统计学习方法
作者：李航

第二章《感知机》讲解

演讲者：王昕毅

基本概念

- 感知机是二类分类的线性分类模型，其输入为实例的特征向量，输出为实例的类别，取+1或者-1二值。
- 感知机对应输入空间中将实例划分为正负两类的分离超平面，属于判别模型
- 感知机旨在求出将训练数据进行线性划分的分离超平面，为此导入基于误分类的损失函数，利用梯度下降对损失函数进行极小化，求得感知机模型
- 感知机分为原始形式和对偶形式

Ⅰ.感知机模型

定义

假设输入空间(特征向量)为$X \subseteq R^n$，输出空间为Y={-1, +1}。输入$x \in X$表示实例的特征向量，对应于输入空间的点；输出y∈Y表示示例的类别。由输入空间到输出空间的函数为

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

称为感知机。其中，参数w叫做权值向量**weight**，b称为偏置**bias**。$w \cdot x$表示w和x的**点积**

$$\sum_{i=1}^{m} w_i x_i = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

sign为符号函数，即

$$f(x) = \begin{cases} +1 & \text{if } x >= 0 \\ -1 & else \end{cases}$$

模型的假设空间

- 感知机模型的假设空间是定义在特征空间中的所有线性分类模型或线性分类器，即函数集合$\{f \mid f(x) = w \cdot x + b\}$

- 线性分类器：模型是参数的线性函数，分类平面是（超）平面
- 非线性分类器：模型分界面可以是曲面或者超平面的组合

感知机使用前提

- 数据是线性可分的，即存在一个超平面能将数据集的正实例点和负实例点完全地划分到超平面的两侧。

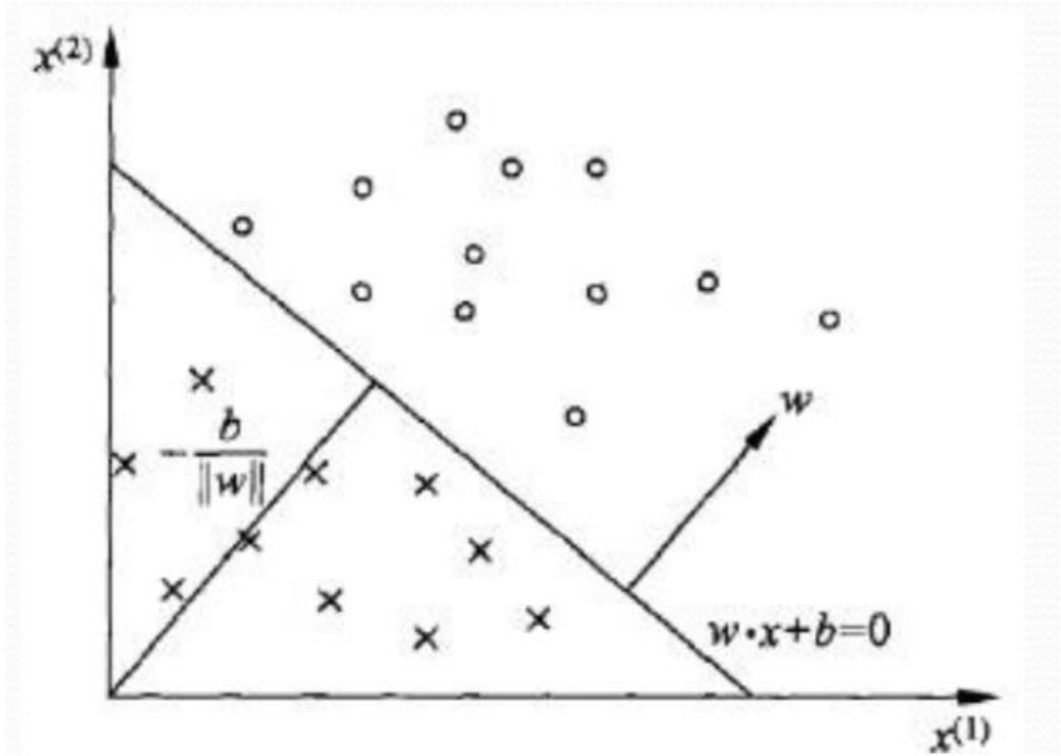
超平面简介

- 感知机有如下解释：线性方程 $w \cdot x + b = 0$ 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。这个超平面将特征空间划分为两个部分，位于两部分的点分别被分为正负两类。因此超平面S称为分离超平面

- 超平面是n维欧氏空间中余维度等于一的线性子空间（也就是必须是(n-1)维度）。这是平面中的直线、空间中的平面之推广（n大于3才被称为“超”平面）

- 性质：1.超平面可以将它所在空间分为两半，法向量指向的一边为正面，另一面为反面。2.超平面维度比被切分空间低一，参考平面中的直线，三维空间的平面

- 感知机模型



II .感知机学习策略

- 前提：数据集是线性可分的
- 目标：求得一个能够将训练集正实例点和负实例点完全正确分开的分离超平面。
- 为了找出这样的超平面，即确定感知机模型参数w，b，需要定义一个损失函数并将其极小化。

损失函数的选择

- 1.误分类点的总数

这样选择，损失函数不是参数w，b的连续可导函数，不易优化。

- 2.误分类点到超平面的总距离

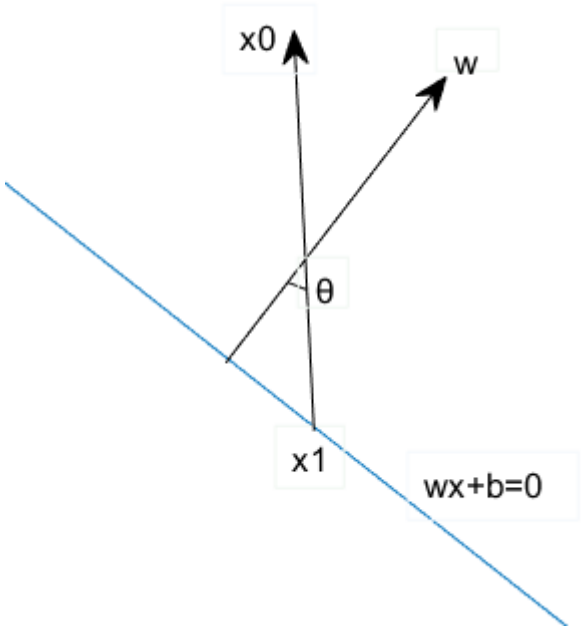
首先写出输入空间R^n中任意一点X_0到超平面S的距离：

$$\frac{1}{\|w\|}\left|w \cdot x_0 + b\right|$$

这里$\|w\|$是w的L_2范数。

$$\frac{1}{\|w\|}|w \cdot x_0 + b|$$

公式推导，其中x_0为空间中任意一点 x_1为超平面上的点，θ为向量x_0x_1与法向量w的夹角



$$d = \left|X_0X_1\right| \times \left|\cos\theta\right|$$

$$\left|\cos\theta\right| = \frac{\left|X_0X_1 \cdot W\right|}{\left|X_0X_1\right| \times \left|W\right|}$$

$$d = \frac{\left|X_0X_1 \cdot W\right|}{\left|X_0X_1\right| \times \left|W\right|} \times \left|X_0X_1\right| = \frac{\left|X_0X_1 \cdot W\right|}{\left|W\right|} = \frac{\left|\left(X_0 - X_1\right) \cdot W\right|}{\left|W\right|} \xrightarrow{\left|W\right| = \left\|W\right\|} \frac{\left|WX_0 - WX_1\right|}{\left\|W\right\|}$$

$$\xrightarrow{x_1\text{是超平面上的点,}wx_1+b=0} \frac{\left|WX_0 + b\right|}{\left\|W\right\|}$$

- 有了单个误分类点到超平面的距离 $\frac{1}{\|w\|}|w \cdot x_0 + b|$ 之后我们来计算误分类点到超平面的距离之和
- 对于误分类的数据(x_i,y_i)来说

$$-y_i(w \cdot x_i + b) > 0$$

成立. 因为当 $w \cdot x_i + b > 0$ 时, $y_i = -1$, 而当 $w \cdot x_i + b < 0$ 时, $y_i = +1$. 因此, 误分类点 x_i 到超平面 S 的距离是

$$-\frac{1}{\|w\|} y_i(w \cdot x_i + b)$$

这样, 假设超平面 S 的误分类点集合为 M, 那么所有误分类点到超平面 S 的总距离为

$$-\frac{1}{\|w\|} \sum_{x_i \in M} y_i(w \cdot x_i + b)$$

不考虑 $\frac{1}{\|w\|}$, 就得到感知机学习的损失函数①.

给定训练数据集

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

其中，$x_i\in\mathcal{X}=\mathbf{R}^n$，$y_i\in\mathcal{Y}=\{+1,-1\}$，$i=1,2,\cdots,N$．感知机 $\text{sign}(w\cdot x+b)$ 学习的损失函数定义为

$$L(w,b)=-\sum_{x_i\in M}y_i(w\cdot x_i+b) \tag{2.4}$$

其中 M 为误分类点的集合．这个损失函数就是感知机学习的经验风险函数．

显然，损失函数 $L(w,b)$ 是非负的．如果没有误分类点，损失函数值是 0．而且，误分类点越少，误分类点离超平面越近，损失函数值就越小．一个特定的样本点的损失函数：在误分类时是参数 w,b 的线性函数，在正确分类时是 0．因此，给定训练数据集 T，损失函数 $L(w,b)$ 是 w,b 的连续可导函数．

感知机学习的策略是在假设空间中选取使损失函数式（2.4）最小的模型参数 w,b，即感知机模型．

为什么可以不考虑$\frac{1}{\|w\|}$？

1.感知机不是求最优解，而是求一个超平面将数据集分开，最终目标是要使误分类点集变为空集，即是说最终目标是损失函数变为0，$\frac{1}{\|w\|}$不影响最终结果。
2. $\frac{1}{\|w\|}$不影响正负，不影响中间过程，感知机是误分类驱动的，这里误分类只是需要判断-y(w·x+b)的正负即可以，不需要计算具体距离

III .感知机学习算法

2.3 感知机学习算法

感知机学习问题转化为求解损失函数式（2.4）的最优化问题，最优化的方法是随机梯度下降法. 本节叙述感知机学习的具体算法，包括原始形式和对偶形式，并证明在训练数据线性可分条件下感知机学习算法的收敛性.

感知机学习算法的原始形式

感知机学习算法是对以下最优化问题的算法．给定一个训练数据集

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

其中，$x_i\in\mathcal{X}=\mathbf{R}^n$，$y_i\in\mathcal{Y}=\{-1,1\}$，$i=1,2,\cdots,N$，求参数$w,b$，使其为以下损失函数极小化问题的解

$$\min_{w,b}L(w,b)=-\sum_{x_i\in M}y_i(w\cdot x_i+b) \tag{2.5}$$

其中M为误分类点的集合．

感知机学习算法是误分类驱动的，具体采用随机梯度下降法（stochastic gradient descent）. 首先，任意选取一个超平面 w_0, b_0，然后用梯度下降法不断地极小化目标函数(2.5). 极小化过程中不是一次使 M 中所有误分类点的梯度下降，而是一次随机选取一个误分类点使其梯度下降.

随机选取一个误分类点 (x_i, y_i)，对 w, b 进行更新：

$$w \leftarrow w + \eta y_i x_i \tag{2.6}$$

$$b \leftarrow b + \eta y_i \tag{2.7}$$

式中 $\eta\ (0 < \eta \leqslant 1)$ 是步长，在统计学习中又称为学习率（learning rate）. 这样，通过迭代可以期待损失函数 $L(w, b)$ 不断减小，直到为 0. 综上所述，得到如下算法：

算法 2.1（感知机学习算法的原始形式）

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in\mathcal{X}=\mathbf{R}^n$，$y_i\in\mathcal{Y}=\{-1,+1\}$，$i=1,2,\cdots,N$；学习率 $\eta(0<\eta\leqslant 1)$；

输出：w,b；感知机模型 $f(x)=\text{sign}(w\cdot x+b)$.

（1）选取初值 w_0,b_0

（2）在训练集中选取数据 (x_i,y_i)

（3）如果 $y_i(w\cdot x_i+b)\leqslant 0$

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

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

（4）转至（2），直至训练集中没有误分类点. ■

这种学习算法直观上有如下解释：当一个实例点被误分类，即位于分离超平面的错误一侧时，则调整 w，b 的值，使分离超平面向该误分类点的一侧移动，以减少该误分类点与超平面间的距离，直至超平面越过该误分类点使其被正确分类.

算法 2.1 是感知机学习的基本算法，对应于后面的对偶形式，称为原始形式. 感知机学习算法简单且易于实现.

例

例 2.1 如图 2.2 所示的训练数据集，其正实例点是 $x_1=(3,3)^{\mathrm{T}}$，$x_2=(4,3)^{\mathrm{T}}$，负实例点是 $x_3=(1,1)^{\mathrm{T}}$，试用感知机学习算法的原始形式求感知机模型 $f(x)=\mathrm{sign}(w\cdot x+b)$．这里，$w=(w^{(1)},w^{(2)})^{\mathrm{T}}$，$x=(x^{(1)},x^{(2)})^{\mathrm{T}}$．

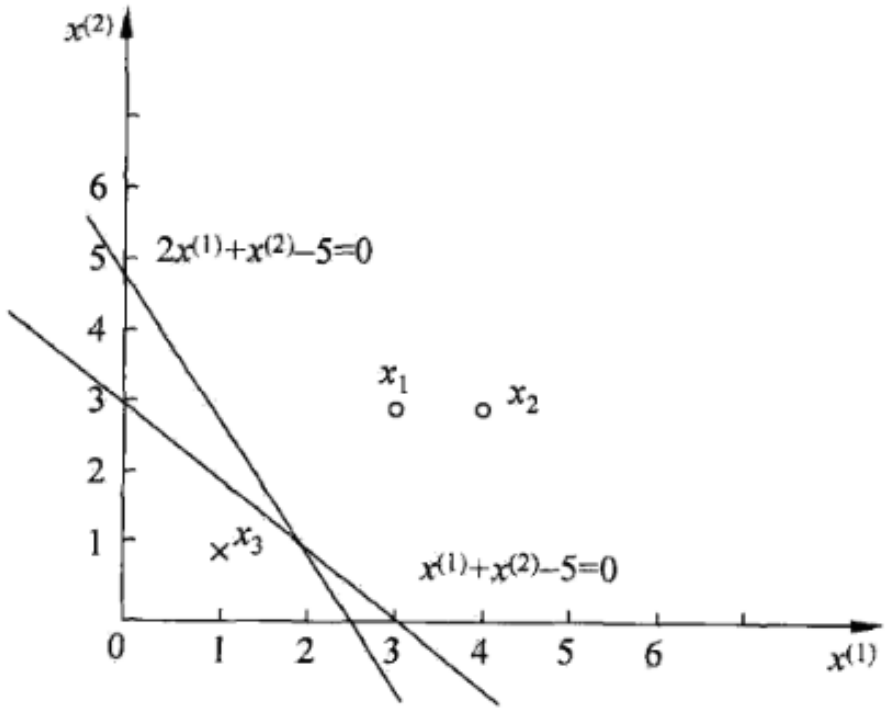


图 2.2 感知机示例

解 构建最优化问题：

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

按照算法 2.1 求解 w，b．$\eta=1$.

（1）取初值 $w_0=0$， $b_0=0$

（2）对 $x_1=(3,3)^{\mathrm{T}}$， $y_1(w_0\cdot x_1+b_0)=0$，未能被正确分类，更新 w,b

$$w_1=w_0+y_1x_1=(3,3)^{\mathrm{T}}，\quad b_1=b_0+y_1=1$$

得到线性模型

$$w_1\cdot x+b_1=3x^{(1)}+3x^{(2)}+1$$

（3）对 x_1,x_2，显然， $y_i(w_1\cdot x_i+b_1)>0$，被正确分类，不修改 w,b；

对 $x_3=(1,1)^{\mathrm{T}}$， $y_3(w_1\cdot x_3+b_1)<0$，被误分类，更新 w,b.

$$w_2=w_1+y_3x_3=(2,2)^{\mathrm{T}}，\quad b_2=b_1+y_3=0$$

得到线性模型

$$w_2\cdot x+b_2=2x^{(1)}+2x^{(2)}$$

如此继续下去，直到

$$w_7=(1,1)^{\mathrm{T}}，\quad b_7=-3$$

$$w_7\cdot x+b_7=x^{(1)}+x^{(2)}-3$$

对所有数据点 $y_i(w_7\cdot x_i+b_7)>0$，没有误分类点，损失函数达到极小.

分离超平面为 $x^{(1)}+x^{(2)}-3=0$

感知机模型为 $f(x)=\mathrm{sign}(x^{(1)}+x^{(2)}-3)$ ■

迭代过程见表 2.1.

表 2.1　例 2.1 求解的迭代过程

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

这是在计算中误分类点先后取 $x_1, x_3, x_3, x_3, x_1, x_3, x_3$ 得到的分离超平面和感知机模型. 如果在计算中误分类点依次取 $x_1, x_3, x_3, x_3, x_2, x_3, x_3, x_3, x_1, x_3, x_3$，那么得到的分离超平面是 $2x^{(1)}+x^{(2)}-5=0$.

可见，感知机学习算法由于采用不同的初值或选取不同的误分类点，解可以不同.

算法收敛性证明--为什么经过有限次迭代一定能找到超平面

现在证明，对于线性可分数据集感知机学习算法原始形式收敛，即经过有限次迭代可以得到一个将训练数据集完全正确划分的分离超平面及感知机模型.

为了便于叙述与推导，将偏置 b 并入权重向量 w，记作 $\hat{w}=(w^{\mathrm{T}},b)^{\mathrm{T}}$，同样也将输入向量加以扩充，加进常数 1，记作 $\hat{x}=(x^{\mathrm{T}},1)^{\mathrm{T}}$. 这样，$\hat{x}\in\mathbf{R}^{n+1}$，$\hat{w}\in\mathbf{R}^{n+1}$. 显然，$\hat{w}\cdot\hat{x}=w\cdot x+b$.

定理 2.1（Novikoff） 设训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$ 是线性可分的，其中 $x_i\in\mathcal{X}=\mathbf{R}^n$，$y_i\in\mathcal{Y}=\{-1,+1\}$，$i=1,2,\cdots,N$，则

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

$$y_i(\hat{w}_{\mathrm{opt}}\cdot\hat{x}_i)=y_i(w_{\mathrm{opt}}\cdot x_i+b_{\mathrm{opt}})\geqslant\gamma \tag{2.8}$$

（2）令 $R=\max\limits_{1\leqslant i\leqslant N}\|\hat{x}_i\|$，则感知机算法 2.1 在训练数据集上的误分类次数 k 满足不等式

$$k\leqslant\left(\frac{R}{\gamma}\right)^2 \tag{2.9}$$

证明 （1）由于训练数据集是线性可分的，按照定义 2.2，存在超平面可将训练数据集完全正确分开，取此超平面为 $\hat{w}_{opt} \cdot \hat{x} = w_{opt} \cdot x + b_{opt} = 0$，使 $\|\hat{w}_{opt}\| = 1$．由于对有限的 $i = 1, 2, \cdots, N$，均有

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

所以存在

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

使

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

为什么令 $\|\hat{w}_{opt}\| = 1$ ？

1.方便（2）的证明

2. $\|\hat{w}_{opt}\| = 1$不影响超平面，例$4X_1+4X_2+2=0$与$2X_1+2X_2+1=0$是同一个平面,对于任意W，$\left\| \frac{w}{\|w\|} \right\| = 1$

（2）感知机算法从 $\hat{w}_0=0$ 开始，如果实例被误分类，则更新权重．令 $\hat{w}_{k-1}$ 是

第 k 个误分类实例之前的扩充权重向量，即

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

则第 k 个误分类实例的条件是

$$y_i(\hat{w}_{k-1}\cdot\hat{x}_i)=y_i(w_{k-1}\cdot x_i+b_{k-1})\leqslant 0 \tag{2.10}$$

若 (x_i,y_i) 是被 $\hat{w}_{k-1}=(w_{k-1}^{\mathrm{T}},b_{k-1})^{\mathrm{T}}$ 误分类的数据，则 w 和 b 的更新是

$$w_k\leftarrow w_{k-1}+\eta y_i x_i$$
$$b_k\leftarrow b_{k-1}+\eta y_i$$

即

$$\hat{w}_k=\hat{w}_{k-1}+\eta y_i\hat{x}_i \tag{2.11}$$

下面推导两个不等式：

（1）

$$\hat{w}_k \cdot \hat{w}_{\text{opt}} \geqslant k\eta\gamma \qquad (2.12)$$

由式 (2.11) 及式 (2.8) 得

$$\begin{aligned}\hat{w}_k \cdot \hat{w}_{\text{opt}} &= \hat{w}_{k-1} \cdot \hat{w}_{\text{opt}} + \eta y_i \hat{w}_{\text{opt}} \cdot \hat{x}_i \\ &\geqslant \hat{w}_{k-1} \cdot \hat{w}_{\text{opt}} + \eta\gamma\end{aligned}$$

由（1）得 $y_i(\hat{w}_{\text{opt}} \cdot \hat{x}_i) = y_i(w_{\text{opt}} \cdot x_i + b_{\text{opt}}) \geqslant \gamma$

由此递推即得不等式 (2.12)

$$\hat{w}_k \cdot \hat{w}_{\text{opt}} \geqslant \hat{w}_{k-1} \cdot \hat{w}_{\text{opt}} + \eta\gamma \geqslant \hat{w}_{k-2} \cdot \hat{w}_{\text{opt}} + 2\eta\gamma \geqslant \cdots \geqslant k\eta\gamma$$

（2）

$$\|\hat{w}_k\|^2 \leqslant k\eta^2 R^2 \qquad (2.13)$$

由式 (2.11) 及式 (2.10) 得

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

（Xi,Yi）是被Wk-1误分类的数据，故 $2\eta y_i \hat{w}_{k-1} \cdot \hat{x}_i < 0$

结合不等式 (2.12) 及式 (2.13) 即得

$$k\eta\gamma \leqslant \hat{w}_k \cdot \hat{w}_{\text{opt}} \leqslant \left\|\hat{w}_k\right\| \left\|\hat{w}_{\text{opt}}\right\| \leqslant \sqrt{k\eta}R$$

$$k^2\gamma^2 \leqslant kR^2$$

于是

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

$$\cos\theta = \frac{\vec{W}1 \cdot \vec{W}2}{\left|\vec{W}1\right| \times \left|\vec{W}2\right|} \leq 1 \rightarrow \left|\vec{W}1\right| \times \left|\vec{W}2\right| \geq$$

$$\vec{W}1 \cdot \vec{W}2 \xrightarrow{|w|=\|w\|} \left\|\vec{W}1\right\| \times \left\|\vec{W}2\right\| \geq \vec{W}1 \cdot \vec{W}2$$

■

定理表明，误分类的次数 k 是有上界的，经过有限次搜索可以找到将训练数据完全正确分开的分离超平面．也就是说，当训练数据集线性可分时，感知机学习算法原始形式迭代是收敛的．但是例 2.1 说明，感知机学习算法存在许多解，这些解既依赖于初值的选择，也依赖于迭代过程中误分类点的选择顺序

感知机算法的对偶形式

对偶形式的基本想法是，将 w 和 b 表示为实例 x_i 和标记 y_i 的线性组合的形式，通过求解其系数而求得 w 和 b . 不失一般性，在算法 2.1 中可假设初始值 w_0, b_0 均为 0. 对误分类点 (x_i, y_i) 通过

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

逐步修改 w, b，设修改 n 次，则 w, b 关于 (x_i, y_i) 的增量分别是 $\alpha_i y_i x_i$ 和 $\alpha_i y_i$，这里 $\alpha_i = n_i \eta$. 这样，从学习过程不难看出，最后学习到的 w, b 可以分别表示为

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{2.14}$$

$$b = \sum_{i=1}^{N} \alpha_i y_i \tag{2.15}$$

这里，$\alpha_i \geqslant 0$，$i = 1, 2, \cdots, N$，当 $\eta = 1$ 时，表示第 i 个实例点由于误分而进行更新的次数. 实例点更新次数越多，意味着它距离分离超平面越近，也就越难正确分类. 换句话说，这样的实例对学习结果影响最大.

下面对照原始形式来叙述感知机学习算法的对偶形式.

算法 2.2（感知机学习算法的对偶形式）

输入：线性可分的数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in\mathbf{R}^n$，$y_i\in\{-1,+1\}$，$i=1,2,\cdots,N$；学习率 η（$0<\eta\leqslant 1$）；

输出：α,b；感知机模型 $f(x)=\text{sign}\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x+b\right)$.

其中 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_N)^{\mathrm{T}}$.

（1）$\alpha\leftarrow 0$，$b\leftarrow 0$

（2）在训练集中选取数据 (x_i,y_i)

（3）如果 $y_i\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x_i+b\right)\leqslant 0$

可以通过查gram矩阵得到 $x_j\cdot x_i$ 的值

$$\alpha_i\leftarrow\alpha_i+\eta$$

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

（4）转至（2）直到没有误分类数据. ■

对偶形式中训练实例仅以内积的形式出现. 为了方便，可以预先将训练集中实例间的内积计算出来并以矩阵的形式存储，这个矩阵就是所谓的 Gram 矩阵（Gram matrix）

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

例 2.2 数据同例 2.1，正样本点是 $x_1=(3,3)^{\mathrm{T}}$，$x_2=(4,3)^{\mathrm{T}}$，负样本点是 $x_3=(1,1)^{\mathrm{T}}$，试用感知机学习算法对偶形式求感知机模型.

解 按照算法 2.2，

（1）取 $\alpha_i=0$，$i=1,2,3$，$b=0$，$\eta=1$

（2）计算 Gram 矩阵

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

（3）误分条件

$$y_i\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x_i+b\right)\leqslant 0$$

参数更新

$$\alpha_i\leftarrow\alpha_i+1,\quad b\leftarrow b+y_i$$

（4）迭代．过程从略，结果列于表 2.2.

（5）

$$w=2x_1+0x_2-5x_3=(1,1)^{\mathrm{T}}$$

$$b=-3$$

分离超平面

$$x^{(1)}+x^{(2)}-3=0$$

感知机模型

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

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

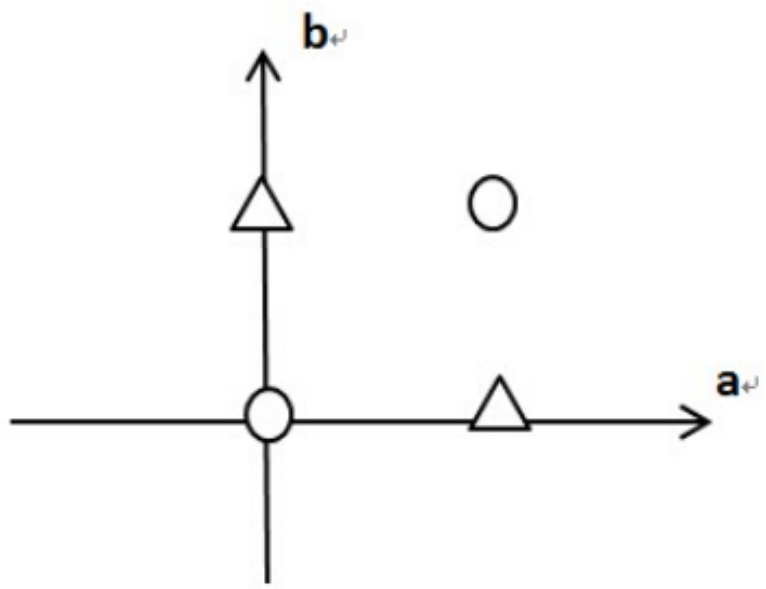
对照例 2.1，结果一致，迭代步骤也是互相对应的.

与原始形式一样，感知机学习算法的对偶形式迭代是收敛的，存在多个解.

课后习题

2.1 Minsky 与 Papert 指出：感知机因为是线性模型，所以不能表示复杂的函数，如异或（XOR）. 验证感知机为什么不能表示异或.

异或逻辑图像



显而易见找不到一条直线将圆和三角形分开，具体证明见下页

对于 a XOR b ，其真值表为：

a	b	y
0	0	0
0	1	1
1	0	1
1	1	0

从数据集线性可分性的角度证明XOR逻辑是非线性的：

设数据集T为：

$$\mathrm{T}=\{(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\}$$

其中：

$$\mathrm{x}=(a,b)^T,y\in\{0,1\}$$

假设T是线性可分的，则存在超平面 $\mathrm{S}:\mathrm{w}\cdot\mathrm{x}+\mathrm{b}=0$ 对T中的4个实例正确分类

则

$$\begin{cases}\mathrm{w}\cdot x_1+\mathrm{b}<0\\ \mathrm{w}\cdot x_2+\mathrm{b}>0\\ \mathrm{w}\cdot x_3+\mathrm{b}>0\\ \mathrm{w}\cdot x_4+\mathrm{b}<0\end{cases}$$

注：这里的大于小于零与y的取值无关，它与参考系有关，表示两个类别，所以也并不要求y的取值是正1和负1.

设 $\mathrm{w}=(u,v)^T$

对上式进行化简

$$\begin{cases}b<0\\ v+b>0\\ u+b>0\\ u+v+b<0\end{cases}$$

四个式子相互之间是矛盾的，所以异或逻辑是非线性问题，得证。

2.2 模仿例题 2.1，构建从训练数据集求解感知机模型的例子.

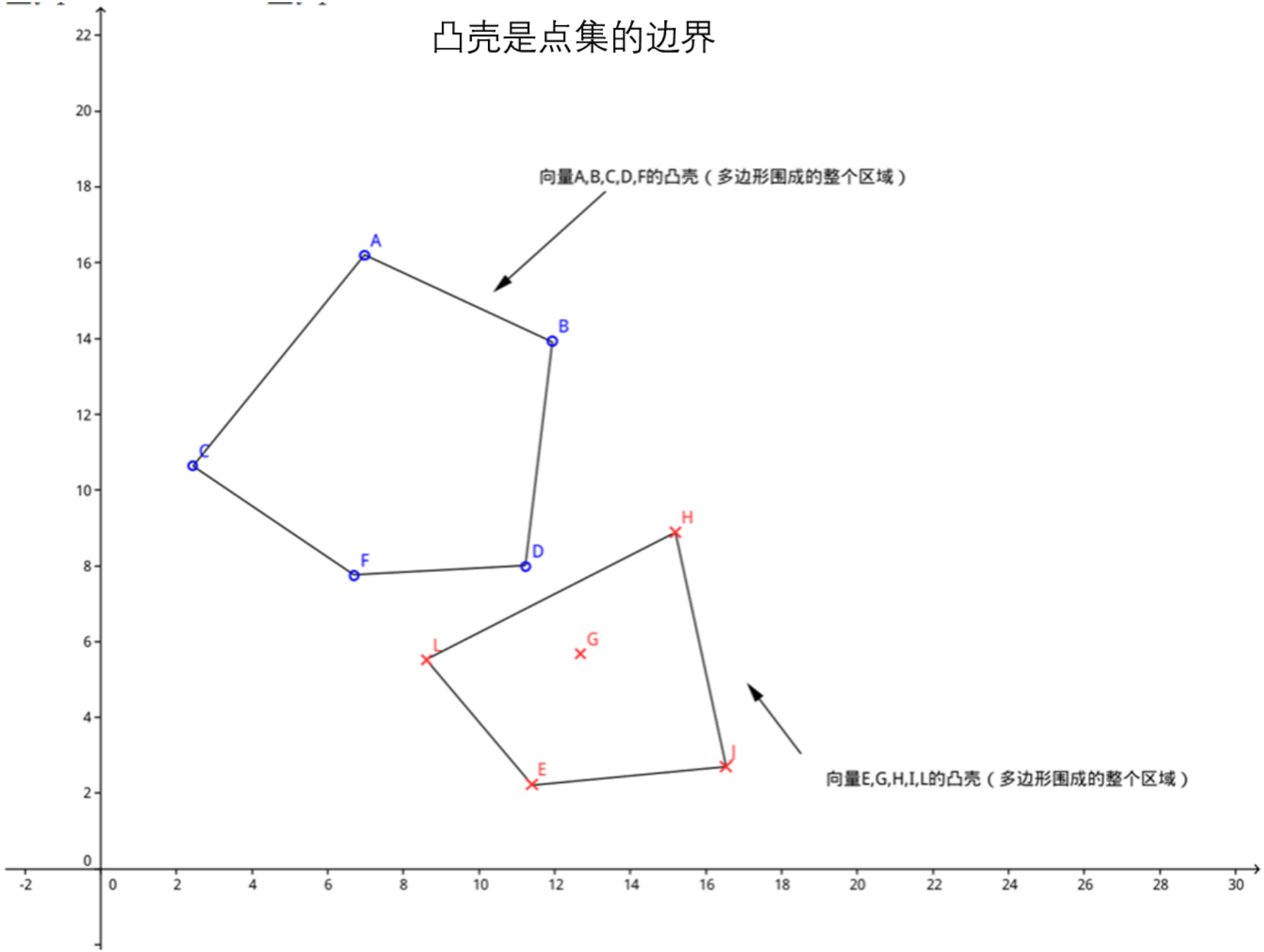
答案：略

2.3 证明以下定理：样本集线性可分的充分必要条件是正实例点集所构成的凸壳[®]与负实例点集所构成的凸壳互不相交.

设集合 $S \subset \mathbf{R}^n$ 是由 $\mathbf{R}^n$ 中的 k 个点所组成的集合，即 $S=\{x_1,x_2,\cdots,x_k\}$. 定义 S 的凸壳 $\mathrm{conv}(S)$ 为

$$\mathrm{conv}(S)=\left\{x=\sum_{i=1}^{k}\lambda_i x_i \,\middle|\, \sum_{i=1}^{k}\lambda_i=1, \lambda_i \geqslant 0, i=1,2,\cdots,k\right\}.$$

凸壳是点集的边界


向量A,B,C,D,F的凸壳（多边形围成的整个区域）
A
B
C
D
F
H
G
L
E
I
向量E,G,H,I,L的凸壳（多边形围成的整个区域）

定理

样本集线性可分的充分必要条件使正实例点集构成的凸壳与负实例点集所构成的凸壳互不相交。

必要性：线性可分->凸壳不相交

设数据集T中的正例点集为S_+，S_+的凸壳为$conv(S_+)$，负实例点集为S_-，S_-的凸壳为$conv(S_-)$，若T是线性可分的，则存在一个超平面:

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

能够将S_+和S_-完全分离。假设对于所有的正例点x_i，有:

$$w \cdot x_i + b = \varepsilon_i$$

易知$\varepsilon_i > 0, i = 1, 2, \cdots, |S_+|$。若$conv(S_+)$和$conv(S_-)$相交，即存在某个元素$s$，同时满足$s \in conv(S_+)$和$s \in conv(S_-)$。对于$conv(S_+)$中的元素$s^+$有

$$w \cdot s^+ = w \cdot \sum_{i=1}^{k} \lambda_i x_i = \sum_{i=1}^{k} \lambda_i (\varepsilon_i - b) = \sum_{i=1}^{k} \lambda_i \varepsilon_i - b$$

同理对于$conv(S_-)$中的元素s^-有$w \cdot s^- + b = \sum_{i=1}^{k} \lambda_i \varepsilon_i < 0$

因此$w \cdot s^+ + b = \sum_{i=1}^{k} \lambda_i \varepsilon_i > 0$，同理对于$S_+$中的元素$s^-$有$w \cdot s^+ + b = \sum_{i=1}^{k} \lambda_i \varepsilon_i < 0$，那么由于$s \in conv(S_+)$且$s \in conv(S_-)$则$w \cdot s + b = \sum_{i=1}^{k} \lambda_i \varepsilon_i > 0$且$w \cdot s + b = \sum_{i=1}^{k} \lambda_i \varepsilon_i < 0$明显推出矛盾，因此$conv(S_+)$和$conv(S_-)$必不相交。从而推出必要性。

充分性：凸壳不相交->线性可分

设数据集T中的正例点集为S_+，S_+的凸壳为$conv(S_+)$，负实例点集为S_-，S_-的凸壳为$conv(S_-)$，且$conv(S_+)$与$conv(S_-)$不相交，定义两个点x_1, x_2的距离为

$$dist(x_1, x_2) = ||x_1 - x_2||_2 = \sqrt{(x_1 - x_2) \cdot (x_1 - x_2)}$$

定义$conv(S_+)$与$conv(S_-)$的距离为

$$dist(conv(S_+), conv(S_-)) = \min ||s_+ - s_-||,\ s_+ \in conv(S_+), s_- \in conv(S_-)$$

设$x_+ \in conv(S_+)$，$x_- \in conv(S_-)$且$dist(x_+, x_-) = dist(conv(S_+), conv(S_-))$。则对于任意正例点$x$有$dist(x, x_-) \geq dist(x_+, x_-)$。同理，对于所有的负例点$x$有$dist(x, x_+) \geq dist(x, x_-)$。存在超平面

此处应为 dist(x_+, x_-)

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

其中

$$w = x_+ - x_-$$
$$b = -\frac{x_+ \cdot x_+ - x_- \cdot x_-}{2}$$

此处应为 x ≠ x_+

则对于所有的正例点x(易知$w \cdot x_+ + b > 0$，因此若x_+属于正例点，则令$x \neq x$)，

$$\begin{aligned} w \cdot x + b &= (x_+ - x_-) \cdot x - \frac{x_+ \cdot x_+ - x_- \cdot x_-}{2} \\ &= x_+ \cdot x - x_- \cdot x - \frac{x_+ \cdot x_+ - x_- \cdot x_-}{2} \\ &= \frac{||x_- - x||_2^2 - ||x_+ - x||_2^2}{2} \\ &= \frac{dist(x, x_-)^2 - dist(x, x_+)^2}{2} \end{aligned}$$

因此对所有的正例点，$w \cdot x + b > 0$成立。同理，对所有 的负例点，$w \cdot x + b < 0$成立。至此，充分性得证。

完

谢谢观看！