



感知机

📖 数据建模与分析

📖 [感知机](#) 是 F.Rosenblatt 提出的一个神经网络，可被视为是一种最简单形式的前馈神经网络，是一种二元线性分类器。

Definition: 假设输入空间 $\mathcal{X} \subseteq \mathbb{R}^n$ ，输出空间 $\mathcal{Y} = \{+1, -1\}$ ，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：

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

称为感知机，其中 $\vec{w} \in \mathbb{R}^n$ 称为权值向量 (weight vector)， $b \in \mathbb{R}$ 叫做偏置 (bias)， $w \cdot x$ 表示内积， sign 表示符号函数，即

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

Definition: (数据集的线性可分性) 给定一个数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 其中 $x_i \in \mathcal{X} = \mathbb{R}^N, y_i \in \mathcal{Y} = \{+1, -1\} i = 1, 2, \dots, N$ ，如果存在超平面 \mathcal{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 $y_i = +1$ 的实例 i ，有 $w \cdot x_i + b > 0$ ，对所有 $y = -1$ 的实例，有 $w \cdot w_i + b < 0$ ，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

损失函数：误分类点到超平面的总距离

对于点 (x_0, y_0) 其到超平面 $w \cdot x + b = 0$ 的距离为

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

这是一个绝对值函数，我们知道绝对值函数不易于优化，对于误分类点，定义给出他们应满足 $-y_i(w \cdot x_i + b) > 0$ ，因此上式也可以写作

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

因此总距离为

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

其中 M 为误分类点的集合，为了使计算方便，我们取 $\|w\| = 1$ 即做归一化处理。

Definition 给定训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 其中 $x_i \in \mathcal{X} = \mathbb{R}^N, y_i \in \mathcal{Y} = \{+1, -1\} i = 1, 2, \dots, N$ 。感知机 $\text{sign}(w \cdot x + b)$ 学习的损失函数定义为

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

其中 M 为误分类点的集合。这个损失函数就是感知机学习的经验风险函数。

显然，损失函数是非负的，如果没有误分类点，损失函数值为 0。而且，误分类点越少，误分类点离超平面越近，损失函数值越小。一个特定的样本点的损失函数：在误分类时是参数 w, b 的线性函数，在正确分类时是 0。因此，给定训练数据集 T ，损失函数 $L(w, b)$ 是 w, b 的连续可导函数。

因此求解如下最优化问题：

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


我们采取随机梯度下降法 ([W Stochastic gradient descent](#)) 来求解这个问题。在每次迭代中，我们随机均匀采样的一个样本索引 $i \in \{1, \dots, n\}$ ，并计算梯度 $\nabla f_i(x)$ 来更新权重，在本问题中，对应如下迭代公式：

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

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

其中 η 为学习率，每次迭代的计算开销为 $\mathcal{O}(1)$ 。随机梯度 $\nabla f_i(x)$ 是对梯度 $\nabla f(x)$ 的无偏估计：

$$\mathbb{E}[\nabla f_i(x)] = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x) = \nabla f(x)$$

 **Warning**

但是，算法的解并不唯一，不同的初值得到的解可能不同

将 b 合并入权重向量 \vec{w} ，记为 $\hat{w} = (\vec{w}^T, b)^T$ ，同时，数据可记为 $\hat{x} = (x^T, 1)^T$ ，超平面则可改写成 $\hat{w} \cdot \hat{x} = 0$ ，对应的迭代公式则为

$$\hat{w}_k = \hat{w}_{k-1} + \eta y_i \hat{x}_i \quad (1)$$

Theorem (Novikoff): 设训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 其中 $x_i \in \mathcal{X} = \mathbb{R}^N, y_i \in \mathcal{Y} = \{+1, -1\} i = 1, 2, \dots, N$ ，则

1. 存在满足条件 $\|\hat{w}_{opt}\| = 1$ 的超平面 $\hat{w}_{opt} \cdot \hat{x} = 0$ 将训练数据集完全正确分开，且存在 $\gamma > 0$ 对所有的 $i = 1, 2, \dots, N$

$$y_i(\hat{w}_{opt} \cdot \hat{x}_i) = y_i(w_{opt} \cdot x_i + b_{opt}) \geq \gamma \quad (2)$$

2. 令 $R = \max_{1 \leq i \leq N} \|x_i\|$ ，则算法在训练数据集上的误分类次数 k 满足不等式

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

Proof.

先证明第一个不等式，由于训练数据集是线性可分的，故存在超平面可将数据集完全正确分开，取超平面 $\hat{w}_{opt} \cdot x = 0$ ，使 $\|\hat{w}_{opt}\| = 1$ ，对于有限的 $i = 1, 2, \dots, N$ ，有：

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

所以存在 $\gamma = \min_i \{y_i(w_{opt} \cdot x_i + b_{opt})\}$ 使得

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



在这证明第二个定理前，先推导两个不等式

$$\begin{aligned}\hat{w}_k \cdot \hat{w}_{opt} &\geq k\eta\gamma \\ \|\hat{w}_k\|^2 &\leq k\eta^2 R^2\end{aligned}$$

(1) 和 (2) 给出

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

递推可得

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

由于

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

有柯西施瓦兹不等式可得

$$\begin{aligned}k\eta\gamma &\leq \hat{w}_k \cdot \hat{w}_{opt} \leq \|\hat{w}_k\| \|\hat{w}_{opt}\| \leq \sqrt{k}\eta R \\ k^2\gamma^2 &\leq kR^2\end{aligned}$$

即

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

Summary

上述收敛性定理表明：

- 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；
- 若需要到唯一分离超平面，需增加约束，如SVM(support vector machine,支持向量机)；

感知机算法的对偶形式的基本想法：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 ，即：

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

对偶形式便于计算。