人工智能的数学基础

华东师范大学 数学科学学院 黎芳(教授) 2019年9月16日

第二章 感知机（perception）

Table of Contents

第二章 感知机（perception） 1
2.1 感知机模型 1
2.2 感知机学习策略 2
2.2.1 数据集的线性可分性 2
2.2.2 感知机学习策略 2
2.3 感知机学习算法 3
2.3.1 感知机学习算法的原始形式 3
2.3.2 算法收敛性 5
2.3.3 感知机学习算法的对偶形式 5
本章小结 7
作业 7

2.1 感知机模型

1957年由Rosenblatt提出，是神经网络与支持向量机的基础.

定义**2.1**（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如下函数

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

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).

感知机是一种线性分类模型，属于判别模型. 感知机的假设空间是定义在特征空间中的所有线性分类模型或线性分类器，即函数集合$\{f | f(x) = w \cdot x + b\}$.

几何解释：

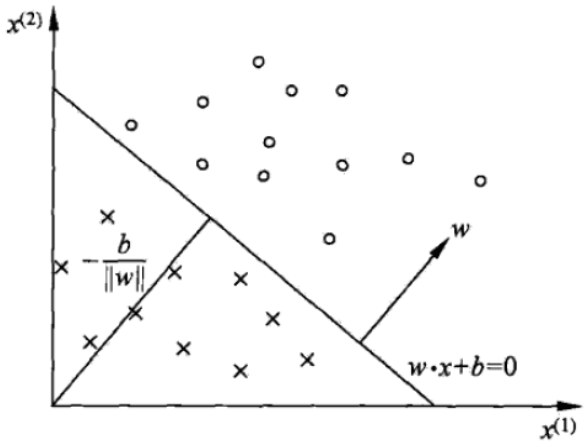


图 2.1 感知机模型

2.2 感知机学习策略

2.2.1 数据集的线性可分性

定义2.2(数据集的线性可分性) 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，即对所有的$y_i = +1$的实例i，有$w \cdot x_i + b > 0$；对所有$y_i = -1$的实例i，有$w \cdot x_i + b < 0$；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.

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

感知机损失函数：$L(w, b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b)$

2.3 感知机学习算法

2.3.1 感知机学习算法的原始形式

求解最优化问题：$\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b)$

随机梯度下降法（stocastic gradient descent）

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：

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

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

随机选择一个误分类点，对w, b进行更新

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

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

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

（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.

算法的直观解释

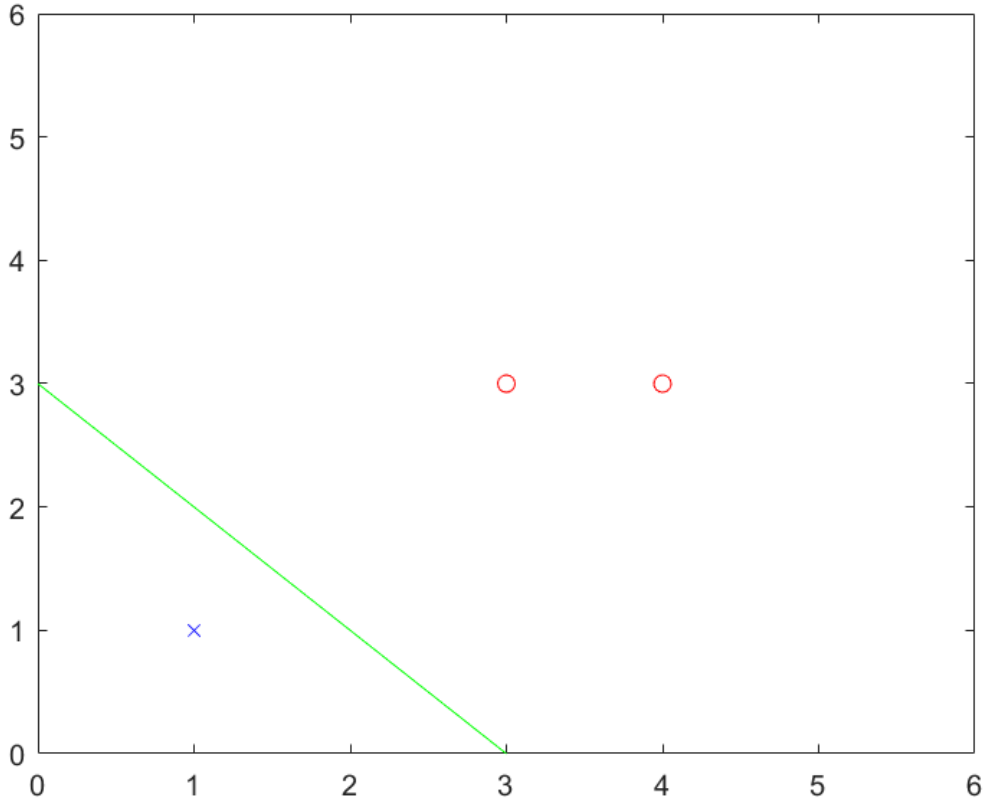
例**2.1** 已知训练数据集中正实例点是$x_1 = (3,3)^T, x_2 = (4,3)^T$，负实例点是$x_3 = (1,1)^T$，试用感知机学习算法的原始形式求感知机模型 $f(x) = \text{sign}(w \cdot x + b)$. 这里，$w = (w^{(1)}, w^{(2)})^{\mathrm{T}}, x = (x^{(1)}, x^{(2)})^{\mathrm{T}}$.

```
w = [0,0]';
b = 0;
eta = 1;
x = [3 4 1; 3 3 1];
y = [1 1 -1];
figure, plot(x(1,1:2),x(2,1:2),'ro')
hold on; plot(x(1,3),x(2,3),'bx')
axis([0 6 0 6])
```

```
x0 = 0:0.1:6;
for i = [1 3 3 3 1 3 3]
    if y(i)*(w'*x(:,i)+b)<=0
        w = w+eta*y(i)*x(:,i)
        b = b+eta*y(i)
    end
end
```

```
w = 2×1
     3
     3
b = 1
w = 2×1
     2
     2
b = 0
w = 2×1
     1
     1
b = -1
w = 2×1
     0
     0
b = -2
w = 2×1
     3
     3
b = -1
w = 2×1
     2
     2
b = -2
w = 2×1
     1
     1
b = -3
```

```
hold on; plot(x0,(-b-w(1)*x0)/w(2),'g')
```



2.3.2 算法收敛性

记$\hat{w}=(w^T,b)^T,\hat{x}=(x^T,1)^T,\Rightarrow\hat{w}\cdot\hat{b}=w\cdot x+b.$

定理**2.1(Novikoff)** 设训练数据$i=I,2,\cdots,N$集是线性$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$是可分的，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$\|\hat{w}_{opt}\|=1$的超平面$\hat{w}_{opt}\cdot\hat{b}_{opt}=w_{opt}\cdot x+b_{opt}$将训练数据集完全正确分开；且存在$\gamma>0,$对所有$i=1,2,\cdots,N,\ y_i(\hat{w}_{\text{opt}}\cdot\hat{x}_i)=y_i(w_{\text{opt}}\cdot x_i+b_{\text{opt}})\geq\gamma.$

(2) 令$R=\max_{1\leq i\leq N}\|\hat{x}_i\|,$ 则感知机算法 2.1 在训练数据集上的误分类次数 k 满足$k\leq\left(\frac{R}{\gamma}\right)^2.$

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.

2.3.3 感知机学习算法的对偶形式

基本想法：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：$\begin{cases} w \leftarrow w+\eta y_i x_i \\ b \leftarrow b+\eta y_i \end{cases}$

w b n i n_i $\alpha = n_i\eta$, w b $\begin{cases} w=\sum_{i=1}^{N}\alpha_i y_i x_i \\ b=\sum_{i=1}^{N}\alpha_i y_i \end{cases}$

设 和 修改了 次，其中第 个实例误分了 次，令 最终学习到的 和 为：

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

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

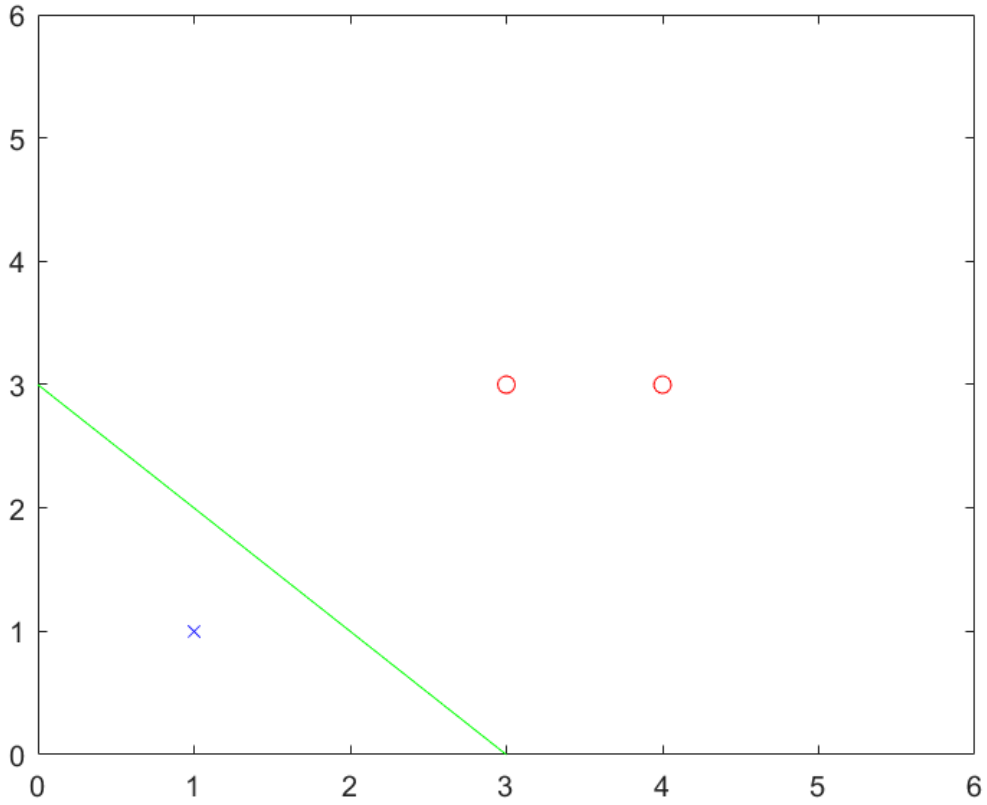
（4）转至（2）直到没有误分类数据.

令$G=[x_i\cdot x_j]_{N\times N}$为Gram矩阵，可以预先存储.

例**2.2** 数据同例2.1，试用感知机学习算法对偶形式求解感知机模型.

```
x = [3 4 1; 3 3 1];
y = [1 1 -1];
figure, plot(x(1,1:2),x(2,1:2),'ro')
hold on; plot(x(1,3),x(2,3),'bx')
axis([0 6 0 6])
x0 = 0:0.1:6;
% initialization
alpha = zeros(3,1);
b = 0;
eta = 1;
% Gram matrix
G = x'*x;
for i = [1 3 3 3 1 3 3]
    if y(i)*sum(G*(alpha.*y')+b)<=0
        alpha(i) = alpha(i)+eta;
        b = b+eta*y(i);
    end
end
% end
w = x*(alpha.*y');
%b = sum(alpha.*y');
```

```
hold on; plot(x0,(-b-w(1)*x0)/w(2),'g')
```



本章小结

1. 感知机是根据输入实例的特征向量x对其进行二类分类的线性分类模型$f(x) = \text{sign}(w \cdot x + b)$. 感知机模型对应于输入空间（特征空间）中的分离超平面$w \cdot x + b = 0$.
2. 感知机学习的策略是极小化损失函数；$\min\limits_{w,b} L(w,b) = -\sum\limits_{x_i \in M} y_i(w \cdot x_i + b)$ 损失函数对应于误分类点到分离超平面的总距离.
3. 感知机学习算法是基于随机梯度下降法的对损失函数的最优化算法，有原始形式和对偶形式. 算法简单且易于实现. 原始形式中, 首先任意选取一个超平面, 然后用梯度下降法不断极小化目标函数. 在这个过程中一次随机选取一个误分类点使其梯度下降.
4. 当训练数据集线性可分时, 感知机学习算法是收敛的. 感知机算法在训练数据集上的误分类次数K满足不等式: $k \leq \left(\frac{R}{\gamma}\right)^2$.

作业

2.1 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.

2.2 模仿例题2.1, 构建从训练数据求解感知机模型的例子.