机器学习
统计学习方法

主讲：蔡 波

武汉大学网络安全学院

第二章 感知机

主讲：蔡 波

武汉大学网络安全学院

第2-1节 感知机

- 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；
- 1957年由ROSENBLATT(罗森布拉特)提出，是神经网络与支持向量机的基础。

感知机

- 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：

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

- 称为感知机，
- 模型参数： w 和 b 为模型参数，$w \in \mathbf{R}^n$ 权值向量，$b \in \mathbf{R}$ 偏置，
- 符号函数

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

感知机

- 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：

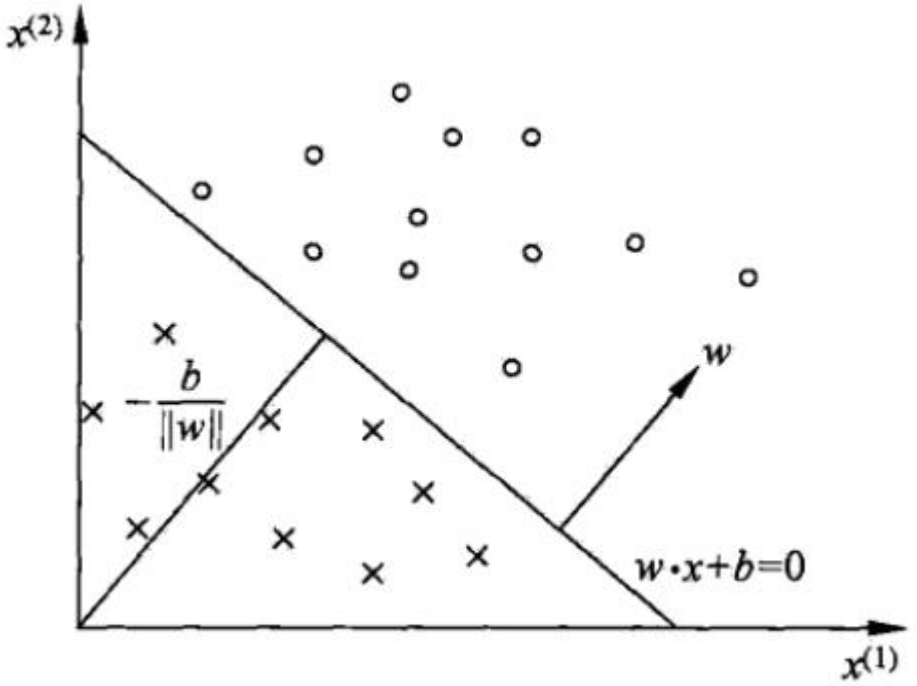


图 2.1　感知机模型

第2-2节 感知机学习策略

- 线性可分
- 给定一个数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$
- 其中 $x_i\in\mathcal{X}=\mathbf{R}^n$， $y_i\in\mathcal{Y}=\{+1,-1\}$， $i=1,2,\cdots,N$ ，如果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。也就是：

  $w\cdot x_i+b>0$时， $y_i=+1$，而当 $w\cdot x_i+b<0$时， $y_i=-1$

- 此时，称数据集线性可分，否则为线性不可分。

- 对于误分类点有： $w\cdot x_i+b>0$时， $y_i=-1$，而当 $w\cdot x_i+b<0$时， $y_i=+1$

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和B连续可导函数，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：$-\frac{1}{\|w\|} \sum_{x_i \in M} y_i(w \cdot x_i + b)$

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
- 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R，损失函数L(W,B)是的连续可导函数
- 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.

第2-3节 感知机学习算法

- 求解最优化问题：求参数W,B，使得 $\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$
-
- 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 函数,损失函数L的梯度：$\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i$ $\nabla_b L(w,b) = -\sum_{x_i \in M} y_i$
- 选取误分类点更新：

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

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

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

（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.

感知机学习算法

- 例：正例：$x_1=(3,3)^{\mathrm{T}}$，$x_2=(4,3)^{\mathrm{T}}$，负例 $x_3=(1,1)^{\mathrm{T}}$，试用感知机学习算法的原始形式求感知机模型 $f(x)=\mathrm{sign}(w\cdot x+b)$

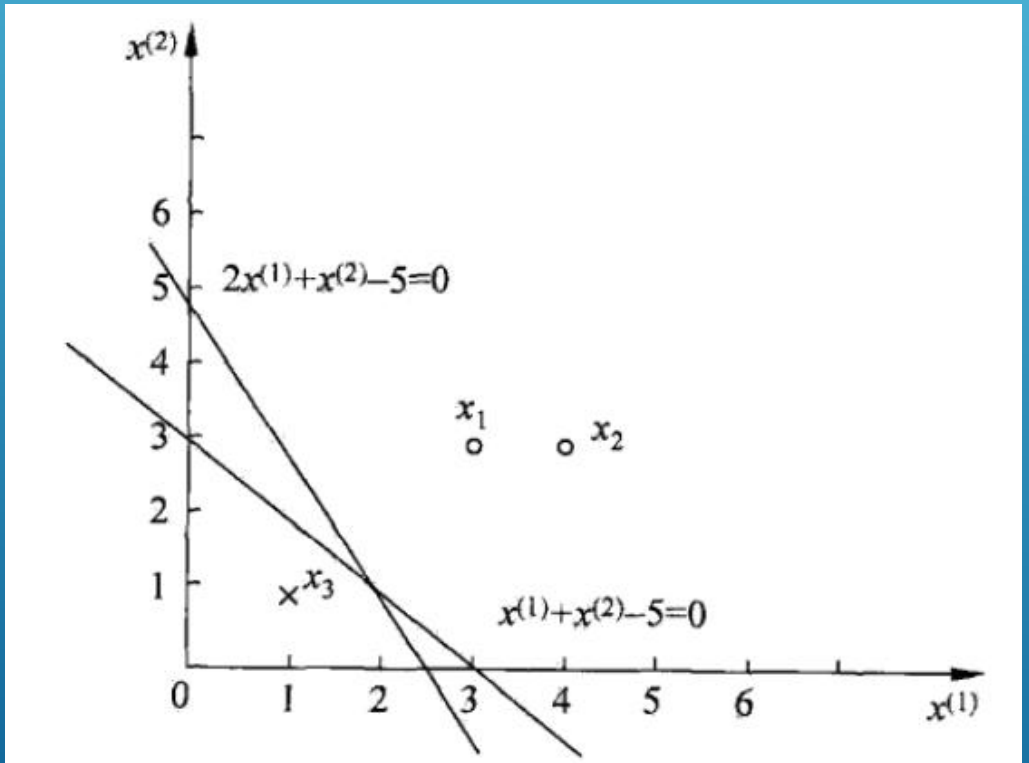


图 2.2　感知机示例

感知机学习算法

- 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. 学习率 = 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 $x_3 = (1,1)^T$，$y_3(w_1 \cdot x_3 + b_1) < 0$，被误分类，更新 w, b.

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

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

- 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.

感知机学习算法

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	$(3,3)^T$	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	$x^{(1)} + x^{(2)} - 3$

感知机学习算法的收敛性

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

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

感知机学习算法的收敛性

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 $\hat{w}_{opt} \cdot \hat{x} = w_{opt} \cdot x + b_{opt} = 0$，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

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

感知机学习算法的收敛性

证明（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

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

感知机学习算法的收敛性

下面推导两个不等式：

（1）

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

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)

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

感知机学习算法的收敛性

（2）

$$\|\hat{w}_k\|^2 \leqslant k\eta^2 R^2 \tag{2.13}$$

由式 (2.11) 及式 (2.10) 得

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

结合不等式 (2.12) 及式 (2.13) 即得

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

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

于是

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

感知机学习算法的收敛性

- 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。

感知机学习算法的对偶形式

我们假设样本点(x_i, y_i)在更新过程中被使用了n_i次。因此，从原始形式的学习过程可以得到，最后学习到的w和b可以分别表示为：

$$w = \sum_{i=1}^{N} n_i \eta \, y_i x_i \tag{1}$$

$$b = \sum_{i=1}^{N} n_i \eta \, y_i \tag{2}$$

考虑n_i的含义：如果n_i的值越大，那么意味着这个样本点经常被误分。什么样的样本点容易被误分？很明显就是离超平面很近的点。超平面稍微一点点移动，这个点就从正变为负，或者从负变为正。如果学过SVM就会发现，这种点很可能就是支持向量。

代入式(1)和式(2)到原始形式的感知机模型中，可得：

$$f(x) = sign(w \cdot x + b) = sign(\sum_{j=1}^{N} n_j \eta \, y_j x_j \cdot x + \sum_{j=1}^{N} n_j \eta \, y_j) \tag{3}$$

此时，学习的目标就不再是w和b，而是n_i，$i = 1, 2, \ldots, N$。

感知机学习算法的对偶形式

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．

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

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

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

（4）转至（2）直到没有误分类数据. ■

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）

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

感知机学习算法的对偶形式例题

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

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

感知机学习算法的对偶形式例题

（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

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

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

2.感知机对偶形式

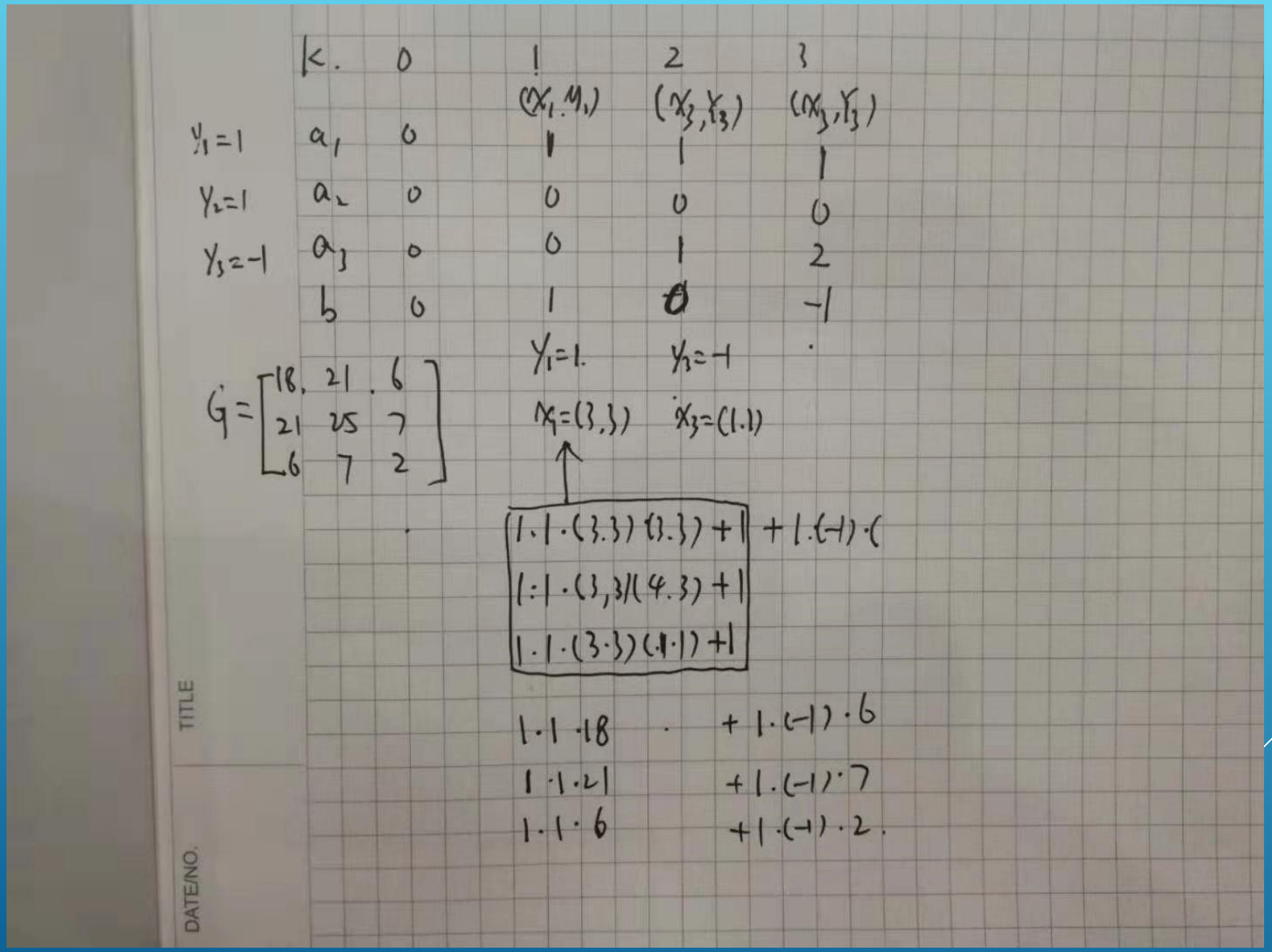
原始形式

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

$$w = n_1 \eta y_1 x_1 + n_2 \eta y_2 x_2 + \cdots + n_N \eta y_N x_N$$

对偶形式

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


k. 0 1 2 3
(x1, y1) (x2, y2) (x3, y3)
y1=1 a1 0 1 1 1
y2=1 a2 0 0 0 0
y3=-1 a3 0 0 1 2
b 0 1 0 -1
y1=1 y3=-1
x1=(3,3) x3=(1,1)
G = [18 21 6; 21 25 7; 6 7 2]
1·1·(3·3)(3·3)+1 + 1·(-1)·(
1·1·(3,3)(4·3)+1
1·1·(3·3)(1·1)+1
1·1·18 + 1·(-1)·6
1·1·21 + 1·(-1)·7
1·1·6 + 1·(-1)·2.
TITLE
DATE/NO.