感知机

张腾

2025 年 3 月 24 日

1 感知机

给定 $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i\in[m]} \subseteq \mathbb{R}^n \times \{1, -1\}$，感知机 (perceptron) 学一个形如 $f(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x})$ 的超平面模型对数据进行分类。

设初始 $\boldsymbol{w}_1 = \boldsymbol{0}$，第 t 轮的样本为 $(\boldsymbol{x}_{i_t}, y_{i_t})$，其中 $(i_1, \dots, i_m)$ 是 $(1, \dots, m)$ 的某个随机排列，若 $y_{i_t}\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t} \le 0$，即预测错误，则更新 $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \eta y_{i_t}\boldsymbol{x}_{i_t}$，其中 $\eta > 0$ 为学习率 (learning rate)。注意

$$y_{i_t}\boldsymbol{w}_{t+1}^\top \boldsymbol{x}_{i_t} = y_{i_t}\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t} + \eta \boldsymbol{x}_{i_t}^\top \boldsymbol{x}_{i_t} > y_{i_t}\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t}$$

即更新后 $\boldsymbol{w}_{t+1}$ 在样本 $(\boldsymbol{x}_{i_t}, y_{i_t})$ 上的预测比原来有所改善，即使更新完还不对，多错几次最终总能对。

考虑如下优化问题

$$\min_{\boldsymbol{w}} \ L(\boldsymbol{w}) = \frac{1}{m}\sum_{i\in[m]} \max\{0, -y_i \boldsymbol{w}^\top \boldsymbol{x}_i\}$$

目标函数次梯度为

$$\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} = -\frac{1}{m}\sum_{i\in[m]} y_i \boldsymbol{x}_i \mathbb{I}(y_i \boldsymbol{w}^\top \boldsymbol{x}_i \le 0)$$

用随机梯度下降 (SGD) 进行求解，设第 t 轮随机采样的样本为 $(\boldsymbol{x}_{i_t}, y_{i_t})$，则更新为

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \eta y_{i_t}\boldsymbol{x}_{i_t}\mathbb{I}(y_{i_t}\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t} \le 0) = \begin{cases} \boldsymbol{w}_t + \eta y_{i_t}\boldsymbol{x}_{i_t}, & \text{if } y_{i_t}\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t} \le 0 \\ \boldsymbol{w}_t, & \text{o.w.} \end{cases}$$

其中 $\eta > 0$ 是 SGD 的步长 (step size)，不难看出感知机就是在用 SGD 优化损失函数 $L(\boldsymbol{w})$。

2 Novikoff's 定理

定理 1 (Novikoff's 定理). 设 $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i\in[m]} \subseteq \mathbb{R}^n \times \{1, -1\}$，假设

1. 对 $\forall i \in [m]$ 存在 $r > 0$ 使得 $\|\boldsymbol{x}_i\| \le r$
2. 存在 $\rho > 0$ 和单位向量 $\boldsymbol{v} \in \mathbb{R}^n$ 使得对于任意 $i \in [m]$ 有 $y_i \boldsymbol{v}^\top \boldsymbol{x}_i \ge \rho$

那么感知机的更新次数不超过 r^2/ρ^2。

证明. 设集合 $\mathcal{I}$ 是感知机进行更新的轮次集合，易知有

$$|\mathcal{I}|\rho \le \sum_{t\in\mathcal{I}} y_{i_t}\boldsymbol{v}^\top \boldsymbol{x}_{i_t} \le \|\boldsymbol{v}\| \left\|\sum_{t\in\mathcal{I}} y_{i_t}\boldsymbol{x}_{i_t}\right\| = \left\|\sum_{t\in\mathcal{I}} \frac{\boldsymbol{w}_{t+1}-\boldsymbol{w}_t}{\eta}\right\| = \left\|\frac{\boldsymbol{w}_{M+1}}{\eta}\right\|$$

其中 $M=\max\{t \mid t\in\mathcal{I}\}$，进一步有

$$\begin{aligned}|\mathcal{I}|\rho &\le \left\|\frac{\boldsymbol{w}_{M+1}}{\eta}\right\| = \sqrt{\left\|\frac{\boldsymbol{w}_{M+1}}{\eta}\right\|^2} = \sqrt{\sum_{t\in\mathcal{I}} \frac{\|\boldsymbol{w}_{t+1}\|^2-\|\boldsymbol{w}_t\|^2}{\eta^2}} = \sqrt{\sum_{t\in\mathcal{I}} \frac{\|\boldsymbol{w}_t+\eta y_{i_t}\boldsymbol{x}_{i_t}\|^2-\|\boldsymbol{w}_t\|^2}{\eta^2}} \\ &= \sqrt{\sum_{t\in\mathcal{I}} \frac{2\eta y_{i_t}\boldsymbol{w}_t^\top \boldsymbol{x}_{i_t}+\eta^2\|\boldsymbol{x}_{i_t}\|^2}{\eta^2}} \le \sqrt{\sum_{t\in\mathcal{I}} \|\boldsymbol{x}_{i_t}\|^2} \le \sqrt{|\mathcal{I}|r^2}\end{aligned}$$

移项整理即可。 ♣

注. 这个界只跟归一化的间隔 r/ρ 有关，与维度 n 无关，同时也与学习率 η 无关，因此一般教材上讲感知机时都是将 η 定为 1 的。

注. 这个界是紧的，考虑这样一个例子，$\boldsymbol{x}_1,\dots,\boldsymbol{x}_m$ 分别为 $\mathbb{R}^m$ 的 m 个单位向量，前 m 轮感知机每次的预测值都 ≤ 0，故会连续更新 m 次，之后得到

$$\boldsymbol{w}_{m+1} = \eta\sum_{i\in[m]} y_i\boldsymbol{x}_i$$

由 $\boldsymbol{x}_1,\dots,\boldsymbol{x}_m$ 之间的两两正交性知

$$\frac{y_j\boldsymbol{w}_{m+1}^\top\boldsymbol{x}_j}{\|\boldsymbol{w}_{m+1}\|} = \frac{y_j\eta\sum_{i\in[m]} y_i\boldsymbol{x}_i^\top\boldsymbol{x}_j}{\|\boldsymbol{w}_{m+1}\|} = \frac{\eta}{\eta\sqrt{m}} = \frac{1}{\sqrt{m}}$$

即 $\boldsymbol{w}_{m+1}$ 可将所有样本正确分类且间隔为 $1/\sqrt{m}$，下面说明 $\rho = 1/\sqrt{m}$，结合 $r^2=1$ 可知界是紧的。若单位向量 $\boldsymbol{v}=[v_i]_{i\in[m]}$ 使得间隔 $\rho > 1/\sqrt{m}$，则对于 $\forall i\in[m]$ 有 $y_i\boldsymbol{v}^\top\boldsymbol{x}_i = y_iv_i \ge \rho > 1/\sqrt{m}$，于是 $\|\boldsymbol{v}\|^2 = \sum_{i\in[m]} v_i^2 > \sum_{i\in[m]} 1/m = 1$，这与 $\boldsymbol{v}$ 是单位向量矛盾。

注. Novikoff's 定理表明只要样本线性可分，感知机都会在有限步内停止；但当间隔 ρ 很小时，收敛可能会很慢，事实上存在一些极端情况，感知机的更新次数达到 $\Omega(2^m)$。设 $\mathbb{R}^m$ 中的 m 个样本为

$$\begin{aligned}
\boldsymbol{x}_1 &= [\ 1, && 0, && 0, && 0, && 0, && \cdots, && 0\ ], && y_1 = 1\\
\boldsymbol{x}_2 &= [\ 1, && -1, && 0, && 0, && 0, && \cdots, && 0\ ], && y_2 = -1\\
\boldsymbol{x}_3 &= [\ -1, && -1, && 1, && 0, && 0, && \cdots, && 0\ ], && y_3 = 1\\
\boldsymbol{x}_4 &= [\ 1, && 1, && 1, && -1, && 0, && \cdots, && 0\ ], && y_4 = -1\\
\boldsymbol{x}_5 &= [\ -1, && -1, && -1, && -1, && 1, && \cdots, && 0\ ], && y_5 = 1\\
&\ \ \vdots\\
\boldsymbol{x}_i &= [\ (-1)^i, && \cdots, && (-1)^i, && (-1)^{i+1}, && 0, && \cdots, && 0\ ], && y_i = (-1)^{i+1}
\end{aligned}$$

易知

$$y_i\boldsymbol{x}_i = [\underbrace{-1,\dots,-1}_{i-1\text{个}},1,0,\dots,0]$$

- 欲使 $\boldsymbol{x}_1$ 分类正确，需要有 $w_1 > 0$，加一次 $\eta y_1 \boldsymbol{x}_1$ 即可
- 在 $w_1 > 0$ 的情况下，欲使 $\boldsymbol{x}_2$ 分类正确，需要有 $w_2 > 0$，同样加一次 $\eta y_2 \boldsymbol{x}_2$ 可使得 $w_2 > 0$，但这会同时将 w_1 减小 η，因此要想 w_1 不发生变化，加一次 $\eta y_2 \boldsymbol{x}_2$ 后需跟着再加一次 $\eta y_1 \boldsymbol{x}_1$
- 在 $w_1, w_2 > 0$ 的情况下，欲使 $\boldsymbol{x}_3$ 分类正确，需要有 $w_3 > 0$，同样加一次 $\eta y_3 \boldsymbol{x}_3$ 可使得 $w_3 > 0$，但这会同时将 w_1、w_2 减小 η，因此加一次 $\eta y_3 \boldsymbol{x}_3$ 后需跟着再加一次 $\eta y_1 \boldsymbol{x}_1$、一次 $\eta y_2 \boldsymbol{x}_2$，而加一次 $\eta y_2 \boldsymbol{x}_2$ 又得再加一次 $\eta y_1 \boldsymbol{x}_1$

如此易知，设欲使 $w_i > 0$ 且同时不改变其他 w_j $(j < i)$ 所需的更新次数为 a_i，由前面的分析

$$
\begin{aligned}
a_1 &= 1 = 2^0 \\
a_2 &= 1 + a_1 = 2 = 2^1 \\
a_3 &= 1 + a_2 + a_1 = 4 = 2^2 \\
a_4 &= 1 + a_3 + a_2 + a_1 = 8 = 2^3 \\
&\vdots \\
a_i &= 1 + a_{i-1} + \cdots + a_1
\end{aligned}
$$

由数学归纳法易证 $a_i = 2^{i-1}$，因此光是将 $\boldsymbol{w}$ 的全部分量变为正数所需的更新次数就至少为 $2^0 + 2^1 + \cdots + 2^{m-1} = 2^m$，而将全部样本分类正确的要求更苛刻，因此所需的更新次数更多，故为 $\Omega(2^m)$。

3 线性不可分

Novikoff's 定理的假设条件是数据线性可分，对于线性不可分的数据有：

定理 2. 设 $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i \in [m]} \subseteq \mathbb{R}^n \times \{1, -1\}$，假设

1. 对 $\forall i \in [m]$ 存在 $r > 0$ 使得 $\|\boldsymbol{x}_i\| \leq r$
2. 对于单位向量 $\boldsymbol{v} \in \mathbb{R}^n$ 和 $\rho > 0$，记 $\epsilon_i = \max\{0, \rho - y_i \boldsymbol{v}^\top \boldsymbol{x}_i\}$ 及 $\delta = \sqrt{\sum_{i \in [m]} \epsilon_i^2}$

则存在映射将所有样本映射到高维空间后线性可分，且感知机的更新次数不超过 $(r+\delta)^2/\rho^2$。

证明. 对 $\forall \boldsymbol{x}_i$，原来的特征保留不变，后面添加 m 位，其中第 i 位为取值待定的 A，其余为零

$$
\boldsymbol{x}_i = [x_{i,1}, \ldots, x_{i,n}] \mapsto \boldsymbol{x}_i' = [x_{i,1}, \ldots, x_{i,n}, 0, \ldots, 0, \underbrace{A}_{\text{第 } n+i \text{ 个分量}}, 0, \ldots, 0]
$$

$\boldsymbol{v}$ 也要映射到 $\mathbb{R}^{n+m}$ 中：

$$
\boldsymbol{v} = [v_1, \ldots, v_n] \mapsto \boldsymbol{v}' = \left[\frac{v_1}{B}, \ldots, \frac{v_n}{B}, \frac{y_1 \epsilon_1}{AB}, \ldots, \frac{y_m \epsilon_m}{AB}\right]
$$

其中 B 是待定参数，根据 $\boldsymbol{v}'$ 为单位向量有

$$
1 = \|\boldsymbol{v}'\|^2 = \sum_{i \in [n]} \frac{v_i^2}{B^2} + \sum_{i \in [m]} \frac{\epsilon_i^2}{A^2 B^2} = \frac{1}{B^2} + \frac{\delta^2}{A^2 B^2} \Longrightarrow B = \sqrt{1 + \frac{\delta^2}{A^2}}
$$

如此映射后有

$$y_i \boldsymbol{v}'^{\top} \boldsymbol{x}'_i = y_i \left(\frac{\boldsymbol{v}^{\top} \boldsymbol{x}_i}{B} + \frac{y_i \epsilon_i}{B} \right) = \frac{y_i \boldsymbol{v}^{\top} \boldsymbol{x}_i + \epsilon_i}{B} \geq \frac{\rho}{B}$$

这意味着 $\boldsymbol{x}'_1, \dots, \boldsymbol{x}'_m$ 线性可分且间隔至少为 ρ/B，注意此时有 $\|\boldsymbol{x}'\|^2 \leq r^2 + A^2$，于是更新次数不超过

$$\frac{r^2 + A^2}{\rho^2/B^2} = \frac{(r^2 + A^2)(1 + \delta^2/A^2)}{\rho^2} = \frac{r^2 + \delta^2 + A^2 + r^2\delta^2/A^2}{\rho^2}$$

上式在 $A^2 = r\delta$ 可使界最紧，为 $(r+\delta)^2/\rho^2$。 ♣

4 大间隔

Novikoff's 定理要求存在超平面 $\boldsymbol{v}^{\top}\boldsymbol{x}$ 间隔为 ρ，但感知机的更新条件是 $y_i \boldsymbol{w}^{\top} \boldsymbol{x}_i \leq 0$，因此最终模型只能做到 $y_i \boldsymbol{w}^{\top} \boldsymbol{x}_i > 0$，间隔是没有保证的，可能会很小。若将更新条件修改成 $y_i \boldsymbol{w}^{\top} \boldsymbol{x}_i / \|\boldsymbol{w}\| < k\rho$，其中 $k \in (0,1)$，这样最终就能得到一个间隔至少为 $k\rho$ 的超平面。下面证明经此改动后，感知机的更新次数不超过 $4r^2/(1-k)^2\rho^2$。因此若想得到间隔至少为 $\rho/3$ 的超平面，更新次数上界就是 $9r^2/\rho^2$；若想得到间隔至少为 $\rho/2$ 的超平面，更新次数上界就是 $16r^2/\rho^2$。

不妨固定学习率 η 为 1，此时同样有

$$\begin{aligned} |\mathcal{I}|\rho &\leq \sum_{t \in \mathcal{I}} y_{i_t} \boldsymbol{v}^{\top} \boldsymbol{x}_{i_t} \leq \|\boldsymbol{v}\| \left\| \sum_{t \in \mathcal{I}} y_{i_t} \boldsymbol{x}_{i_t} \right\| = \left\| \sum_{t \in \mathcal{I}} (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) \right\| = \|\boldsymbol{w}_{M+1}\| = \sqrt{\|\boldsymbol{w}_{M+1}\|^2} \\ &= \sqrt{\sum_{t \in \mathcal{I}} (\|\boldsymbol{w}_{t+1}\|^2 - \|\boldsymbol{w}_t\|^2)} = \sqrt{\sum_{t \in \mathcal{I}} (\|\boldsymbol{w}_t + y_{i_t} \boldsymbol{x}_{i_t}\|^2 - \|\boldsymbol{w}_t\|^2)} = \sqrt{\sum_{t \in \mathcal{I}} (2 y_{i_t} \boldsymbol{w}_t^{\top} \boldsymbol{x}_{i_t} + \|\boldsymbol{x}_{i_t}\|^2)} \end{aligned}$$

Novikoff's 定理的证明中，$y_{i_t} \boldsymbol{w}_t^{\top} \boldsymbol{x}_{i_t}$ 直接放缩成零，从而得到 $|\mathcal{I}|\rho \leq \sqrt{|\mathcal{I}|r^2}$。现在 $y_{i_t} \boldsymbol{w}_t^{\top} \boldsymbol{x}_{i_t}$ 的上界变成了 $k\rho\|\boldsymbol{w}_t\|$，需要更细致的处理。

首先若 $\|\boldsymbol{w}_{M+1}\| < 2r^2/(1-k)\rho$，注意 $2/(1-k) > 2$，于是

$$|\mathcal{I}| \leq \frac{\|\boldsymbol{w}_{M+1}\|}{\rho} < \frac{2r^2}{(1-k)\rho^2} < \frac{4r^2}{(1-k)^2\rho^2}$$

故不妨设 $\|\boldsymbol{w}_{M+1}\| \geq 2r^2/(1-k)\rho$，下面求 $\|\boldsymbol{w}_{M+1}\|$ 的上界，注意

$$\|\boldsymbol{w}_{t+1}\|^2 = \|\boldsymbol{w}_t + y_{i_t} \boldsymbol{x}_{i_t}\|^2 = \|\boldsymbol{w}_t\|^2 + 2 y_{i_t} \boldsymbol{w}_t^{\top} \boldsymbol{x}_{i_t} + \|\boldsymbol{x}_{i_t}\|^2 \leq \|\boldsymbol{w}_t\|^2 + 2k\rho\|\boldsymbol{w}_t\| + r^2 < (\|\boldsymbol{w}_t\| + k\rho)^2 + r^2$$

于是

$$\|\boldsymbol{w}_{t+1}\| - \|\boldsymbol{w}_t\| - k\rho \leq \frac{r^2}{\|\boldsymbol{w}_{t+1}\| + \|\boldsymbol{w}_t\| + k\rho}$$

若 $\|\boldsymbol{w}_t\|$ 和 $\|\boldsymbol{w}_{t+1}\|$ 中至少有一个 $\geq 2r^2/(1-k)\rho$，则

$$\|\boldsymbol{w}_{t+1}\| \leq \|\boldsymbol{w}_t\| + k\rho + \frac{r^2}{2r^2/(1-k)\rho + k\rho} \leq \|\boldsymbol{w}_t\| + k\rho + \frac{(1-k)\rho}{2} = \|\boldsymbol{w}_t\| + \frac{(1+k)\rho}{2}$$

$\|\boldsymbol{w}\|$ 初始为 $\|\boldsymbol{w}_1\| = 0$，最终为 $\|\boldsymbol{w}_{M+1}\| \geq 2r^2/(1-k)\rho$，因此必存在某个轮次，不妨设为第 t' 轮，满足 $\|\boldsymbol{w}_{t'}\| < 2r^2/(1-k)\rho$ 且对 $\forall t > t' : \|\boldsymbol{w}_t\| \geq 2r^2/(1-k)\rho$，于是有

$$\|\boldsymbol{w}_{t'+1}\| \leq \|\boldsymbol{w}_{t'}\| + \frac{(1+k)\rho}{2},\ \ldots,\ \|\boldsymbol{w}_{M+1}\| \leq \|\boldsymbol{w}_M\| + \frac{(1+k)\rho}{2}$$

上面的不等式个数不超过 $|\mathcal{I}|$ 个，累加可得

$$\|\boldsymbol{w}_{M+1}\| \leq \|\boldsymbol{w}_{t'}\| + |\mathcal{I}|\frac{(1+k)\rho}{2} < \frac{2r^2}{(1-k)\rho} + |\mathcal{I}|\frac{(1+k)\rho}{2}$$

回代易知有

$$|\mathcal{I}|\rho \leq \|\boldsymbol{w}_{M+1}\| < \frac{2r^2}{(1-k)\rho} + |\mathcal{I}|\frac{(1+k)\rho}{2} \Longrightarrow |\mathcal{I}| < \frac{4r^2}{(1-k)^2\rho^2}$$