感知机原理

【参考资料】

李航 《统计学习方法》

1. 基本形式

感知机模型的形式为

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

其中$sign$是符号函数，即

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

感知机是线性判别模型。对于感知机来说，数据集需要满足线性可分性，若样本标签$y_i \in \mathcal{Y}=\{+1,-1\}$，则对所有$y_i=+1$的实例$i$，有$w \cdot x_i+b>0$，对所有$y_i=-1$的实例$i$，有$w \cdot x_i+b<0$。

感知机的损失函数为

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

其中M为误分类点的集合。

显然，损失函数$L(w,b)$是非负的：若没有误分类点，则损失为0；误分类点越少，以及误分类点离分离超平面越近，则损失越小。

2. 学习算法

感知机学习算法是误分类驱动的，具体采用随机梯度下降法。在极小化损失函数的过程中，不是一次使误分类集合M中所有误分类点的梯度下降，而是一次随机选取一个误分类点使其梯度下降。

假设误分类点集合M是固定的，那么损失函数$L(w,b)$由

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

给出。

随机选取一个误分类点(x_i,y_i)，对w，b进行更新：

$$\begin{gathered}w \leftarrow w+\eta y_i x_i \\ b \leftarrow b+\eta y_i\end{gathered}$$

其中$0 < \eta \leqslant 1$为学习率。

迭代持续进行直至训练集中没有误分类点。

3. 收敛性

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

$$y_i\left(\hat{w}_{\text{opt}} \cdot \hat{x}_i\right)=y_i\left(w_{\text{opt}} \cdot x_i+b_{\text{opt}}\right) \geqslant \gamma \tag{3.1}$$

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

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

【证明】

（1）

取分离超平面为$\hat{w}_{\text{opt}} \cdot \hat{x}=w_{\text{opt}} \cdot x+b_{\text{opt}}=0$，使$\|\hat{w}_{\text{ opt }}\|=1$，由于对有限的$i=1,2,\cdots,N$，均有

$$y_i\left(\hat{w}_{\text{ opt }} \cdot \hat{x}_i\right)=y_i\left(w_{\text{ opt }} \cdot x_i+b_{\text{ opt }}\right)>0$$

所以存在

$$\gamma=\min_i\left\{y_i\left(w_{\text{opt}} \cdot x_i+b_{\text{opt}}\right)\right\}$$

使

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

（2）

为方便起见，将偏置b并入权重向量w，记作$\hat{w}=\left(w^{\mathrm{T}}, b\right)^{\mathrm{T}}$，同样也将输入向量加以扩充，加进常数1，记作$\hat{x}=\left(x^{\mathrm{T}}, 1\right)^{\mathrm{T}}$。这样，$\hat{x} \in \mathbf{R}^{n+1}$，$\hat{w} \in \mathbf{R}^{n+1}$。显然$\hat{w} \cdot \hat{x}=w \cdot x+b$。

感知机算法从$\hat{w}_0=0$开始，如果实例被误分类，则更新权重。令$\hat{\boldsymbol{w}}_{k-1}$是第$i$个误分类实例之前的扩充权重向量，即

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

则第k个误分类实例的条件是

$$y_i\left(\hat{w}_{k-1} \cdot \hat{x}_i\right)=y_i\left(w_{k-1} \cdot x_i+b_{k-1}\right) \leqslant 0 \tag{3.2}$$

若(x_i, y_i)是被$\hat{w}_{k-1}=\left(w_{k-1}^{\mathrm{T}}, b_{k-1}\right)^{\mathrm{T}}$误分类的数据，则$w$和$b$的更新是

$$\begin{aligned} w_k & \leftarrow w_{k-1}+\eta y_i x_i \\ b_k & \leftarrow b_{k-1}+\eta y_i \end{aligned}$$

即

$$\hat{w}_k = \hat{w}_{k-1} + \eta y_i \hat{x}_i \tag{3.3}$$

下面推导两个不等式

- $$\hat{w}_k \cdot \hat{w}_{\text{opt}} \geqslant k\eta\gamma \tag{3.4}$$

  由式（3.1）和（3.3）得

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

  由此递推即得不等式（3.4）

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

- $$\|\hat{w}_k\|^2 \leqslant k\eta^2 R^2 \tag{3.5}$$

  由式（3.2）及（3.3）得

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

结合不等式（3.4）和（3.5）即得

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

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

于是

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

定理表明，误分类的次数k是有上界的，经过有限次搜索可以找到将训练数据集完全正确分开的分离超平面。也就是说，当数据集线性可分时，感知机学习算法原始形式迭代是收敛的。

当训练集线性不可分时，感知机学习算法不收敛，迭代结果会发生震荡。