

第二章

感知机

张志忠 副研究员

华东师范大学计算机学院

回顾

- 1、学习是什么？
- 2、统计学习的流程？
- 3、学习三要素：模型（假设空间：一次函数、二次函数）、策略（损失函数）和算法（梯度下降）
- 4、经验风险与结构风险

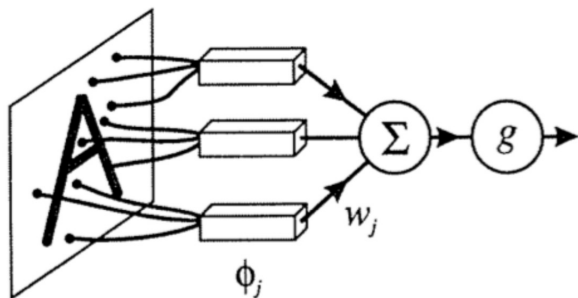
简要的历史



图灵测试:

如果一台机器能够通过电传设备与人类展开对话而能不被辨别出其机器身份，那么称这台机器具有智能。

Perceptrons, 1958



感知机算法:

第一个神经元模型，神经网络的雏形

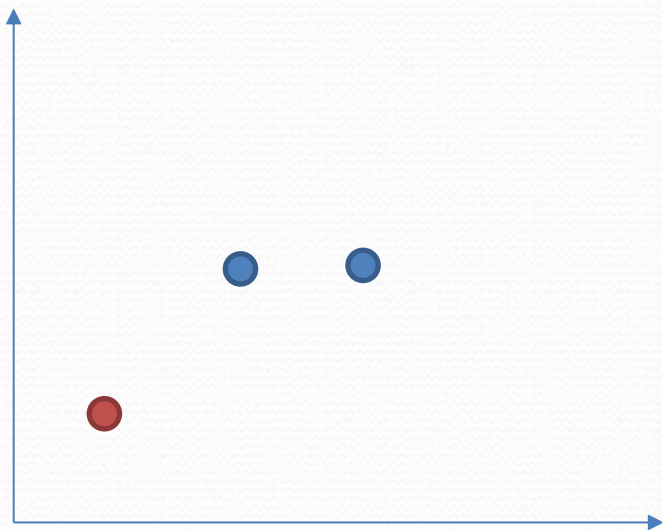
它是一个二类分类的线性分类模型

目录

1. 感知机模型
2. 感知机学习策略
3. 感知机学习算法

引子

零件编号	宽度(wid) z_1	长度(len) z_2	检验类别
1	3	3	正品
2	4	3	正品
3	1	1	次品



模型?

策略?

算法?

一、感知机(Perceptron)

- ❧ **输入**为实例的特征向量，**输出**为实例的类别，取+1和-1；
- ❧ 感知机对应于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；
- ❧ 导入基于误分类的损失函数；
- ❧ 利用梯度下降法对损失函数进行极小化；
- ❧ 感知机学习算法具有简单而易于实现的优点，分为原始形式和对偶形式；

感知机模型

∞ 定义(感知机):

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

∞ 输入 $x \in \mathcal{X}$: 表示实例的特征向量, 对应于输入空间 (特征空间) 的点, 输出 $y \in \mathcal{Y}$ 表示实例的类别, 由输入空间到输出空间的函数:

$$f(x) = \text{sign}(w \cdot x + b) \quad \text{模型}$$

∞ 称为感知机,

∞ 模型参数: w b , 内积, 权值向量, 偏置,

∞ 符号函数:

$$\text{sign}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

感知机模型

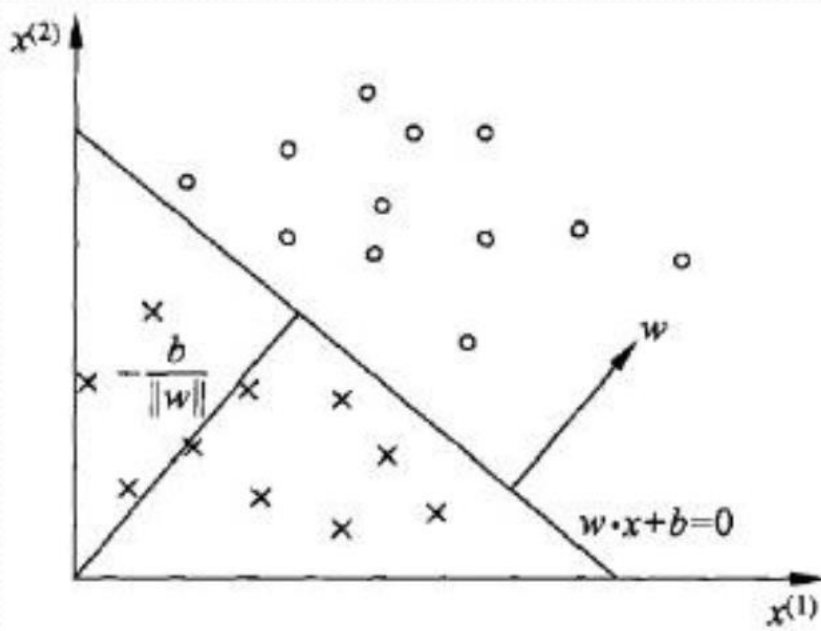
∞ 感知机几何解释:

∞ 线性方程:

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

∞ 对应于超平面S, w 为法向量, b 截距, 分离正、负类:

∞ 分离超平面:



二、感知机学习策略

∞ 数据集的线性可分性 (非常重要)

定义 2.2 (数据集的线性可分性) 给定一个数据集

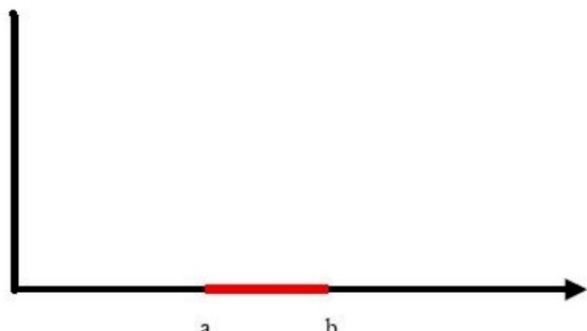
$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

其中, $x_i \in \mathcal{X} = \mathbf{R}^n$, $y_i \in \mathcal{Y} = \{+1, -1\}$, $i = 1, 2, \dots, N$, 如果存在某个超平面 S

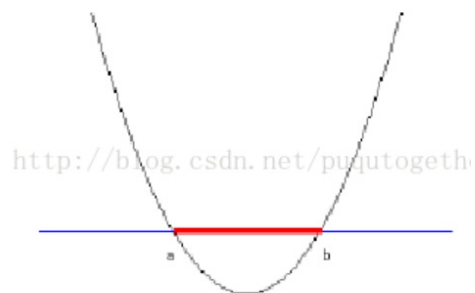
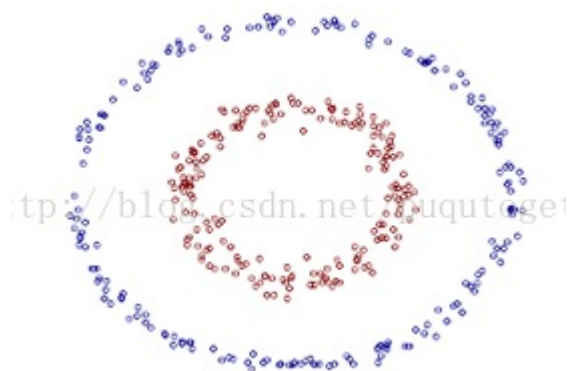
$$w \cdot x + b = 0$$

能够将数据集的正实例点和负实例点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 $w \cdot x_i + b > 0$, 对所有 $y_i = -1$ 的实例 i , 有 $w \cdot x_i + b < 0$, 则称数据集 T 为线性可分数据集 (linearly separable data set); 否则, 称数据集 T 线性不可分。

二、线性不可分的情况



我们可能需要换一个更为强大的模型



<http://blog.csdn.net/puqutogeth>

<http://blog.csdn.net/puqutogeth>

感知机学习策略

∞ 如何定义损失函数？

∞ 自然选择：误分类点的数目，但损失函数不是 w, b 连续可导，不宜优化。

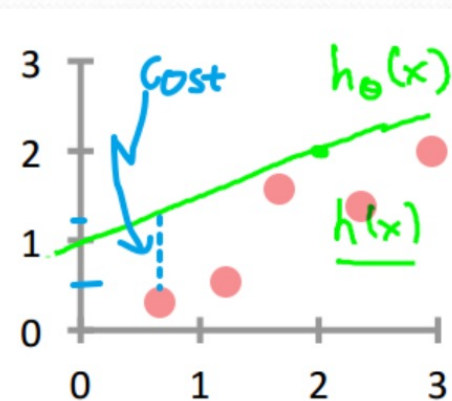
∞ 另一选择：误分类点到超平面的总距离：

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

误分类点：
$$-y_i(w \cdot x_i + b) > 0$$

误分类点距离：
$$-\frac{1}{\|w\|} y_i(w \cdot x_i + b)$$

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



→ $\theta_0 = 1$

→ $\theta_1 = 0.5$

感知机学习策略

∞ 损失函数:

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

∞ M为误分类点的数目

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,

∞ 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:

∞ 选取误分 $\nabla_w L(w, b) = - \sum_{x_i \in M} y_i x_i$ $\nabla_b L(w, b) = - \sum_{x_i \in M} y_i$

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

感知机学习算法

∞ 感知机学习算法的原始形式:

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

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

学习率 η ($0 < \eta \leq 1$);

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 $y_i(w \cdot x_i + b) \leq 0$

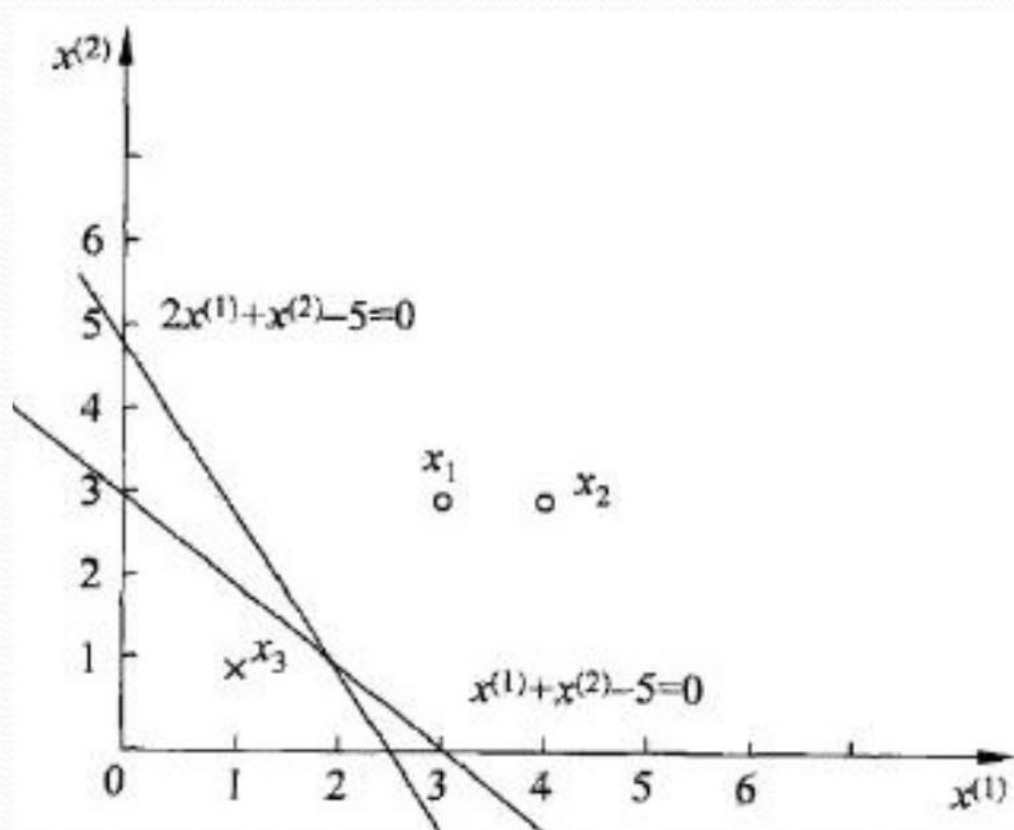
$$w \leftarrow w + \eta y_i x_i$$

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

(4) 转至 (2), 直至训练集中没有误分类点

零件编号	宽度(wid) z_1	长度(len) z_2	检验类别
1	3	3	正品
2	4	3	正品
3	1	1	次品

∞例：正例： $x_1 = (3, 3)^T$, $x_2 = (4, 3)^T$ $x_3 = (1, 1)^T$;



感知机学习算法

∞ 解：构建优化问题：

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

∞ 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) x_2 , 显然, $y_1(w_1 \cdot x_i + b_1) > 0$, 被正确分类,

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

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

感知机学习算法

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

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

∞ 算法的收敛性：证明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。

∞ 将**b**并入权重向量**w**，记作： $\hat{w} = (w^T, b)^T$

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

∞ 定理：

设训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 是线性

可分的，其中 $x_i \in \mathcal{X} = \mathbf{R}^n$ ， $y_i \in \mathcal{Y} = \{-1, +1\}$ ， $i = 1, 2, \dots, N$ ，

感知机学习算法

则

(1) 存在满足条件 $\|\hat{w}_{\text{opt}}\| = 1$ 的超平面 $\hat{w}_{\text{opt}} \cdot \hat{x} = w_{\text{opt}} \cdot x + b_{\text{opt}} = 0$:

且存在 $\gamma > 0$, 对所有 $i = 1, 2, \dots, N$

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

感知机学习算法

证明: (1)

由线性可分, 存在超平面: $\hat{w}_{opt} \cdot \hat{x} = w_{opt} \cdot x + b_{opt} = 0$

使 $\|\hat{w}_{opt}\| = 1$, 由有限的点, 均有:

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

存在

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

使:

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

感知机学习算法

∞ (2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$ ，感知机算法在训练集的误分类次数 k 满足不等式，

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

∞ 证明：令 \hat{w}_{k-1} 是第 k 个误分类实例之前的扩充权值向量，即：

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

∞ 第 k 个误分类实例的条件是：

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

∞ 则 w 和 b 的更新： $w_k \leftarrow w_{k-1} + \eta y_i x_i$ 即： $\hat{w}_k = \hat{w}_{k-1} + \eta y_i \hat{x}_i$

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

感知机学习算法

∞ (2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$ ，感知机算法在训练集的误分类次数 k 满足不等式，

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

∞ 推导两个不等式：

∞ (1)

$$\hat{w}_k \cdot \hat{w}_{\text{opt}} \geq k\eta\gamma$$

∞ 由：

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

∞ 得：

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

感知机学习算法

∞ (2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$ ，感知机算法在训练集的误分类次数 k 满足不等式，

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

(2)

$$\|\hat{w}_k\|^2 \leq k\eta^2 R^2$$

错分项 < 0

∞ 则：

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

感知机学习算法

∞ (2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$ ，感知机算法在训练集的误分类次数 k 满足不等式，

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

结合两个不等式：

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

$$k^2\gamma^2 \leq kR^2$$

得：

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

感知机学习算法

- ∞ 定理表明：
- ∞ 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。
- ∞ 感知机算法存在许多解，既依赖于初值，也依赖迭代过程中误分类点的选择顺序。
- ∞ 为得到唯一分离超平面，需要增加约束，如SVM。
- ∞ 线性不可分数据集，迭代震荡。

感知机学习算法


∞ 感知机算法的对偶形式：

∞ 回顾 SVM 对偶形式：

∞ 基本想法：

∞ 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 ，对误分类点：

$$\begin{aligned}w &\leftarrow w + \eta y_i x_i \\b &\leftarrow b + \eta y_i\end{aligned}$$


最后学习到的 w, b

$$\begin{aligned}w &= \sum_{i=1}^N \alpha_i y_i x_i \\b &= \sum_{i=1}^N \alpha_i y_i\end{aligned}$$

感知机学习算法

∞ 感知机学习算法的对偶形式:

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

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

学习率 η ($0 < \eta \leq 1$);

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 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$

感知机学习算法

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

(2) 在训练集中选取数据 (x_i, y_i)

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

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

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

(4) 转至 (2) 直到没有误分类数据.

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

感知机学习算法

例：正样本点是 $x_1 = (3,3)^T$ ， $x_2 = (4,3)^T$ ，负样本点是 $x_3 = (1,1)^T$

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

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

(3) 误分条件

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

参数更新

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

感知机学习算法

例：正样本点是 $x_1 = (3,3)^T$ ， $x_2 = (4,3)^T$ ，负样本点是 $x_3 = (1,1)^T$

(4) 迭代。过程从略，结果列于表 2.2.

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

(5) $w = 2x_1 + 0x_2 - 5x_3 = (1,1)^T$ 分离超平面 $x^{(1)} + x^{(2)} - 3 = 0$
 $b = -3$

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