第2章 感知机

感知机(Perceptron)

- 感知机(Rosenblatt，1957)
  - 二类分类的线性分类模型
    - 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

- 感知机学习：旨在求出将训练数据进行线性划分的分离超平面
  - 引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
- 特点
  - 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；

1 感知机模型

感知机模型

定义2.1(感知机)假设输入空间(特征空间)$\mathcal{X} \subseteq \mathbf{R}^n$，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

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

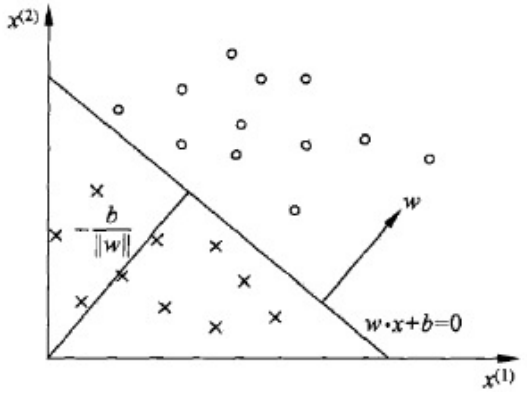
称为**感知机**。其中，w和b为感知机模型参数，$w \in \mathbf{R}^n$叫作**权值**(weight) 或权值向量(weight vector) , $b \in \mathbf{R}$叫作**偏置**(bias), $w \cdot x$表示内积。$sign$是符号函数

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

感知机是一种线性分类模型，属于判别模型。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(linear classification model) 或线性分类器(linear classifier) , 即函数集合$\{f \mid f(x) = w \cdot x + b\}$

感知机几何解释

- 特征空间$\mathbf{R}^n$中的超平面S
  - $w \cdot x + b = 0$
  - w为法向量，b截距
- 超平面将特征空间划分成两类
  - 指定超平面正向
  - 大于0，正类(圆圈)
  - 小于0，负类(叉)



2 感知机学习策略

线性可分性

- 如果存在某个超平面S
  - 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，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

- 如何定义损失函数？
  - 自然选择：误分类点的数目，但损失函数对参数不是连续可导，不适合优化
  - 另一选择：误分类点到超平面的总距离
    - $\frac{1}{||w||}$，其中，$||w||$是w的L_2范数
    - 误差应该有大小的区别
- 对于误分数据集合M，$x_i \in M$
  - $w \cdot x_i + b > 0$，$y_i = -1$。x_i到S的距离 $-\frac{1}{||w||} y_i (w \cdot x_0 + b)$
  - $w \cdot x_i + b < 0$，$y_i = +1$。x_i到S的距离 $-\frac{1}{||w||} y_i (w \cdot x_0 + b)$
- 所有误分类点到超平面S的总距离：$-\frac{1}{||w||} \sum_{x_i \in M} y_i (w \cdot x_0 + b)$
- 损失函数$L(w, b)$可以定义为：$L(w, b) = -\sum_{x_i \in M} y_i (w \cdot x_0 + b)$

感知机学习策略

- 损失函数： $L(w,b) = -\sum_{x_i \in M} y_i(w \cdot x_0 + b)$
  - 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
  - 误分类点越少，误分类点离超平面越近，损失函数值就越小
  - 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

3 感知机学习算法

可导损失函数

迭代梯度下降最小化损失函数

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_0，b_0，然后不断极小化目标函数$L(w,b)$
  - 计算$L(w,b)$梯度：

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

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

  - 随机选取一个误分类点(x_i, y_i)对w, b进行更新(负的梯度)

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

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

$\eta(0 < \eta \leqslant 1)$是步长，学习率(learning rate) 。通过迭代期待损失函数不断减小，直到为0

感知机学习算法的原始形式

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

(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 $y_i(w \cdot x_i + b) \leqslant 0$, $\quad w \leftarrow w + \eta y_i x_i$, $\quad b \leftarrow b + \eta y_i$

(4) 转至(2), 直至训练集中没有误分类点

【注】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

w是向量， $w \cdot x_i$计算量不够最优

感知机学习算法

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, $w = \left(w^{(1)}, w^{(2)}\right)^{\mathrm{T}}, x = \left(x^{(1)}, x^{(2)}\right)^{\mathrm{T}}$ 。

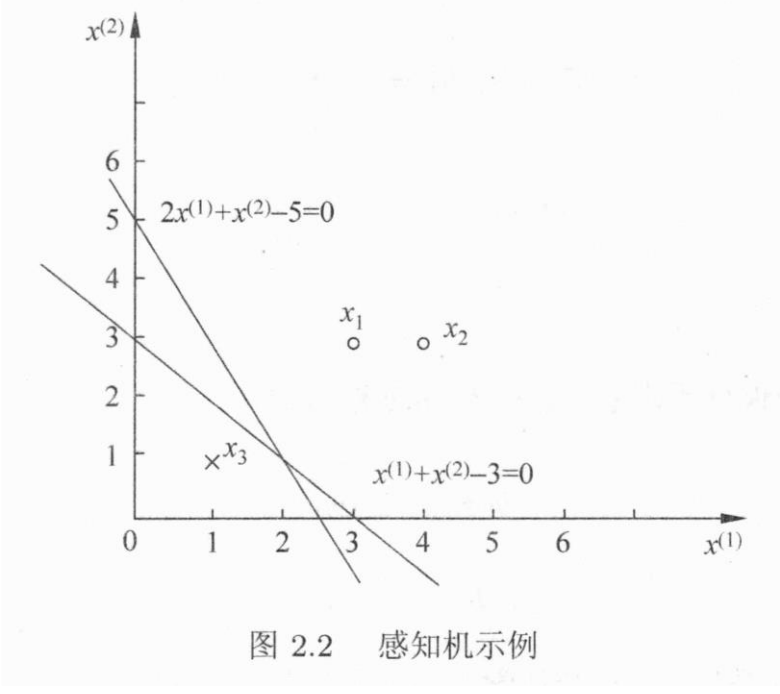


图 2.2　感知机示例

算法的收敛性

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

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

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

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

Novikoff定理

- 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 – 基本思想

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, 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 $\boldsymbol{\alpha_i = n_i \eta}$，$\boldsymbol{n_i}$**是点** $\boldsymbol{(x_i, y_i)}$ **被误分类的次数**。这样, 最后学习到 的 w, b 可以分别表示为

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

这里$\alpha_i \geqslant 0, i = 1,2, \cdots, N$，当$\eta = 1$时，$\alpha_i$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，越难正确分类，该实例点对学习结果影响最大

此时，感知机学习的目标变成了学习α_i或者n_i

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 $\eta(0<\eta\leqslant 1)$;

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

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

(4) 转至 (2) 直到没有误分类数据。

【注】从式(3)中可以看出对偶形式中训练实例**仅以内积（ $x_j\cdot x_i$ ）的形式出现**。为了**加速计算**，可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，该矩阵为Gram矩阵(Gram matrix)

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

【注】内积形式可以推广到更一般性的空间，如核方法构成的线性空间