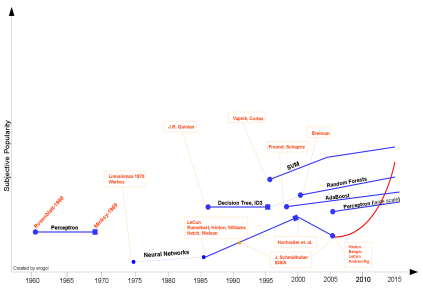
感知机

刘杰
人工智能学院
南開大學 Nankai University

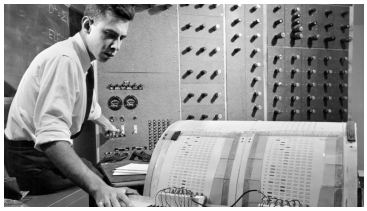
主要内容

- 背景
- 模型定义
- 学习策略
- 算法及分析

感知机（Perceptron）的诞生



感知机（Perceptron）的诞生

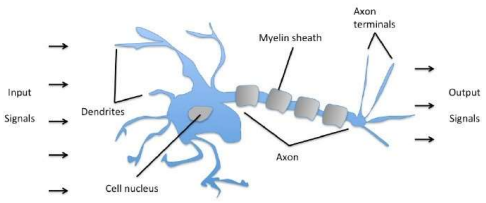


The perceptron's algorithm was invented in 1957 at the ***Cornell Aeronautical Laboratory*** *by* ***Frank Rosenblatt****, funded by the United States Office of Naval Research. It was intended to be a machine, rather than a program, and while its first implementation was in software for the IBM 704, it was subsequently implemented in custom-built hardware as the "****Mark 1 perceptron****". This machine was designed for image recognition: it had an array of 400 photocells, randomly connected to the "****neurons****". Weights were encoded in potentiometers, and weight updates during learning were performed by electric motors.*

Frank Rosenblatt '50, Ph.D. '56, works on the "perceptron" – what he described as the first machine "capable of having an original idea."

He was right – but it took half a century to prove it.

感知机（Perceptron）模型



感知机（Perceptron）模型

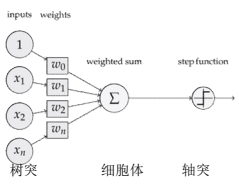
输入：实例的特征向量

权重：权重是在模型训练期间计算的值。初始我们用一些初始值开始权值，这些值针对每个训练误差进行更新。我们用$[w_1, w_2, w_3, \ldots w_n]$。

偏置：偏置神经元允许分类器向左或向右移动决策边界,有助于更快、更高质量地训练模型。

加权：加权求和是我们在与每个特征值相关联的每个权重$[w_n]$相乘后得到的值的总和$[x_n]$。

输出：实例的类别，取+1和-1；



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：

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

称为感知机，

模型参数：权值向量 *w*，偏置*b*;

符号函数：

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

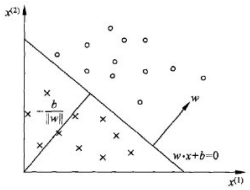
感知机模型

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：



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

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：

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

选取误分类点更新：

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

感知机学习算法

输入：训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$
其中 $x_i \in \mathcal{X} = \mathbf{R}^n, y_i \in = \{-1, +1\}, i = 1, 2, ..., N$
学习率 $\eta(0 < \eta \leq 1)$;

输出：w, b 感知机模型 $f(x) = sign(w \cdot x + b)$

（1）选取初值 w_0, b_0

（2）在训练集中选取数据 (x_i, y_i)

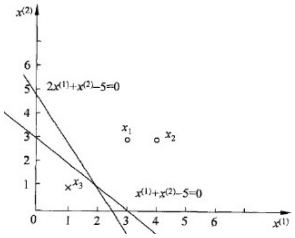
（3）如果 $y_i(w \cdot x_i + b) \leq 0$

$w_l \leftarrow w_{l-1} + \eta y_i x_i$
$b_l \leftarrow b_{l-1} + \eta y_i$

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



感知机学习算法

解：构建优化问题：

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

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$w, b$

得线性模型：　　$w_1 \cdot x + b_1 = 3x^{(1)} + 3x^{(2)} + 1$

(3)x_2 ，显然，$y_i(w_i \cdot x_i + b_1) > 0$，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：　　$w_7 = (1,1)^T, b_7 = -3$

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

分离超平面：　　$x^{(1)} + x^{(2)} - 3 = 0$

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

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_1	$(0,0)^T$	−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	$x^{(1)}+x^{(2)}-3$

感知机算法收敛性分析

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：　$\hat{w} = (w^T, b)^T$

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

定理：

设训练数据 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ 是线性可分的，

其中：

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

感知机算法收敛性分析

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

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

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

感知机算法收敛性分析

证明：(1) $\hat{w}_{opt} \cdot \hat{x} = w_{opt} \cdot x + b_{opt} = 0$

由数据线性可分，

使$\|\hat{w}_{opt}\| = 1$，对有限的点，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：

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

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

感知机算法收敛性分析

(2) 令 $R = \max_{1 \geq i \geq N} \|\hat{x}_i\|$ ，算法在训练集上的误分类次数*k*满足不等式 $k \leq (\frac{R}{\gamma})^2$

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：

$$y_1(\hat{w}_{k-1} \cdot \hat{x}_i) = y_1(w_{k-1} \cdot x_1 + b_{k-1}) \leq 0$$

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

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

感知机算法收敛性分析

推导两个不等式：

①由：$\hat{w}_k \cdot \hat{w}_{opt} \geq k\eta\gamma$

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

得： $\geq \hat{w}_{k-1} \cdot \hat{w}_{opt} + \eta\gamma$

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

感知机算法收敛性分析

②

则：$\|\hat{w}_k\|^2 \geq k\eta^2 R^2$

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

感知机算法收敛性分析

结合两个不等式：$k\eta\gamma \leq \hat{w}_k \cdot \hat{w}_{opt} \leq \|\hat{w}_k\| \|\hat{w}_{opt}\| \leq \sqrt{k}\eta R$

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

得： $$k \leq (\frac{R}{\gamma})^2$$

感知机算法收敛性分析

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。

感知机算法是不稳定的：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*，对误分类点：

$$\begin{aligned}w &\leftarrow w + \eta y_i x_i \\ b &\leftarrow b + \eta y_i\end{aligned} \Longrightarrow \begin{aligned}w &= \sum_{i=1}^{N} \alpha_i y_i x_i \\ b &= \sum_{i=1}^{N} \alpha_i y_i \\ \alpha &= n_i \eta\end{aligned}$$

感知机学习对偶形式

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

学习率是 $\eta(0<\eta\leq 1)$;

输出：α,b; 感知机模型 $f(x)=sign(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x+b)$

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

感知机学习对偶形式

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

(2) 在训练集中选取数据 (x_i,y_i)

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

$\alpha_i\leftarrow\alpha_i+\eta$

$b\leftarrow b+\eta y_i$

(4) 转至(2)直到没有误分类数据。

Gram矩阵 $G=[x_i\cdot x_j]_{N\times N}$

感知机学习对偶形式

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

参数更新 $\alpha_i\leftarrow\alpha_i+1, b\leftarrow b+y_i$

感知机学习对偶形式

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

表2.2

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

$b=-3$

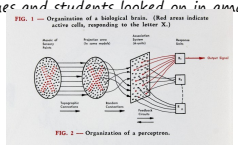
感知机模型 $f(x)=sign(x^{(1)}+x^{(2)}-3)$

感知机的局限

Frank Rosenblatt, left, and Charles W. Wightman work on part of the unit that became the first perceptron in December 1958.

Marvin Minsky, who was a grade behind Rosenblatt at the Bronx High School of Science, was an MIT professor whose research into neural networks left him deeply skeptical of Rosenblatt's claims.

At conferences, Rosenblatt and Minsky publicly debated the perceptron's viability, as their colleagues and students looked on in amazement.



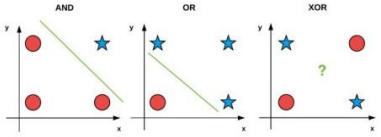
感知机的局限

异或运算

x_0	x_1	$x_0\&x_1$
0	0	0
0	1	0
1	0	0
1	1	1

x_0	x_1	$x_0\|x_1$
0	0	0
0	1	1
1	0	1
1	1	1

x_0	x_1	$x_0\wedge x_1$
0	0	0
0	1	1
1	0	1
1	1	0




谢谢!
31