第二章

感知机

袁春 清华大学深圳研究生院
李航 华为诺亚方舟实验室

目录

一、感知机

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；
- 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(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:

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

- 称为感知机,
- 模型参数：w x，内积，权值向量，偏置,
- 符号函数:

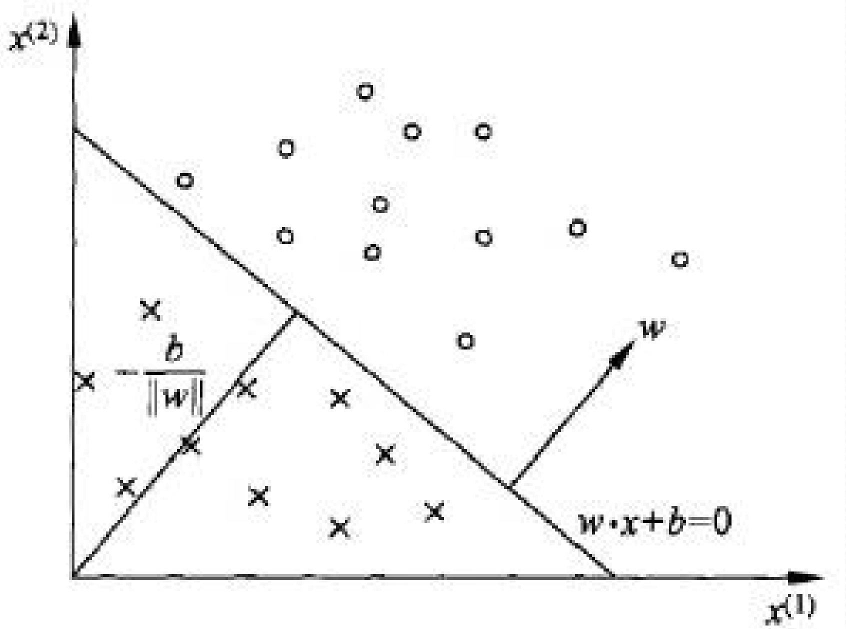
$$\operatorname{sign}(x) = \begin{cases} +1, & x \geqslant 0 \\ -1, & x < 0 \end{cases}$$

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

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：

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

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

三、感知机学习算法

- 概念
- 相关学术资源
- 应用
- 发展历程
- 国内外研究者
- 与数据挖掘的关系
- 相关学术期刊和会议
- 与统计学习的关系

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

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

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

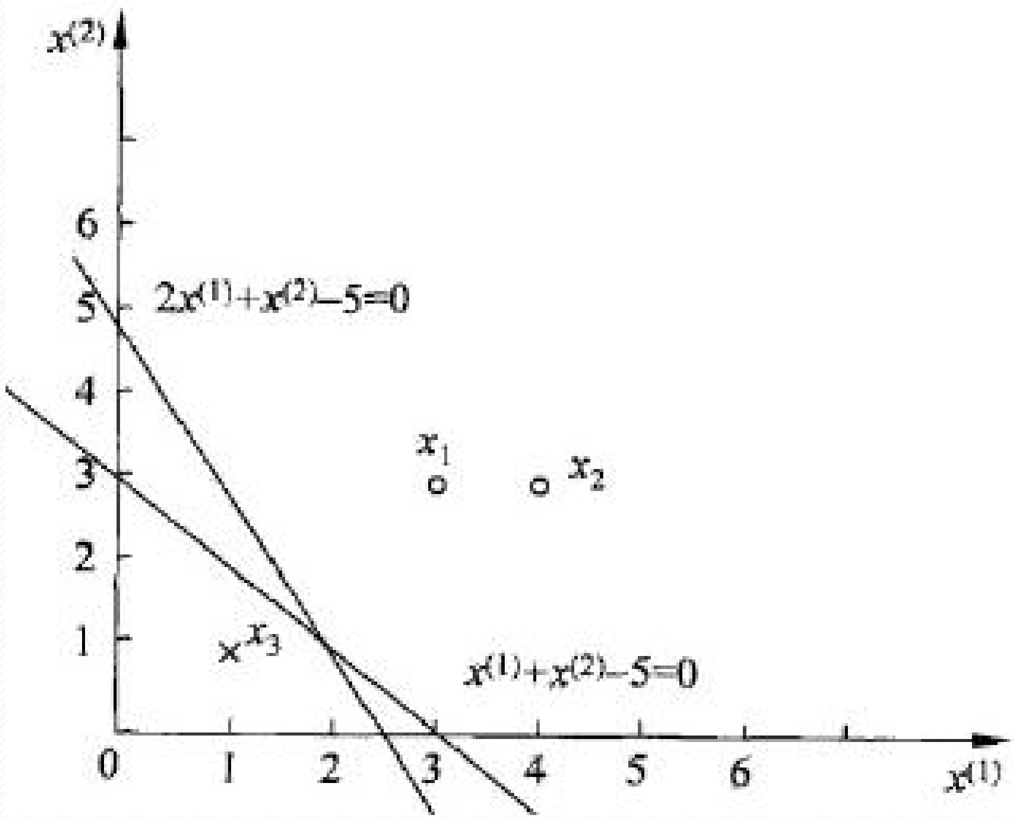
$$w\leftarrow w+\eta y_i x_i$$

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

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



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

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

- 得线性模型：$w_1 \cdot x + b_1 = 3x^{(1)} + 3x^{(2)} + 1$
- （3）x_2，显然，$y_i(w_1 \cdot x_i + b_1) > 0$，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，

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

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：$w_7 = (1,1)^{\mathrm{T}}$， $b_7 = -3$

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

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

迭代次数	误分类点	w	b	$w \cdot x + b$
0		0	0	0
1	x_1	$(3,3)^{\mathrm{T}}$	1	$3x^{(1)} + 3x^{(2)} + 1$
2	x_3	$(2,2)^{\mathrm{T}}$	0	$2x^{(1)} + 2x^{(2)}$
3	x_3	$(1,1)^{\mathrm{T}}$	-1	$x^{(1)} + x^{(2)} - 1$
4	x_3	$(0,0)^{\mathrm{T}}$	-2	-2
5	x_1	$(3,3)^{\mathrm{T}}$	-1	$3x^{(1)} + 3x^{(2)} - 1$
6	x_3	$(2,2)^{\mathrm{T}}$	-2	$2x^{(1)} + 2x^{(2)} - 2$
7	x_3	$(1,1)^{\mathrm{T}}$	-3	$x^{(1)} + x^{(2)} - 3$
8	0	$(1,1)^{\mathrm{T}}$	-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：$\hat{w}=(w^{\mathrm{T}},b)^{\mathrm{T}}$

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

- 定理：

- 设训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$ 是线性可分的，其中 $x_i\in\mathcal{X}=\mathbf{R}^n$，$y_i\in\mathcal{Y}=\{-1,+1\}$，$i=1,2,\cdots,N$，

感知机学习算法

则

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

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

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

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 $R=\max\limits_{1\leqslant i\leqslant N}\|\hat{x}_i\|$，感知机算法在训练集的误分类次数k满足不等式，

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

- 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：

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

- 第k个误分类实例的条件是：

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

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

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

感知机学习算法

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

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

- 推导两个不等式:
- (1)

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

- 由:

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

- 得:

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

感知机学习算法

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

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

（2）

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

- 则：

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

感知机学习算法

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

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

结合两个不等式：

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

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

得：

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

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xi和标记yi的线性组会的形式，通过求解其系数而求得w和b，对误分类点：

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

$\xrightarrow{\text{最后学习到的 } w,b}$

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

感知机学习算法

- 感知机学习算法的对偶形式：

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

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：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 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_N)^{\mathrm{T}}$

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

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

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

  （4） 转至（2）直到没有误分类数据.

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

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

参数更新

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

感知机学习算法

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

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

$b=-3$

分离超平面 $x^{(1)}+x^{(2)}-3=0$

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