统计学习第二章

2022年8月9日 星期二 23:10

S2. 感知机

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

"." 表示向量的内积，b为偏差 bias.

训练集：$T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3) \ldots (x_n, y_n)\}$

$x_i \in \mathcal{X} = R^n$，$y_i \in \mathcal{Y} = \{-1, +1\}$

感知机器学习策略

○输入空间 R^n 中任一点 x_0 到超平面 S 的距离：

$\frac{1}{\|w\|} |w \cdot x_0 + b|$，其中 $\|w\|$ 为二范数.

对于误分类数据：

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

误分类点到超平面的距离：

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

x_i 到超平面S距离：$-\frac{1}{\|w\|} \sum y_i (w \cdot x_i + b)$

不考虑 $\frac{1}{\|w\|}$，得损失函数：

$L(w, b) = -\sum y_i (w \cdot x_i + b)$

感知机学习算法

由上述知，目标为 $\min_{w, b} L(w, b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$

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

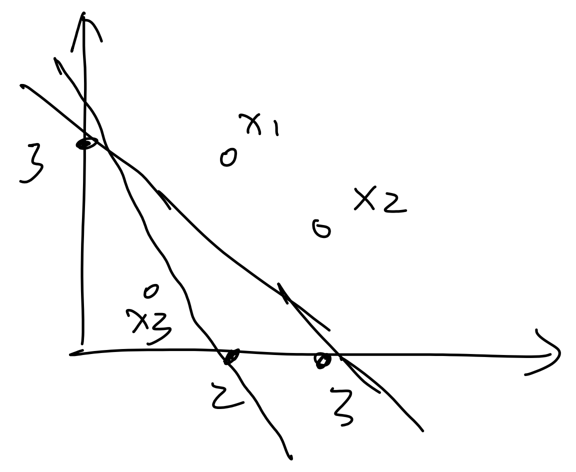
$\nabla_b L(w, b) = -\sum_{x_i \in M} y_i$

随机选取一个误分类点 (x_i, y_i)，对 w, b 更新：

$w := w + \eta y_i x_i$　其中 η 为学习率（也可以为 α.，$0 < \eta \le 1$）

$b := b + \eta y_i$

eg: 正实例：$x_1 = (3, 3)^T$，$x_2 = (4, 3)^T$，负实例：$x_3 = (1, 1)^T$



「解」：目标：$\min_{w, b} L(w, b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$

(1) 令 $w_0 = 0$，$b_0 = 0$，$\eta = 1$.

对于 $x^{(1)}$，$w \cdot x_1 + b = 0$.

未被正确分类

(2) 令 $w_1 = w_0 + x_1 y_1 = (3, 3)^T$

$b_1 = b_0 + y_1 = 1$.（w为向量，b为数）

对于 $x^{(3)}$，$w \cdot x_3 + b = 7 > 0$，未被正确分类.

(3) $w_2 = w_1 - x_3 y_3 = (2, 2)^T$，$b_2 = 0$

(4) $w_3 = w_2 - x_3 y_3 = (1, 1)^T$，$b_3 = b_2 - y_3 = -1$

...

总结：不断为误分类点进行迭代.

○算法的收敛性：算法原始形式收敛！

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

$\hat{w} \cdot \hat{x} = w \cdot x + b$

$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ b \end{bmatrix} \cdot [x_1\ x_2\ \cdots\ x_n, 1]$

(w^T, x)

$\hat{w}^T \cdot x$（内积）

☆Novikoff 定理：

设 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$

是线性可分的

(1) 存在 $\|w_{opt}\| = 1$ 的超平面

$\hat{w}_{opt} \cdot \hat{x} = w_{opt} + b_{opt} = 0$ 将训练数据集完全分开.

且存在 $\gamma > 0$，对所有 $i \in (1, N)$

$y_i (\hat{w}_{opt} \cdot \hat{x}_i) = y_i (w_{opt} \cdot x_i + b) \ge \gamma$.

(2) 令 $R = \max_{1 \le i \le N} \|\hat{x}_i\|$，则感知机误分类次数 k 满足：

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

具体证明略（Tips：(1) 用下确界法；(2) 用迭代法.）

• 感知学习机的对偶形式：

双对把 w和 b 写成：

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

输入：$T = \{(x_1, y_1), (x_2, y_2) \ldots, (x_N, y_N)\}$，学习率 η $(0 < \eta \le 1)$

输出：模型：$f(x) = \mathrm{sign}\left(\sum_{j=1}^{N} \alpha_j y_j x_j \cdot x + b\right)$

其中 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$

(1) $\alpha := 0$，$b := 0$

(2) 选取 T 上 (x_i, y_i)

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

则 $\begin{cases} \alpha_i := \alpha_i + \eta \\ b := b + \eta y_i \end{cases}$

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

为方便计算，将预先训练集中的内积计算出来以矩阵的形式存储，即为 Gram matrix.