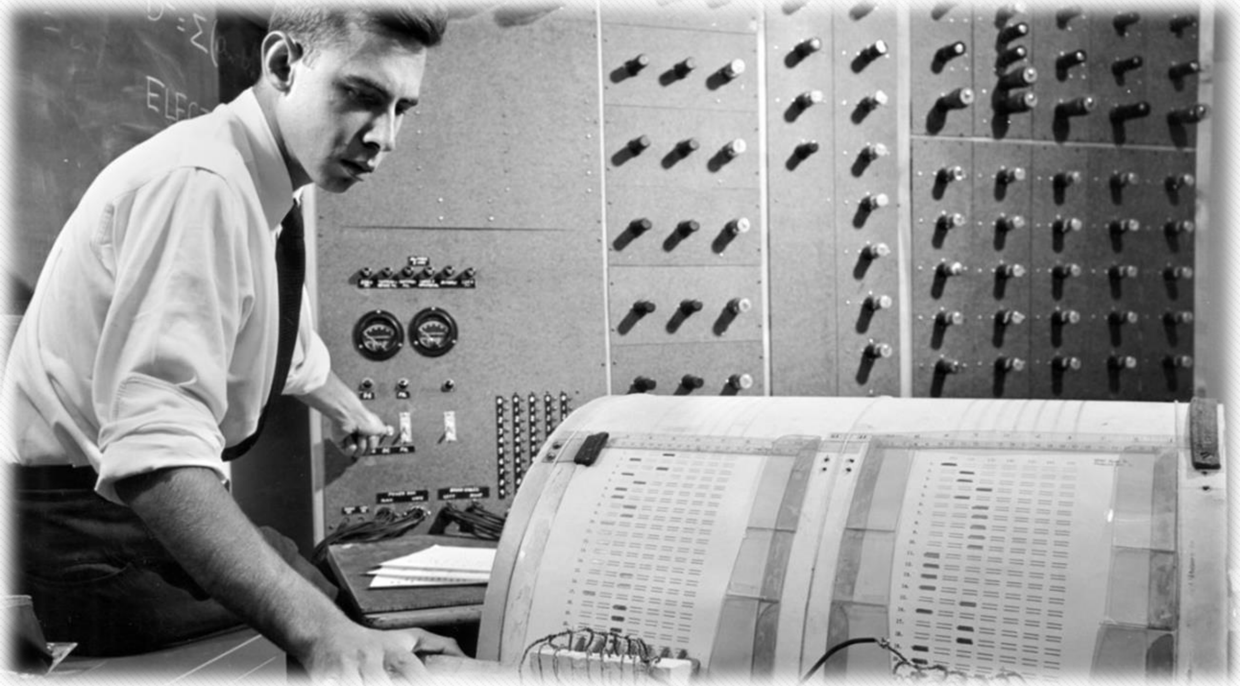
第二章：感知机
Ch2: Perceptron



感知机(Perceptron)

1943年，神经科学家麦卡洛克(W.S. McCulloch) 和数学家皮兹（W. Pitts）建立了神经网络和数学模型（M-P模型）。M-P模型是按照生物神经元的结构和工作原理构造出来的一个抽象和简化了的模型，即所谓的“模拟大脑”，人工神经网络的大门由此开启

1949年，Hebb规则：神经网络的学习过程最终是发生在神经元之间的突触部位，突触的联结强度随着突触前后神经元的活动而变化，变化的量与两个神经元的活性之和成正比

1957年，结合M-P模型和Hebb规则，**Frank Rosenblatt**提出感知机模型

Mark 1 Perceptron

- 首台人工神经网络机器
- 400个光电管阵列
- 可对20x20像素的图片正确分类

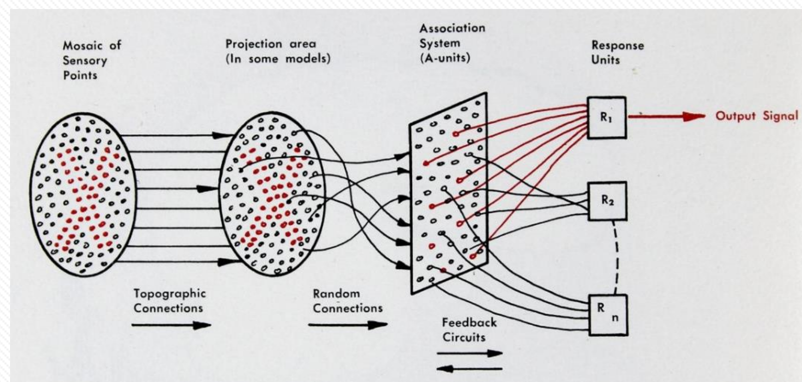


FIG. 2 — Organization of a perceptron.

https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

感知机(Perceptron)

1958 New York Times...

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of $100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.

"海军透露了一种电子计算机的雏形，它将能够走路、说话、看、写、自我复制并感知到自己的存在……据预测，不久以后，感知器将能够识别出人并叫出他们的名字，立即把演讲内容翻译成另一种语言并写下来。"

感知机模型定义

假设：输入空间(特征空间)是 $\mathrm{X} \subseteq \mathbf{R}^n$，输出空间是 $\mathrm{Y} = \{+1, -1\}$

输入$\boldsymbol{x} \in \boldsymbol{X}$ 表示实例的**特征向量**（feature vector），对应于输入空间（特征空间）的点，输出 $\boldsymbol{y} \in \boldsymbol{Y}$ 表示实例的**类别**

感知机定义为由输入空间到输出空间的函数：

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

$\boldsymbol{w} \subseteq \mathbf{R}^n$ **权值向量**（weight vector）

$b \subseteq \mathrm{R}$ 偏置（bias）

$w \cdot x$ 向量内积

$\mathrm{sign}(x)$ 为符号函数：$sign(x) = \begin{cases} +1, x \geq 0 \\ -1, x < 0 \end{cases}$

$$w = (w^{(1)}, w^{(2)}, \ldots, w^{(n)})^T$$

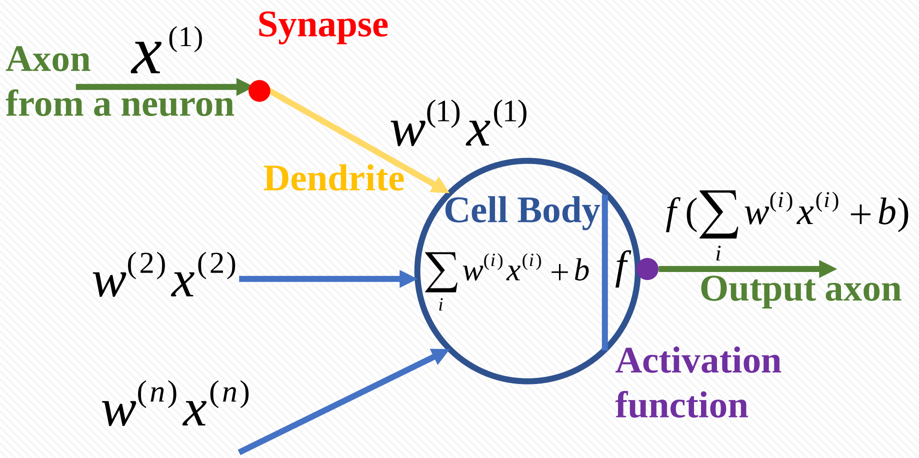
$$x = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})^T$$

感知机是二分类的线性分类器模型，是神经网络与支持向量机的基础

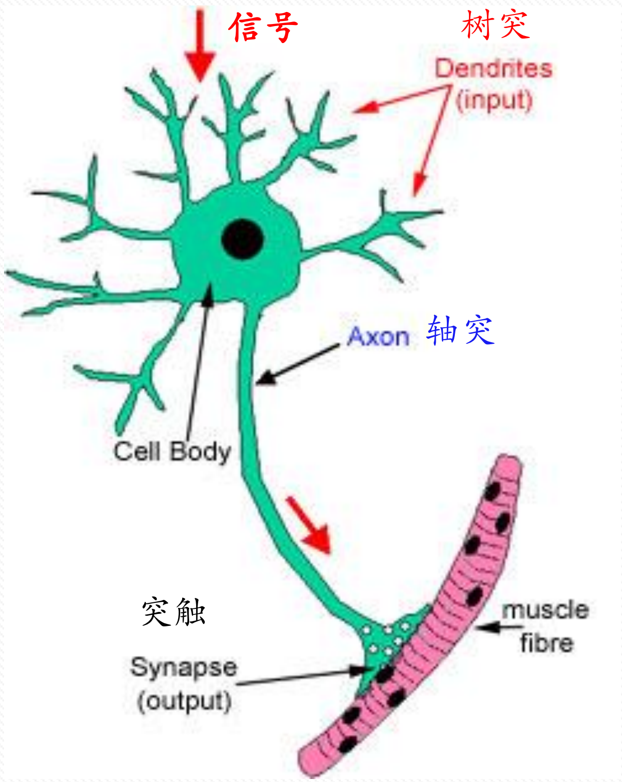
感知机即人工神经元

$$w = (w^{(1)}, w^{(2)}, ..., w^{(n)})^T$$

$$x = (x^{(1)}, x^{(2)}, ..., x^{(n)})^T$$



感知机



神经元

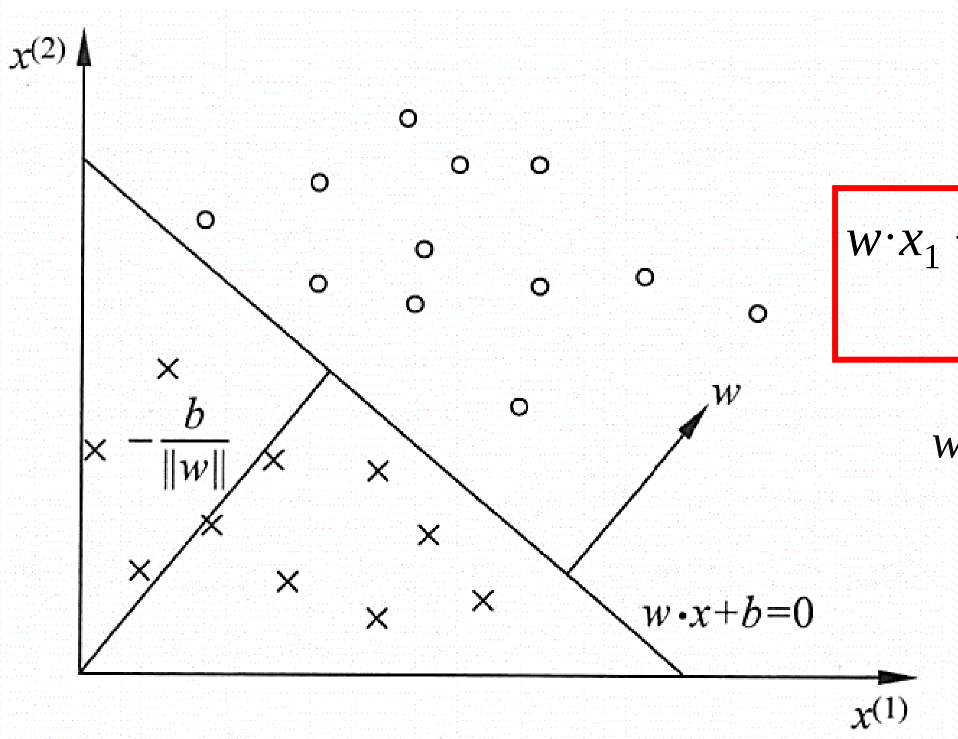
$\boxed{\text{形似且神似！}}$

感知机模型几何解释

线性方程：$w \cdot x + b = 0$ 对应于**分离超平面** S （separating hyperplane）

分离正类 $w \cdot x + b > 0$ 和负类 $w \cdot x + b < 0$

w为法向量，b为截距。

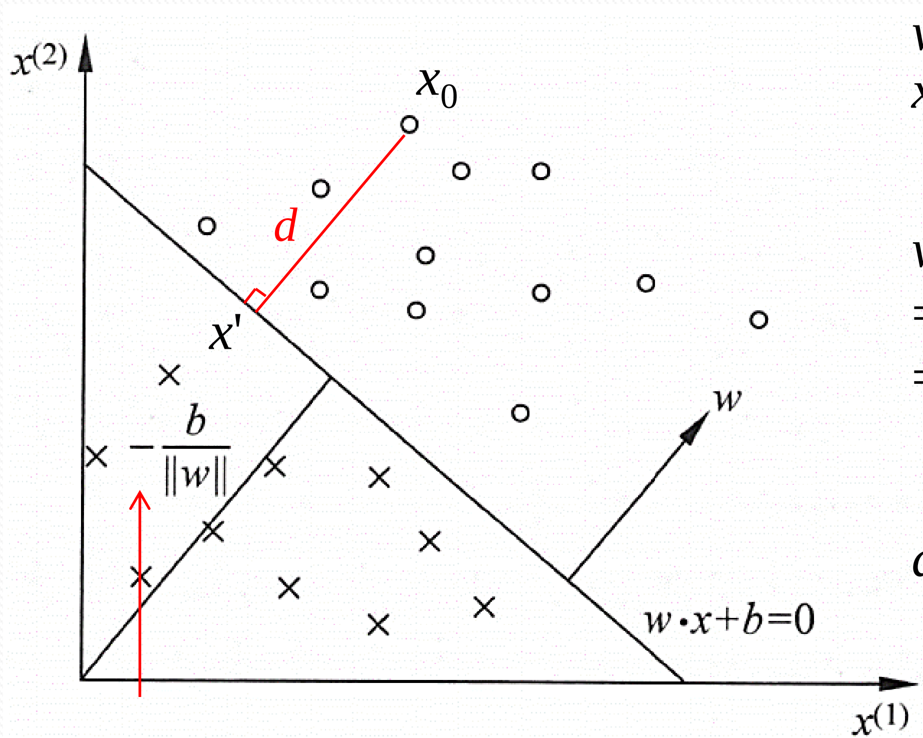


$$\boxed{\begin{gathered} w \cdot x_1 + b = w \cdot x_2 + b = 0 \\ w \cdot (x_1 - x_2) = 0 \end{gathered}}$$

w 为超平面**法向**

感知机模型几何解释

任意点x_0到超平面$w \cdot x + b = 0$**的距离**d



$w \cdot x' + b = 0$

$x_0 = x' + \lambda w$

$wx_0 + b = w(x' + \lambda w) + b$

$= wx' + b + w \cdot \lambda w$

$= w \cdot \lambda w$

$d = |x_0 - x'| = |\lambda w| = \frac{|wx_0 + b|}{||w||}$

原点到超平面距离

感知机机器学习

学习：已知数据，求模型参数

给定训练数据集：

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

其中，$x_i \in \mathrm{X} \subseteq \mathbf{R}^{\mathbf{n}}, y_i \in \mathrm{Y}=\{+1,-1\}, i=1,2,\cdots,N$

求感知机模型：

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

如果感知机模型存在，则称数据集**线性可分**（linearly separable）

机器学习=模型+策略+**算法**

感知机学习策略

损失函数：

自然选择：误分类点的数目？损失函数不是w, b 连续可导，不宜优化（ × ）

另一选择：误分类点到超平面的总距离 （√）

误分类点：$y_i(w\cdot x_i+b)<0$（按定义，正确分类时 y_i 和$w\cdot x_i+b$同号）

误分类点总距离： $\sum_{x_i\in M}\frac{|w\cdot x_i+b|}{||w||}=\sum_{x_i\in M}\frac{|y_i(w\cdot x_i+b)|}{||w||}=\sum_{x_i\in M}\frac{-y_i(w\cdot x_i+b)}{||w||}$

进一步简化： $\sum_{x_i\in M}\frac{-y_i(w\cdot x_i+b)}{||w||}$=-$\frac{1}{||w||}\sum_{x_i\in M}y_i(w\cdot x_i+b)$（$M$为误分类点总数）

感知机学习损失函数： $L(w,b)$=-$\sum_{x_i\in M}y_i(w\cdot x_i+b)$（可通过缩放使 $||w||=1$）

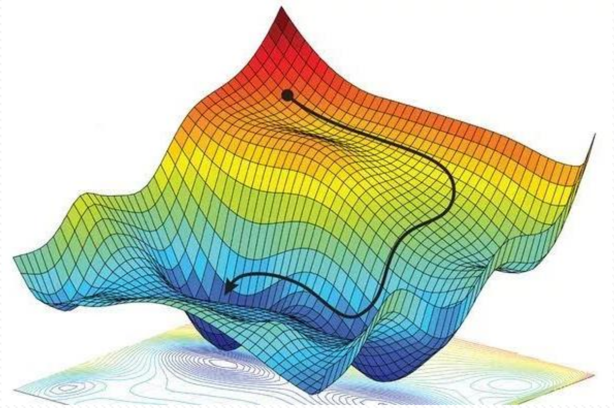
感知机学习策略： $\min_{w,b}L(w,b)$=-$\sum_{x_i\in M}y_i(w\cdot x_i+b)$

机器学习=模型+策略+算法

感知机学习算法

求解最优化问题：

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



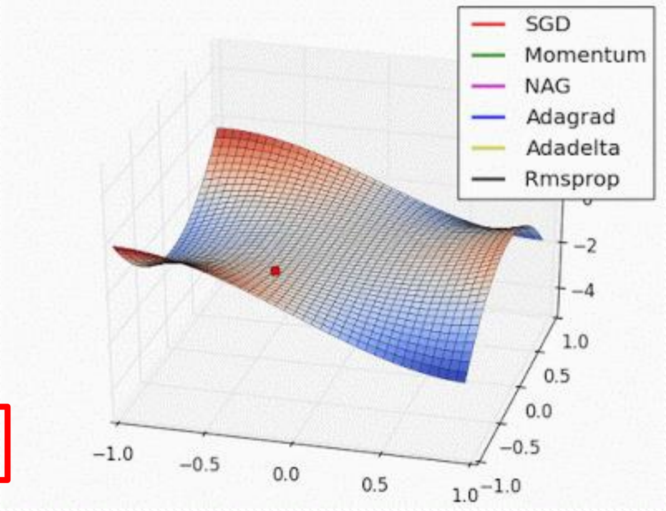
随机梯度下降法 (stochastic gradient descent)

首先任意选择一个超平面，w，b，然后不断极小化目标函数

损失函数L的梯度：$\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i, \nabla_b L(w,b) = -\sum y_i$

随机选取任意一个误分类点更新：

$$w \leftarrow w + \eta y_i x_i, b \leftarrow b + \eta y_i$$ **(η为学习率)**



机器学习=模型+策略+算法

感知机学习算法原始形式

感知机算法原始形式

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$

输出：w, b；感知机模型$f(x)=sign(w\cdot x+b)$

（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$y_i(w\cdot x_i+b)\leq 0$

$w = w + \eta y_i x_i$；

$b = b + \eta y_i$（η为学习率）

（4）转至（2）直到 M 为空

直观解释：当一个实例点被误分类，即位于分类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的距离，直至超平面越过该误分类点使其被正确分类

感知机学习原始算法Python实现

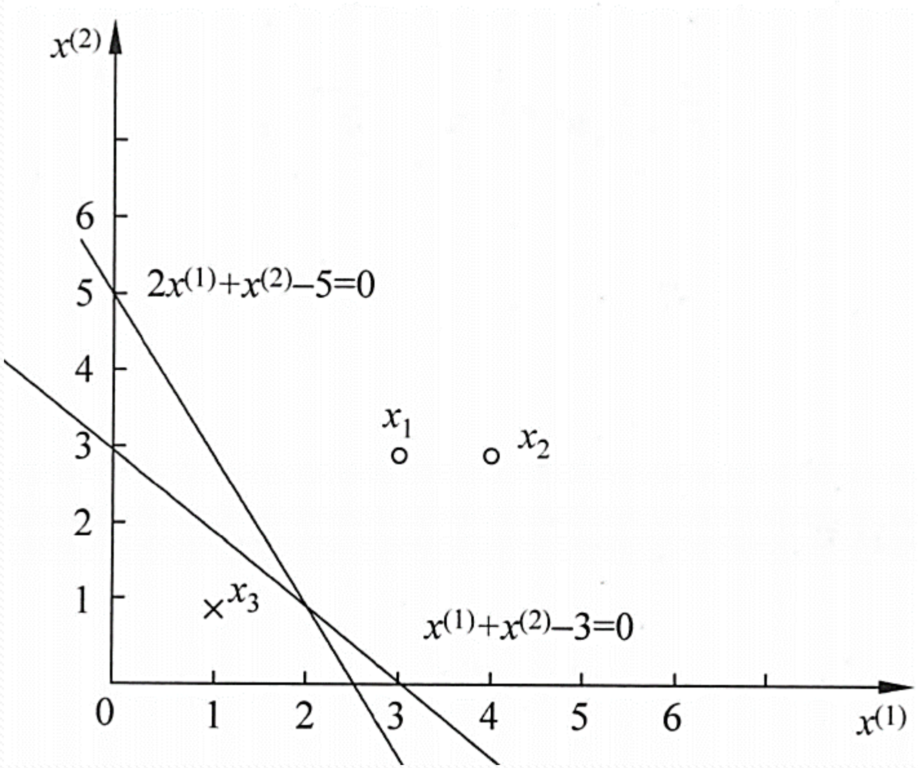
```
def trainPerceptron(data_mat, label_mat, eta):
        """
        data_mat: 训练数据
        lable_mat: 特征标签
        eta: 学习速率
        Return
        w: 权值
        b: 偏置
        """
    m, n = data_mat.shape
    w = np.zeros(n)
    b = 0
    flag = True
    while flag:
        for i in range(m):
            if label_mat[i] * (np.dot(w, data_mat[i]) + b) <= 0:
                w = w + eta * label_mat[i] * data_mat[i].T
                b = b + eta * label_mat[i]
                print("weight, bias: ", end="")
                print(w, end="  ")
                print(b)
                flag = True
                break
            else:
                flag = False
    return w, b
```

感知机学习算法-实例

例题2-1

已知正例：$x_1 = (3,3)^T, x_2 = (4,3)^T$，负例 $x_3 = (1,1)^T$

求感知机模型: $f(x) = sign(w \cdot x + b)$



感知机学习算法-实例

解： 构建优化问题：$\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b)$

正例：$x_1 = (3,3)^T, x_2 = (4,3)^T$
负例：$x_3 = (1,1)^T$

求解：w，b，$\eta=1$

（1）取初值 $w_0 = 0$，$b_0 = 0$

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

$$w_1 = w_0 + y_1 x_1 = (3,3)^T, \quad b_1 = b_0 + y_1 = 1$$

得线性模型： $w_1 \cdot x + b_1 = 3x^{(1)} + 3x^{(2)} + 1$

（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 $x_3 = (1,1)^T$，$y_3(w_1 \cdot x_3 + b_1) < 0$，被误分类，

$$w_2 = w_1 + y_3 x_3 = (2,2)^T, \quad b_2 = b_1 + y_3 = 0$$

得到线性模型： $w_2 \cdot x + b_2 = 2x^{(1)} + 2x^{(2)}$

如此继续下去： $w_7 = (1,1)^T$，$b_7 = -3$

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

分离超平面： $x^{(1)} + x^{(2)} - 3 = 0$

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

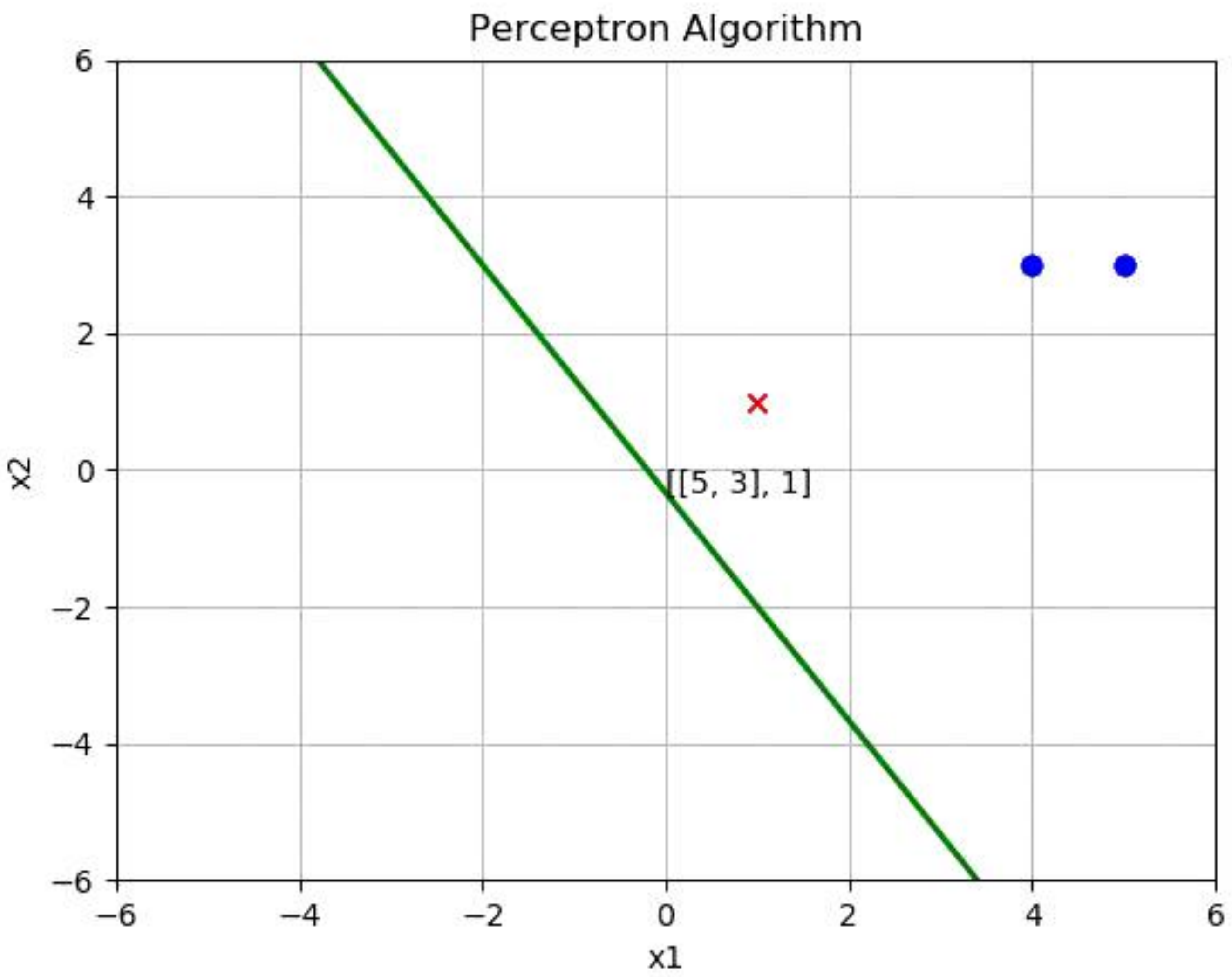
问题：分离超平面唯一吗？

迭代次数	误分类点	w	b	$w \cdot x + b$
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	$x^{(1)} + x^{(2)} - 1$
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	$x^{(1)} + x^{(2)} - 3$

感知机学习算法（计算机实例演示）

训练数据集T

$x^{(1)}$	$x^{(2)}$	y
5	3	1
4	3	1
1	1	-1



问：会不会存在无法收敛的场合？

算法收敛性

证明经过**有限次**迭代可以得到一个将训练数据集（注：**线性可分**）完全正确划分的分离超平面及感知机模型。

将b并入权重向量w，记做 $\hat{w}=(w^{\mathrm{T}},b)^{\mathrm{T}}$；将1并入输入向量$x$，记做 $\hat{x}=(x^{\mathrm{T}},1)^{\mathrm{T}}$

则有，$\hat{x}\in\mathbf{R}^{n+1}$，$\hat{w}\in\mathbf{R}^{n+1}$ $\quad \hat{w}\cdot\hat{x}=w\cdot x+b$

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

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

算法收敛性

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 $\hat{w}_{\text{opt}} \cdot \hat{x} = w_{\text{opt}} \cdot x + b_{\text{opt}} = 0$，使 $\|\hat{w}_{\text{opt}}\| = 1$。由于对有限的 $i = 1, 2, \cdots, N$，均有

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

所以存在

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

使

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

算法收敛性

证明(2)　（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

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

感知机算法从$w_0=0, b_0=0$开始，即$\hat{w}=0$ 。设第k次被超平面$w_{k-1}\cdot x+b_{k-1}=0$ 判为误分类的点为（x_i, y_i）满足：

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

则对w_k，b_k更新，

$$\boxed{\begin{aligned}w_k&\leftarrow w_{k-1}+\eta y_i x_i\\ b_k&\leftarrow b_{k-1}+\eta y_i\end{aligned}}\Rightarrow\hat{w}_k=\hat{w}_{k-1}+\eta y_i\hat{x}_i$$

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

证明1

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

递推有，$\hat{w}_k\bullet\hat{w}_{\rm opt}\geqslant\hat{w}_{k-1}\bullet\hat{w}_{\rm opt}+\eta\gamma\geqslant\hat{w}_{k-2}\bullet\hat{w}_{\rm opt}+2\eta\gamma\geqslant\cdots\geqslant k\eta\gamma$

即，$\hat{w}_k\bullet\hat{w}_{\rm opt}\geqslant k\eta\gamma$

算法收敛性

证明(2)

$$\|a+b\|^2=(a+b)\cdot(a+b)=a^2+2ab+b^2=\|a\|^2+2ab+\|b\|^2$$

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

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

$$y_i(\hat{w}_{k-1}\bullet\hat{x}_i)\leqslant 0$$

柯西不等式

$$k\eta\gamma\leqslant\boxed{\hat{w}_k\bullet\hat{w}_{\mathrm{opt}}\leqslant\|\hat{w}_k\|\ \|\hat{w}_{\mathrm{opt}}\|}\leqslant\sqrt{k}\eta R$$

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

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

误分类的次数k是有上界的，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。

感知机学习算法对偶形式

感知机算法对偶形式

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

具体说来，如果对误分类点i逐步修改w，b修改了n次，则w，b关于i 的增量分别为 $\alpha_i y_i x_i$ 和 $\alpha_i y_i$ ，这里$\alpha_i = n_i \eta$ ，n_i是点（x_i ,y_i）被误分的次数。则最终求解到的参数w和b分别表示为：

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

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

最后学习得到的w和b

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

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

就是用α去记录每个$y_i x_i$要加多少次，最后一次加上去

感知机学习算法对偶形式

感知机学习算法的对偶形式算法

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 $\eta\,(0<\eta\leqslant 1)$；

输出：α,b；感知机模型 $f(x)=\operatorname{sign}\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x+b\right)$，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 $y_i\left(\sum_{j=1}^{N}\alpha_j y_j \boxed{x_j\cdot x_i}+b\right)\leqslant 0$

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

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

(4) 转至 (2) 直到没有误分类数据。

$G=[x_i\cdot x_j]_{N\times N}$ （**Gram 矩阵**）

感知机学习对偶算法Python实现

```
def trainModel(data_mat, label_mat, eta):
    m, n = data_mat.shape
    alpha = np.zeros(m)
    b = 0
    flag = True
    while flag:
        for i in range(m):
            if (label_mat[i] * (np.sum(alpha * label_mat * np.dot(data_mat, data_mat[i].T)) + b)) <= 0:
                alpha[i] = alpha[i] + eta
                b = b + eta * label_mat[i]
                print("alpha, bias: ", end="")
                print(alpha, end="  ")
                print(b)
                flag = True
                break
            else:
                flag = False
    w = np.dot(data_mat.T, alpha * label_mat)
    return w, b
```

感知机学习算法对偶形式-实例

例题2-2

已知正例：$x_1=(3,3)^T, x_2=(4,3)^T$，负例 $x_3=(1,1)^T$，求感知机 :$f(x)=sign(w\cdot x+b)$

解： 按照算法 2. 2

(1) 取α_i= 0, i = 1, 2, 3, b =0, η= 1;

(2) 计算 Gram 矩阵

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

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

（4）迭代见下表。

（5）最后$w=2x_1+0x_2-5x_3=(1,1)^{\mathrm{T}}\quad b=-3$

分离超平面 $x^{(1)}+x^{(2)}-3=0$　感知机模型 $f(x)=\mathrm{sign}(x^{(1)}+x^{(2)}-3)$

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	1	2	2	2
α_2	0	0	0	0	0	0	0	0
α_3	0	0	1	2	3	3	4	5
b	0	1	0	−1	−2	−1	−2	−3

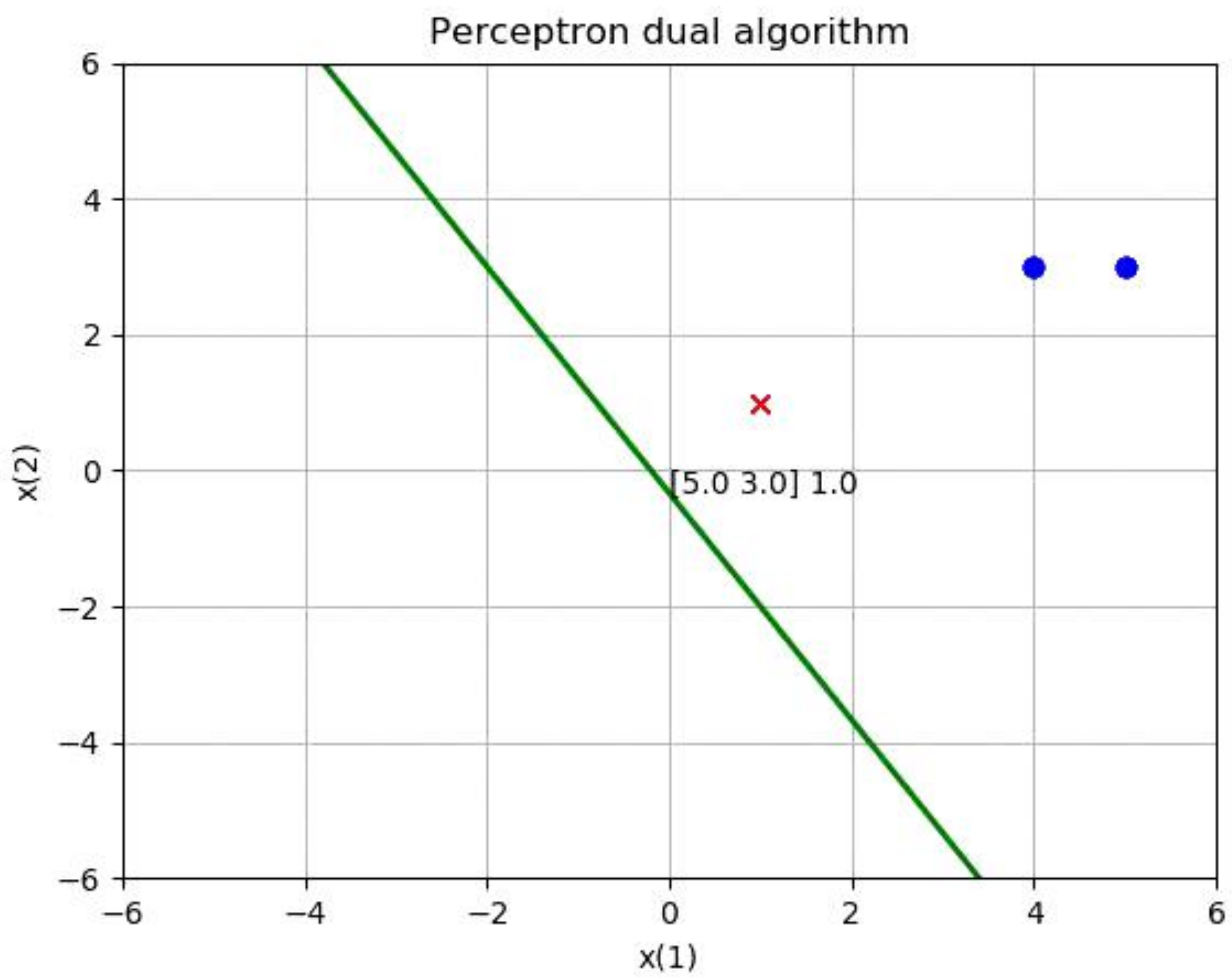
$$w=\sum_{i=1}^{N}\alpha_i y_i x_i$$

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

感知机学习算法（计算机实例演示）

训练数据集 T

$x^{(1)}$	$x^{(2)}$	y
5	3	1
4	3	1
1	1	-1



上机试验

上机内容： 感知机算法

时间：4月28日，星期四，5-6节

地点：T2102/线上