第2-2节　支持向量机基础

—— 单层感知机算法

三、 感知机模型

1957年，Rosnblatt提出的感知机，是神经网络和支持向量机的基础，其结构如图2.3所示。

图2.3 实现线性二分类的感知机模型

相应的数学模型如下： $f(x) = w \bullet x + b$

3.1 感知机的分类原理

1) 输入空间：输入样本 $x \in R^n$，为样本的特征向量；

2) 输出空间：$y = \{+1, -1\}$，表示样本的类别标签；

3) 分类器函数：$f(x) = w \cdot x + b$

如果：所有类别标签 $y_i = +1$ 的样本 x_i，$w \cdot x_i + b > 0$

所有类别标签 $y_j = -1$ 的样本 x_j，$w \cdot x_j + b < 0$

则数据集 X 称为线性可分数据集。其中：w 和 b 为感知机分类器的参数，w 为权值向量，b 叫做偏置。

感知机的线性分界面方程如下：有如下：

$$w \bullet x + b = 0$$

其几何解释为：对应于特征空间R^n中的一个超平面 S，其中 w 是超平面的法向量，b 是超平面的截距，如图2.4所示。

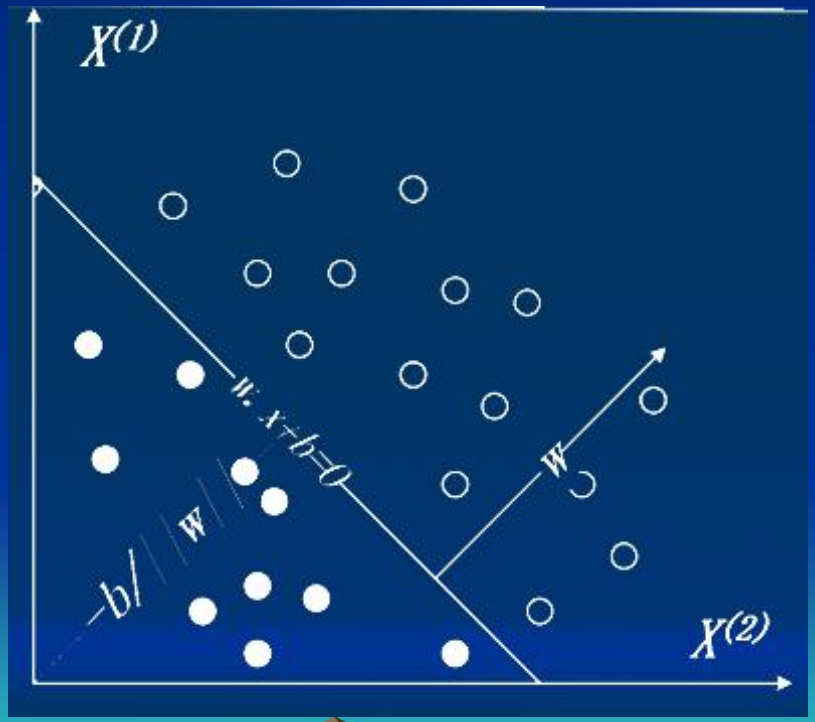

图2.4 感知机实现线性分界面方程的几何示意图

3.2 感知机的学习策略

目标：找到一个线性分界面，将训练样本中不同两个类别的点完全正确分开。

即确定参数w, b，将误分类点的分类错误总和最小化。即求出所有误分类点到该几何平面S 的总距离，然后使该总距离最短，从而将样本点完全区分开来。单个误分类点 x_i 到该超平面 S 的距离，公式如下：

$$\frac{1}{\|w\|}|w \bullet x_i + b| \qquad \text{其中：}\|w\|\text{是}w\text{的}L_2\text{范数}$$

上述式子也可以改写为：

$$\frac{1}{\|w\|}[-y_i(w \bullet x_i + b)]$$ 实际输出与类别标签相反

所有误分类点到超分界面 S 的距离总和为：

$$-\frac{1}{\|w\|}\sum_{x_i \in M}[-y_i(w \bullet x_i + b)]$$ M为所有误分类点的集合

则超分界面 S 的求取，通过选取使得上式最小的 w, b 值获得。

3.3 感知机学习算法

感知机学习算法的原始形式：

给定一个数据集：

$$T = \{ (x_1,y_1), (x_2,y_2), ..., (x_N,y_N) \}$$

求参数 w，b 使其为以下损失函数极小化问题的解：

$$\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$$

最优化方法采用随机梯度下降法。则

在一次求解过程中，损失函数对于w，b 梯度为：

$$\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i$$

$$\nabla_b L(w,b) = -\sum_{x_i \in M} y_i$$

随机选取一个误分类点(x_i, y_i)，对w, b进行更新

$$w \leftarrow w + \eta y_i x_i$$

$$b \leftarrow b + \eta y_i$$

- η是参数修改步长，$(0 < \eta \leq 1)$, 又称为学习率。

3.4 感知机训练步骤：

1) w, b参数初始化，即随机选取w_0, b_0

2) 在训练集中选取数据 (x_i, y_i)

3) 如果 $y_i(w \bullet x_i + b) \leq 0$，说明 (x_i, y_i) 是误分点，需要对相应的 w, b 的值进行修正：

$$w \leftarrow w + \eta y_i x_i$$

$$b \leftarrow b + \eta y_i$$

4) 计算 $y_i(w \bullet x_i + b)$ 的值，如果还是 ≤ 0，就返回步骤 3) 重新更新 w, b 的值，直至该值 > 0为止

5) 返回步骤 2)，重复步骤 3)、4)的训练

6) 训练集完成一次训练后，再返回步骤 2)进行下一次迭代，直到训练集中没有误分类点

7) 输出最终的 w, b 值，获得超分界面方程。

例2.1

如图 2.5 所示的训练数据集，1类的点是$x_1=(3,3)^T$，$x_2=(4,3)^T$；-1类的点是$x_3=(1,1)^T$。采用感知机学习算法求上述三个数据点线性分界面的方程。

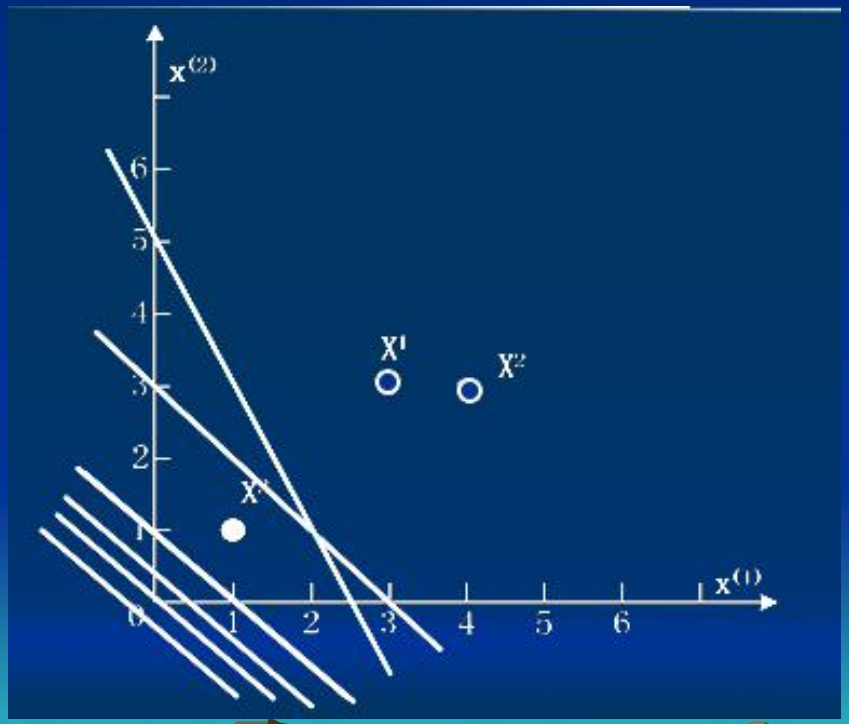

图2.5 感知机线性分类示例

1) 初始化。取 $w_0 = b_0 = 0$, $\eta = 1$

2) 对$x_1=(3,3)^T$，$y_1 (w_0 \bullet x_1 + b_0) = 0$ ，未能正确分类，需要更新 w, b 的值：

$$w_1=w_0+y_1 \bullet x_1=(3,3)^T$$

$$b_1=b_0+y_1 = 1$$

得到线性分界面方程：$w_1 \bullet x + b_1 = 3x^{(1)}+3x^{(2)}+1$

3) 重新计算 $y_1 (w_1 \bullet x_1 + b_0) = 19 > 0$，满足要求，停止训练数据点$x_1$

4) 对$x_2=(4,3)^T$，$y_2(w_1 \bullet x_2 + b_1) = 22 > 0$，正确分类，无需训练

5) 对$x_3=(1,1)^T$，$y_3(w_1 \bullet x_3 + b_1) = -7 < 0$，误分类，更新$w$, b 的值

$$w_2=w_1+y_3 \bullet x_3=(2,2)^T$$

$$b_2=b_1+y_3 = 0$$

得到线性分界面方程：$w_2 \bullet x + b_2 = 2x^{(1)}+2x^{(2)}$

6) 重新计算 $y_3(w_2 \bullet x_3 + b_2) = -4 < 0$，不满足分类要求，返回到第 5) 步再训练。最后得到满足数据点 x_3 的分界面方程 - 2

7) 第二次迭代，再对$x_1=(3,3)^T$计算分类函数的值，为 -2 < 0 ，误分类，需训练更新w , b 的值

$$w_6=w_5+y_1 \bullet x_1=(3,3)^T$$

$$b_6=b_5+y_1 = -1$$

得到线性分界面方程：$w_6 \bullet x + b_6 = 3x^{(1)}+3x^{(2)}-1$

8) 重新计算 $y_1 (w_6 \bullet x_1 + b_6) = 17 > 0$，满足分类要求，停止训练数据点$x_1$

9) 对$x_2=(4,3)^T$，$y_2(w_6 \bullet x_2 + b_6) = 20 > 0$，正确分类，无需训练

10) 对$x_3=(1,1)^T$，$y_3(w_6 \bullet x_3 + b_6) = -5 < 0$，误分类，更新$w$, b 的值

$$w_7=w_6+y_3 \bullet x_3=(2,2)^T$$

$$b_7=b_6+y_3 = -2$$

得到线性分界面方程：$w_7 \bullet x + b_7 = 2x^{(1)}+2x^{(2)}-2$

11) 重新计算 $y_3(w_7 \bullet x_3 + b_7) = -2 < 0$，不满足分类要求，返回到第 10) 步再训练。最后得到满足数据点 x_3 的分界面方程 $x^{(1)}+x^{(2)}-3$

12) 第三次迭代，所有的数据点都满足了$y_i(w \bullet x_i + b) \geq 0$ 的要求。则停止所有训练，上述过程如表2.1所示。

表2.1 图2.5 示例的迭代过程

迭代次数	误分类点	w	b	w • x+b
0	0	0	0	0
1	x_1	$(3,3)^T$	1	$3x^{(1)}+3x^{(2)}+1$
2	x_3	$(2,2)^T$	0	$2x^{(1)}+2x^{(2)}$
3	x_3	$(1,1)^T$	-1	$x^{(1)}+x^{(2)}$
4	x_3	$(0,0)^T$	-2	-2
5	x_1	$(3,3)^T$	-1	$3x^{(1)}+3x^{(2)}-1$
6	x_3	$(2,2)^T$	-2	$2x^{(1)}+2x^{(2)}-2$
7	x_3	$(1,1)^T$	-3	$x^{(1)}+x^{(2)}-3$
8	0	$(1,1)^T$	-3	$x^{(1)}+x^{(2)}-3$

3.5 感知机算法的收敛性分析

为了便于推导，将偏置 b 并入权重向量 w，记作 $\dot{w}=(w^{\mathrm{T}}, b)^{\mathrm{T}}$；输入向量也进行扩充，加进常数 1 作为偏置 b 的输入，记作 $\dot{x}=(x^{\mathrm{T}}, 1)^{\mathrm{T}}$。这样，单层感知机的模型可改写为 $\dot{w}\cdot\dot{x}$。

定理2.1 (Novikoff)

设训练数据集 $T = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \}$ 是线性可分的，其中 $x_i \in \mathbf{R}^n$， $y_i \in \{ -1, +1 \}$。则

(1) 存在满足条件 $\| \dot{w}_{opt} \| = 1$ 的超平面 $\dot{w}_{opt} \cdot \dot{x} = w_{opt} \cdot x + b = 0$ 将训练数据集完全正确分开，且存在 $\gamma > 0$，对所有 $i = 1, 2, \ldots, N$

$$y_i (\dot{w}_{opt} \cdot \dot{x}_i) = y_i (w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

证明：

由于训练数据集是线性可分的，则存在超平面将其完全正确分开，取此超平面为 $\dot{w}_{opt} \cdot \dot{x} = w_{opt} \cdot x + b = 0$ 。由于对有限的 $i = 1, 2, \ldots, N$ ，均有：

$$y_i (\dot{w}_{opt} \cdot \dot{x}_i) = y_i (w_{opt} \cdot x_i + b_{opt}) > 0$$

所以存在

$$\gamma = \min \{ y_i (w_{opt} \cdot x_i + b_{opt}) \}$$

使

$$y_i (\dot{w}_{opt} \cdot \dot{x}_i) = y_i (w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

定理2.1 (Novikoff)

(2) 令 $R = \max \|\dot{x}_i\|$，则感知机算法在训练集上的误分类次数 k 满足不等式：

$$k \leq \left(\frac{R}{\gamma}\right)^2$$

证明：

令 $\dot{w}_{k-1}$ 是第 k 个误分类实例之前的权重向量：

$$\dot{w}_{k-1} = (w_{k-1}^{\mathrm{T}},\ b_{k-1})^{\mathrm{T}}$$

则第 k 个误分类实例的条件是

$$y_i(\dot{w}_{k-1} \cdot \dot{x}_i) = y_i(w_{k-1} \cdot x_i + b_{k-1}) \leq 0$$

第 k 步更新过程为：

$$\dot{w}_k = \dot{w}_{k-1} + \eta y_i \dot{x}_i$$

下面推导两个不等式：

(1) $\dot{w}_k \cdot \dot{w}_{opt} \geq k\eta\gamma$

由 $\dot{w}_k$ 和 $\dot{w}_{opt}$ 的表达式可得：

$$\dot{w}_k \cdot \dot{w}_{opt} = (\dot{w}_{k-1} + \eta y_i \dot{x}_i) \cdot \dot{w}_{opt}$$

$$= \dot{w}_{k-1} \cdot \dot{w}_{opt} + \eta y_i \dot{x}_i \cdot \dot{w}_{opt}$$

$$\geq \dot{w}_{k-1} \cdot \dot{w}_{opt} + \eta\gamma$$

由此递推可得：

$$\dot{w}_k \cdot \dot{w}_{opt} \geq \dot{w}_{k-1} \cdot \dot{w}_{opt} + \eta\gamma \geq \dot{w}_{k-2} \cdot \dot{w}_{opt} + 2\eta\gamma \geq \cdots \geq k\eta\gamma$$

(2) $\|\dot{w}_k\| \le k\eta^2 R^2$

由 $\dot{w}_k$ 的表达式有：

$$\begin{aligned}
\|\dot{w}_k\|^2 &= \|\dot{w}_{k-1}\|^2 + 2\eta\, y_i\, \dot{w}_{k-1}\cdot \dot{x}_{k-1} + \eta^2\|\dot{x}_i\|^2 \\
&\le \|\dot{w}_{k-1}\|^2 + \eta^2\|\dot{x}_i\|^2 \\
&\le \|\dot{w}_{k-1}\|^2 + \eta^2 R^2 \\
&\le \|\dot{w}_{k-2}\|^2 + 2\eta^2 R^2 \le \cdot\ \cdot\ \cdot \\
&\le k\ \eta^2 R^2
\end{aligned}$$

组合不等式(1)，可得

$$k\eta\gamma \le \dot{w}_k \cdot \dot{w}_{\text{opt}} \le \|\dot{w}_k\| \|\dot{w}_{\text{opt}}\| \le \sqrt{k}\eta R$$

于是：　　$k^2\gamma^2 \le kR^2$

即：

$$k \le \left(\frac{R}{\gamma}\right)^2$$

定理2.1表明，误分类的次数 k 式有上界的。单层感知机经过有限次学习可以得到将训练数据集完全正确分开的超平面。

3.6 感知机学习算法的对偶形式

假设 w 和 b 关于数据(x_i, y_i)修改的增量分别为$\alpha_i y_i x_i$和$\alpha_i y_i$，最后学习得到的 w 和 b 可以表示为：

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

$$b = \sum_{i=1}^{N} \alpha_i y_i$$

算法2.2 感知机学习算法的对偶形式：

输入：线性可分的训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$，其中 $x_i\in R^n$，$y_i\in\{-1,+1\}$，学习率 $\eta(0<\eta\leq 1)$。

输出：感知机模型

$$f(x)=sign\left[\sum_{j=1}^{N}\propto_j y_j x_j\cdot x+b\right]$$

训练过程变为：

(1) α_i、b 初始化为 0；

(2) 选取训练数据 (x_i, y_i)；

(3) 如果

$$y_i\left[\sum_{j=1}^{N} \propto_j y_j x_j \cdot x_i + b\right] \leq 0$$

则修改：$\alpha_i = \alpha_i + \eta$

$$b = b + \eta y_i$$

(4) 转步骤 (2)，选取下一个数据，直至没有误差。

内积Gram矩阵：

由于对偶问题中，数据仅仅以内积形式出现，因此可以提前计算出来，迭代优化时直接调用。内积构成的矩阵，就是内积对称的Gram (格拉姆) 矩阵：

$$G = [\, x_i \cdot \quad x_j \,]_{N \times N}$$

例2.2

数据同例2.1，正样本点是 $x_1=(3,3)^T$，$x_2=(4,3)^T$；负样本点是$x_3=(1,1)^T$。试用感知机学习算法的对偶形式求线性分界面的方程。

(1) 初始化 $\alpha_i=0$，$i=1,2,3$；$b=0$；$\eta=1$

(2) 计算Gram矩阵

$$G=\begin{bmatrix}18 & 21 & 6\\ 21 & 25 & 7\\ 6 & 7 & 2\end{bmatrix}$$

(3) 误分类条件：

$$y_i\left[\sum_{j=1}^{N}\propto_j y_j x_j\cdot x_i+b\right]\leq 0$$

参数更新：

$$\alpha_i=\alpha_i+1$$

$$b=b+y_i$$

(4) 迭代，过程从略。

结果列于表2.2。

表2.2 例2.2 求解的迭代过程

k	0	1	2	3	4	5	6	7
		x_1	x_3	x_3	x_3	x_1	x_3	x_3
α_1	0	1	1	1	2	2	2	2
α_2	0	0	0	0	0	0	0	0
α_3	0	0	1	2	2	3	4	5
b	0	1	0	−1	0	−1	−2	−3

(4) $w = 2x_1 + 0x_2 - 5x_3 = (1, 1)^T$

$b = -3$

超平面分界面方程： $x^{(1)} + x^{(2)} - 3 = 0$

感知机模型为：

$$f(x) = \text{sign}(x^{(1)} + x^{(2)} - 3)$$

与原始形式学习得到的结果一致。