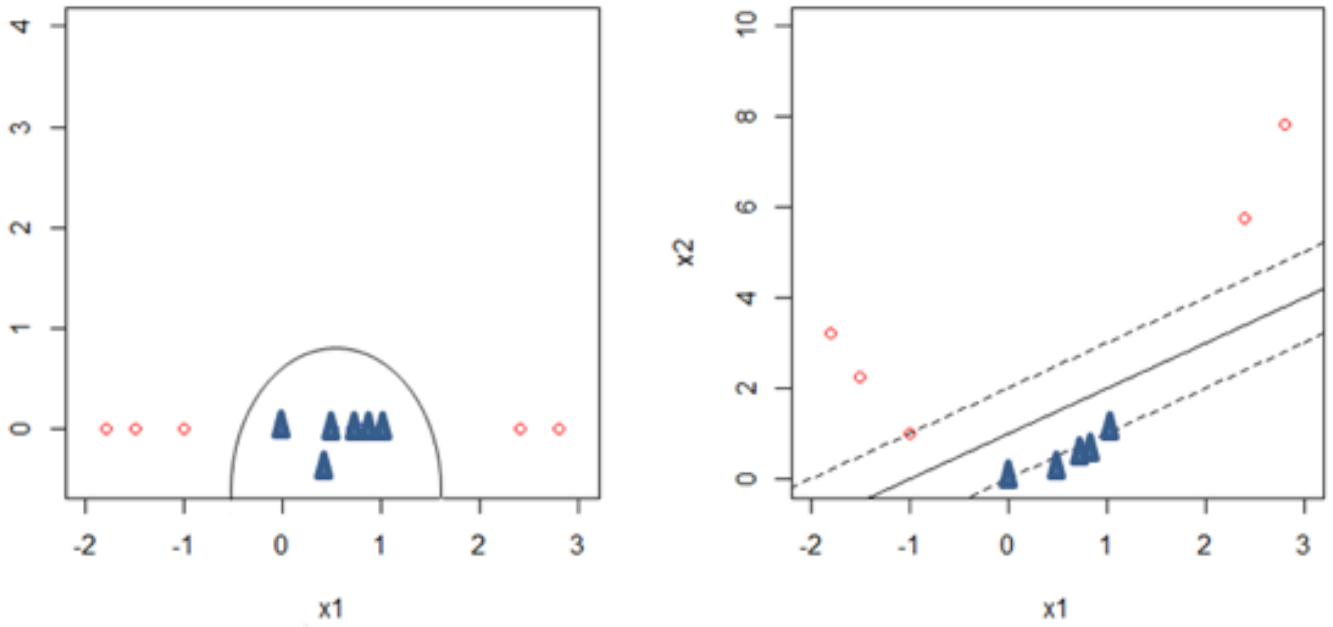


图 11: 左：在一维空间中，观测点无法用一个线性边界分割开来；右：将空间扩展到一个二维空间中，此时可以用一个线性超平面将观测点分割开来

尝试将自变量的二次项添加到超平面中，即此时的超平面是：

非线性支持向量机

- 可以将x看作是一个变量，看作是另一个变量，这样它就变成了一个二维空间的问题，如图11（右）。
- 把这个问题扩展到p维空间中
  对于 p 维观测点 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$，使用 $2p$ 个特征 $x_{i1}, x_{i1}^2, x_{i2}, x_{i2}^2, \ldots x_{ip}^2$ 来构造线性支持向量机：

$$
\begin{aligned}
&\max_{\beta_0,\beta_1,\beta_2,\varepsilon} M \\
\text{s.t.}\ &\sum_{k=1}^{2}\sum_{j=1}^{p}\beta_{kj}^2 = 1 \\
&y_i f(x_i) = y_i(\beta_0 + \beta_1^T x_i + \beta_2^T x_i^2) \geq M(1-\varepsilon_i) \\
&\sum_{i=1}^{n}\varepsilon_i \leq C, \quad \varepsilon_i \geq 0, i = 1, 2, \cdots, n
\end{aligned}
\tag{32}
$$

非线性支持向量机

- 可以考虑使用观测点的不同多项式，如二次、三次甚至是更高阶多项式，或者不同观测点的交互项来扩大特征空间，进而在这个扩大的特征空间中构造一个线性超平面
- 该方法比较简单、容易理解，但是该方法使得特征空间的维度扩展过快，计算量是比较惊人的。

非线性支持向量机

- 设 X 是输入空间（欧式空间 R^n 的子集或者离散集合)， H 为特征空间（希尔伯特空间)，如果存在一个从X 到 H 的映射 $\phi(x) : X \to H$使得对所有的$x, z \in X$, 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。

- **定理4 正定核的充要条件：**设 $K : \mathrm{X} \times \mathrm{X} \to R$ 是对称函数，则 $K(x, z)$ 为正定核函数的充要条件是对任意 $x_i \in X \quad , \quad i = 1, \cdots, n, \quad K(x, z)$ 对应的 Gram 矩阵

$$K = [K(x_i, x_j)]_{n \times n}$$

总是半正定的。

非线性支持向量机

表 2: 最常使用的核函数

名称	表达式	参数
线性核函数	$K(x_i, x_k) = x_i^\top x_k$	
多项式核函数	$K(x_i, x_k) = (1 + x_i^T x_k)^d$	$d \geq 1$ 为多项式幂次数
高斯核函数	$K(x_i, x_k) = \exp(-\frac{1}{2\sigma^2}\|x_i - x_k\|^2)$	$\sigma > 0$ 为高斯核的带宽
sigmoid核函数	$K(x_i, x_k) = \tanh(vx_i^T x_k + \eta)$	Tanh为双曲正切函数 $v > 0, \quad \eta < 0$

假设 $K_1(x, z)$ 和 $K_2(x, z)$ 是核函数，则通过如下方式构造的新函数也是核函数：

构造方法	说明
$K(x, z) = cK_1(x, z)$	其中 $c > 0$ 是常数
$K(x, z) = f(x)K_1(x, z)f(z)$	$f(\cdot)$ 是任意函数
$K(x, z) = q(K_1(x, z))$	$q(\cdot)$ 是非负系数的多项式函数
$K(x, z) = \exp(K_1(x, z))$	
$K(x, z) = K_1(x, z) + K_2(x, z)$	
$K(x, z) = K_1(x, z)K_2(x, z)$	

非线性支持向量机

核支持向量机（*kernel SVM*）：基于核方法的非线性支持向量机

基本思想：首先通过一个非线性的变换（映射$\phi(\cdot)$）将原始输入空间（欧式空间R^n或离散集合）变换为一个特征空间（希尔伯特空间H），而特征空间往往是更高维度的，甚至是无穷维的；然后在特征空间里用线性支持向量机方法从训练数据集中学习模型，特征空间H里训练得到的超平面对应于原始输入空间R^n中的超曲面。在变换后的特征空间中支持向量机问题变成：

$$\begin{aligned} &\min_{\beta_0,\beta} \|\beta\| \\ &\text{s.t. } y_i\left(\beta_0+\phi\left(x_i\right)^{\top}\beta\right)\geq 1-\varepsilon_i \\ &\sum_{i=1}^{n}\varepsilon_i\leq C,\quad \varepsilon_i\geq 0, i=1,2,\cdots,n \end{aligned} \tag{33}$$

非线性支持向量机

特征空间中的超平面所对应的模型为$f(x)=\beta_0+\phi(x_i)^{\top}\beta$，其对偶问题是

$$\begin{aligned}\min\ & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{n}a_i \\ \text{s.t.}\ & \sum_{i=1}^{N}\alpha_i y_i=0 \\ & 0\le\alpha_i\le\lambda, i=1,2,\cdots,n\end{aligned} \tag{34}$$

求解问题(34)得到的分类决策函数为：

$c(x)=\text{sign}\left(\beta_0+\sum_{i\in S}\hat{\alpha}_i y_i K(x,x_i)\right)$

非线性支持向量机

- 关于核函数的选择一直以来都是支持向量机研究的热点，但是学者们通过大量的研究并没有形成定论，即没有最优核函数。通常情况下，径向核函数是非线性支持向量机使用最多的。
- 采用核函数而不是直接扩展特征空间的方式的优势在于，使用核函数，仅需要计算$C_n^2 = \frac{n(n-1)}{2}$ 个成对组合的$K\left(x_i, x_k\right)$，而若采取直接扩展特征空间的方式，是没有明确的计算量的。
- 线性支持向量机某种角度上可以看做是核支持向量机的特例，因为取线性核函数即为线性支持向量机。

非线性支持向量机

核支持向量机写成“损失函数+惩罚函数”形式：

$$\min_{f\in H}\frac{1}{n}\sum_{i=1}^{n}\left[1-y_i f\left(x_i\right)\right]+\tau\|f\|_H^2 \tag{35}$$

根据希尔伯特空间可再生核*(reproducing kernel Hilbert space)*理论(Wahba，1990)，式（35）的解是由可再生核函数组成的有限维表达式，即

$$\hat{f}(x)=\hat{\beta}_0+\sum_{i=1}^{n}\hat{\alpha}_i K\left(x,x_i\right)$$

其中 $\hat{\beta}_0$ 和 $\hat{\alpha}$ 是如下式子的解

$$\min_{\beta_0,\alpha}\frac{1}{n}\sum_{i=1}^{n}\left[1-y_i\left(\hat{\beta}_0+\sum_{j=1}^{n}\hat{\alpha}_j K\left(x_i,x_j\right)\right)\right]_+ +\tau\alpha^T K\alpha$$

非线性支持向量机

- 支持向量机统计意义上的解释:

  因为 $\frac{1}{n}\sum_{i=1}^{n}\left[1-y_i f\left(x_i\right)\right]_+$ 是期望 hinge 损失 $E\left(\left[1-yf(x)\right]_+\right)$ 的经验版本, 期望 hinge 损失的最小值正好是贝叶斯决策规则:

$$\underset{f}{\operatorname{argmin}} E\left(\left[1-yf(x)\right]_+\right)=\operatorname{sign}\left(p_+(x)-p_-(x)\right)$$

  其中, $p_+(x)=\Pr(Y=1\mid X=x)$ 和 $p_-(x)=\Pr(Y=-1\mid X=x)$

- 还有没有其他的损失函数也能得到比较好的分类结果呢?
  - **定理5** 如果损失函数 $l(t)$ 满足条件:

    A1. $l(t)<l(-t),\forall t>0$,

    A2. $l'(0)\neq 0$ 存在,

    A3. 条件风险函数 $E(l(yf(x))\mid x)$ 具有全局最小值,

    则损失函数$l(t)$是Fisher一致的。
  - **定理6** 令 $l(t)$ 是凸的损失函数,在零点处可导且 $l'(0)<0$, 则损失函数 $l(t)$ 是 Fisher一致的。
- Logistic损失函数和Huberized hinge损失函数都是Fisher一致的。

非线性支持向量机

在高维数据的分类问题中，也就是说自变量个数可能大于样本数时，前面介绍的线性支持向量机和核支持向量机就不再适用。

稀疏支持向量机

- Bradley 和Mangasarian(1998)提出的L_1SVM:

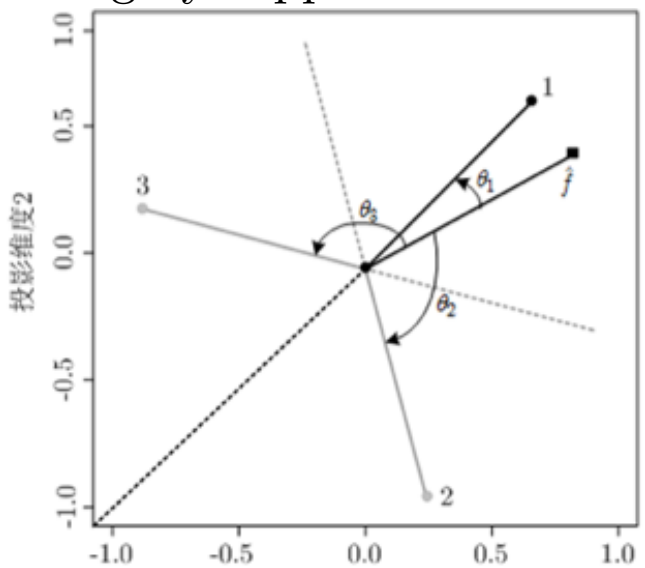
$$\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n}\left[1-y_i\left(\beta_0+\beta^T x_i\right)\right]_+ + \tau\|\beta\|_1 \qquad (36)$$

- 稀疏支持向量机一般化的目标函数：

$$\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n} l\left(y_i\left(\beta_0+\beta^T x_i\right)\right) + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right) \qquad (37)$$

多分类问题

- 将多分类转化为一系列二分类问题
  - “一对一”（one-versus-one）
  - “一对余类”(one-versus-the-rest或者one-versus-all)
- 构建单一目标函数同时优化求解多分类问题
  - Liu等（2011）：强化MSVM（Reinforced Multicategory Support Vector Machines）模型
  - Zhang等（2014）：基于角度间隔的多分类支持向量机（AMSVM Angle-based multicategory support vector machines, AMSVM）



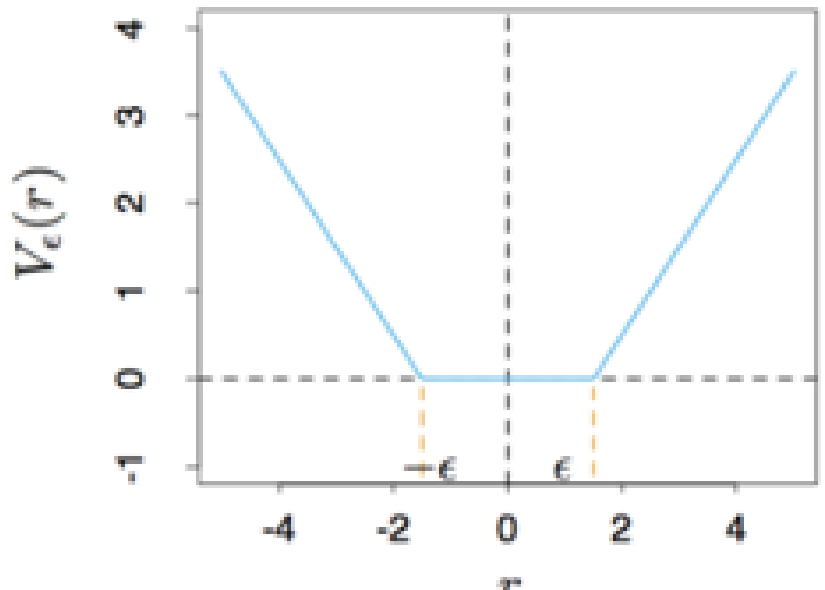
支持向量回归（SVR）

- 给定训练数据集$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，　$y_i \in R$，希望学习得到

$$f(x) = \beta_0 + \beta^T x_i \tag{35}$$

- ϵ-不敏感损失*(ϵ-insensitive loss)*

$$V(r) = \begin{cases} 0, & \text{若 } |r| < \epsilon, \\ |r| - \epsilon, & \text{其他.} \end{cases}$$



支持向量回归

- SVR的问题就可以写成

$$\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{n} V\left(f\left(x_i\right) - y_i\right) \tag{10.36}$$

- 引入松他变量 ξ_i 和 $\tilde{\xi}_i$，可将式（10.36）重写为

$$\begin{aligned} \min_{\beta_0,\beta,\xi,\xi} & \frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{n} \left(\xi_i + \tilde{\xi}_i\right) \\ \text{s.t.} \quad & f\left(x_i\right) - y_i \leq \epsilon + \xi_i \\ & y_i - f\left(x_i\right) \leq \epsilon + \tilde{\xi}_i \\ & \xi_i \geq 0, \tilde{\xi}_i \geq 0, i = 1, 2, \cdots, n \end{aligned} \tag{37}$$

- 引入拉格朗日乘子，得到式（37）的拉格朗日函数：

$$\begin{aligned} & L\left(\beta, \beta_0, \alpha, \tilde{\alpha}, \xi, \tilde{\xi}, \mu, \tilde{\mu}\right) \\ & = \tfrac{1}{2}\|\beta\|^2 + \lambda \textstyle\sum_{i=1}^{n} \left(\xi_i + \tilde{\xi}_i\right) - \sum_{i=1}^{n} \mu_i \xi_i - \sum_{i=1}^{n} \tilde{\mu}_i \tilde{\xi}_i \end{aligned}$$

支持向量回归

- 将 $L\left(\beta, \beta_0, \alpha, \tilde{\alpha}, \xi, \tilde{\xi}, \mu, \tilde{\mu}\right)$ 对 β_0, β, ξ 和 $\tilde{\xi}$ 的偏导为零可得

$$\beta = \sum_{i=1}^{n} \left(\tilde{\alpha}_i - \alpha_i\right) x_i \tag{39}$$

$$0 = \sum_{i=1}^{n} \left(\tilde{\alpha}_i - \alpha_i\right) \tag{40}$$

$$\lambda = \alpha_i + \mu_i \tag{41}$$

$$\lambda = \tilde{\alpha}_i + \tilde{\mu}_i \tag{42}$$

- 将式（39）-（42）代入式（38）可得SVR的对偶问题

$$\max_{\alpha, \tilde{\alpha}} \sum_{i=1}^{n} y_i \left(\tilde{\alpha}_i - \alpha_i\right) - \left(\tilde{\alpha}_i + \alpha_i\right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\tilde{\alpha}_i - \alpha_i\right) \left(\tilde{\alpha}_j - \alpha_j\right) x_i^T x_j$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \left(\tilde{\alpha}_i - \alpha_i\right)$$

$$0 \leq \alpha_i, \tilde{\alpha}_i \leq \lambda \tag{43}$$

支持向量回归

- 上述过程需要满足KKT条件：

$$\begin{cases} \alpha_i \left(f\left(x_i\right) - y_i - \epsilon - \xi_i \right) = 0 \\ \tilde{\alpha}_i \left(y_i - f\left(x_i\right) - \epsilon - \tilde{\xi}_i \right) = 0 \\ \alpha_i \tilde{\alpha}_i = 0, \xi_i \xi_i = 0 \\ \left(\lambda - \alpha_i\right) \xi_i = 0, \left(\lambda - \tilde{\alpha}_i\right) \tilde{\xi}_i = 0 \end{cases} \tag{44}$$

- 将式（39）代入式（35），求得SVR的解

$$f(x) = \beta_0 + \sum_{i=1}^{n} \left(\tilde{\alpha}_i - \alpha_i\right) x_i^T x_i \tag{45}$$

- 进一步求得

$$\beta_0 = y_i + \epsilon - \sum_{i=1}^{n} \left(\tilde{\alpha}_i - \alpha_i\right) x_i^T x_i \tag{46}$$

支持向量回归

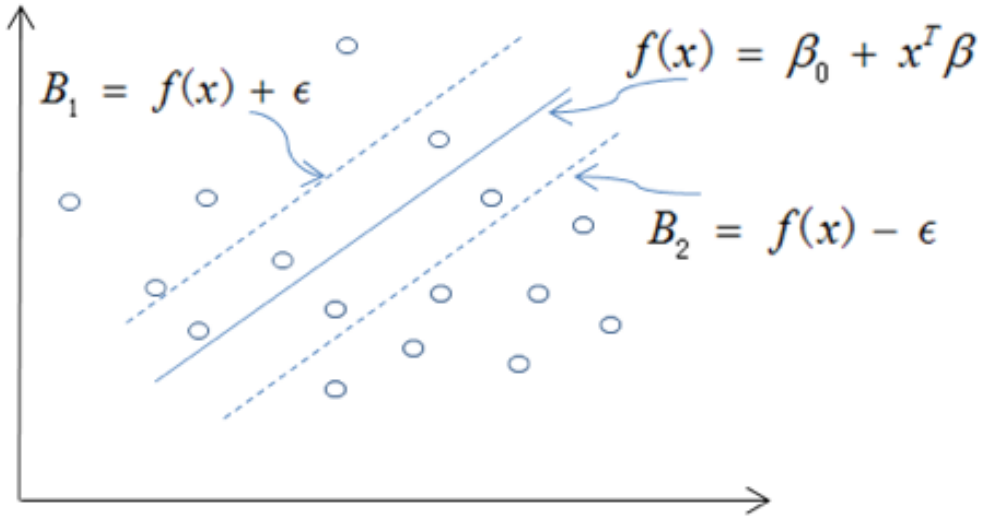


图 12: 支持向量回归示意图

支持向量回归

- 非线性SVR

$$f(x) = \beta_0 + \beta^T \phi(x_i) \tag{47}$$

- 支持向量回归所选的超平面是如下最优化问题的解：

$$\min_{\beta_0, \beta} \frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{n} V(f(x_i) - y_i) \tag{48}$$

- 最优化问题（48）的解可以表示为：

$$f(x) = \sum_{i=1}^{n} (\tilde{\alpha}_i - \alpha_i) K(x, x_i) + \beta_0 \tag{49}$$

其中 $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ 是核函数。

习题

1. 请证明定理1。
2. 请证明定理2。
3. 请分析R软件ISLR包中的股票数据Smarket，以股票的涨跌方向 Direction 为因变量，以 Lag1-Lag5 以及 Volume 为自变量，进行如下分析：

（1）分析一下该数据集的股票涨跌天数分别是多少以及它们的比例是多少？

（2）请以 2005 年数据为训练集， 2005 年及之后的数据为测试集。用训练集数据进行建模，先分析当 cost=1 时的建模结果，然后利用交叉验证方法选取最优的 cost参数，并分析最优的模型结果。

（3）利用得到的最优模型对测试集进行预测，分析预测准确率。

习题

4. 请分析 R 软件 ISLR 包中的 Auto 数据集：

（1）将 Auto 数据集中的 mpg 按照中位数划分为两类，新增一个变量 grade，并用0和1分别表示。

（2）从该数据集随机抽取292个样本作为训练集，剩下的作为测试集。

（3）利用 maximal margin classifier 进行建模，利用交叉验证选取最优的模型，分析该最优模型的结果，并利用该最优模型对测试集进行预测分析。

（4）请利用 radial kernel 的 SVM 对训练集进行建模，利用交叉验证选择最优的模型，分析该最优模型的结果，并利用最优模型对测试集进行预测分析。