感知机

2021-09-11

1. 感知机模型简介

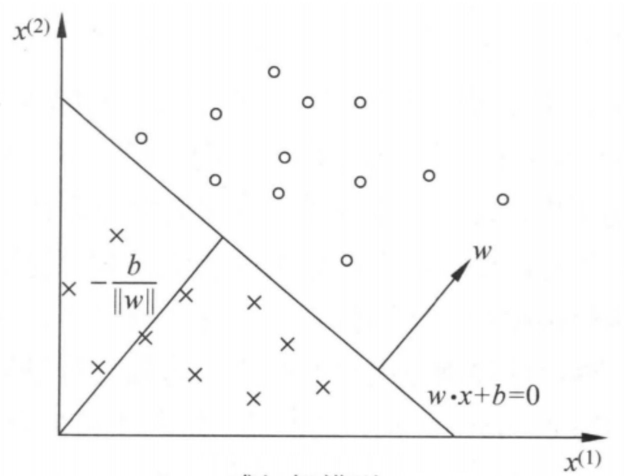
感知机（perceptron）是二分类的线性分类模型，其输入为实例的特征向量，输出为实例的类别，取值为$\{+1,-1\}$。感知机对应于输入空间（特征空间）中将实例划分为正负两类的分离超平面，属于判别模型。

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如下函数称为**感知机**:

$$f(x)=\operatorname{sign}(w \cdot x+b) \tag{13}$$

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), $w \cdot x$ 表示 w 和 x 的内积。 sign 是符号函数, 即

$$\operatorname{sign}(x)=\begin{cases}+1, & x \geqslant 0 \\ -1, & x<0\end{cases} \tag{14}$$



感知机模型

2. 感知机学习策略

空间$\mathbf{R}^n$中任意一点x_0到超平面S的距离：

$$\frac{1}{\|w\|}|w \cdot x_0+b| \text{其中，} \|w\| \text{是} w \text{的} L_2 \text{范数} \tag{15}$$

对误分类的数据(x_i, y_i)来说：

$$-y_i(w \cdot x+b)>0 \tag{16}$$

因此，误分类点x_i到超平面S的距离是：

$$-\frac{1}{\|w\|}|w \cdot x_i + b| \tag{17}$$

假设超平面S的误分类点集合为M，那么所有误分类点到超平面S的总距离为：

$$-\frac{1}{\|w\|}\sum_{x_i \in M} y_i\left(w \cdot x_i + b\right) \tag{18}$$

根据以上推导，可得出感知机的**损失函数**：

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

$$L(w, b) = -\sum_{x_i \in M} y_i\left(w \cdot x_i + b\right) \text{其中，} M \text{为误分类点的集合} \tag{19}$$

3. 感知机学习算法

3.1. 原始形式

优化目标：

$$\min_{w,b} L(w, b) = -\sum_{x_i \in M} y_i\left(w \cdot x_i + b\right) \text{其中，} M \text{为误分类点的集合} \tag{20}$$

采用梯度下降法进行优化，损失函数的梯度为：

$$\begin{gathered}\nabla_w L(w, b) = -\sum_{x_i \in M} y_i x_i \\ \nabla_b L(w, b) = -\sum_{x_i \in M} y_i\end{gathered} \tag{21}$$

随机选取一个误分类点(x_i, y_i)，对w, b进行更新：

$$\begin{gathered}w \leftarrow w + \eta y_i x_i \\ b \leftarrow b + \eta y_i \\ \text{其中，} \eta(0 < \eta \leq 1) \text{为学习率}\end{gathered} \tag{22}$$

算法1.1（感知机学习的原始形式）

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

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

(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 $y_i(w \cdot x_i + b) \leq 0$ ：

$w \leftarrow w + \eta y_i x_i$

$b \leftarrow b + \eta y_i$

(4) 转至(2)，直至训练集中没有误分类点

3.2. 对偶形式

基本想法是：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。不失一般性，在算法1.1中可假设初始值w_0, b_0均为0。对分类点(x_i, y_i)通过

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

逐步修改w, b。

设修改n次，则w, b关于(x_i, y_i)的增量分别是$\alpha_i y_i x_i$和$\alpha_i y_i$，这里$\alpha_i = n_i \eta$。从学习过程中不难看出，最后学习到的w, b可以分别表示为：

$$\begin{aligned} w &= \sum_{i=1}^{N} \alpha_i y_i x_i \\ b &= \sum_{i=1}^{N} \alpha_i y_i \end{aligned} \tag{23}$$

其中，$\alpha_i \geq 0$，$i = 1, 2, \cdots, N$，当$\eta = 1$时，α_i表示第i个实例点由于误分类而进行更新的次数。

由α_i定义可知，$\alpha_{i+1} = \eta(n_i + 1) = \eta n_i + \eta = \alpha_i + \eta$

算法1.2（感知机学习的对偶形式）

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

输出：w, 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，$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$。

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

(2) 在训练集中选取数据（x_i, y_i）

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

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

(4) 转至(2)，直至训练集中没有误分类点

对偶形式中训练实例仅以内积的形式出现。为了方便，可以预先将训练集中实例间的内积计算出来并以矩阵的形式存储，该矩阵称为Gram矩阵：

$$\boldsymbol{G} = [x_i \cdot x_j]_{N \times N} \tag{24}$$