

第四章：线性模型-感知机

线性模型的定义和基本形式（回忆）

设特征空间的维数为 d , 则线性模型的数学表示为:

$$f(X) = \sum_{k=1}^d w_k x_k + b = W^T X + b$$

加长向量表示:

$$f(X) = W'^T X' = \sum_{k=1}^{d+1} w_k x_k$$
$$W' = (w_1, w_2, \dots, w_d, w_{d+1})^T$$
$$X' = (x_1, x_2, \dots, x_d, 1)^T$$

任务: 根据训练样本确定所求权向量

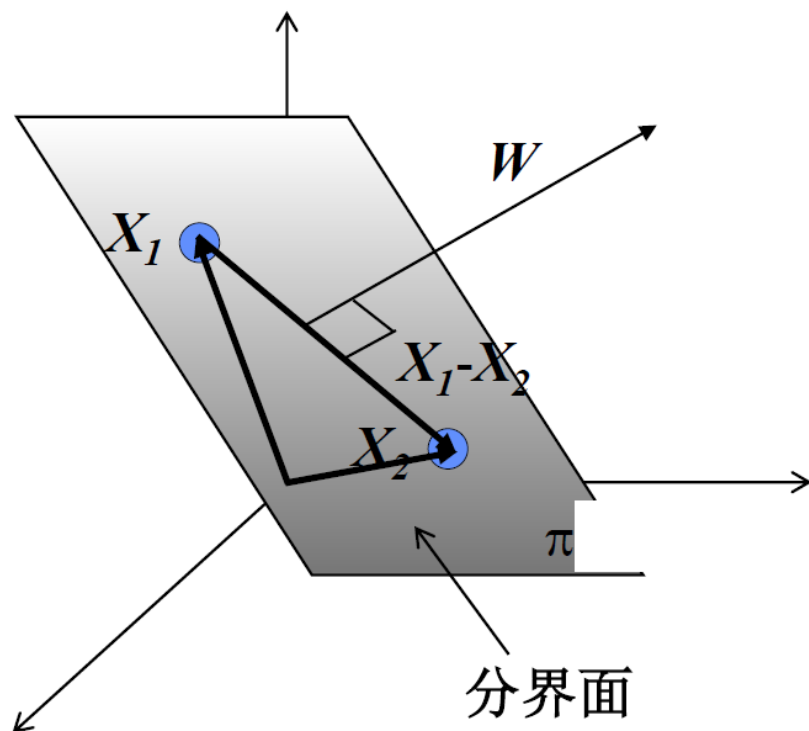
线性模型的几何意义

对于二分类学习任务下的线性模型，决策面（分界面）方程 $(f(X) = 0)$ 的几何意义：

当特征空间是一维空间时，它是一个点；
当特征空间是二维空间时，它是一条直线；
当特征空间是三维空间时，它是一个平面；
当特征空间是多维空间时，它是一个超平面。

线性模型的几何意义

权向量的重要性质：



设 X_1 和 X_2 是分界面上任意两点，显然有：
 $W^T X_1 + b = W^T X_2 + b = 0$

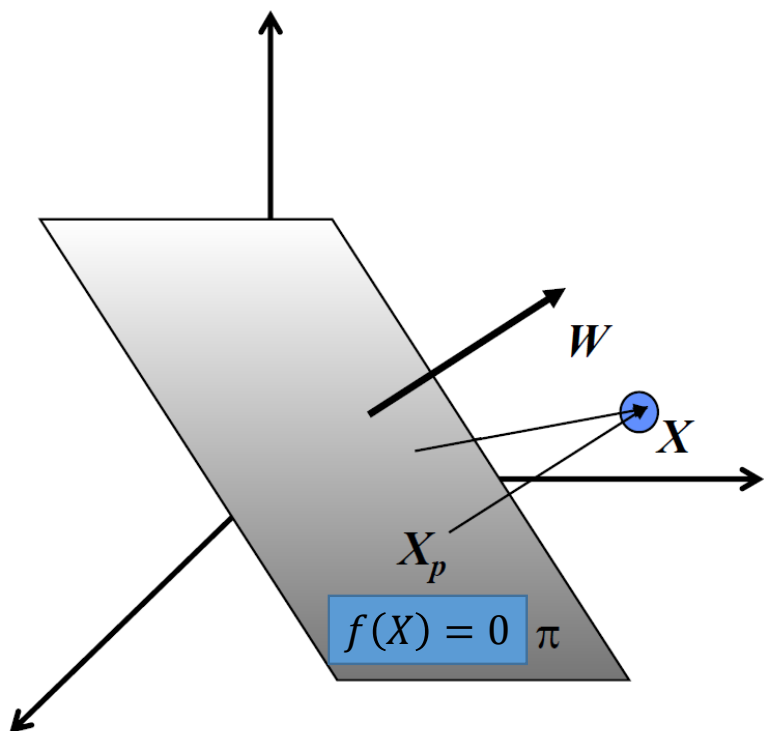
→ $W^T (X_1 - X_2) = 0$

因 $X_1 - X_2$ 是分界面上任意一个向量，
所以，权向量 W 与分界面 π 正交。

即，权向量和分界面法线方向一致。

线性模型的几何意义

分界面函数在 X 处取值的几何意义：



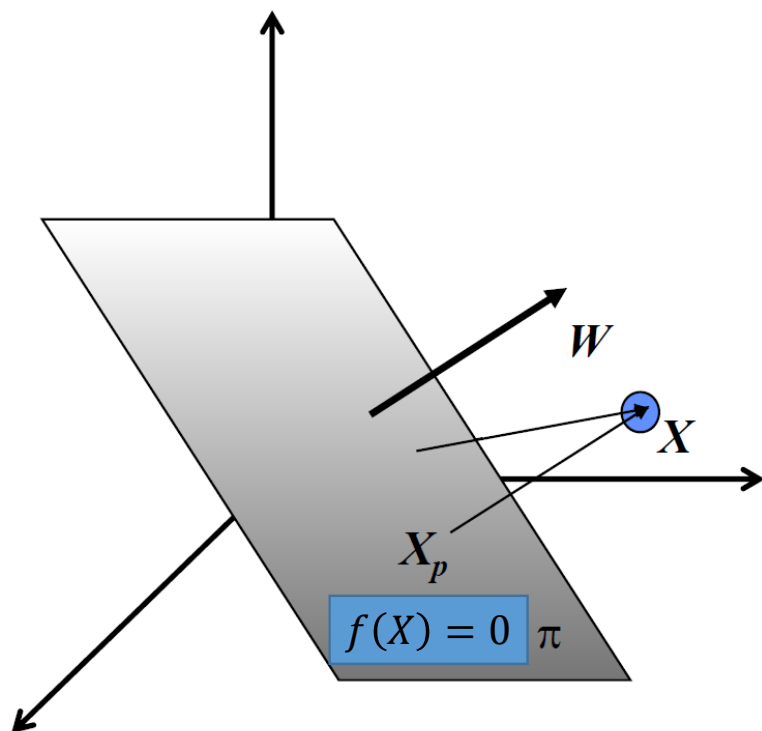
特征空间中一点 X 到分界面的距离
自 X 引垂线交分界面于 X_p
则所求距离由 $X - X_p$ 的模给出
根据定义 $X - X_p$ 与分界面正交，所以它
和权向量方向一致或者相反。

$$X - X_p = r \frac{W}{\|W\|}$$

r 是一个代数量，其大小等于所求距离，
符号反映所处位置。

线性模型的几何意义

分界面函数在 X 处取值的几何意义：



$$W^T X - W^T X_p = r \frac{W^T W}{\|W\|} = r \|W\|$$
$$(W^T X + b) - (W^T X_p + b) = r \|W\|$$

$$f(X) = W^T X + b$$

$$f(X_p) = W^T X_p + b = 0$$

$$f(X) = r \|W\|$$

$$r = \frac{f(X)}{\|W\|}$$

结论： $f(X)$ 给出了特征空间中一个点 X 到分界面距离的度量

感知机 (perceptron)

- 输入为样本的特征向量，输出为样本的类别，取+1和-1
- 对应于输入空间中将样本划分为正负两类的分界面，属于线性模型
- 导入基于误分类的损失函数
- 可以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
- 感知机能够容易地实现逻辑与、或、非运算
- 1957年由Rosenblatt提出，是神经网络与支持向量机的基础

感知机模型定义

- 假设输入空间（特征空间）是 $X \in R^n$ ，输出空间 $Y = \{+1, -1\}$
- 输入 $X \in R^n$ 表示样本的特征向量，对应于输入空间（特征空间）的点，输出表示样本的类别，由输入空间到输出空间的函数：

$$f(X) = \text{sign}(W^T X + b)$$

称为感知机。

- 模型参数： W (weight) 和 b (bias)
- 符号函数：

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

数据集的线性可分性定义

- 给定一个数据集 $T = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$, 其中 $X_i \in R^n$, $Y_i \in Y = \{+1, -1\}$, $i = 1, 2, \dots, n$
- 如果存在某个分界面 $S: W^T X + b = 0$ 能够将数据集的正样本和负样本完全正确地划分到分界面的两侧, 即对于所有 $Y_i = +1$ 的样本 i , 有 $W^T X + b > 0$; 对于所有 $Y_i = -1$ 的样本 i , 有 $W^T X + b < 0$ 。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

感知机学习策略

- ▣ 假设训练数据集是线性可分的，感知机学习的目标是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界面
- ▣ 为了找出这样的分界面，即确定感知机模型参数 W 和 b ，需要确定一个学习策略，即定义（经验）损失函数并将损失函数极小化
- ▣ 损失函数（loss function）或代价函数（cost function）：度量预测错误程度的函数

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 的连续可导函数，不宜优化
- 另一选择：误分类点到分界面的总距离
- 输入空间任一点 X 到分界面的距离：

$$\frac{|W^T X + b|}{\|W\|}$$

对于误分类数据 (X_i, Y_i) : $-Y_i(W^T X_i + b) > 0$, 误分类点距离: $-\frac{Y_i(W^T X_i + b)}{\|W\|}$

总距离: $-\frac{\sum_{X_i \in M} Y_i(W^T X_i + b)}{\|W\|}$, 其中 M 为分界面 S 的误分类点集合

感知机学习策略

□ 不考虑 $\frac{1}{\|W\|}$ ，损失函数定义为：

$L(W, b) = -\sum_{X_i \in M} Y_i(W^T X_i + b)$ ，其中 M 为分界面 S 的误分类点集合

□ 该损失函数就是感知机学习的经验风险函数

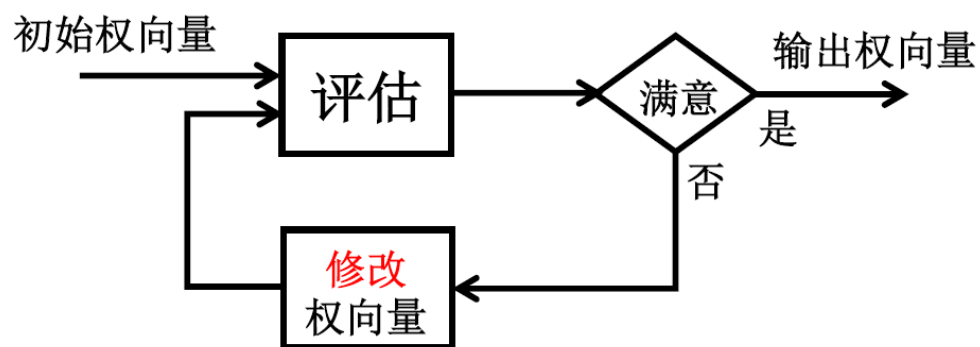
□ $L(W, b) \geq 0$ ，没有错误分类点（ M 为空集），达到极小值

□ 误分类点越少，误分类点离分界面越近，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 和 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 和 b 的连续可导函数

感知机学习算法-固定增量的感知机算法

□ 借助于优化技术：固定增量的感知机算法



- 如果所选权向量不能做到对所有输入训练样本正确分类，那么应该如何对其进行修改？
- 迭代算法收敛吗？它在什么情况下收敛？如何判断其收敛性？

固定增量的感知机算法

- ① 赋初值：迭代步数 $k = 0$ ，固定比例因子 $0 \leq \rho \leq 1$ ， $W(0)$ ，连续正确分类计数器 $N_c = 0$ ；
- ② 读入训练样本集合 $T = \{(X_0, Y_0), (X_1, Y_1), \dots, (X_{n-1}, Y_{n-1})\}$ ；
- ③ 取样本 $X = X_{[k]_n}$ ， $Y = Y_{[k]_n}$ ， $[k]_n = k \bmod (n)$ ；计算 $f(X) = W(k)^T X$ ；
- ④ 修正权向量：
 - 当 $Y = 1$ 时，
若 $f(X) \leq 0$ ，则 $W(k+1) = W(k) + \rho X$ ， $N_c = 0$ ；
否则 $W(k+1) = W(k)$ ， $N_c += 1$ ；
 - 当 $Y = -1$ 时，
若 $f(X) \geq 0$ ，则 $W(k+1) = W(k) - \rho X$ ， $N_c = 0$ ；
否则 $W(k+1) = W(k)$ ， $N_c += 1$ ；
- ⑤ 若 $N_c \geq n$ ，算法结束；否则 $k = k + 1$ ，返回第3步。

固定增量的感知机算法

权向量更新规则

当 $Y = 1$ 时, $f(X) \leq 0$,

$$W(k+1) = W(k) + \rho X$$

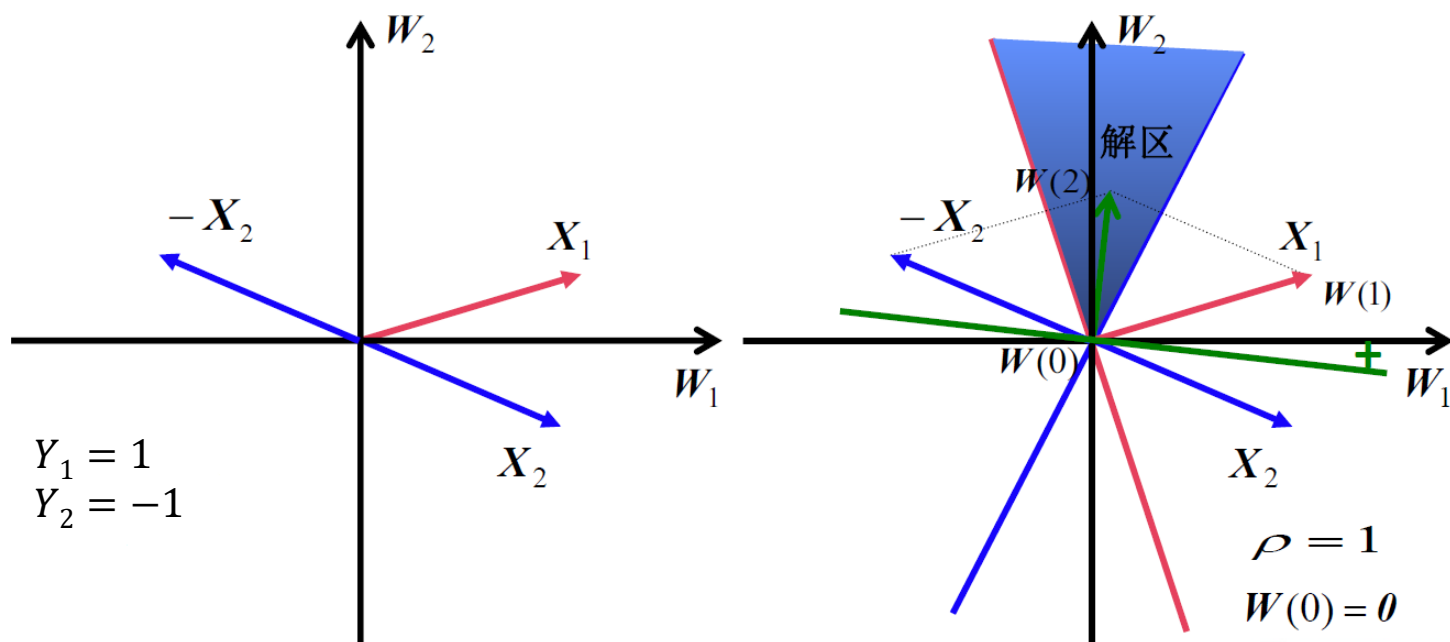
$$W(1) = X_1$$

当 $Y = -1$ 时, $f(X) \geq 0$,

$$W(k+1) = W(k) - \rho X$$

$$\begin{aligned} W(2) &= W(1) - X_2 \\ &= X_1 - X_2 \end{aligned}$$

举例：取 $W(0) = 0$, $\rho = 1$



固定增量的感知机算法

权向量更新规则

当 $Y = 1$ 时, $f(X) \leq 0$,
 $W(k+1) = W(k) + \rho X$

当 $Y = -1$ 时, $f(X) \geq 0$,
 $W(k+1) = W(k) - \rho X$

修改权向量会影响下一次迭代的结果!

$$\begin{aligned} f^{k+1}(X) &= W(k+1)^T X \\ &= (W(k) + \rho X)^T X \\ &= W(k)^T X + \rho X^T X \end{aligned}$$

$$\rho X^T X \geq 0$$

$$f^k(X) = W(k)^T X \leq 0$$

$$\Rightarrow f^{k+1}(X) = f^k(X) + \rho X^T X \Rightarrow f^{k+1}(X) = f^k(X) + \text{正数}$$

固定增量的感知机算法

权向量更新规则

当 $Y = 1$ 时, $f(X) \leq 0$,
 $W(k+1) = W(k) + \rho X$

当 $Y = -1$ 时, $f(X) \geq 0$,
 $W(k+1) = W(k) - \rho X$

$$\begin{aligned} f^{k+1}(X) &= W(k+1)^T X \\ &= (W(k) - \rho X)^T X \\ &= W(k)^T X - \rho X^T X \end{aligned}$$

$$\rho X^T X \geq 0$$

$$f^k(X) = W(k)^T X \geq 0$$

$$\Rightarrow f^{k+1}(X) = f^k(X) - \rho X^T X \Rightarrow f^{k+1}(X) = f^k(X) - \text{正数}$$

固定增量的感知机算法（修正）

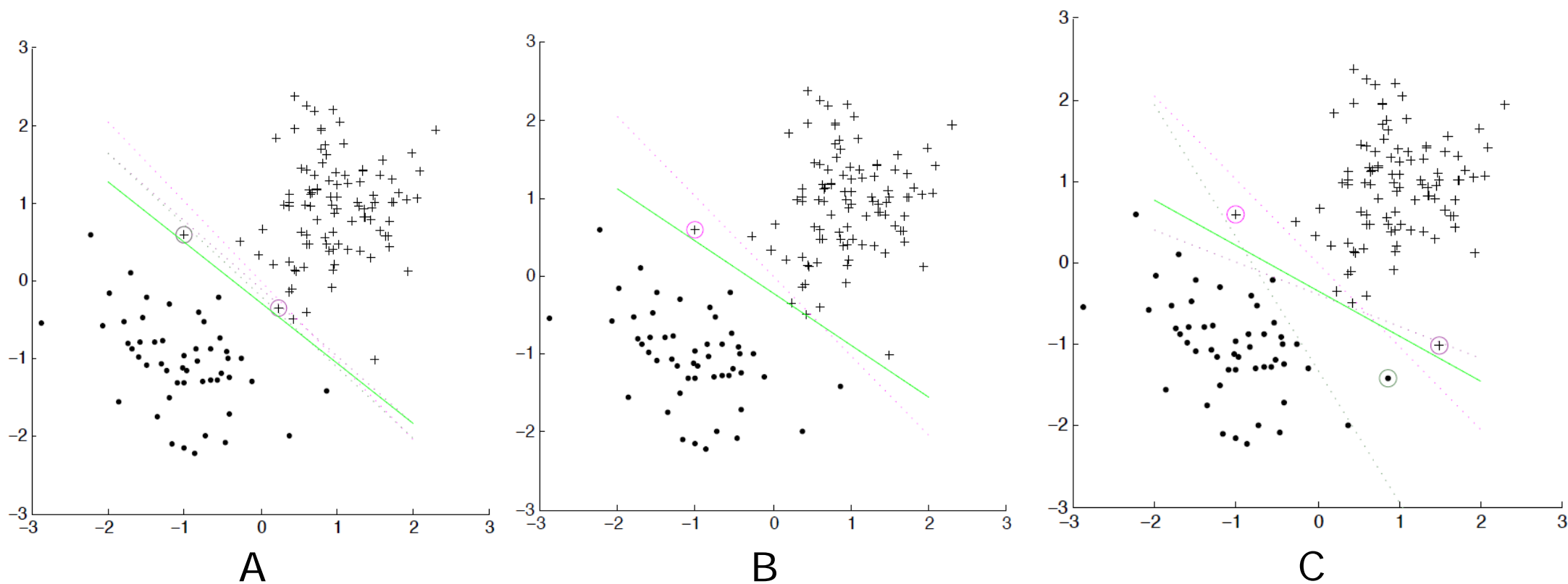
- ① 赋初值：迭代步数 $k = 0$ ，固定比例因子 $0 \leq \rho \leq 1$ ， $W(0)$ ，连续正确分类计数器 $N_c = 0$ ；
- ② 读入训练样本集合 $T = \{(X_0, Y_0), (X_1, Y_1), \dots, (X_{n-1}, Y_{n-1})\}$ ，做如下规范化处理：若 $Y_i = -1$ ，则 $X_i = -X_i$ ；
- ③ 取样本 $X = X_{[k]_n}$ ， $[k]_n = k \bmod n$ ；计算 $f(X) = W(k)^T X$ ；
- ④ 修正权向量：
若 $f(X) \leq 0$ ，则 $W(k+1) = W(k) + \rho X$ ， $N_c = 0$ ；
否则 $W(k+1) = W(k)$ ， $N_c += 1$ ；
- ⑤ 若 $N_c \geq n$ ，算法结束；否则 $k = k + 1$ ，返回第3步。

固定增量的感知机算法:讨论

- 该算法何处使用了损失函数作为分类判断的依据？
- 固定比例因子 ρ 的选择问题？

固定增量的感知机算法:讨论

不同固定比例因子 ρ 与算法收敛速度的关系



感知机算法学习能力

什么是感知机算法无法解决的？

- 不可分类的问题
- 非线性可分，但线性不可分的问题
- 多类分类问题直接解决不了

➡ 可以处理的问题为线性可分的两类问题



问题：所有线性可分两类问题都可以完美解决？

感知机算法学习能力

收敛性定理

如果训练样本集 χ 是线性可分的，则基于固定增量法则的感知器算法经过有限次迭代后将收敛到正确的解权向量

几个假定

- ✓ 采用加长向量表示
- ✓ 执行规范化处理
- ✓ 选择步长因子 $\rho = 1$

感知器算法权向量的更新规则


对 $X \in \chi$ ，若 $f(X) = W^T X \leq 0$ ，则 $W + X \rightarrow W$ ，否则 $W \rightarrow W$

感知机算法学习能力

收敛性定理的证明

经过规范化处理后， χ 线性可分是指：

$$\exists W^*, \text{ 使对 } \forall X \in \chi, \text{ 有 } f(X) = W^{*T}X > 0$$

对预先任意给定的正常数 C_p ，总可找到大正数 C ，使  换个说法

$$(CW^*)^T X > C_p, \quad \forall X \in \chi$$

特别地，若令 $M = \text{Max}(\|X\|^2)$

则总可以找到大的正数 C ，使

$$(CW^*)^T X = W_s^T X > M, \quad \forall X \in \chi$$



根据定义，为解权向量

$$W(k+1) = W(k)$$



故：只要证明每进行一次权向量的更新，都使 $W(k+1)$ 接近 W_s 一个正的非无穷小量就可以了



$$W(k+1) = W(k) + X$$

感知机算法学习能力

收敛性定理的证明

考察：更新前后权向量与 W_s 之间欧式距离的变化

$W(k+1)$ 与 W_s 的欧式距离 $\|W_s - W(k+1)\|$

$W(k)$ 与 W_s 的欧式距离 $\|W_s - W(k)\|$

$$\begin{aligned} & \|W_s - W(k)\|^2 - \|W_s - W(k+1)\|^2 \\ &= (W_s - W(k))^T (W_s - W(k)) - (W_s - W(k+1))^T (W_s - W(k+1)) \\ &= -2W_s^T W(k) + W(k)^T W(k) + 2W_s^T W(k+1) - W(k+1)^T W(k+1) \\ &= -2W(k)^T X + 2W_s^T X - X^T X \\ &\geq M \end{aligned}$$

所以，每一次更新都使权向量接近 W_s 一个正的非无穷小量
即经过有限次迭代后 W 将收敛于 W_s

感知机算法学习能力

收敛性定理表明

- 误分类的次数是有上界的，当训练数据集线性可分时，感知机学习算法迭代是收敛的
- 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
- 线性不可分数据集，迭代震荡

固定增量的感知机算法:小结

- 算法简单，且当训练样本集线性可分时，算法经有限次迭代后可收敛到解权向量，但算法的收敛速度不理想
- 有办法改善吗？
- 回顾：求解解权向量的步骤
 - 第一步：采集训练样本，构成预分类的样本集合
 - 第二步：选用或确定一个损失函数 L
 - ✓ $L = L(W, b)$
 - ✓ 损失函数的极值解和最优分类判决相对应
 - 第三步：求损失函数的极值解得到待求的解权向量

基于梯度下降法的感知机算法（原始形式）

- 求解最优化问题

$$\min_{W,b} L(W,b) = - \sum_{x_i \in M} Y_i (W^T X_i + b)$$

- 随机梯度下降法

- 首先任意选择一个超平面 (W_0, b_0) ，然后不断极小化目标函数

- 假设误分类点集合 M 是固定的，那么损失函数的梯度：

$$\nabla_W L(W,b) = - \sum_{X_i \in M} Y_i X_i, \quad \nabla_b L(W,b) = - \sum_{x_i \in M} Y_i$$

- 更新权重：

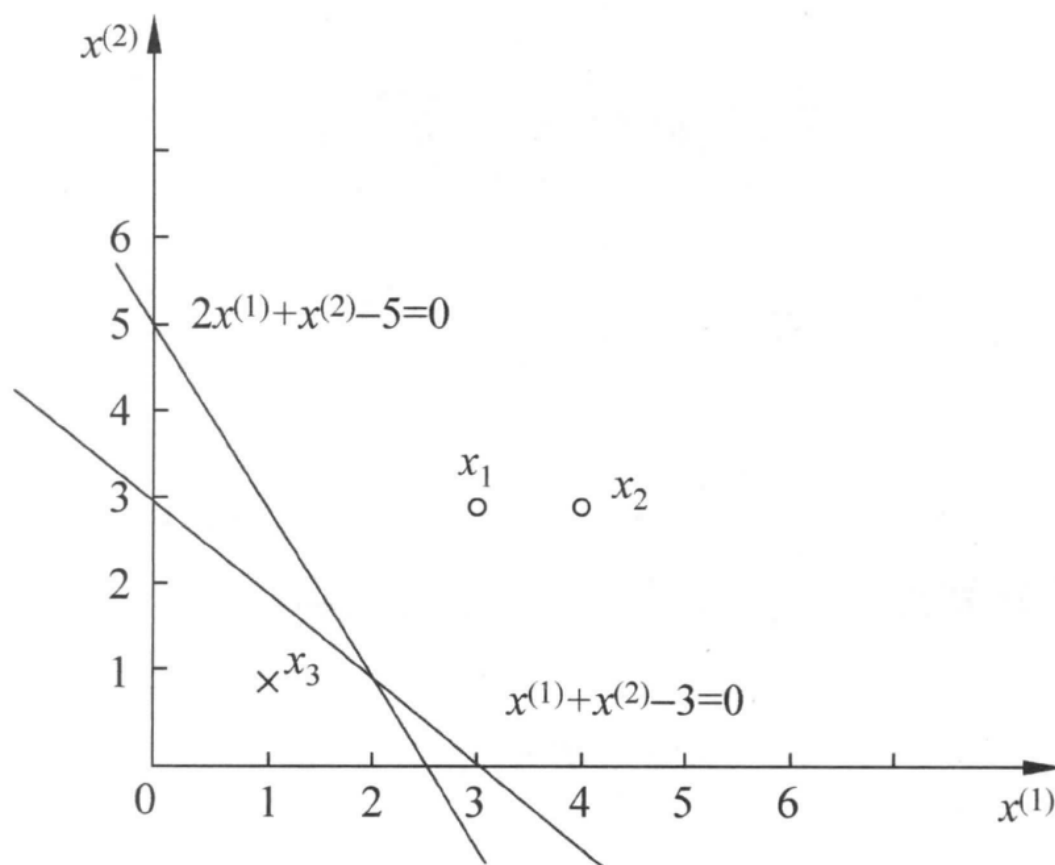
$$W \leftarrow W + \eta Y_i X_i, \quad b \leftarrow b + \eta Y_i$$

基于梯度下降法的感知机算法（原始形式）

- 输入：训练数据集 $T = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ ，其中 $X_i \in R^n$ ， $Y_i \in Y = \{+1, -1\}$ ， $i = 1, 2, \dots, n$ ，学习率 η ($0 < \eta \leq 1$)
- 输出： W, b ；感知机模型 $f(X) = \text{sign}(W^T X + b)$
 - 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 $Y_i (W^T X + b) \leq 0$ ， $W \leftarrow W + \eta Y_i X_i$ ， $b \leftarrow b + \eta Y_i$
 - ④ 转至第2步，直至训练集中没有误分类点

基于梯度下降法的感知机算法（原始形式）

□ 例题：正样本： $X_1 = (3,3)^T$, $X_2 = (4,3)^T$, 负样本： $X_3 = (1,1)^T$



基于梯度下降法的感知机算法（原始形式）

□ 例题：正样本： $X_1 = (3,3)^T$, $X_2 = (4,3)^T$, 负样本： $X_3 = (1,1)^T$

□ 求解： $W, b, \eta = 1$ （这里 $W = (W^{(1)}, W^{(2)})^T$, $X = (X^{(1)}, X^{(2)})^T$ ）

① 取初值 $W_0 = 0$, $b_0 = 0$

② 对 X_1 , $Y_i(W_0^T X + b_0) = 0$, 未能正确分类, 更新 W, b

$$W_1 = W_0 + \eta Y_1 X_1 = (3,3)^T, b_1 = b_0 + \eta Y_1 = 1$$

得到线性模型： $W_1^T X + b_1 = 3X^{(1)} + 3X^{(2)} + 1$

③ 对 X_1 和 X_2 , $Y_i(W_1^T X + b_1) > 0$, 正确分类, 不修改 W, b ; 对 X_3 , $Y_i(W_1^T X + b_1) < 0$, 未能正确分类, 更新 W, b

$$W_2 = W_1 + \eta Y_3 X_3 = (2,2)^T, b_2 = b_1 + \eta Y_3 = 0$$

得到线性模型： $W_2^T X + b_2 = 2X^{(1)} + 2X^{(2)}$

④ 如此继续下去, 直到 $W_7 = (1,1)^T, b_7 = -3$

$$W_7^T X + b_7 = X^{(1)} + X^{(2)} - 3$$

基于梯度下降法的感知机算法（原始形式）

- 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^T	b	$W^T X + b$
0		0	0	0
1	X_1	(3,3)	1	$3X^{(1)} + 3X^{(2)} + 1$
2	X_3	(2,2)	0	$2X^{(1)} + 2X^{(2)}$
3	X_3	(1,1)	-1	$X^{(1)} + X^{(2)} - 1$
4	X_3	(0,0)	-2	-2
5	X_1	(3,3)	-1	$3X^{(1)} + 3X^{(2)} - 1$
6	X_3	(2,2)	-2	$2X^{(1)} + 2X^{(2)} - 2$
7	X_3	(1,1)	-3	$X^{(1)} + X^{(2)} - 3$
8	0	(1,1)	-3	$X^{(1)} + X^{(2)} - 3$

感知机算法（对偶形式）

- 基本思想：将 W, b 表示为样本 X_i 和标签 Y_i 的线性组合形式，通过求解其系数而得到 W, b ，对误分类点：

$$W \leftarrow W + \eta Y_i X_i,$$

$$b \leftarrow b + \eta Y_i$$

最后学习得到的 W, b

$$W = \sum_{i=1}^n a_i Y_i X_i$$

$$b = \sum_{i=1}^n a_i Y_i$$

感知机算法（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 $X_i \in R^n$ ， $Y_i \in Y = \{+1, -1\}$ ， $i = 1, 2, \dots, n$ ，学习率 η ($0 < \eta \leq 1$)

□ 输出： a, b ；感知机模型 $f(X) = \text{sign}(\sum_{j=1}^n a_j Y_j X_j \cdot X + b)$ ，其中 $a = (a_1, a_2, \dots, a_n)^T$

① $a \leftarrow 0, b \leftarrow 0$

② 在训练集中选取数据 (X_i, Y_i)

③ 如果 $Y_i (\sum_{j=1}^n a_j Y_j X_j \cdot X_i + b) \leq 0$ ， $a_i \leftarrow a_i + \eta$ ， $b \leftarrow b + \eta Y_i$

④ 转至第2步，直至训练集中没有误分类点

感知机算法（对偶形式）

- 对偶形式中训练样本仅以内积的形式出现
- 可以预先将训练集中样本间的内积计算出来并以矩阵形式存储
- Gram矩阵： $G = [X_i \cdot X_j]_{n \times n}$

感知机算法（对偶形式）

□ 例题：正样本： $X_1 = (3,3)^T, X_2 = (4,3)^T$, 负样本： $X_3 = (1,1)^T$

□ 求解

① 取 $a_i = 0, i = 1,2,3, b = 0, \eta = 1$

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

$$Y_i \left(\sum_{j=1}^n a_j Y_j X_j \cdot X_i + b \right) \leq 0$$

参数更新： $a_i \leftarrow a_i + \eta, b \leftarrow b + \eta Y_i$

感知机算法（对偶形式）

□ 例题：正样本： $X_1 = (3,3)^T, X_2 = (4,3)^T$ ，负样本： $X_3 = (1,1)^T$

④ 迭代。过程从略，结果列于下表。

k	0	1	2	3	4	5	6	7
		X_1	X_3	X_3	X_3	X_1	X_3	X_3
a_1	0	1	1	1	2	2	2	2
a_2	0	0	0	0	0	0	0	0
a_3	0	0	1	2	2	3	4	5
b	0	1	0	-1	0	-1	-2	-3

⑤ $W = 2X_1 + 0X_2 - 5X_3 = (1,1)^T, b = -3$

□ 分离超平面： $X^{(1)} + X^{(2)} - 3 = 0$

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

讨论

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。
2. 证明 W_s 不在解区的边界上。