感知机

1.感知机模型

- 输入空间：$\mathcal{X} \subseteq \mathbf{R^n}$
- 输出空间：$\mathcal{Y} \subseteq \{+1, -1\}$
- 决策函数：$f(x) = sign(w \cdot x + b)$

其中w叫做权值或权值向量，b叫做偏置。$sign(x)$为符号函数。

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

感知机是一个线性分类器，属于判别模型，感知机模型的**假设空间**是所有的**线性分类模型**或**线性分类器**。即$\{f|f(x) = w \cdot x + b\}$。

- 感知机中的线性方程将空间切分为两个部分，从而数据点被分为两类。这个线性方程我们称为分离超平面。

2.策略

为了将找出将数据分开的超平面$w \cdot x + b = 0$，确定超平面的模型参数w, b,我们需要确定一个损失函数。

- 损失函数是误分类点的总数，但是这样的损失函数无法对w, b求导，不易优化。
- 误分类点到超平面的距离之和。
  - 平面的一般方程：

    $$AX + BY + CZ + D = 0$$

  - 平面外一点(x_0, y_0, z_0)到平面的距离:

    $$\frac{|Ax_0 + By_0 + Cz_0 + D|}{\sqrt{A^2 + B^2 + C^2}}$$

    或：

    $$\frac{|w \cdot x + b|}{||w||}$$

- 对于误分类的点(x_i, y_i):

  $$-y_i(w \cdot x_i + y_i) > 0$$

- 假设误分类点的集合为M，则这些点到超平面的距离和为：

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

  忽略$\frac{1}{||w||}$,我们可以得到感知机的损失函数：

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

3.算法

3.1 原始形式：

- 梯度下降,损失函数$L(w,b)$的梯度为：

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

于是我们选取误分类点进行梯度更新：

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

其中η为学习率。

- 算法

> 输入：$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$
> $x_i\in\mathcal{X}=\mathbf{R^n}, y_i\in\mathcal{Y}=\{-1,+1\}, i=1,2,\ldots,N;\quad 0<\eta\leqslant 1$
>
> 输出：$w,b; f(x)=sign(w\cdot x+b)$
>
> 1. 选取初值w_0,b_0
> 2. 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

3.2 对偶形式：

假设一个样本(x_i,y_i)点在更新的过程中使用了n_i次，最后能学习到的w,b可以表示为：

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

其中n_i如果值越大，代表这个样本经常被误分。他就越靠近超平面。于是感知机模型变为：

$$f(x)=sign(w\cdot x+b)=sign(\sum_{j=1}^{N} n_j\eta y_j x_j\cdot x+\sum_{j=1}^{N} n_j\eta y_j)$$

此时，需要更新的参数不再是w,b，而是$n_i, i=1,2,3,\ldots,N$。

- 算法：

> 输入：$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$
> $x_i\in\mathcal{X}=\mathbf{R^n, y_i\in\mathcal{Y}=\{-1,+1\}, i=1,2,\ldots,N; 0<\eta\leqslant 1}$
>
> 输出：
>
> $$\alpha,b;f(x)=sign\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x+b\right)$$
> $$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_N)^T$$
>
> 1. $\alpha\leftarrow 0, b\leftarrow 0$
> 2. 训练集中选取数据(x_i,y_i)
> 3. 如果$y_i\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x+b\right)\leqslant 0$
>
> $$\alpha_i\leftarrow\alpha_i+\eta(n_i=n_i+1)$$
> $$b\leftarrow b+\eta y_i$$
>
> 4. 转至(2)，直至训练集中没有误分类点

其中$\alpha_i=n_i\eta$。

样本点中的特征向量以内积的方式存在于感知机的算法中，如果提前计算好内积(Gram矩阵)，就会大大提升计算速度。

3.3 算法的收敛性：

设训练数据集$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$是线性可分的，其中
$x_i\in\mathcal{X}=\mathbf{R^n, y_i\in\mathcal{Y}=\{-1,+1\}, i=1,2,\ldots,N; 0<\eta\leqslant 1}$，则：

(1)存在满足条件的$||\hat{w}_{opt}||=1$的超平面$\hat{w_{opt}}\cdot\hat{x}=w_{opt}\cdot x+b_{opt}=0$将训练数据集完全分开，且存在$\gamma>0$,对于所有输入：

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

(2)令$R=\max\limits_{1\leq i\leq N}||\hat{x_i}||$,则误分类的次数$k$:

$$k\leq(\frac{R}{\gamma})^2$$

上式表明，只要数据集线性可分，就一定会找到一个超平面将数据正确分类。即表明算法收敛。