机器学习与数据科学

Machine Learning and Data Science

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

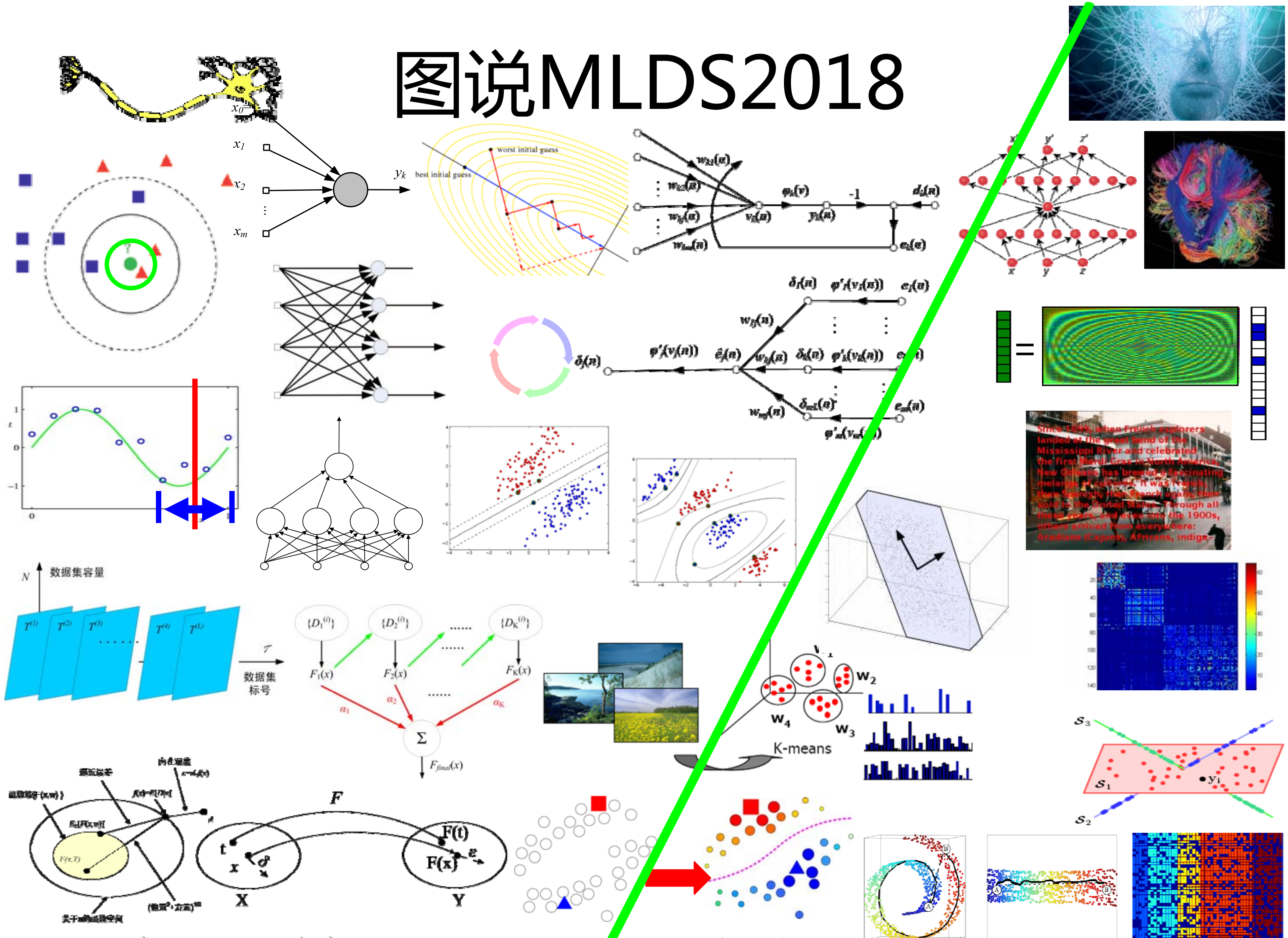
信息与通信工程学院 网络搜索教研中心

北京邮电大学

专题二：线性模型

- **内容提要**
  - 引言
  - 线性分类
  - 线性回归
  - 逻辑斯蒂回归

图说MLDS2018

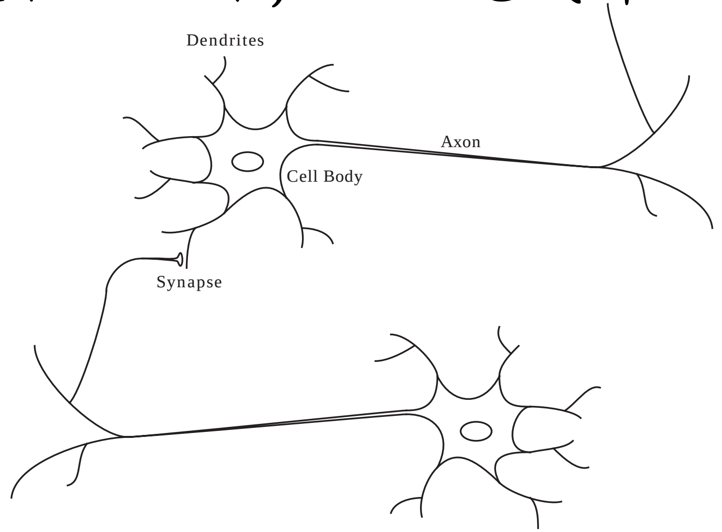


专题二：线性模型

- **内容提要**
  - 引言
  - 线性分类 → 感知器
  - 线性回归 → **Adaline**
  - 逻辑斯蒂回归

引言: 生物神经元的结构

- **神经元构成**
  - 树突：接受从其他神经元传入信息的神经元纤维
  - 胞体：接受外来的信息，对各种信息进行汇总，并进行阈值处理，产生神经冲动
  - 轴突：连接其他神经元的树突和细胞体，以及完成神经元之间的信息传递



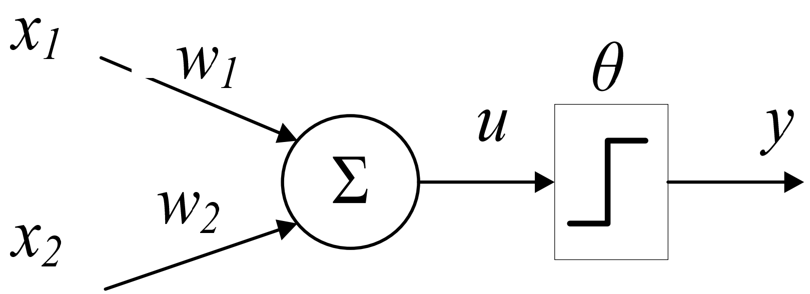
[美]**Dennis Coon, John O. Mitterer** 著，郑钢译，心理学导论——思想与行为的认识之路（**Gateways to Mind and Behavior**）**(11th Edition)**，中国轻工业出版社，**2007.06**

- 突触的特点：单方向的传递信息，且强度可变、具有学习功能

Artificial neuron...

McCulloch-Pitts 神经元模型

- 对输入信号加权求和，再与阈值比较以确定输出

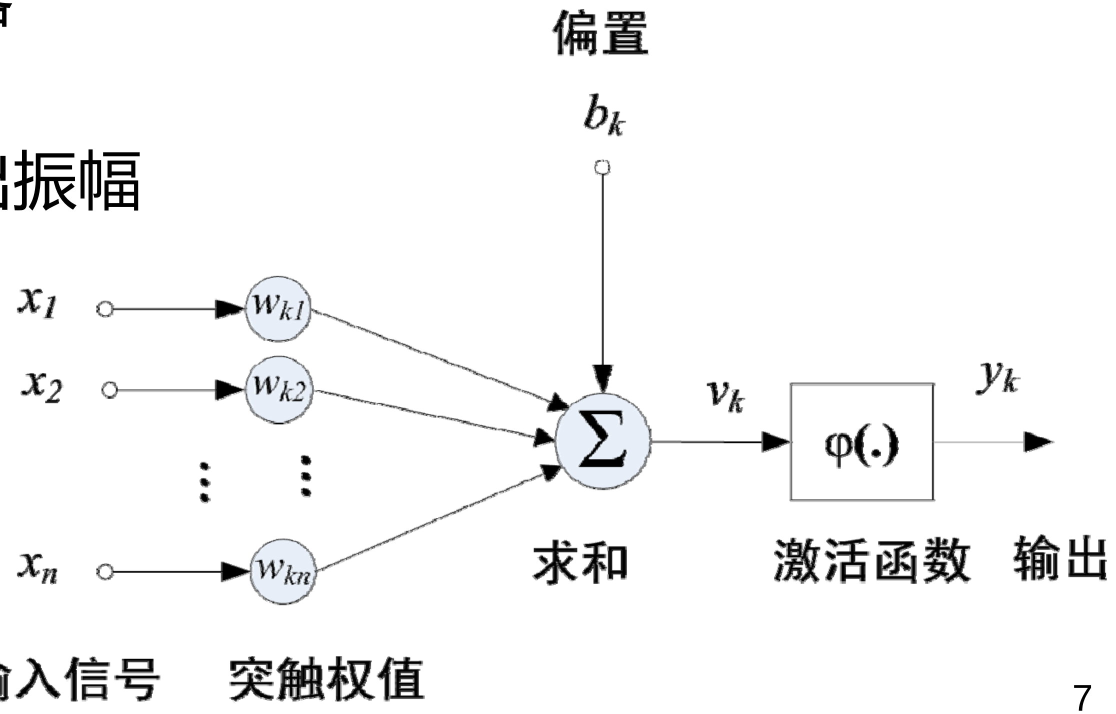


  – [1] McCulloch W.S. and Pitts W., “A logical calculus of the ideas immanent in nervous activity”. Bulletin of Mathematical Biophysics, vol.5, 1943, pp.115-133.

- 标志神经网络和人工智能学科的诞生

神经元模型

- 神经元是神经网络的基本信息处理单位
  - 突触权值：
    - 对输入信号加权
  - 加法器
    - 构成线性组合器
  - 激活函数
    - 限制神经元输出振幅



神经元模型的数学表达

- **数学表达**

偏置(bias)

第 *i* 个输入

$$y = \varphi\left(b + \sum_i x_i w_i \right)$$

第*i*个输入上的权值

输出

所有输入连接的指标

- **神经元的排列和突触的强度确立了神经网络的结构**
  – 突触单方向的传递信息，且强度可变、具有学习功能

神经元模型的数学表达

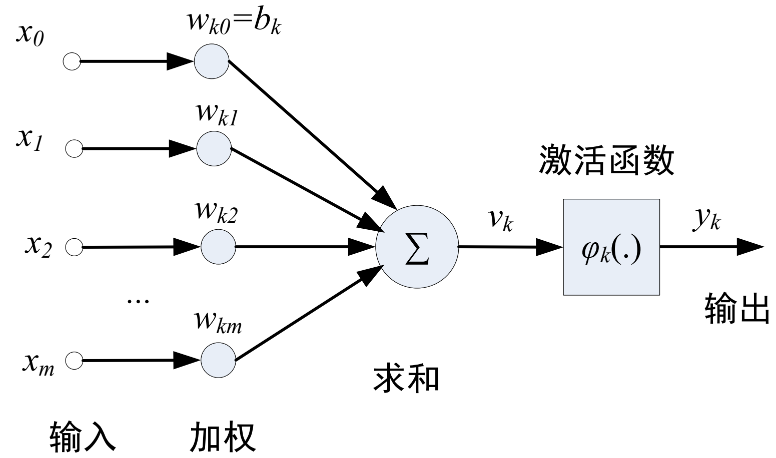
- 数学表达

$$y_k = \varphi\left(\sum_{i=0,1,\ldots,n} x_i w_{ki}\right) = \varphi\left(\mathbf{w}^T\mathbf{x}\right)$$

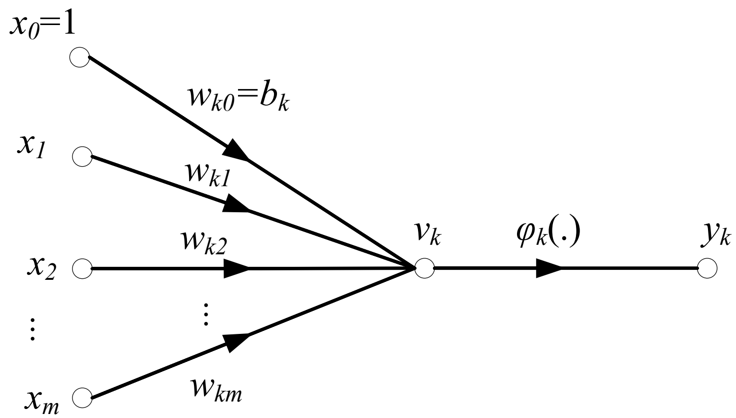
所有输入连接的指标, 从0开始

神经元的表示

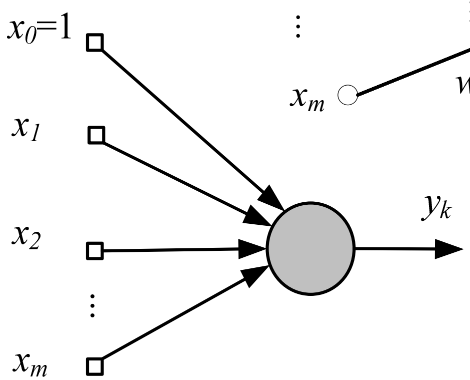
- 框图
  - 提供功能描述



- 信号流图
  - 提供信号流的完备描述
    - 信号沿箭头流动
    - 汇聚，即线性求和



- 体系结构图
  - 描述网络的布局
    - 输入节点
    - 计算节点



- **内容提要**
  - 引言
  - 线性分类 → 感知器
  - 线性回归 → **Adaline**
  - 逻辑斯蒂回归

Frank Rosenblatt

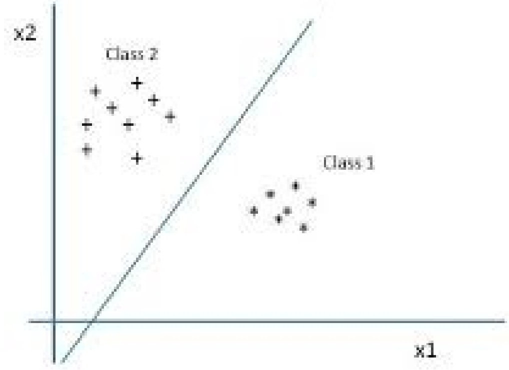
线性可分

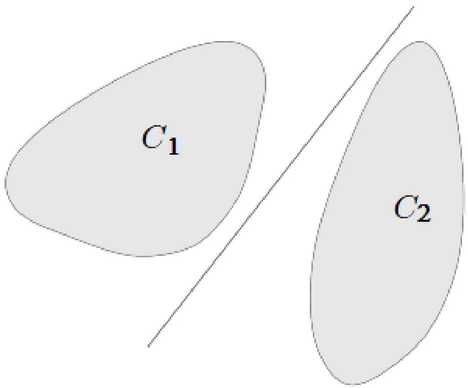
- 两个类别C_1 和 C_2是线性可分的
  - 如果存在一个权值向量 *w* 满足
    - 对于来自类别 C_1 的输入向量 $\mathbf{x}$ ：

$$\mathbf{w}^T\mathbf{x} > 0$$

    - 对于来自类别 C_2 的输入向量 $\mathbf{x}$ ：

$$\mathbf{w}^T\mathbf{x} \le 0$$

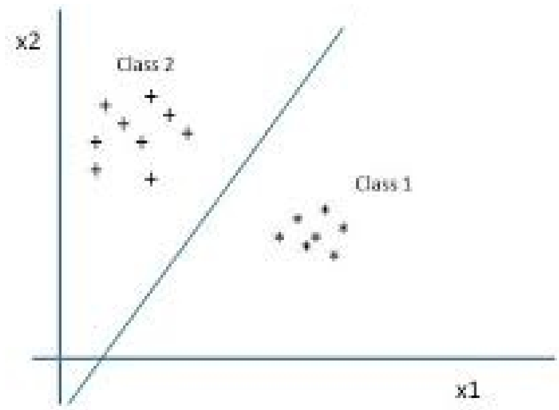




Farkas 定理: A w <0, c^T w >0有解的充要条件是：A^T y = c, y>=0 无解。

线性分类

- 寻找一个线性判别函数 $g(x) = w^T x + b$，对于来自类别C_1 和 C_2的样本 x
  - 如果 **$g(x) > 0$,** 我们把x 判定为类别**1**
  - 如果 **$g(x) < 0$,** 我们把x判定为类别**2**

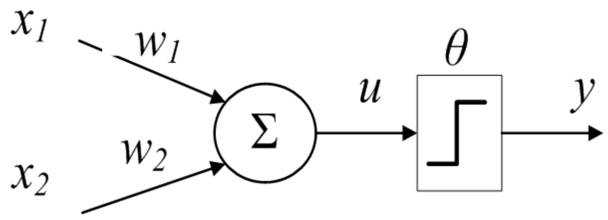


- **内容提要**
  - 引言
  - 线性分类 → 感知器
  - 线性回归 → **Adaline**
  - 逻辑斯蒂回归

Frank Rosenblatt

感知器(Perceptron)的简史

- 1943年 McCulloch和Pitts 提出了M-P神经元模型

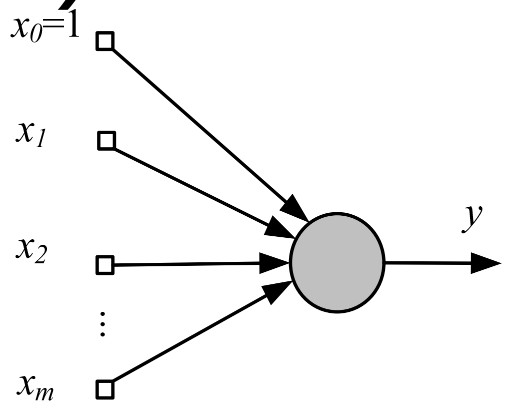


- 1958年Frank Rosenblatt 第1次给出了面向计算的神经网络
  - @ Cornell Aeronautical Laboratory (1957-1959)

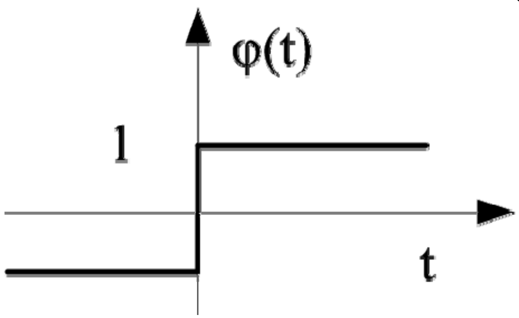
– 主要贡献:
  - 在算法上，提出用于解决模式分类问题的神经网络训练规则(或学习规则)
  - 在理论上，证明: 只要求解问题的权值存在，通常会收敛到正确的权值上

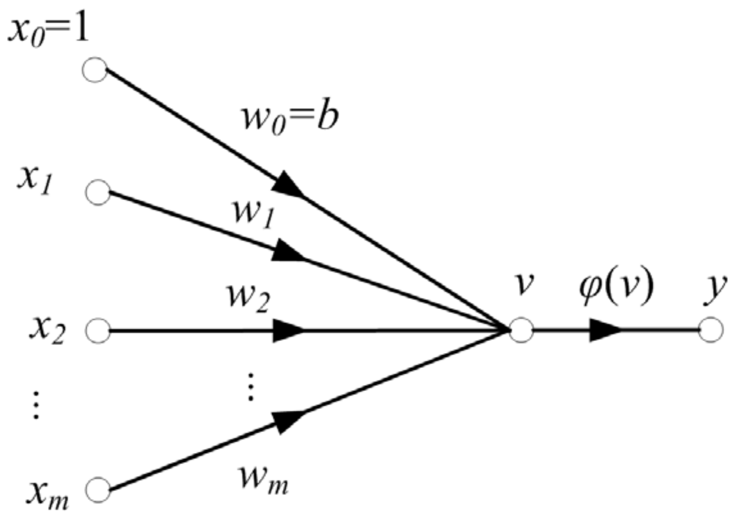
[1] Frank Rosenblatt. The perceptron: probabilistic model for information storage and organization in the brain [J]. Psychological Review,1958,65(6):386-408.
[2] Albert B.J. Novikoff. On convergence proofs on perceptrons. In Proc. of Symposium on the Mathematical Theory of Automata,12, 615–622, 1962.

感知器 (Perceptron)



- 即1个MP神经元
  - 非线性神经元
  - 数学表达 $y=\varphi\left(\mathbf{w}^T\mathbf{x}\right)=\varphi\left(\sum_{i=0}^{m} w_i x_i\right)$
    - 其中 $\varphi(t)=\begin{cases}1 & t>0\\ -1 & t\le 0\end{cases}$

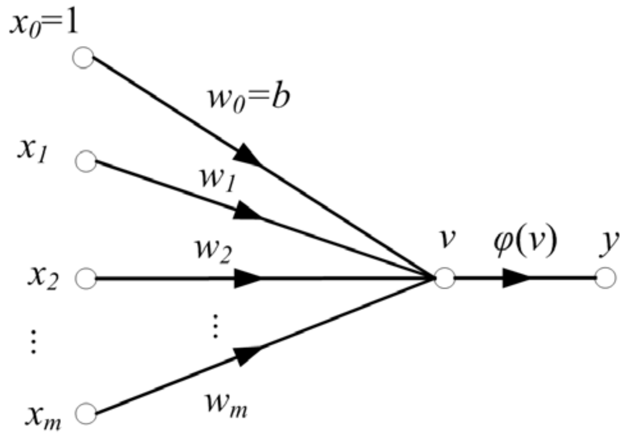




感知器训练算法

- 假设输入数据是线性可分的
  - 感知器模型

$$y=\varphi\left(\mathbf{w}^T\mathbf{x}\right)=\operatorname{sgn}\left(\mathbf{w}^T\mathbf{x}\right)$$



其中

$$\operatorname{sgn}\left(\mathbf{w}^T\mathbf{x}\right)=\begin{cases}+1, & \mathbf{w}^T\mathbf{x}>0\\ -1, & \mathbf{w}^T\mathbf{x}\le 0\end{cases}$$

- 感知器的训练
  - 不断调整权值，直到分类正确
  - 感知器准则函数 $E(\mathbf{w})=-\sum_{i\in I} y_i\left(\mathbf{w}^T\mathbf{x}_i\right)$
  - 其中 $I=\left\{i: y_i\neq\operatorname{sgn}\left(\mathbf{w}^T\mathbf{x}_i\right)\right\}$

即发生分类错误的样本下标集合

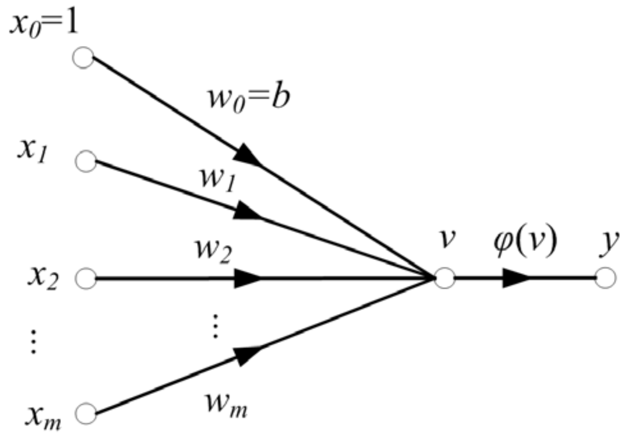
感知器训练算法

- 假设输入数据是线性可分的
  - 感知器模型

$$y=\varphi\left(\mathbf{w}^T\mathbf{x}\right)=\mathrm{sgn}\left(\mathbf{w}^T\mathbf{x}\right)$$

其中

$$\mathrm{sgn}\left(\mathbf{w}^T\mathbf{x}\right)=\begin{cases}+1, & \mathbf{w}^T\mathbf{x}>0\\ -1, & \mathbf{w}^T\mathbf{x}\le 0\end{cases}$$



- 感知器的训练
  - 不断调整权值，直到分类正确
  - 权值更新规则 $\mathbf{w}\leftarrow\mathbf{w}+\eta y_i\cdot\mathbf{x}_i \quad \text{if} \quad y_i\mathbf{w}^T\mathbf{x}_i<0$
  - 其中 $\eta>0.$

$y_i\mathbf{w}^T\mathbf{x}_i<0$ 等价于发生分类错误即 $\hat{y}_i\neq y_i$,$\hat{y}_i=\mathrm{sgn}\left(\mathbf{w}^T\mathbf{x}_i\right)$

感知器训练算法

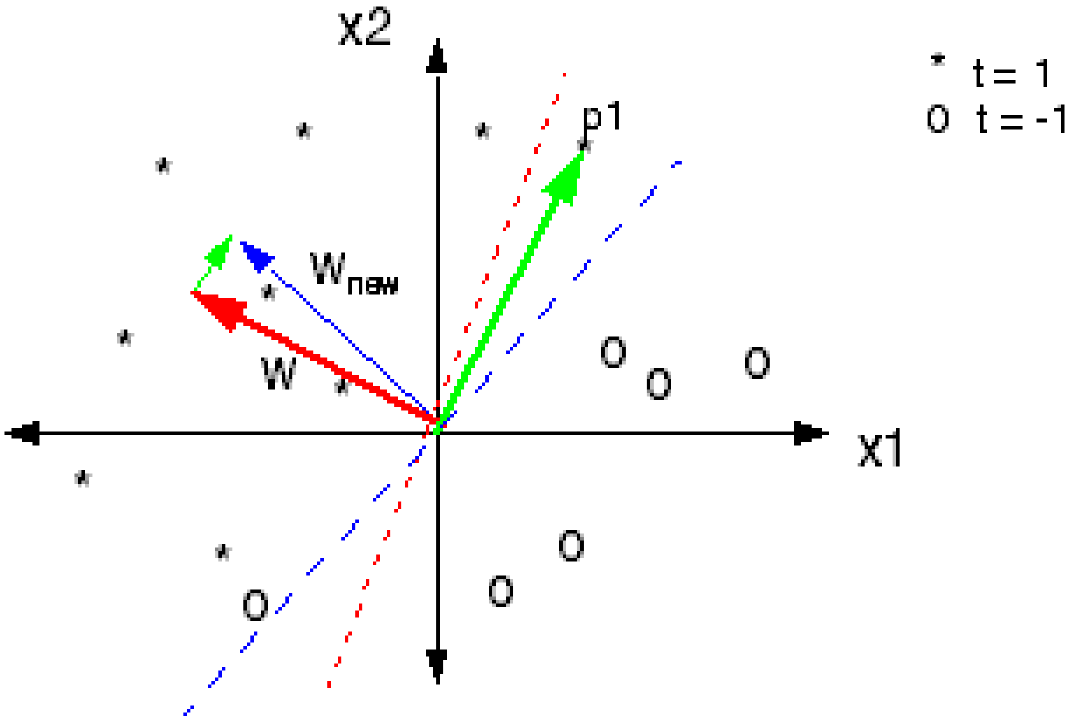
Algorithm 1 Perceptron Algorithm

Input: train data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, learning rate $\eta = 1$
Output: weights $\mathbf{w}$
Initialize weights $\mathbf{w}$ randomly
Repeat
  For $i = 1$ to N
    If $y_i \mathbf{w}^\top \mathbf{x}_i \leq 0$ then
      $\mathbf{w} \leftarrow \mathbf{w} + \eta \cdot y_i \mathbf{x}_i$
    End-If
  End-For
Until no sample is misclassified

感知器训练过程示意图

- 类别1的样本被错分到类别 2：

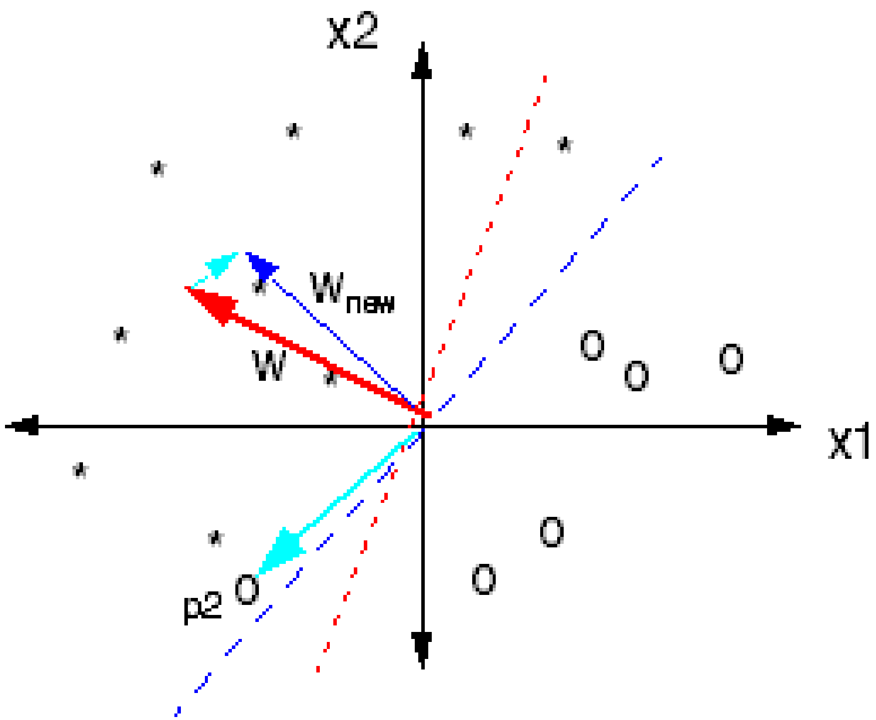
$$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \cdot \mathbf{x}_i \quad \Rightarrow \quad \mathbf{w} \leftarrow \mathbf{w} + \eta \cdot \mathbf{x}_i$$



感知器训练过程示意图

- 类别2的样本被错分到类别 1：

$$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \cdot \mathbf{x}_i \quad \Rightarrow \quad \mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \mathbf{x}_i$$



关于感知器算法的几个问题

- 感知器算法会收敛吗？
  - 若收敛，何时收敛？
  - 感知器的解唯一吗？

- 若收敛，那么所得到的解是**最优解**吗？
  - If “No” ：如何判断最优？
  - If “Yes”：如何定义最优？
  - If No & Yes, 如何找到最优解？

感知器算法收敛定理

- **感知器收敛定理：**设训练样本序列为

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_T, y_T),$$

所有样本满足$\|\mathbf{x}_t\| \le r$，其中$r > 0$；假设训练样本是线

性可分的，即存在$\rho > 0$和向量$\mathbf{v} \ne 0$，对于所有样本，

总有$\dfrac{y_t \mathbf{v}^T \mathbf{x}_t}{\|\mathbf{v}\|} \ge \rho$成立。那么，感知器算法中的权值更新

次数上界为r^2 / ρ^2。

[1] Albert B.J. Novikoff. On convergence proofs on perceptrons. In Proc. of Symposium on the Mathematical Theory of Automata,12, 615–622, 1962.

感知器算法收敛定理的证明

- **证明:** 设 I 为感知器算法在 T 次迭代过程中发生权值更新的下标集合，M 是权值更新的总次数。那么，我们有下列不等式：

$$M\rho \le \frac{\mathbf{v}^T}{\|\mathbf{v}\|}\sum_{t\in I} y_t\mathbf{x}_t \le \boxed{\left\|\sum_{t\in I} y_t\mathbf{x}_t\right\| \le \sqrt{\sum_{t\in I}\|\mathbf{x}_t\|^2}} \le \sqrt{Mr^2},$$

因此，我们有：$\sqrt{M} \le \frac{r}{\rho}$，即 $M \le \frac{r^2}{\rho^2}$。

感知器训练算法得到的解

- 感知器算法得到的解向量w并不唯一
  - 感知器训练算法收敛在任意一个在训练数据能够获得零错误率的向量w
    - **感知器算法得到的"最优解"是一个可行解**
- 最优权值向量
  - 最优决策超平面
  - 与两个类别的数据分布均保持最远距离