

## 感知器模型

目标：针对线性可分的二分类问题，将数据正确分类

数据集  $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$   
 样本个数  $N$   
 一个样本  $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$  为第  $i$  个输入样本  
 特征个数  $d$   
 一个标签  $y_i \in \{-1, 1\}$

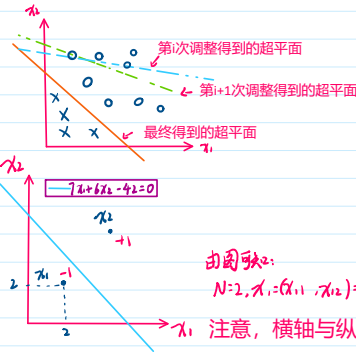
模型公式  $y = f(w^T x + b)$   
 $w$  权重，个数与输入变量  $x$  维度相同  
 $f(x)$ ：如符号函数  $\text{sign}(x) = \begin{cases} +1, x > 0 \\ -1, x < 0 \end{cases}$  以上图为例： $y_i = f(w^T x_i + b) = \text{sign}(7x_{i1} + 6x_{i2} - 4) = \text{sign}(14 + 12 - 4) = \text{sign}(16) = +1$

符号函数  
 当  $w^T x + b > 0$  时，输出预测值  $y = +1$   
 当  $w^T x + b < 0$  时，输出预测值  $y = -1$

与真实值的关系  
 正确分类时，标签与预测值符号相同  
 错误分类时，标签与预测值符号不同

损失函数  
 对于所有分类错误的样本  $M$ ：将  $y(w^T x + b)$  作为损失函数  
 由于分类错误时  $y(w^T x + b) < 0$ ，我们需要这个式子越接近 0 越好，即最大化  
 然而损失函数一般都是求最小值，所以需要变号，因此将  $-y(w^T x + b)$  作为损失函数  
 数，最终表达式为  $L(w, b) = -\sum_{x \in M} y(w^T x + b)$ ，其中  $M$  为分类错误的样本集合

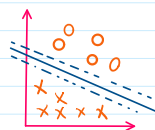
存在问题  
 超平面不唯一：只要能样本分开的，都是符合要求的超平面  
 噪声敏感：某些超平面的位置太靠边，如果这些边界稍微加点噪声或多添加一些样本，很可能会错误分类  
 超平面的表示方法不唯一：从数学的角度观察， $7x_1 + 6x_2 - 4 = 0 \Rightarrow 14x_1 + 12x_2 - 8 = 0$ ，这样就导致超平面的  $w, b$  表示方法不唯一  
 无法解决非线性问题



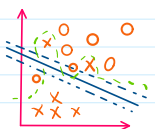
## 支持向量机的分类

支持向量机就是用来解决上述各种问题的算法

线性可分支持向量机（硬间隔最大化，选择最优平面）  
 样本是线性可分的



线性支持向量机（加入松弛因子，战略放弃一部分样本）  
 样本近似性可分



非线性支持向量机（数据升维，核函数）  
 样本非线性可分

