# 第 2 章 学习回答是或否

## 2.1 感知机假设集

**一个简单的假设集合：感知机（Perceptron）**

对于输入特征向量

$$\mathbf{x} = (x_1, x_2, \ldots, x_d) \quad \text{——表示客户的各项特征，}$$

计算加权“分数”并

- 批准信贷 若 $\sum_{i=1}^{d} w_i x_i >$ 阈值
- 拒绝信贷 若 $\sum_{i=1}^{d} w_i x_i <$ 阈值

输出：$y \in \{+1\ (\text{好客户}),\ -1\ (\text{坏客户})\}$，0 被忽略。

线性公式 $h \in \mathcal{H}$ 定义为

$$h(\mathbf{x}) = \text{sign}\Big(\sum_{i=1}^{d} w_i X_i - \text{阈值}\Big)$$

历史上称为感知机假设。

**命题 2.1.1 (感知机的向量形式)**

设输入特征向量为 $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top$，引入齐次坐标

$$\tilde{\mathbf{x}} = (1, x_1, x_2, \ldots, x_d)^\top.$$

令参数向量

$$\mathbf{w} = (w_0, w_1, \ldots, w_d)^\top, \quad w_0 = -\text{threshold}.$$

则感知机假设可写成

$$h(\mathbf{x}) = \text{sign}\Big(\Big(\sum_{i=1}^{d} w_i x_i\Big) + (-\text{threshold})(+1)\Big) = \text{sign}(\mathbf{w}^\top \tilde{\mathbf{x}})$$

每一个向量 $\mathbf{w}$ 对应一个假设 $h$；几何上 $\mathbf{w}$ 定义了一个穿过原点的超平面。

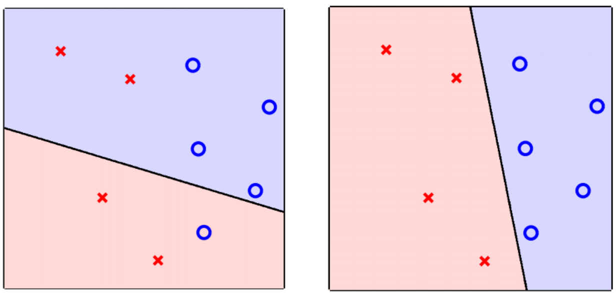

图 2.1.1: 感知机在二维空间上的分类

**命题 2.1.2 (感知机在二维空间中的几何意义)**

设顾客特征 $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$，标签 $y \in \{\circ\,(+1),\ \times\,(-1)\}$。则感知机假设

$$h(\mathbf{x}) = \mathrm{sign}(w_0 + w_1 x_1 + w_2 x_2)$$

对应平面上的直线（在 $\mathbb{R}^d$ 中为超平面）。直线的一侧为正类，另一侧为负类；不同的直线即为不同的顾客分类器。因此，感知机即线性二分类器。

## 2.2 感知机学习算法（PLA）

**算法 2.2.1:** 感知机学习算法 PLA（含 Cyclic 实现）

**输入:** 数据集 $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$，其中 $y_n \in \{+1, -1\}$
**输出:** 权重向量 $\mathbf{w}_{\mathrm{PLA}}$ 作为最终假设 $g$
初始化 $\mathbf{w}^0 \leftarrow \mathbf{0}$;
    选择遍历顺序 $\pi$：$\pi \leftarrow (1, 2, \ldots, N)$ ;  // 朴素循环
    或预先生成随机排列 $\pi$;
**repeat**  // 直到完整一圈无误分
    **foreach** $i = 1, \ldots, N$(按顺序 $\pi$) **do**
        **if** $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}_{\pi(i)}) \neq y_{\pi(i)}$ **then**
            $\mathbf{w} \leftarrow \mathbf{w} + y_{\pi(i)} \mathbf{x}_{\pi(i)}$;
**until** 一个完整循环无更新;
**return** $\mathbf{w}_{\mathrm{PLA}} \leftarrow \mathbf{w}$;

笔记 为便于可视化，可令所有 $x_i \gg x_0$（即取 $x_0 = 1$ 作为基准）。

## 2.3 感知机学习算法（PLA）的保证

**命题 2.3.1 (PLA：$\mathbf{w}_t$ 与 $\mathbf{w}_f$ 越来越对齐)**

若数据集 $D$ 线性可分，则存在完美权重 $\mathbf{w}_f$ 使得

$$y_n = \mathrm{sign}(\mathbf{w}_f^\top \mathbf{x}_n), \quad \forall n.$$

此时

$$y_{n(t)}\, \mathbf{w}_f^\top \mathbf{x}_{n(t)} \geq \min_n y_n\, \mathbf{w}_f^\top \mathbf{x}_n > 0.$$

更新一次后

$$\begin{aligned}
\mathbf{w}_f^\top \mathbf{w}_{t+1} &= \mathbf{w}_f^\top (\mathbf{w}_t + y_{n(t)} \mathbf{x}_{n(t)}) \\
&= \mathbf{w}_f^\top \mathbf{w}_t + y_{n(t)} \mathbf{w}_f^\top \mathbf{x}_{n(t)} \\
&\geq \mathbf{w}_f^\top \mathbf{w}_t + \min_n y_n \mathbf{w}_f^\top \mathbf{x}_n > \mathbf{w}_f^\top \mathbf{w}_t.
\end{aligned}$$

故 $\mathbf{w}_t$ 在每次更新后都与 $\mathbf{w}_f$ 更加对齐。

**命题 2.3.2 (PLA：$w_t$ 的范数增长受限)**

PLA 仅在误分时更新，此时

$$\operatorname{sign}(\mathbf{w}_t^\top \mathbf{x}_{n(t)}) \neq y_{n(t)} \iff y_{n(t)}\mathbf{w}_t^\top \mathbf{x}_{n(t)} \leq 0.$$

于是

$$\begin{aligned}\|\mathbf{w}_{t+1}\|^2 &= \|\mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)}\|^2 \\ &= \|\mathbf{w}_t\|^2 + 2y_{n(t)}\mathbf{w}_t^\top \mathbf{x}_{n(t)} + \|y_{n(t)}\mathbf{x}_{n(t)}\|^2 \\ &\leq \|\mathbf{w}_t\|^2 + 0 + \max_n \|\mathbf{x}_n\|^2.\end{aligned}$$

从 $\mathbf{w}_0 = \mathbf{0}$ 开始，经 $T$ 次误分修正后

$$\frac{\mathbf{w}_f^\top}{\|\mathbf{w}_f\|}\frac{\mathbf{w}_T^\top}{\|\mathbf{w}_T\|} \geq \sqrt{T} \cdot \text{constant}$$

**命题 2.3.3 (PLA 误分次数上界)**

定义

$$R^2 = \max_n \|\mathbf{x}_n\|^2, \qquad \rho = \min_n y_n \frac{w_f^T x_n}{\|w_f\|}$$

则 PLA 的误分总次数 $T$ 满足

$$T \leq \frac{R^2}{\rho^2}.$$

**证明** 首先，注意到内积 $\frac{w_f^T}{\|w_f\|}\frac{w_t}{\|w_t\|}$ 由 Cauchy-Schwarz 不等式限制，其最大值为 1，当 $w_t$ 与 $w_f$ 共线时取等。

其次，考虑感知机算法 (PLA) 经过 $T$ 次错误更正，内积增量约为 $\sqrt{T} \cdot c$，其中 $c$ 为正常数（由更正的平均效应决定）。

由于内积的上限为 1，且初始内积至少为 $\rho$（定义 $\rho = \min_n y_n \frac{w_f^T x_n}{\|w_f\|}$），更正次数 $T$ 受限于 $\frac{1}{c^2}$。

结合 $R^2 = \max_n \|x_n\|^2$ 和感知机收敛性，得出 $T \leq \frac{R^2}{\rho^2}$。 ■

## 2.4 不可分数据

**更多关于 PLA 的要点**

**保证** 只要数据线性可分且每次修正的都是误分，$\mathbf{w}_f^\top \mathbf{w}_t$ 会快速增长，而 $\|\mathbf{w}_t\|$ 增长缓慢。因此 PLA 的超平面越来越与 $\mathbf{w}_f$ 对齐，算法必然停机。

**优点** 实现简单、速度快、适用于任意维度 $d$。

**缺点**

- 要求数据集线性可分才能停机；
- 事先并不知道该性质（若已知 $\mathbf{w}_f$ 就无需 PLA）；
- 无法事先确定停机时间（上界依赖于未知的 $\rho$），尽管实践中很快。

**带噪声容忍的线性分类**

**假设场景** 仅存在少量标记噪声，即大多数情况下 $y_n = f(\mathbf{x}_n)$。此时期望所得假设 $g$ 在数据集 $D$ 上与真实函数 $f$ 足够接近，亦即

$$y_n \approx g(\mathbf{x}_n) = \operatorname{sign}(\mathbf{w}^\top \mathbf{x}_n).$$

**核心难题** 求解最小化 0/1 损失的问题

$$\mathbf{w}_g \leftarrow \arg\min_{\mathbf{w}} \sum_{n=1}^{N} [\![ y_n \neq \operatorname{sign}(\mathbf{w}^\top \mathbf{x}_n) ]\!]$$

该问题已被证明是 NP-hard 问题。因此，需要一种能在多项式时间内给出“近似最优”权重的替代策略。

**算法 2.4.2:** Pocket 算法

**输入:** 数据集 $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$，其中 $y_n \in \{+1, -1\}$；最大迭代次数 $T_{\max}$

**输出:** 口袋权重 $\mathbf{w}_{\text{Pocket}}$

初始化 $\mathbf{w}_0 \leftarrow \mathbf{0}$，$\hat{\mathbf{w}} \leftarrow \mathbf{0}$;

**for** $t = 0$ **to** $T_{\max} - 1$ **do**

  随机选取误分样本 $(\mathbf{x}_{n(t)}, y_{n(t)})$;

  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)}$;

  **if** $\mathbf{w}_{t+1}$ 的误分率低于 $\hat{\mathbf{w}}$ **then**

    $\hat{\mathbf{w}} \leftarrow \mathbf{w}_{t+1}$;

**return** $\mathbf{w}_{Pocket} \leftarrow \hat{\mathbf{w}}$;

笔记 [优势] 实现简单、运行高效，可在非严格线性可分场景下获得近似最优的分类超平面。

## 2.5 总结

笔记 [学习回答是或否]

- 感知机假设集：$\mathbb{R}^d$ 中的超平面（线性分类器）。
- 感知机学习算法（PLA）：通过不断纠正错误来迭代改进。
- PLA 的保证：若数据线性可分，则最终不再犯错。
- 非线性可分数据：把“相对最佳”的权重保留在口袋里（Pocket 算法）。