第2章 感知机模型 —— 适用条件 / 解决什么问题 / 三要素

针对二分类问题，并假设为线性可分的

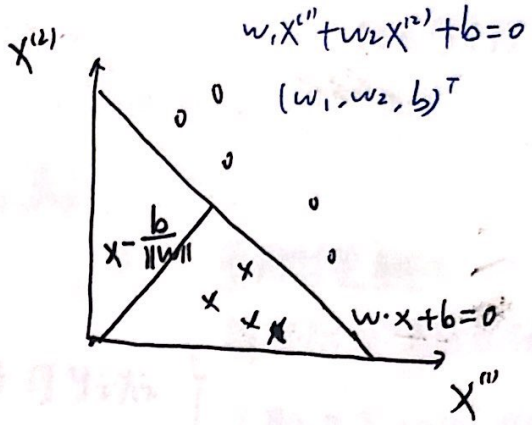
## 2.1 符号表示

输入空间 $\mathcal{X} \subseteq \mathbb{R}^n$

输入变量 $x \in \mathcal{X}$

输出空间 $\mathcal{Y} = \{+1, -1\}$

输出变量 $y \in \{+1, -1\}$




假设空间 → 符号函数

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

$$= \begin{cases} +1, & w \cdot x + b \geq 0 \\ -1, & w \cdot x + b < 0 \end{cases}$$

∵ $x$ 是个 $n$ 维向量

∴ $w$ 也是 $n$ 维向量

$w \cdot x$ 是内积的意思

$= w_1x^{(1)} + w_2x^{(2)} + \cdots + w_nx^{(n)}$

## 2.2 感知机学习策略

输入 $n$ 维则用 $n-1$ 维超平面去分

损失函数：误分类点到超平面的总距离

★ 求 $x_i$ 到一条直线的距离

$$\frac{|w \cdot x_i + b|}{\|w\|}$$ 二范式

↳ $\|w\| = \sqrt{w_1^2 + \cdots + w_n^2}$

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

↓解出 $w$ 与 $b$

推导来源：

$$\sum_{x_i \in M} \frac{|w \cdot x_i + b|}{\|w\|}$$

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

要使其最小，∵ $\|w\|$ 为正数

## 2.3 感知机学习算法（目的：求出 $w, b$，使其为以下损失函数极小化的问题

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

算法 2.1（随机梯度下降法）

输入：训练数据集 $T = [(x_1, y_1), \cdots (x_N, y_N)]$，学习率 $\eta$ $(0 < \eta \leq 1)$

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 $y_i (w \cdot x_i + b) \leq 0$

$$w := w_{前} + \eta y_i x_i$$
$$b := b + \eta y_i$$

要同时更新
每次都更靠近正确的
（取决于初始的点与更新顺序）

4. 转至 2，直到训练集中没有误分类的点

输出：$w, b$

感知机模型对偶形式

$$f(x) = \text{sign}\left(\underbrace{\sum_{j=1}^{N} \alpha_j y_j x_j}_{w} \cdot x + b\right)$$

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

由 $(\alpha_j, b)$ 决定的 $j$ 由 $1 \sim N$

算法 2.2

输入：训练数据集 $T$，学习率 $\eta$

1. 初值 $\alpha := 0$ $b := 0$

2. 在训练集中选取数据 $(x_i, y_i)$

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

$$\alpha_i := \alpha_i + \eta$$
$$b := b + \eta y_i$$

4. 转至 2，直至训练集中没有误分类点

★对偶算法与梯度下降算法比较.

① 对偶算法每次更新的时候只用更新 $\alpha_i$ 与 $b$ 两个数

而

梯度下降算法每次都要更新一整个 $w$ 向量与 $b$ 这个数

② 在判断是否分类错误时.

对偶算法算的是 $x_i \cdot x_j$ 的内积，可以预先算好 → 以 Gram 矩阵 → 计算量小

而

梯度下降算法需要计算一整个 $w$ 向量.

证：

定理 2.1 (Novikoff) 算法的收敛性

① 前提：记 $\hat{w}=(w^T,b)^T$，$\hat{w}\in R^{n+1}$ $\hat{w}\cdot\hat{x}=w\cdot x+b$

$\hat{x}=(x^T,1)^T$，$\hat{x}\in R^{n+1}$

② $\hat{w}_{opt}$ 意义解释

设训练数据集线性可分，则超平面表示为 $w\cdot x+b=0$

找到的这个超平面为 $w_{opt}\cdot x+b_{opt}=0$

则 $\hat{w}_{opt}=(w_{opt}^T, b_{opt})^T$，$\hat{x}=(x^T,1)^T$ $\Rightarrow \hat{w}_{opt}\cdot\hat{x}=w_{opt}\cdot x+b_{opt}=0$

$\because \hat{w}_{opt}\cdot\hat{x}=0$

$\therefore 2\hat{w}_{opt}\hat{x}=0$

$2\hat{w}_{opt}$ 抽象为 $\hat{w}$ 新的

为了让 $\hat{w}_{opt}$ 比较唯一

我们约束 $\|\hat{w}_{opt}\|=1$

---

念伽马

$\therefore$ ① 我们要证明 存在 $\|\hat{w}_{opt}\|=1$ 的超平面 $\hat{w}_{opt}\cdot\hat{x}=0$

使得训练集完全分开，且存在 $\gamma>0$，使得

$y_i(\hat{w}_{opt}\cdot\hat{x}_i)\geq\gamma$

证 ②①

$\because w_{opt}$ 的平面使所有数据分开

$\therefore$ 有 $y_i(\hat{w}_{opt}\cdot\hat{x}_i)>0$ (对所有实例来说)

$\therefore$ 让 $\gamma=\min_i y_i(\hat{w}_{opt}\cdot\hat{x}_i)$ 即 $\gamma\leq y_i(\hat{w}_{opt}\cdot\hat{x}_i)$

② 令 $R = \max_{1 \le i \le N} \|\hat{x}_i\|$，$k \le \left(\frac{R}{\gamma}\right)^2$

$\hat{x}_i$ 可以理解为这个向量长度的表示。

$k$：修正的次数

$\gamma$：上一个定理中的 $\gamma$

分2步

1) 要证 $\hat{w}_k \cdot \hat{w}_{opt} \ge k\eta\gamma$ → 念 qā mà

念 yī tà

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

由上个定理有 $y_i(\hat{w}_{opt} \cdot \hat{x}_i) \ge \gamma$

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

$\therefore \ge \hat{w}_{k-1} \cdot \hat{w}_{opt} + \eta\gamma$

$\ge \hat{w}_{k-2} \cdot \hat{w}_{opt} + \eta r + \eta r$

$\vdots$

$\ge \hat{w}_0 \cdot \hat{w}_{opt} + k\eta r$

∵假设 $\hat{w}_0 = (0, \cdots, 0)^T$

$\therefore = k\eta r$

a与b都是向量

$\|a+b\|^2 = (a+b)^T(a+b)$

$= \|a\|^2 + 2a \cdot b + \|b\|^2$

2) 要证 $\|\hat{w}_k\|^2$ 要小于某个值

$= \|\hat{w}_{k-1} + \eta y_i \hat{x}_i\|^2$

$= \|\hat{w}_{k-1}\|^2 + 2\eta \cdot y_i \hat{w}_{k-1} \cdot \hat{x}_i + \eta^2 \|\hat{x}_i\|^2$

★重要

对i来说这是个误分类点

$\therefore 2\eta y_i \hat{w}_{k-1} \cdot \hat{x}_i < 0$

∵假设 $\|\hat{x}_i\|^2 \le R^2$

$\therefore \le \|\hat{w}_{k-1}\|^2 + \eta^2 R^2$

$\vdots$

$\le \|\hat{w}_0\|^2 + k\eta^2 R^2$ ∵ $\hat{w}_0$ 为 $(0 \cdots 0)^T$

$\therefore = k\eta^2 R^2$

得证

∴由上面2个证明可得出

$$\begin{cases} \hat{w}_k \cdot \hat{w}_{opt} \geq k\eta\gamma & —① \\ \|\hat{w}_k\|^2 \leq k\eta^2 R^2 & —② \end{cases}$$

★来推 $k \leq (\frac{R}{\gamma})^2$

★根据柯西不等式

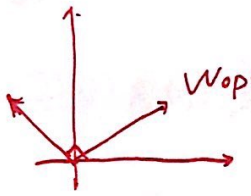
$$k\eta\gamma \leq \hat{w}_k \cdot \hat{w}_{opt} \leq \|\hat{w}_k\| \underbrace{\|\hat{w}_{opt}\|}_{=1} \leq \sqrt{k\eta^2 R^2}$$

即 $k\eta\gamma \leq \sqrt{k\eta^2 R^2} \Rightarrow k \leq (\frac{R}{\gamma})^2$

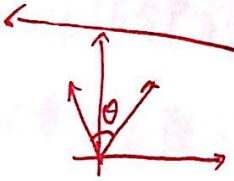
★理解①,②两式

①表示 $\hat{w}_k \cdot \hat{w}_{opt}$ 在逐渐增大 ∴k在↑

②表示 $\|\hat{w}_k\|$ 长度可以被限制住




①表示




中间角度越来越小

★即①②联合起来大概表示为 $w_k$ 在逐渐靠近 $w_{opt}$

(重要)