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

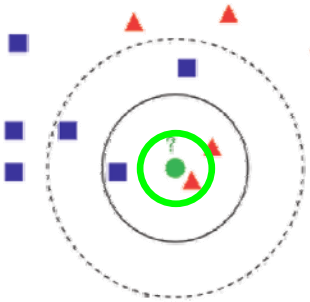
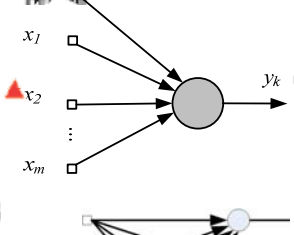
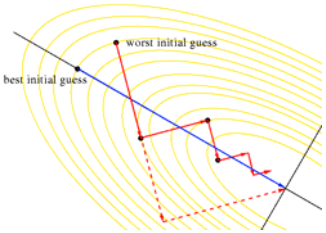
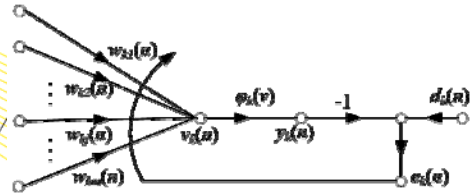
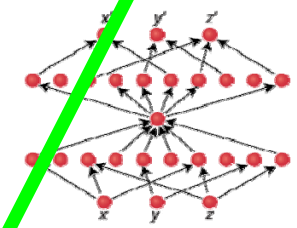
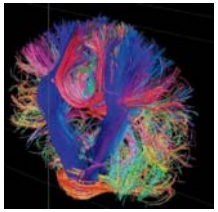
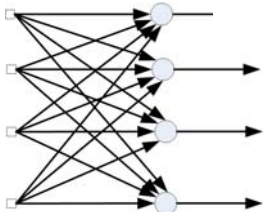
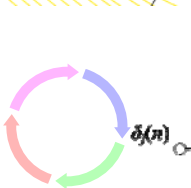
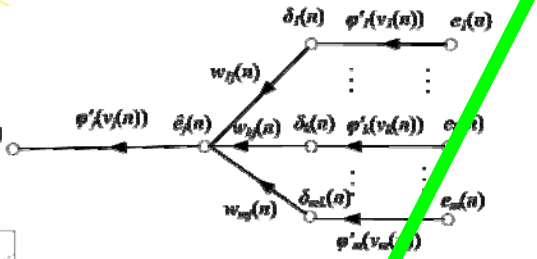
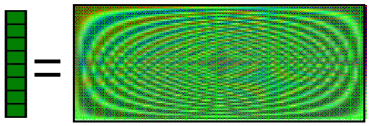
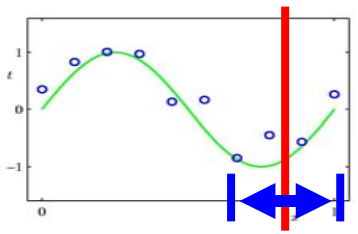
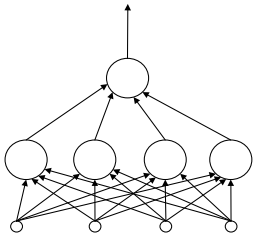
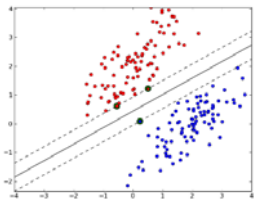
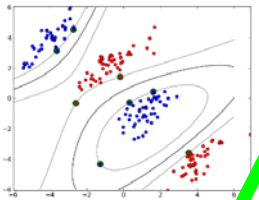
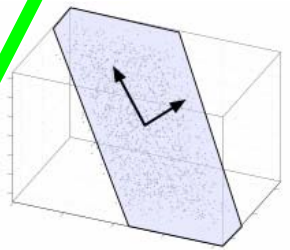
北京邮电大学

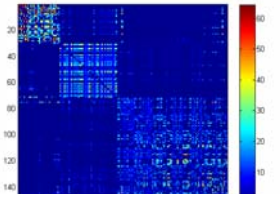
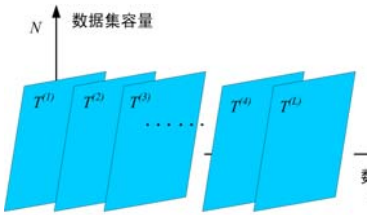
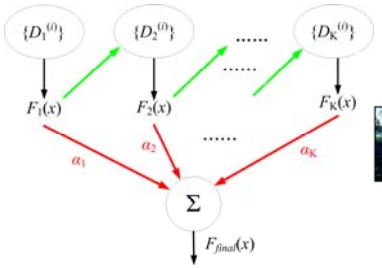
专题 二：线性模型

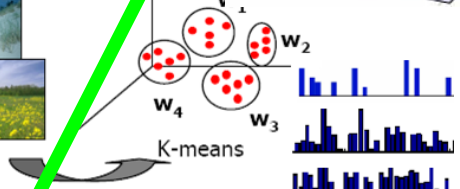
- **内容提要**
  - 引言
  - 线性分类
  - 线性回归
  - 逻辑斯蒂回归

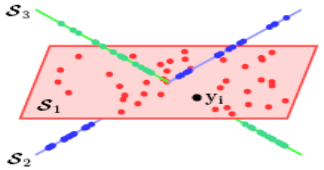
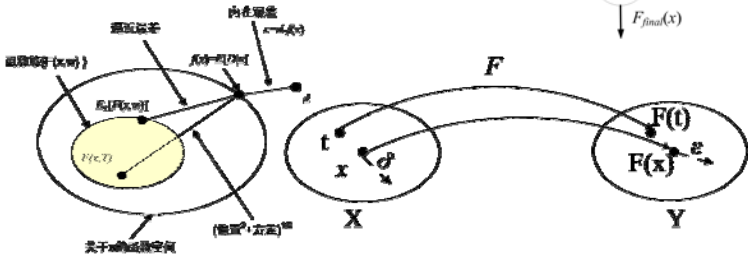
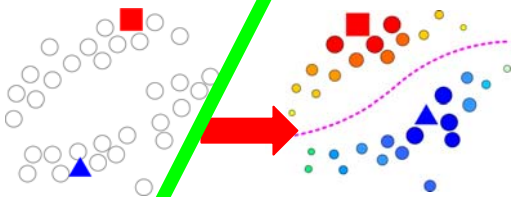
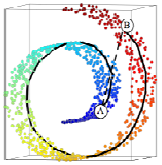
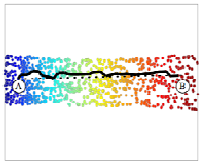
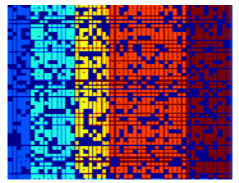
图说MLDS2018





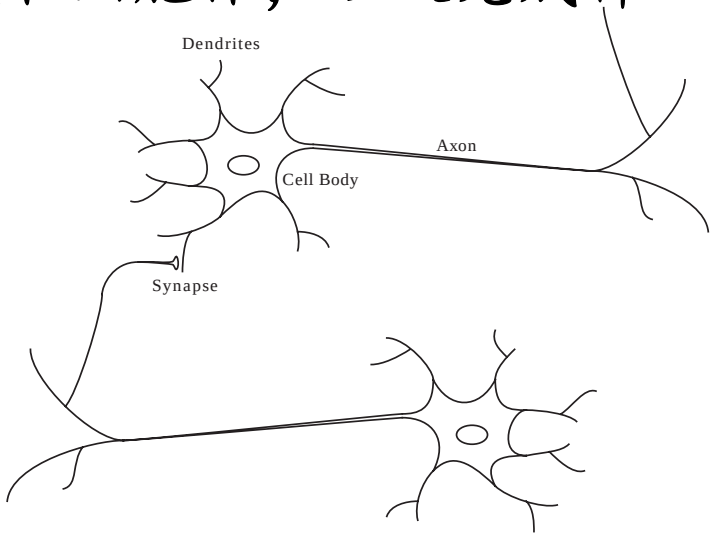







专题 二：线性模型

- **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



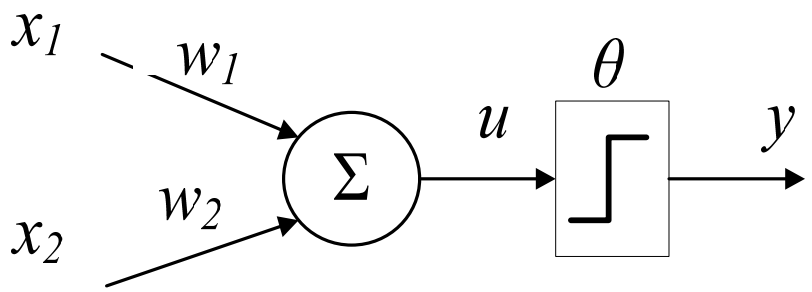
[美]Dennis Coon, John O. Mitterer 著，郑钢译，心理学导论——
思想与行为的认识之路（Gateways to Mind and Behavior）
(11th Edition)，中国轻工业出版社，2007.06

- 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

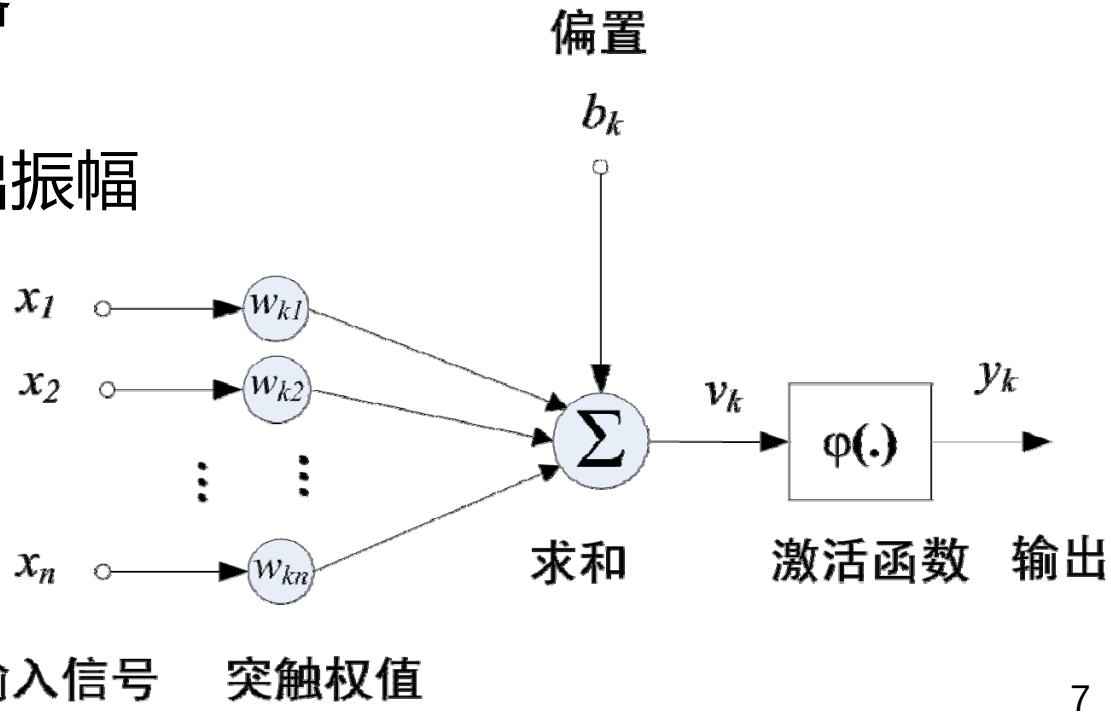


  - **[1] McCulloch W.S. and Pitts W., “A logical calculus of the ideas immanent in nervous activity”. Bulletin of Mathematical Biophysics, vol.5, 1943, pp.115-133.**

- 标志神经网络和人工智能学科的诞生

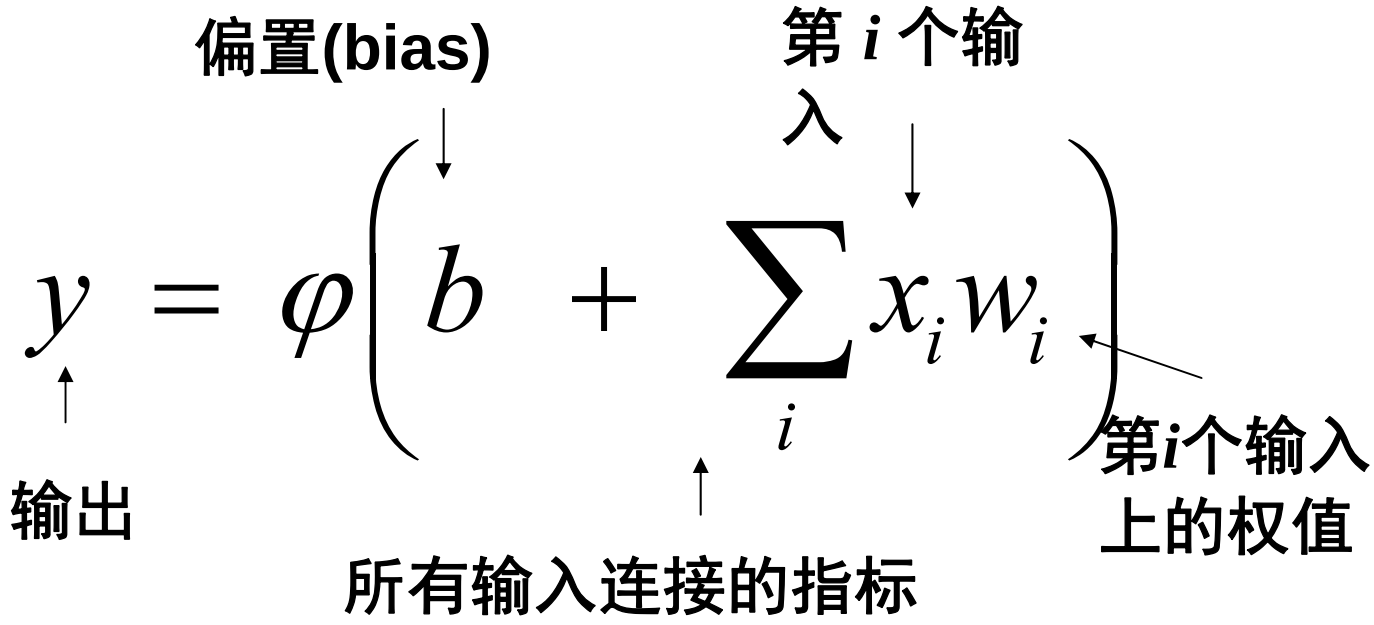
神经元模型

- 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**



- **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

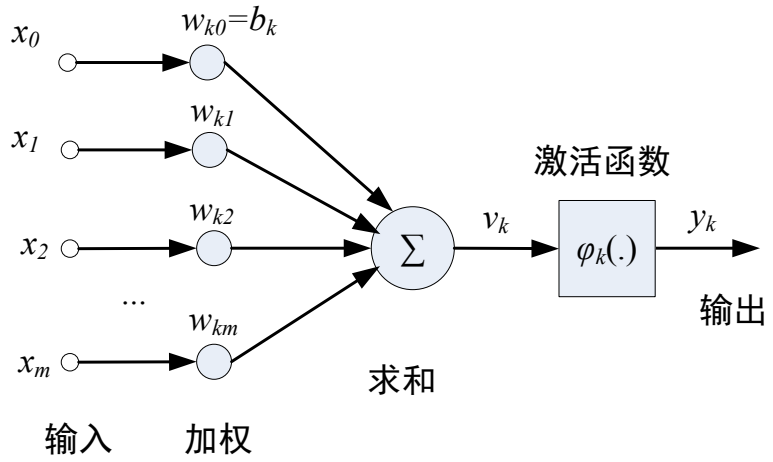
$$y_k = \varphi\left(\sum_{i=0,1,\ldots,n} x_i w_{ki}\right) = \varphi\left(\mathbf{w}^T\mathbf{x}\right)$$

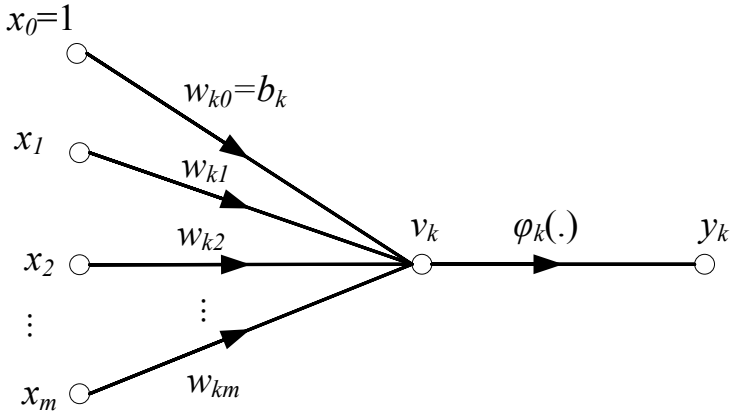
↑

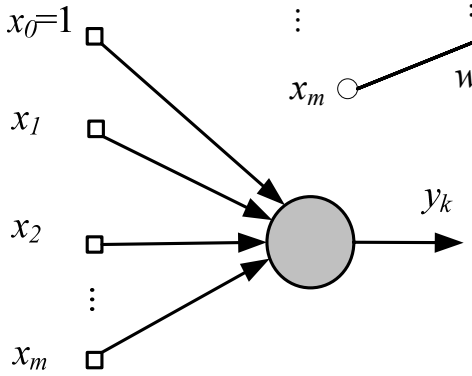
所有输入连接的指标, 从0开始

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

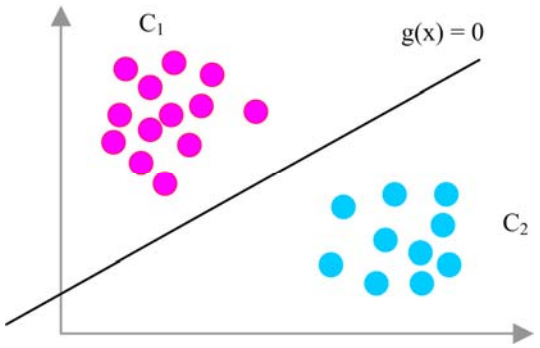
线性可分

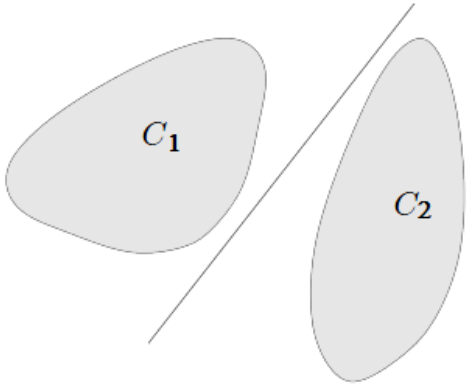
- 两个类别C_1 和 C_2是线性可分的
  - 如果存在一个权值向量 $\mathbf{w}$ 满足
    - 对于来自类别 C_1 的输入向量 $\mathbf{x}$ ：

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

    - 对于来自类别 C_2 的输入向量 $\mathbf{x}$ ：

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

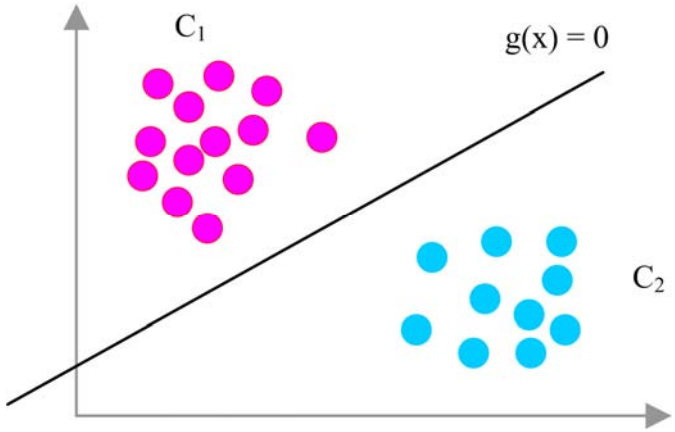




Farkas 定理: A w <0, c^T w >0有解的充要条件是：
A^T y = c, y>=0 无解。

线性分类

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

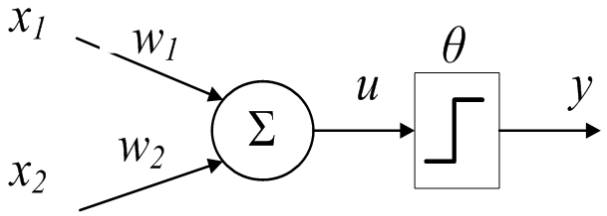


- **内容提要**
  - 引言
  - 线性分类 → 感知器
  - 线性回归 → **Adaline**
  - 逻辑斯蒂回归

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)

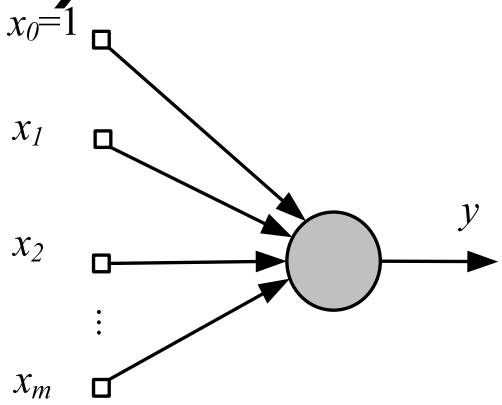
  - 主要贡献:
    - 在算法上，提出用于解决模式分类问题的神经网络训练规则(或学习规则)
    - 在理论上，证明: 只要求解问题的权值存在，通常会收敛到正确的权值上

[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.

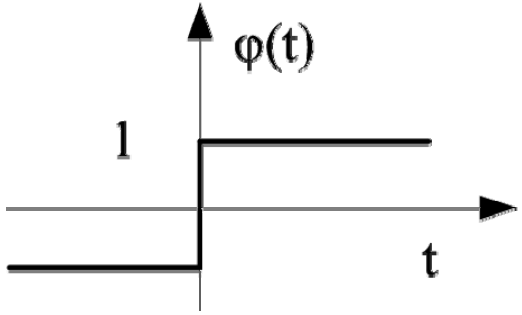
感知器 (Perceptron)

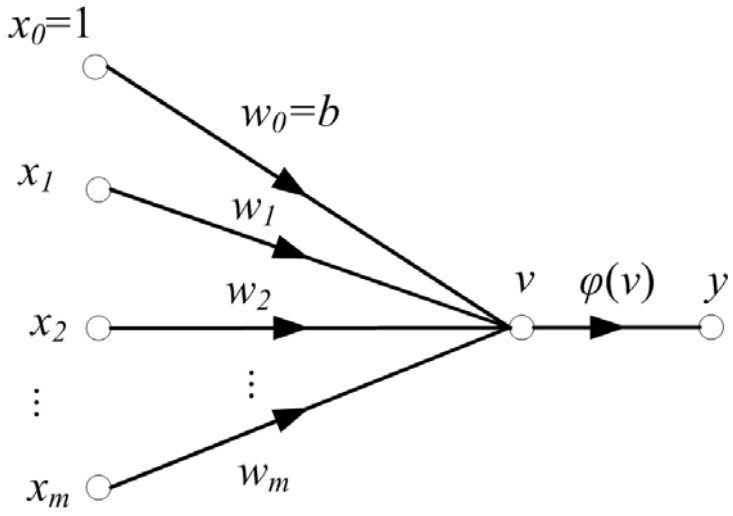
- 即1个MP神经元
  - 非线性神经元



- 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 $\varphi(t) = \begin{cases} 1 & t > 0 \\ -1 & t \le 0 \end{cases}$

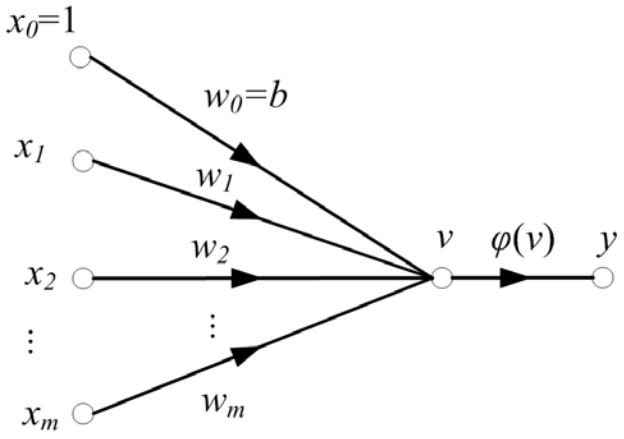




感知器训练算法

- 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 $E\left(\mathbf{w}\right)=-\sum_{i\in I} y_i\left(\mathbf{w}^T\mathbf{x}_i\right)$

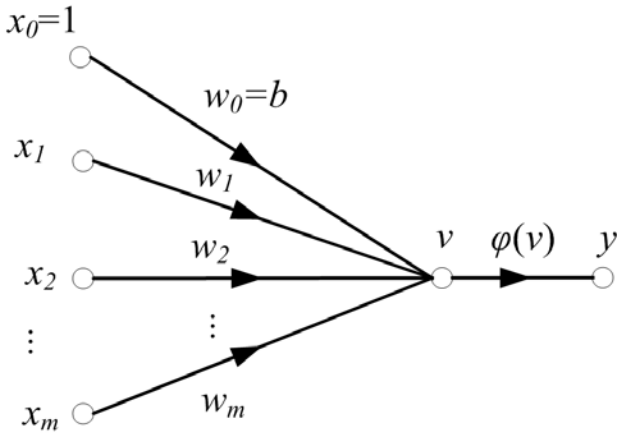
- 其中 $I=\left\{i: y_i\neq \operatorname{sgn}\left(\mathbf{w}^T\mathbf{x}_i\right)\right\}$

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　　$\mathbf{w}\leftarrow\mathbf{w}+\eta y_i\cdot\mathbf{x}_i \quad \text{if} \quad y_i\mathbf{w}^T\mathbf{x}_i<0$

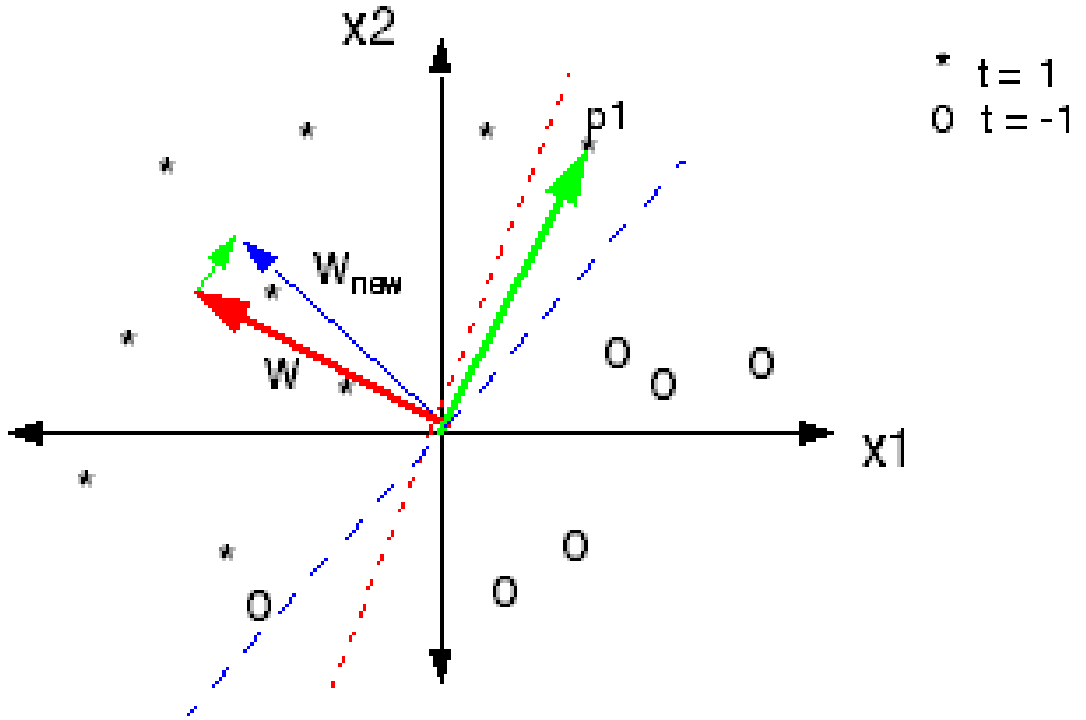
    - 其中 $\eta>0.$

$y_i\mathbf{w}^T\mathbf{x}_i<0$ 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：

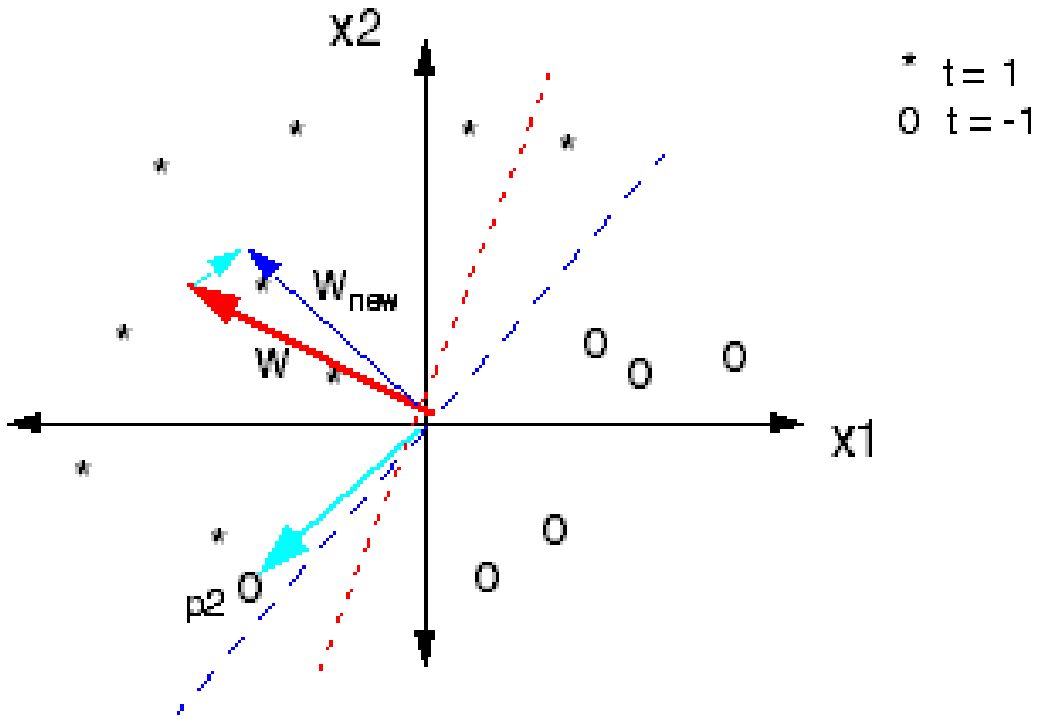
$$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \cdot \mathbf{x}_i \quad \Rightarrow \quad \mathbf{w} \leftarrow \mathbf{w} + \eta \cdot \mathbf{x}_i$$



感知器训练过程示意图

- 类别2的样本被错分到类别 1：

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



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

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 $\frac{y_t \mathbf{v}^T \mathbf{x}_t}{\|\mathbf{v}\|} \ge \rho$ 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

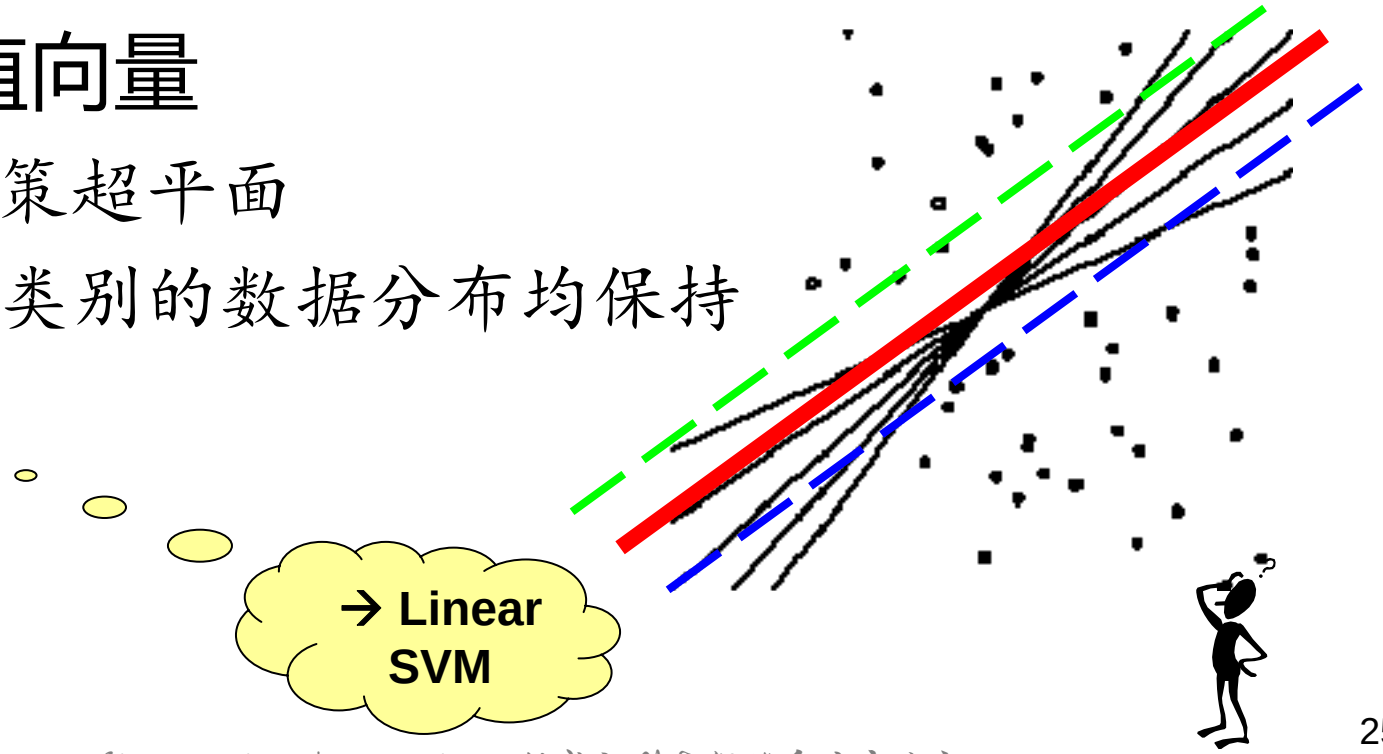
- 证明：设 I 为感知器算法在 T 次迭代过程中发生权值更新的下标集合，M 是权值更新的总次数。那么，我们有下列不等式：

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

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
  - 感知器训练算法可收敛在任意一个在训练数据上获得零错误率的向量**w**
    - **感知器算法得到的"最优解"只是一个可行解**
- 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



- **内容提要**
  - 引言
  - 线性分类 → 感知器 $F(\mathbf{x})=\varphi(\mathbf{w}^T\mathbf{x})=\mathrm{sgn}(\mathbf{w}^T\mathbf{x})$
  - 线性回归 → **Adaline** $F(\mathbf{x})=\varphi(\mathbf{w}^T\mathbf{x})=\mathbf{w}^T\mathbf{x}$
  - 逻辑斯蒂(Logistic)回归

线性回归问题

- **回归模型**
  - 设X是随机输入向量, Y是实数值随机标量, 联合分布密度p(x,y), 寻找一个函数 $f(x)$, 给定输入X的值预测Y

$$y = f(\mathbf{x}) + \varepsilon$$

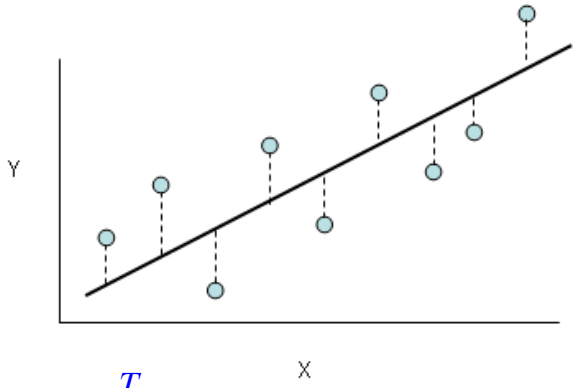
  - 联合分布密度p(x,y)未知，给定训练数据集D

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$$

寻找回归函数 f，使得

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, i = 1, ...N$$

- 线性回归(Linear Regression)



  - 假设函数**f(x)**为 线性函数
    - 即输入的线性组合

$$f(\mathbf{x}) = w_0 x_0 + w_1 x_1 + ... + w_m x_m = \mathbf{w}^T \mathbf{x}$$

最小二乘法

- 最小二乘法 (Least Square)
  - 给定训练数据集 $D=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_N,y_N)\}$
    寻找回归函数f，使得 $y_i=f(\mathbf{x}_i)+\varepsilon_i,\quad i=1,...,N$
  - 选择目标函数为：$\varepsilon(\mathbf{w})=\frac{1}{2}\sum_{i=1}^{N}\left(f(\mathbf{x}_i)-y_i\right)^2=\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i-y_i\right)^2$
  - 构造最优化问题：$\min_{\mathbf{w}}\varepsilon(\mathbf{w})$
- 线性最小二乘回归:
  - 选用最小二乘法则，且回归函数f为线性函数

$$\Rightarrow \quad \min_{\mathbf{w}}\varepsilon(\mathbf{w})\Rightarrow\min_{\mathbf{w}}\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i-y_i\right)^2$$

线性最小二乘(LLS: Linear Least Square)问题

线性最小二乘问题的闭式解

- 把训练数据表达成向量-矩阵形式

$$X=\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right),\mathbf{y}=\left(y_1,y_2,\ldots,y_N\right)^T$$

$$\Rightarrow \arg\min_{\mathbf{w}}\varepsilon\left(\mathbf{w}\right)=\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i-y_i\right)^2=\frac{1}{2}\left\|X^T\mathbf{w}-\mathbf{y}\right\|_2^2$$

$$=\frac{1}{2}\left(X^T\mathbf{w}-\mathbf{y}\right)^T\left(X^T\mathbf{w}-\mathbf{y}\right)$$

- 寻找最优值 (i.e. 对 w 求梯度, 令其为0)

$$\nabla\varepsilon\left(\mathbf{w}\right)=X\left(X^T\mathbf{w}-\mathbf{y}\right)\doteq 0$$

$$\Rightarrow XX^T\mathbf{w}=X\mathbf{y} \Rightarrow \mathbf{w}_o=\left(XX^T\right)^{-1}X\mathbf{y}$$

伪逆

$$\mathbf{w}_o=\arg\min_{\mathbf{w}}\varepsilon\left(\mathbf{w}\right)=X^+\mathbf{y}$$

线性最小二乘问题的3种解释

- 线性最小二乘的代数解释
- 线性最小二乘的几何解释
- 最小二乘法的统计学解释
  - 源于3个基本假设下的最大似然估计(MLE)

线性最小二乘问题的代数解释

- 把训练数据表达成向量-矩阵形式

$$X=\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right),\mathbf{y}=\left(y_1,y_2,\ldots,y_N\right)^T$$

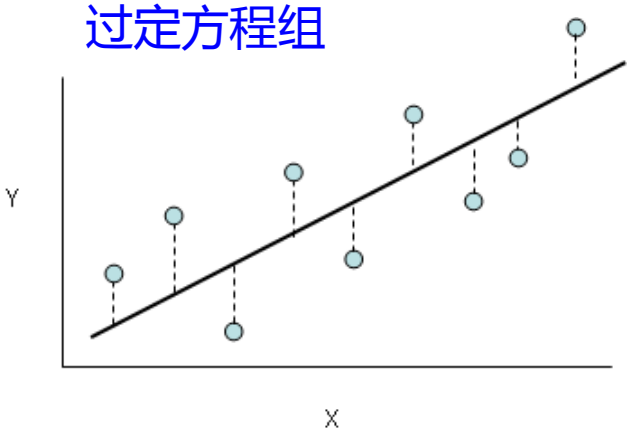
若构造线性方程组 $X^T\mathbf{w}=\mathbf{y}$

则 该方程组可能无解, 所以寻找误差最小意义下的解

$$\Rightarrow \min_{\mathbf{w}}\varepsilon\left(\mathbf{w}\right)=\frac{1}{2}\left\|X^T\mathbf{w}-\mathbf{y}\right\|_2^2=\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i-y_i\right)^2$$

- 说明:
  - 线性方程组**Ax=b**何时无解?
  - 何时唯一解? 增广矩阵满秩
  - 何时无穷多解? 增广矩阵欠秩

过定方程组



线性最小二乘问题的几何解释

- 把训练数据表达成向量-矩阵形式

$$X=\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right),\quad \mathbf{y}=\left(y_1,y_2,\ldots,y_N\right)^T$$

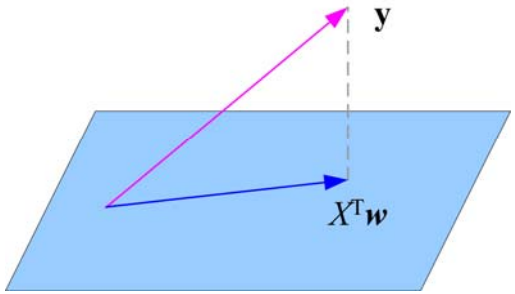
则

$$\arg\min_{\mathbf{w}}\varepsilon\left(\mathbf{w}\right)=\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i-y_i\right)^2=\frac{1}{2}\left\|X^T\mathbf{w}-\mathbf{y}\right\|_2^2$$

– 几何意义：

- 在X的行空间里找 y 的最佳投影

$$\hat{\mathbf{y}}=X^T\left(XX^T\right)^{-1}X\mathbf{y}$$



- 再由线性方程解出 w

$$X^T\mathbf{w}=\hat{\mathbf{y}}\quad\Rightarrow\quad\mathbf{w}_o=\left(XX^T\right)^{-1}X\mathbf{y}$$

最小二乘问题的统计理论解读

- 最小二乘法的统计理论解释
  - 下述3个基本假设下，线性回归模型的最大似然估计(MLE)即导出最小二乘法则
    - 加性误差模型(Additive error model)
    - 独立同分布(i. i. d.: independent identity distribution)
    - 误差服从高斯分布: $y_i - \mathbf{w}^T \mathbf{x}_i = \varepsilon_i \sim N\left(0, \sigma^2\right)$
  - 最大似然估计 $L\left(D;\mathbf{w}\right) = p\left(D;\mathbf{w}\right)$

$$\Rightarrow \arg\max_{\mathbf{w}} L\left(D;\mathbf{w}\right) = p\left(\varepsilon_1, \varepsilon_2, \varepsilon_3, ..., \varepsilon_N;\mathbf{w}\right)$$

$$= \prod_{i=1}^{N} p\left(\varepsilon_i;\mathbf{w}\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2}\right)$$

最小二乘问题的统计理论解读

- 最小二乘法的统计理论解释
  - 下述3个基本假设下，线性回归模型的最大似然估计(MLE)　　….. MLE → MAP / Bayesian Estimation
    - 加性误差模型(Additive error model)
    - 独立同分布(i. i. d.: independent identity distribution)
    - 误差服从高斯分布: $y_i - \mathbf{w}^T\mathbf{x}_i = \varepsilon_i \sim N\left(0, \sigma^2\right)$
  - 最大似然估计

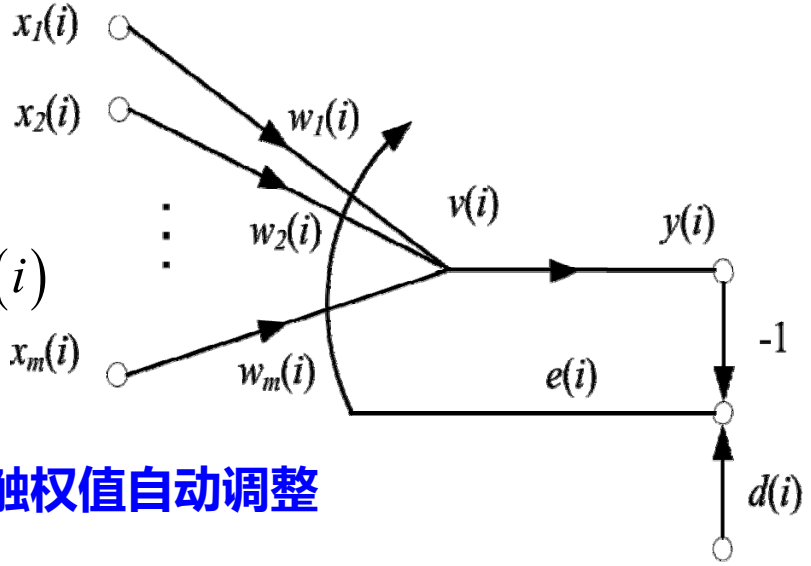
$$\arg\max_{\mathbf{w}} L\left(D;\mathbf{w}\right) = \arg\min_{\mathbf{w}} \; -\ln L\left(D;\mathbf{w}\right)$$

$$= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2$$

- **内容提要**
  - 引言
  - 线性分类 → 感知器 $F(\mathbf{x})=\varphi(\mathbf{w}^T\mathbf{x})=\mathrm{sgn}(\mathbf{w}^T\mathbf{x})$
  - 线性回归 → **Adaline** $F(\mathbf{x})=\varphi(\mathbf{w}^T\mathbf{x})=\mathbf{w}^T\mathbf{x}$
  - 逻辑斯蒂回归

自适应线性单元(Adaline)

- 在1960年Widrow & Hoff提出了自适应线性单元
  - 构成自适应滤波器的基础
    - **滤波过程**
      - **两个信号的计算**
        - » 输出信号 $y(i)$
        - » 误差信号 $e(i)=d(i)-y(i)$

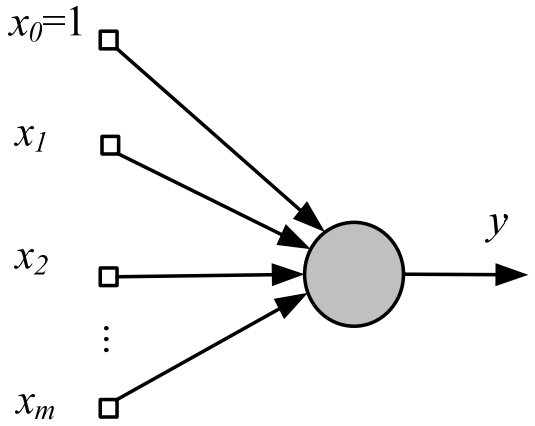


    - **自适应过程**
      - **根据误差信号 $e(i)$ 对神经元突触权值自动调整**

  - **Adaline训练算法:**
    - 最小均方(LMS: Least Mean Square)算法
    - 与感知器在本质上是相关的

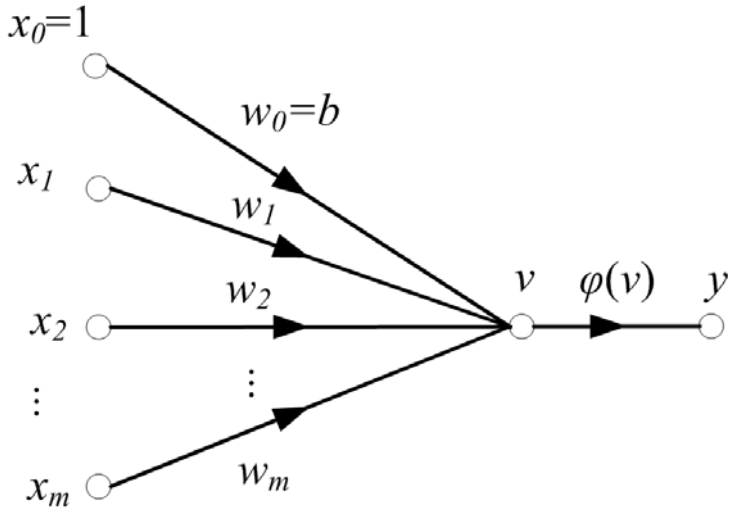
线性神经元模型

- 线性回归模型
  - 1个线性神经元



- 数学表达 $y = \varphi\left(\sum_{i=0}^{m} w_i x_i\right) = \varphi\left(\mathbf{w}^T\mathbf{x}\right) = \mathbf{w}^T\mathbf{x}$

其中 $\varphi(v) = v$



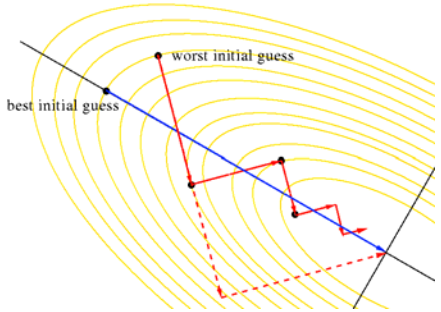
Adaline的训练算法

- 定义瞬时误差函数和瞬时能量:

$$\varepsilon(\mathbf{w}) = \frac{1}{2}e^2 = \frac{1}{2}\left(\mathbf{w}^T\mathbf{x}_i - y_i\right)^2, \quad e = y_i - \mathbf{w}^T\mathbf{x}_i$$

- 计算随机梯度:

$$\nabla\varepsilon(\mathbf{w}) = \left(\mathbf{w}^T\mathbf{x}_i - y_i\right)\mathbf{x}_i$$



- 权值更新法则:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla\varepsilon(\mathbf{w})$$

  - 基于随机梯度下降**(Stochastic Gradient Descent)**
  - 线性最小二乘问题的在线版本
    - 应用于自适应信号处理

自适应线性单元: Adaline

Algorithm 3 Adaline Algorithm

Input: train data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, learning rate $\eta > 0$
Output: weights $\mathbf{w}$
Initialize weights $\mathbf{w}(0)$ at $t = 0$ randomly
Repeat
  For $i = 1$ to N

$$\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta \cdot (\mathbf{w}^\top \mathbf{x}_i - y_i)\mathbf{x}_i$$

  End-For
Until stop criterion is satisfied.

学习速率的设定

- 学习速率
  - 学习速率在计算过程中保持不变：$\eta = \eta_0$
    - 收敛速度慢
      - 对数尺度
  - 学习速率随时间改变
    - 搜索然后收敛进度：$\eta(t) = \frac{\eta_0}{1 + (t/\tau)}$
    - 随机逼近进度：

$$\eta(t) = \frac{c}{t}$$

改写感知器训练算法

Algorithm 2 Perceptron Algorithm

Input: train data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, learning rate $\eta = 0.50$

Output: weights $\mathbf{w}$

Initialize weights $\mathbf{w}(0)$ at $t = 0$ 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

$$\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta \cdot (\text{sign}(\mathbf{w}^\top \mathbf{x}_i) - y_i)\mathbf{x}_i$$

End-For

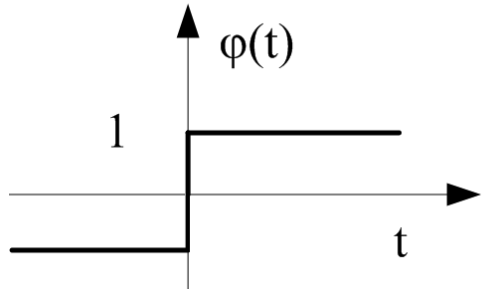
Until stop criterion is satisfied.

- **内容提要**
  - 引言
  - 线性分类 → 感知器 $F(\mathbf{x})=\varphi(\mathbf{w}^T\mathbf{x})=\mathrm{sgn}(\mathbf{w}^T\mathbf{x})$
  - 线性回归 → **Adaline** $F(\mathbf{x})=\varphi(\mathbf{w}^T\mathbf{x})=\mathbf{w}^T\mathbf{x}$
  - 逻辑斯蒂回归 $F(\mathbf{x})=\varphi(\mathbf{w}^T\mathbf{x})=?$

三种模型的比较

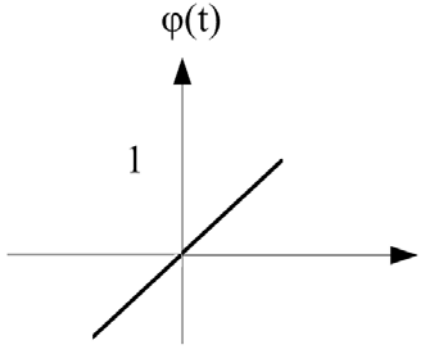
- 线性分类 → 感知器

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



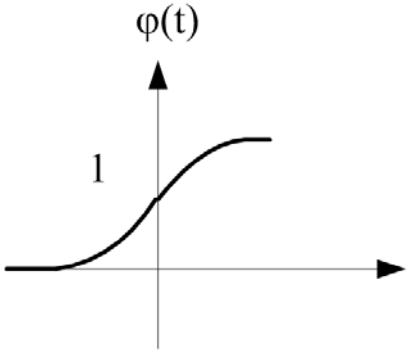
- 线性回归 → Adaline

$$F(\mathbf{x})=\varphi\left(\mathbf{w}^T\mathbf{x}\right)=\mathbf{w}^T\mathbf{x},\quad \varphi(t)=t$$



- Logistic回归

$$F(\mathbf{x})=\varphi\left(\mathbf{w}^T\mathbf{x}\right)=\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}},\quad \varphi(t)=\frac{1}{1+e^{-t}}$$



逻辑斯蒂(Logistic)回归

- 给定训练数据为 $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ ，其中类别标签为1和0，即

$$y_i = \begin{cases} 1 & \text{class } C_1 \\ 0 & \text{class } C_2 \end{cases}$$

其中，y用{0,1}编码

– 使用 $F(\mathbf{x}) = \varphi(\mathbf{w}^T\mathbf{x}) = \dfrac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$ 近似 $P(y=1|\mathbf{x})$

则 $1-F(\mathbf{x}) = 1-\varphi(\mathbf{w}^T\mathbf{x}) = \dfrac{e^{-\mathbf{w}^T\mathbf{x}}}{1+e^{-\mathbf{w}^T\mathbf{x}}}$ 近似 $P(y=0|\mathbf{x})$

数据点$\mathbf{x}_i$所对应的输出为 $\hat{y}_i = F(\mathbf{x}_i) = \varphi(\mathbf{w}^T\mathbf{x}_i) = \dfrac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_i}}$

– 逻辑回归，也被称为广义线性模型，因为：

$$\ln\frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \ln\exp(\mathbf{w}^T\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$

参数w的最大似然估计

- 似然函数为

$$L(\mathbf{y}\,|\,X,\mathbf{w})=\prod_{i=1}^{N}\left[F(\mathbf{w},\mathbf{x}_i)\right]^{y_i}\left[1-F(\mathbf{w},\mathbf{x}_i)\right]^{1-y_i}$$

$$=\prod_{i=1}^{N}\left[\varphi\left(\mathbf{w}^T\mathbf{x}_i\right)\right]^{y_i}\left[1-\varphi\left(\mathbf{w}^T\mathbf{x}_i\right)\right]^{1-y_i}$$

– 取负对数，则

$$\varepsilon(\mathbf{w})=-\ln L(\mathbf{y}\,|\,X,\mathbf{w})$$

$$=-\sum_{i=1}^{N}y_i\log\left[\varphi\left(\mathbf{w}^T\mathbf{x}_i\right)\right]+(1-y_i)\log\left[1-\varphi\left(\mathbf{w}^T\mathbf{x}_i\right)\right]$$

- 最大似然估计: $\max_{\mathbf{w}} L(\mathbf{y}\,|\,X,\mathbf{w})$ ➡ $\min_{\mathbf{w}} \varepsilon(\mathbf{w})$

逻辑斯蒂回归的第2种导出方式

其中y用{-1,1}编码

- Logistic 回归模型

$$F(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}},$$

  - 模型输出0到1之间的一个实数，可解释为概率
  - 给定数据(x,y), 其中y为二值变量

$$P(y|\mathbf{x}) = \begin{cases} f(\mathbf{x}), & y=+1 \\ 1-f(\mathbf{x}), & y=-1 \end{cases}$$

  - 目标: 寻找函数F(x)，以最佳地近似函数 f(x)

似然函数

- 对于每一对(x,y)，我们用F(x) 近似函数f(x)

  - 注意到：

$$P(y|\mathbf{x}) \approx \begin{cases} F(\mathbf{x}), & y=+1 \\ 1-F(\mathbf{x}), & y=-1 \end{cases}$$

$$F(\mathbf{x})=\varphi(\mathbf{w}^T\mathbf{x}),\ 1-F(\mathbf{x})=1-\varphi(\mathbf{w}^T\mathbf{x})=\varphi(-\mathbf{w}^T\mathbf{x})$$

$$\Rightarrow\ P(y|\mathbf{x}) \approx \varphi(y\mathbf{w}^T\mathbf{x})$$

- 给定 i. i. d. 训练数据集 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^N$，则似然函数为

$$L(\mathbf{y}|X,\mathbf{w})=\mathbf{P}(\mathbf{y}|X,\mathbf{w})=\prod_{i=1}^N P(y_i|\mathbf{x}_i,\mathbf{w})=\prod_{i=1}^N \varphi(y_i\mathbf{w}^T\mathbf{x}_i)$$

最大似然估计(MLE)

- 寻找参数w, 使得似然函数最大化

$$\max_{\mathbf{w}} \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w}) \quad \Rightarrow \quad \min_{\mathbf{w}} -\ln\left(\prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})\right)$$

$$= -\sum_{i=1}^{N} \ln\left(\frac{1}{1+e^{-y_i \mathbf{w}^T \mathbf{x}_i}}\right)$$

$$\Rightarrow \quad \varepsilon(\mathbf{w}) = \sum_{i=1}^{N} \ln\left(1+e^{-y_i \mathbf{w}^T \mathbf{x}_i}\right)$$

两种导出方式所得到的目标函数$\varepsilon(\mathbf{w})$是等价的！

模型求解: 迭代优化

- 寻找参数w, 最小化下述目标函数

$$\min_{\mathbf{w}} \varepsilon(\mathbf{w}) = \sum_{i=1}^{N} \log\left(1+e^{-y_i \mathbf{w}^T \mathbf{x}_i}\right)$$

– 计算梯度向量: $\nabla\varepsilon(\mathbf{w}) = -\sum_{i=1}^{N} \frac{y_i \mathbf{x}_i}{1+e^{y_i \mathbf{w}^T \mathbf{x}_i}}$

– **Hessian矩阵**

  • 半正定

$$H = \sum_{i=1}^{N} \frac{y_i^2 e^{y_i \mathbf{w}^T \mathbf{x}_i}}{\left(1+e^{y_i \mathbf{w}^T \mathbf{x}_i}\right)^2} \mathbf{x}_i \mathbf{x}_i^T$$

➡ **凸优化问题(Convex Program)**

梯度下降法求解Logistic回归

Algorithm 4 Logistic Regression Algorithm

Input: training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, learning rate $\eta > 0$
Output: weight vector $\mathbf{w}$
Initialize weights $\mathbf{w}(0)$ at $t = 0$ randomly
Repeat
Compute the gradient

$$\nabla \mathcal{E}(\mathbf{w}) = -\sum_{i=1}^{N} \frac{y_i \mathbf{x}_i}{1+\exp(y_i \mathbf{w}(t)^\top \mathbf{x}_i)}$$

Update the weights: $\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta \cdot \nabla \mathcal{E}(\mathbf{w})$
Until stop criterion is satisfied.

梯度下降(GD: Gradient Descent)

随机梯度法求解Logistic回归

Algorithm 5 Logistic Regression Algorithm (Online)

Input: training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, learning rate $\eta > 0$
Output: weight vector $\mathbf{w}$
Initialize weights $\mathbf{w}(0)$ at $t = 0$ randomly
Repeat
  Compute the gradient

$$\nabla \mathcal{E}(\mathbf{w}) = -\frac{y_i \mathbf{x}_i}{1+\exp(y_i \mathbf{w}(t)^\top \mathbf{x}_i)}$$

  Update the weights: $\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta \cdot \nabla \mathcal{E}(\mathbf{w})$
Until stop criterion is satisfied.

随机梯度下降(SGD: Stochastic Gradient Descent)

- **内容提要**
  - 引言
  - 线性分类 → 感知器 $F(\mathbf{x})=\varphi\left(\mathbf{w}^T\mathbf{x}\right)=\mathrm{sgn}\left(\mathbf{w}^T\mathbf{x}\right)$
  - 线性回归 → **Adaline** $F(\mathbf{x})=\varphi\left(\mathbf{w}^T\mathbf{x}\right)=\mathbf{w}^T\mathbf{x}$
  - 逻辑斯蒂回归 $F(\mathbf{x})=\varphi\left(\mathbf{w}^T\mathbf{x}\right)=\dfrac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$

线性模型的拓展

- 1. 线性模型的核化扩展
  - **Kernel ABC…**
- 2. 最大似然估计被替换为最大后验概率估计或贝叶斯估计
  - **Ridge 回归 / Lasso / … Bayesian Estimation**
- 3. 感知器准则函数可以被应用于结构化学习问题
  - **Structural Learning**
- 4. 两类分类问题被推广到多类分类问题
  - **Multi-Class Classification**
- 5. 逻辑回归模型被推广到多类分类问题
  - **Multi-Class Logistic Regression**
- 6. 多类分类问题中考虑多任务学习思想
  - **Multi-Task Learning**
- 7. 感知器学习错误的界(bound)
  - **Novikoff 1962; V. & C. 1974; Freund & Schapire 1999; Mohri 2013**