感知机（PLA）

主要参考资料：

> 《统计学习方法》第二章 感知机
> 《机器学习》第三章 线性模型
> 《模式识别与机器学习》第四章 分类的线性模型

---

1. 目的与假设

- 目的：解决二分类问题
- 假设：**训练集线性可分**

2. 模型（函数假设）

- 输入：实例的特征向量
- 输出：实例的类别，**取+1和-1**
- 假设空间：

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

其中：

$$sign(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

- 模型性质:
  - 判别模型
  - 特征空间中的N维超平面，如下图所示:

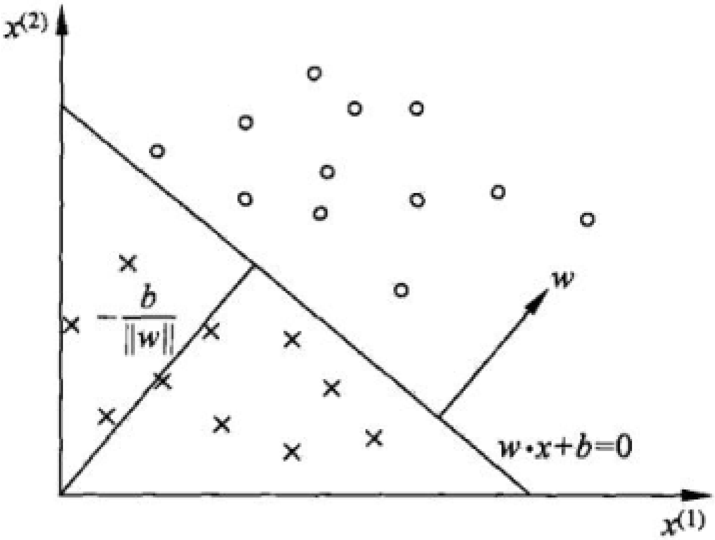


图 2.1 感知机模型

需要注意:
    - 哪一边的实例是正例、哪边是反例
    - 关于法向量的一点思考[1]
- 备注：输入向量的某些值可能是连续的，也可能是离散的（如 1, 2…）

3. 策略（损失函数）

a. 基于误分类点的总数

- 不是参数 w, b 的连续可导函数，不容易优化

b. 基于误分类点到超平面的总距离

- R^n 中的任意一点 x_0 到超平面 S 的距离（推导[2]）：

$$\frac{1}{\|w\|}|w \cdot x_0 + b|$$

- 误分类点 (x_i, y_i) 的性质:

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

- 误分类点 (x_i, y_i) 到超平面的距离:

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

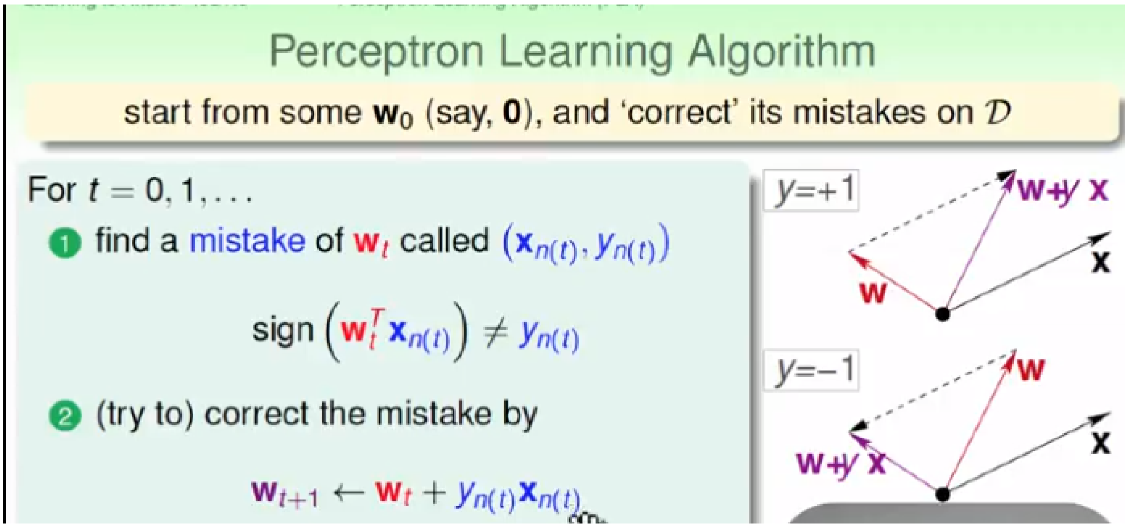
- 假设超平面 S 的误分类点集合为 M，那么所有误分类点到超平面 S的总距离为：

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

- 感知机的**损失函数**（经验风险函数）：

$$L(w, b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b) \geq 0$$

- **不考虑 $\frac{1}{||w||}$ 的原因：**
    - 求导之后可知，$||w||$ 只影响步长，并不影响方向，步长可以再用 η 调节（这种调节真的不会导致步长突然太大么？）
    - 也可以考虑为损失函数中的 w, b 是集合距离中 $\frac{w}{||w||}$ 和 $\frac{b}{||w||}$ 的替换，最后可以再乘回去，不会影响超平面（但是在实际过程中这两个向量并不是归一化的）
    - 可以从神经网络中感知机模型达到阈值的角度思考这个问题



    - 这样可以简便运算（这条承认）

- 备注：
    - 存在误分类点时损失函数是误分类点到平面的总距离，不存在误分类点时损失函数是0，因此损失函数对于 w, b 是连续可导的（**函数对于参数可导也是需要严谨地证明的**）
    - 其实也可以考虑正确分类的点到超平面的间隔尽可能大（SVM）

4. 算法（优化方法）

a. 目标：

- 损失函数最小化

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

b. 梯度下降（Gradient Descent）[3]

- 因为负梯度方向函数下降最快[4]
- 假设误分类点集合 M 是固定的，那么损失函数 $L(w,b)$ 的梯度由

  $$\nabla_w L(w,b) = \nabla_w\left(-\sum_{x_i \in M} y_i(w^T x_i + b)\right)$$

  给出，又因为

  $$\frac{\partial w^T x}{\partial w} = x$$

  所以

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

  同理

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

- 感知机学习算法中采用的是随机梯度下降 （Stochastic Gradient Descent）

b. 感知机学习算法的原始形式

算法 2.1（感知机学习算法的原始形式）

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

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

（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 $y_i(w\cdot x_i+b)\leqslant 0$

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

（4）转至（2），直至训练集中没有误分类点. ■

- 备注：
  - $\eta y_i x_i$的变化规律是什么？
  - 感知机学习算法由于采用不同的初值或者选取不同的误分类点，解可以不同
  - 感知机模型只求正确分类就好，这里求的是全局最优么还是局部最优？损失函数是凸函数么？
  - 感知机学习算法原始形式算法收敛性证明[5]

c. 感知机学习算法的对偶形式[6]

- 思路：在感知机算法的原始形式中，假设 $w_0 = 0, b_0 = 0$, 因为

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

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

  所以**对这个误分类点**修改 n_i 次后，w, b关于 (w_i, y_i) 的增量分别是 $\alpha_i y_i x_i$ 和 $\alpha_i y_i$, 其中 $\alpha_i = n_i \eta$,因此有

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

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

  其中 N 是样本点数，**当 $\eta = 1$ 时 α 表示的是第 i 个实例点由于误分而进行的更新次数，更新次数越多，说明它距离超平面越近，就越难以正确分类，对学习结果影响越大。**
- 对偶形式

  算法 2.2（感知机学习算法的对偶形式）

  输入：线性可分的数据集 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，其中 $x_i \in \mathbf{R}^n$，$y_i \in \{-1, +1\}$，$i = 1, 2, \cdots, N$；学习率 η（$0 < \eta \leqslant 1$）；

  输出：α, b；感知机模型 $f(x) = \mathrm{sign}\left(\sum_{j=1}^{N} \alpha_j y_j x_j \cdot x + b\right)$.

  其中 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^{\mathrm{T}}$.

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

  （2）在训练集中选取数据 (x_i, y_i)

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

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

  （4）转至（2）直到没有误分类数据.

  其中内积可以提前计算并存储成Gram矩阵形式[7]

5. 代码实例

6. 相关参考

- “机器学习技法”，林轩田

---

1. 关于法向量的一点思考：

$$w \cdot x + b = \|w\| \cdot \|x\| \, cos\theta + b = \|w\| \, (\|x\| \cos\theta) + b$$

$\|x\| \, cos\theta$ 其实是原点到平面的距离 ↩

2. 距离公式的推导： ↩
3. 几种常见梯度下降方法的比较： ↩
4. 负梯度方向函数下降最快的证明：泰勒展开式 ↩
5. 感知机学习算法原始形式算法收敛性证明 ↩
6. 为什么叫做“对偶形式”： ↩
7. Gram矩阵 ↩