2020.2.18 - 2020.6.2 9:50-12:15@6C102雨课堂+腾讯会议
《模式识别与机器学习》

清华大学

第二章
线性学习机器

2020.2.18

Xuegong Zhang 1

多选题 10分 设置

以下陈述正确的是：

- A 模式识别既指对样本分类的任务，也指实现分类的一些方法
- B 机器学习是一大类方法的总称，包括模式识别方法
- C 机器学习除了能够分类外，还可以用于其他任务
- D 机器学习是把任务分解成规则、由计算机自动执行规则进行决策
- E 在很多情况下，模式识别与机器学习基本是同义词

提交

一个简单的例子

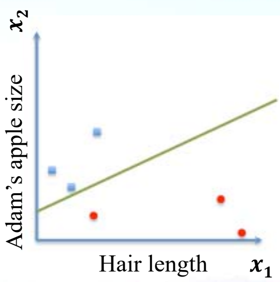
- 如何教小孩从照片识别性别？

一个简单的例子

- 如何教小孩从照片识别性别？

Hair length

The **Adam's apple**, or laryngeal prominence, is a feature of the human neck, and is the lump or protrusion that is formed by the angle of the thyroid cartilage surrounding the larynx seen especially in males.



For the Green line:

$$w_1x_1 + w_2x_2 + w_0 = 0$$

For each blue dots:

$$w_1x_1 + w_2x_2 + w_0 > 0$$

For each red dots:

$$w_1x_1 + w_2x_2 + w_0 < 0$$

线性分类器

$$g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0$$

特征

权值

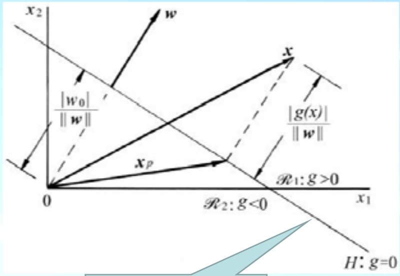
线性判别函数

$$y = \mathrm{sgn}\left(\sum_{i=1}^{d} w_i x_i + w_0\right)$$

决策规则

$$y = \begin{cases} +1 & \Rightarrow \text{class } A \;\; or \;\; \mathbf{x} \in \omega_1 \\ -1 & \Rightarrow \text{class } B \;\; or \;\; \mathbf{x} \in \omega_2 \end{cases}$$

分类标签



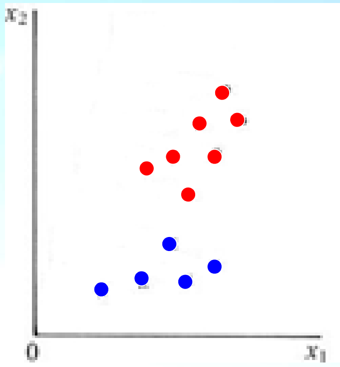
决策面（线）

2.1 Fisher线性判别(FLD）

Fisher's Linear Discriminant

a.k.a. Fisher's Linear Discriminant Analysis
(FLD or LDA)

- 假设数据分布如图所示，如何求解线性分类器？



Sir Ronald Aylmer Fisher (R.A. Fisher)

(17 February 1890 – 29 July 1962)

- British statistician and geneticist　20世纪最伟大的统计学家之一
  - "a genius who almost single-handedly created the foundations for modern statistical science"
  - "the single most important figure in 20th century statistics"
  - "the greatest of Darwin's successors".
- Some of the stuff he invited or popularized
  - ANOVA (analysis of variance)
  - Maximum likelihood
  - Fisher's z-distribution (F distribution)
  - Fisher's method for data fusion (meta-analysis)
  - The 0.05 cutoff of p-value, the notion of hull hypothesis
  - Fisher's exact test
  - Fisher's Discriminant Analysis (in 1936)
  - ……
  - *The Genetical Theory of Natural Selection* (1930)
  - *The Design of Experiments* (1935)

From Wikipedia

不能推断因果关系!

Sir Ronald Aylmer Fisher (R.A. Fisher)

(17 February 1890 – 29 July 1962)

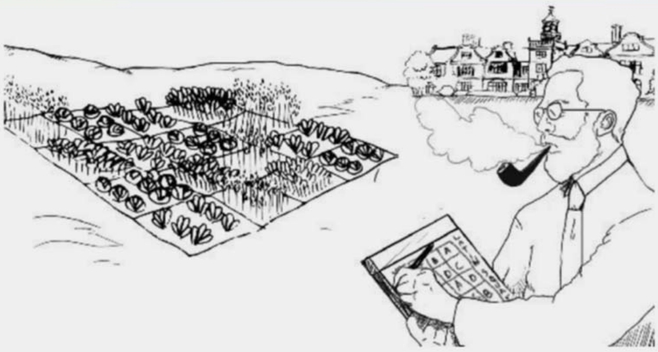


FIGURE 4.3. R. A. Fisher with one of his many innovations: a Latin square experimental design, intended to ensure that one plot of each plant type appears in each row (fertilizer type) and column (soil type). Such designs are still used in practice, but Fisher would later argue convincingly that a randomized design is even more effective. (*Source:* Drawing by Dakota Harr.)

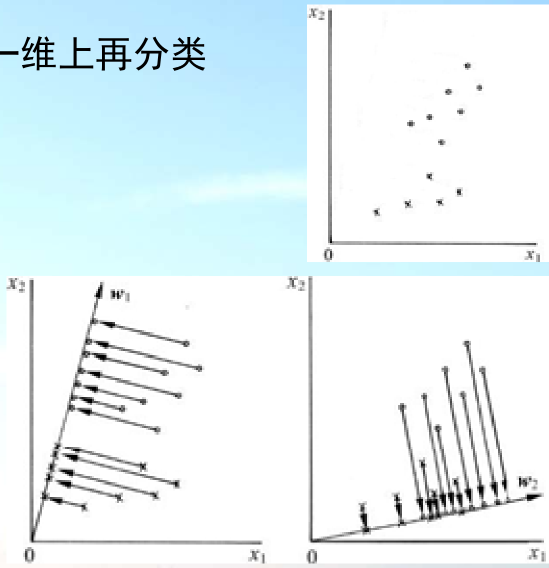
Judea Pearl, *The Book of Why*, 2018

Suggestion: Find stories about Fisher's role in denying the harm of smoking on health, and learn about the topics of causal inference from the story.

Fisher判别准则

- 求解最佳投影方向，把样本投影到一维上再分类
  - 类内：样本越紧密越好
  - 类间：两类离越远越好

- 样本集：$\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$,
  其中，第一类 (ω_1): $\mathcal{X}_1 = \{\boldsymbol{x}_1^1, \cdots, \boldsymbol{x}_{N_1}^1\}$,
  第二类 (ω_2): $\mathcal{X}_2 = \{\boldsymbol{x}_1^2, \cdots, \boldsymbol{x}_{N_2}^2\}$
- 投影： $\mathcal{X} \rightarrow \mathcal{Y}$: $y_i = \boldsymbol{w}^T \boldsymbol{x}_i$, $i = 1, \cdots, N$



Fisher判别准则

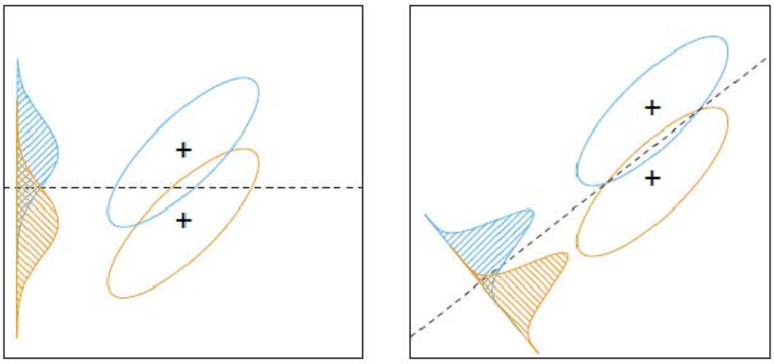


FIGURE 4.9. *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer

这里的类内和类间有什么特殊的关系?

考查样本的类内和类间离散度

求解最佳投影方向，把样本投影到一维上再分类
- **类内：样本越紧密越好**
- **类间：两类离越远越好**

- 在$\mathcal{X}$空间：

类均值向量 $\boldsymbol{m}_i = \frac{1}{N_i}\sum_{x_j \in \mathcal{X}_i} \boldsymbol{x}_j,\ i = 1,2$

类内离散度矩阵 within-class scatter matrix

$$\boldsymbol{S}_i = \sum_{x_j \in \mathcal{X}_i} (\boldsymbol{x}_j - \boldsymbol{m}_i)(\boldsymbol{x}_j - \boldsymbol{m}_i)^T, \qquad i = 1,2$$

总类内离散度矩阵 $\boldsymbol{S}_w = \boldsymbol{S}_1 + \boldsymbol{S}_2$

类间离散度矩阵 between-class scatter matrix

$$\boldsymbol{S}_b = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$$

考查样本的类内和类间离散度

求解最佳投影方向，把样本投影到一维上再分类
- 类内：样本越紧密越好
- 类间：两类离越远越好

- 在$\mathcal{Y}$空间：

  类均值 $\tilde{m}_i = \frac{1}{N_i}\sum_{y_j \in \mathcal{Y}_i} y_j,\quad i=1,2$

  类内离散度

  $$\tilde{S}_i = \sum_{y_j \in \mathcal{Y}_i} (y_j - \tilde{m}_i)(y_j - \tilde{m}_i)^T, \qquad i=1,2$$

  总类内离散度 $\tilde{S}_w = \tilde{S}_1 + \tilde{S}_2$

  类间离散度矩阵 $\tilde{S}_b = (\tilde{m}_1 - \tilde{m}_2)^2$

Xuegong Zhang 13

考查样本的类内和类间离散度

求解最佳投影方向，把样本投影到一维上再分类
- 类内：样本越紧密越好
- 类间：两类离越远越好

- 在$\mathcal{Y}$空间：

  类均值 $\tilde{m}_i = \frac{1}{N_i}\sum_{y_j \in \mathcal{Y}_i} y_j,\quad i=1,2$

  类内离散度

  $$\tilde{S}_i = \sum_{y_j \in \mathcal{Y}_i} (y_j - \tilde{m}_i)(y_j - \tilde{m}_i)^T, \qquad i=1,2$$

  总类内离散度 $\tilde{S}_w = \tilde{S}_1 + \tilde{S}_2$

  类间离散度矩阵 $\tilde{S}_b = (\tilde{m}_1 - \tilde{m}_2)^2$

- Fisher准则

$$\max_{\boldsymbol{w}} \ J_F(\boldsymbol{w}) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1 + \tilde{S}_2}$$

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i$$

这里是doubleU

Xuegong Zhang 14

- Fisher准则：

$$\max_{\boldsymbol{w}} \ J_F(\boldsymbol{w}) = \frac{(\widetilde{m}_1 - \widetilde{m}_2)^2}{\tilde{S}_1 + \tilde{S}_2}$$

求解最佳投影方向，把样本投影到一维上再分类
- 类内：样本越紧密越好
- 类间：两类离越远越好

- 代入$y = \boldsymbol{w}^T\boldsymbol{x}$，得最优投影方向

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmax}} \ J_F(\boldsymbol{w})$$

- Fisher准则：

$$J_F(\boldsymbol{w}) = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$$

存在多解，导致求解困难

思考：最大化$J_F(\boldsymbol{w})$会遇到什么问题？

雨课堂随机点名，
视频会议中语音回答

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmax}} \ J_F(\boldsymbol{w})$$

$$J_F(\boldsymbol{w}) = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$$

求解：

- 问题：改变$\boldsymbol{w}$的幅度, $J_F(\boldsymbol{w})$不会改变 → 无唯一解
- 不妨令分母$\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w} = c \neq 0$，最大化分子$\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}$, 即：

$$\max \boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}$$

$$s.t. \ \boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w} = c$$

拉格朗日乘子法

--- 带有等式约束的优化问题

怎样求解？

雨课堂弹幕抢答

$$\max \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}$$
$$s.t.\ \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} = c$$

- 拉格朗日乘子法求最优：
- 定义拉格朗日函数

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - c)$$

令 $\frac{\partial L}{\partial w} = 0$，可得

$$\boldsymbol{S}_w^{-1} \boldsymbol{S}_b \boldsymbol{w}^* = \lambda \boldsymbol{w}^*$$

即：w^*为$S_w^{-1}S_b$矩阵的本征向量(eigenvector)

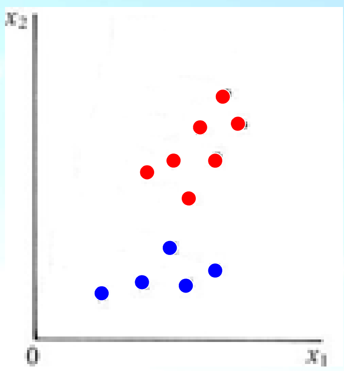
eigen：自我

只考虑投影的方向，得

$$\boldsymbol{w}^* \propto \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

Xuegong Zhang 17

- 不忘初心：求分类器的目标实现了吗？



请在雨课堂中弹幕回答

Xuegong Zhang 18

$$\boldsymbol{w}^* \propto \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

- 有了投影方向，还需要确定决策的分界点

$$y = \mathrm{sgn}\left(\sum_{i=1}^{n} w_i x_i + w_0\right) = \mathrm{sgn}(\mathbf{w}^T\mathbf{x} + w_0), \qquad y = \begin{cases} +1 & \Rightarrow \mathbf{x} \in \omega_1 \\ -1 & \Rightarrow \mathbf{x} \in \omega_2 \end{cases}$$

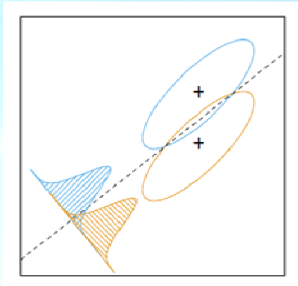
- 如何选择w_0？
  - 根据对数据的不同认识，可以有多种选择方法，比如

$$w_0 = -\frac{1}{2}(\tilde{m}_1 + \tilde{m}_2)$$

$$w_0 = -\tilde{m}$$

$$w_0 = -\frac{1}{2}(\tilde{m}_1 + \tilde{m}_2) + \frac{1}{N_1 + N_2 - 2}\ln\frac{P(\omega_1)}{P(\omega_2)}$$

  - 可以根据对错误率的要求来选择（见下周课）



- 完美！
- 不过，这算机器学习吗？

不完全算：
学习的成分不够强大，而是完全通过解析的方法

D

请在雨课堂中弹幕回答

- 完美！
- 不过，这算机器学习吗？

D

人 “看”到东西 ➔ 认出东西、产生想法

观察 ➔ 判断

↓ 观测 ➔ 分类决策

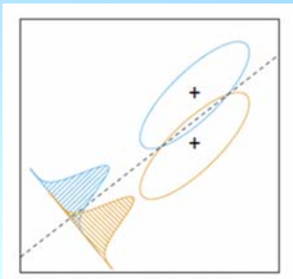
量化观测 ➔ 类别标签

机器 $x \in R^d$ ➔ $y \in \{\omega_1, \omega_2, \cdots\}$

模式识别

如果机器通过实例学会识别，而不是在程序里写好怎样识别，那就是机器学习。

基于数据的机器学习



两种视角看FLD：

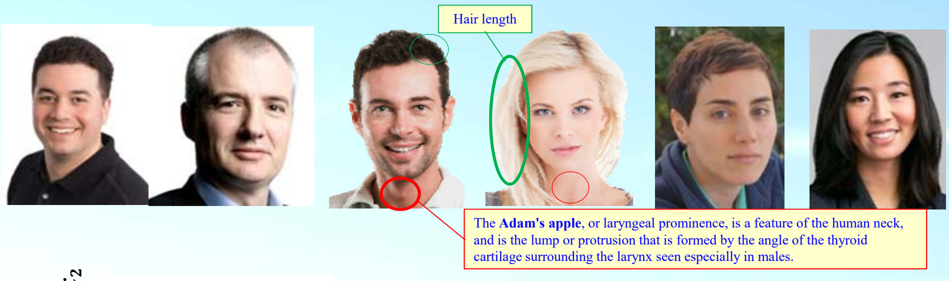
- 解析求解线性分类器
- 提取一维最优特征 + 在一维上用阈值分类器

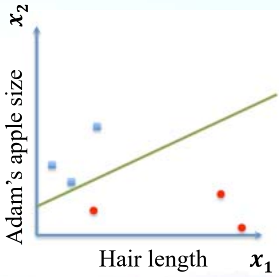
休息1分钟

2.1 感知器
Perceptron

一个简单的例子

- 如何教小孩从照片识别性别？





For the Green line:

$w_1x_1 + w_2x_2 + w_0 = 0$

For each blue dots:

$w_1x_1 + w_2x_2 + w_0 > 0$

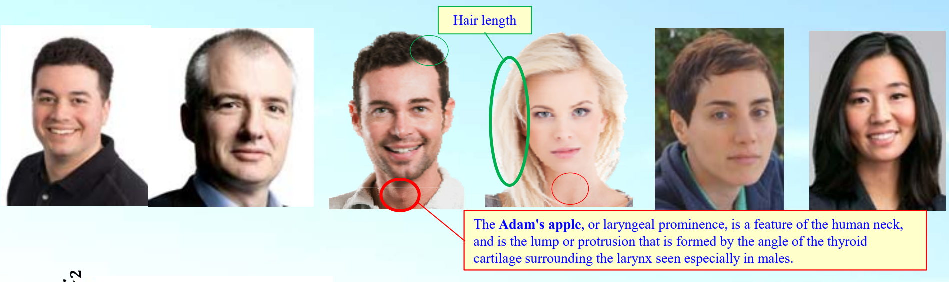
For each red dots:

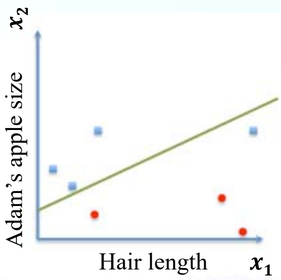
$w_1x_1 + w_2x_2 + w_0 < 0$

Xuegong Zhang 25

一个简单的例子

- 如何教小孩从照片识别性别？





For the Green line:

$w_1x_1 + w_2x_2 + w_0 = 0$

For each blue dots:

$w_1x_1 + w_2x_2 + w_0 > 0$

For each red dots:

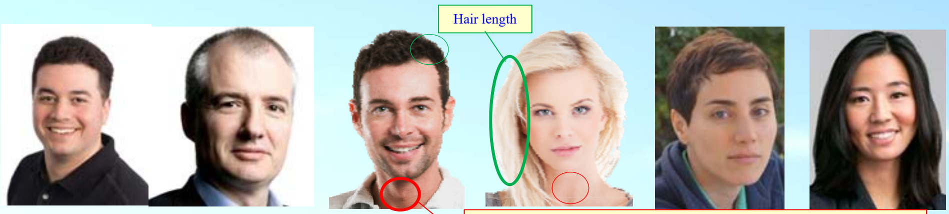
$w_1x_1 + w_2x_2 + w_0 < 0$

D

Xuegong Zhang 26

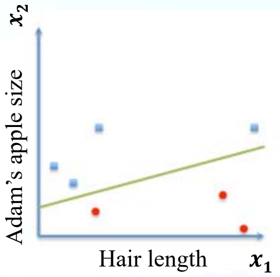
一个简单的例子

- 如何教小孩从照片识别性别？



Hair length

The **Adam's apple**, or laryngeal prominence, is a feature of the human neck, and is the lump or protrusion that is formed by the angle of the thyroid cartilage surrounding the larynx seen especially in males.



For the Green line:

$w_1 x_1 + w_2 x_2 + w_0 = 0$

For each blue dots:

$w_1 x_1 + w_2 x_2 + w_0 > 0$

For each red dots:

$w_1 x_1 + w_2 x_2 + w_0 < 0$

Xuegong Zhang 27

感知器（ Perceptron）

Frank Rosenblatt, *The Perceptron – a perceiving and recognizing automaton*, Report 85-460-1, Cornell Aeronautical Laboratory, Jan. 1957

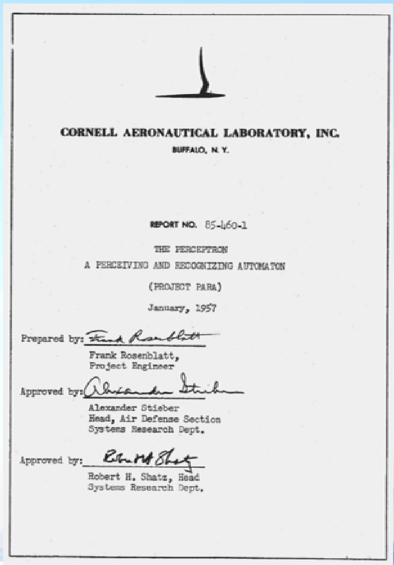
- 感知器：第一台学习机器

$$y = \mathrm{sgn}\left(\sum_{i=1}^{d} w_i x_i + w_0\right)$$

- 为什么把它叫做学习机器？ D

请在雨课堂中弹幕回答

因为在当时就是一台机器!


CORNELL AERONAUTICAL LABORATORY, INC.
BUFFALO, N. Y.

REPORT NO. 85-460-1

THE PERCEPTRON
A PERCEIVING AND RECOGNIZING AUTOMATON
(PROJECT PARA)
January, 1957

Prepared by: Frank Rosenblatt
Frank Rosenblatt, Project Engineer

Approved by: Alexander Stieber
Alexander Stieber, Head, Air Defense Section, Systems Research Dept.

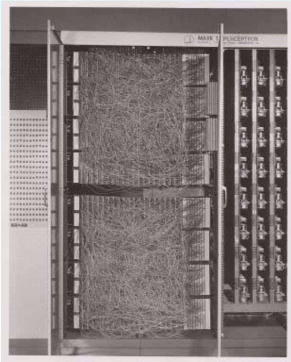
Approved by: Robert M. Shatz
Robert M. Shatz, Head, Systems Research Dept.

Xuegong Zhang

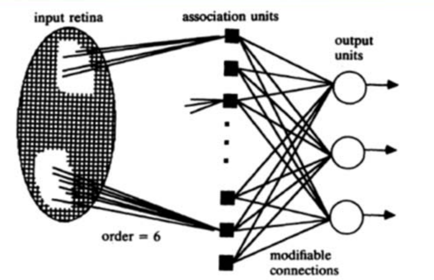
感知器

- 为什么把它叫做学习机器？
  ① 因为它是一台机器
  ② 因为它会学习！

$$y = \mathrm{sgn}(\sum_{i=1}^{n} w_i x_i + w_0)$$



https://en.wikipedia.org/wiki/Perceptron

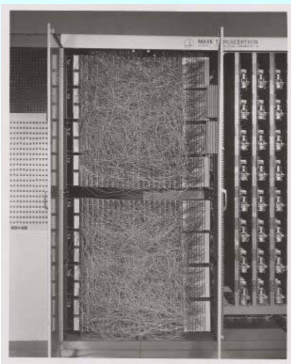


M. Olazaran, A sociological study of the official history of the perceptrons controversy, *Social Studies of Science*, 1996

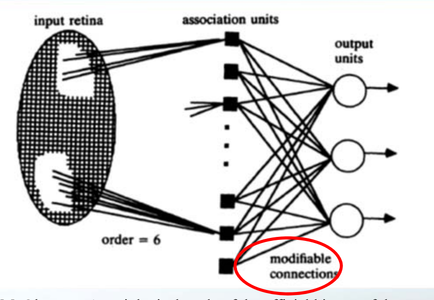
感知器

- 为什么把它叫做学习机器？
  ① 因为它是一台机器
  ② 因为它会学习！
  – 它不是编好程序的冯诺依曼计算机，是一台根据训练数据自我调整的学习机器

$$y = \mathrm{sgn}(\sum_{i=1}^{n} w_i x_i + w_0)$$



https://en.wikipedia.org/wiki/Perceptron



M. Olazaran, A sociological study of the official history of the perceptrons controversy, *Social Studies of Science*, 1996

感知器

- 用数据 $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ 训练线性机器

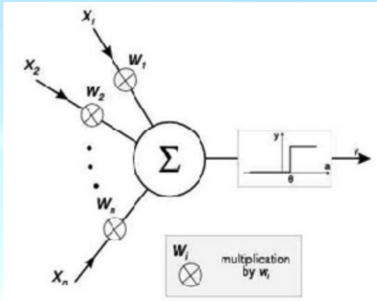
$$y = \mathrm{sgn}(\sum_{i=1}^{d} w_i x_i + w_0)$$

  - 目标：最优化目标函数 $J(w)$

    $J(\boldsymbol{w})$ = 训练错误数

  - 学习算法
    - 梯度下降法

do $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla J$
until $(\nabla J < threshold)$



Xuegong Zhang 31

感知器

- 用数据 $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ 训练线性机器

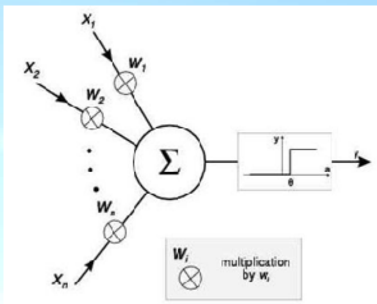
$$y = \mathrm{sgn}(\sum_{i=1}^{d} w_i x_i + w_0)$$

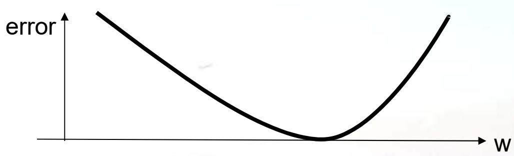
  - 目标：最优化目标函数 $J(w)$

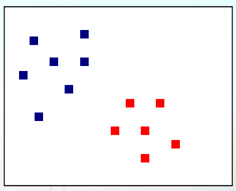
    $J(\boldsymbol{w})$ = 训练错误数

  - 学习算法
    - 梯度下降法

do $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla J$
until $(\nabla J < threshold)$







Xuegong Zhang 32

感知器

- 用数据 $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ 训练线性机器

$$y = \mathrm{sgn}(\sum_{i=1}^{d} w_i x_i + w_0)$$

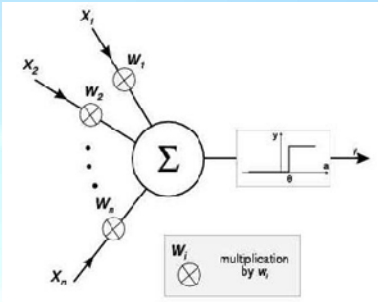
  – 目标：最优化目标函数 $J(w)$

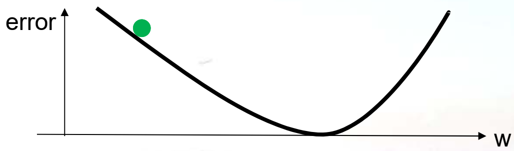
$J(\boldsymbol{w})$ = 训练错误数

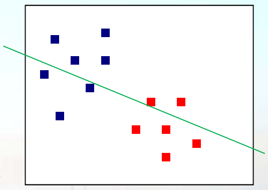
  – 学习算法

    • 梯度下降法

do $\mathbf{w}(t+1) \ = \mathbf{w}(t) - \eta \nabla J$
until $(\nabla J < threshold)$







感知器

- 用数据 $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ 训练线性机器

$$y = \mathrm{sgn}(\sum_{i=1}^{d} w_i x_i + w_0)$$

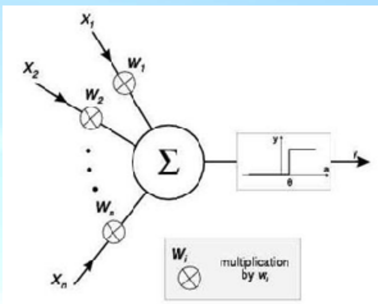
  – 目标：最优化目标函数 $J(w)$

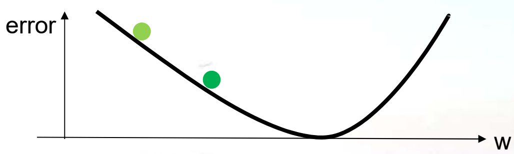
$J(\boldsymbol{w})$ = 训练错误数

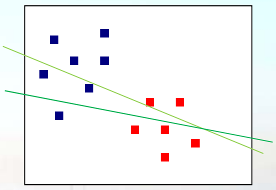
  – 学习算法

    • 梯度下降法

do $\mathbf{w}(t+1) \ = \mathbf{w}(t) - \eta \nabla J$
until $(\nabla J < threshold)$







感知器

- 用数据 $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ 训练线性机器

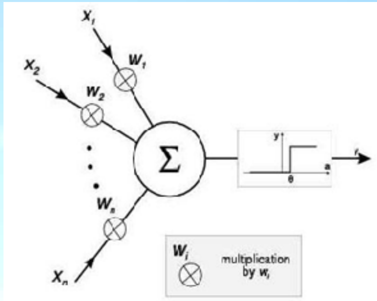
$$y = \mathrm{sgn}(\sum_{i=1}^{d} w_i x_i + w_0)$$

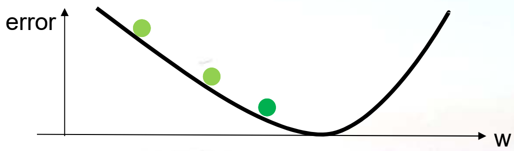
  – 目标：最优化目标函数 $J(w)$

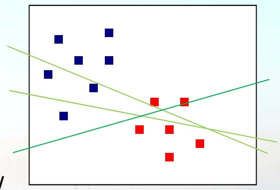
$J(\boldsymbol{w})$ = 训练错误数

  – 学习算法
    - 梯度下降法

do $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla J$
until $(\nabla J < threshold)$







Xuegong Zhang 35

感知器

- 用数据 $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ 训练线性机器

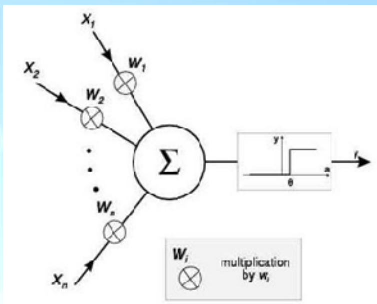
$$y = \mathrm{sgn}(\sum_{i=1}^{d} w_i x_i + w_0)$$

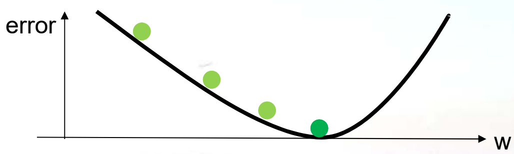
  – 目标：最优化目标函数 $J(w)$

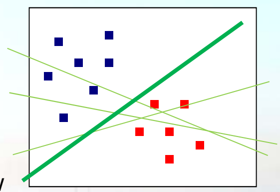
$J(\boldsymbol{w})$ = 训练错误数

  – 学习算法
    - 梯度下降法

do $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla J$
until $(\nabla J < threshold)$







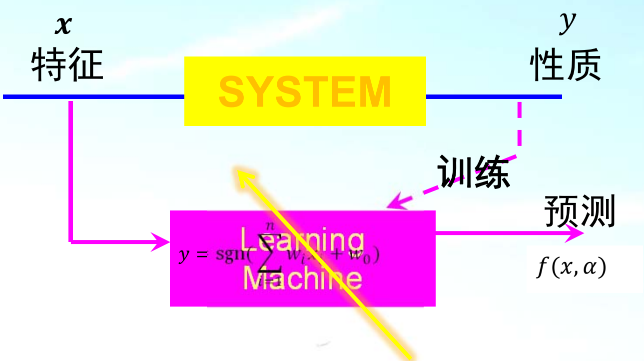
Xuegong Zhang 36

机器学习的基本要素

- 怎样造一个学习机器？
  - 它需要老师
    - → 我们设计它（特征和模型）
  - 它需要训练/学习材料
    - → 训练数据
  - 我们需要为它树立学习的目标
    - → 目标函数、学习准则
  - 我们需要告诉它怎样学
    - → 学习/训练算法

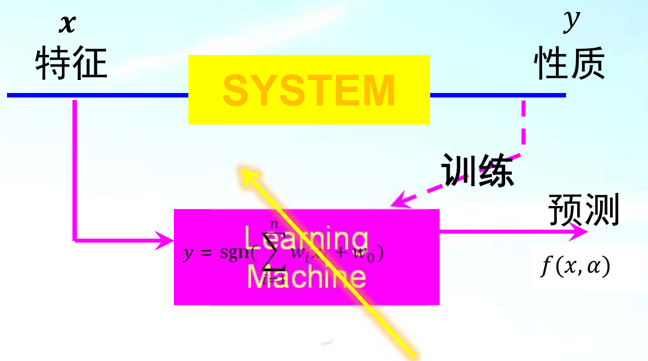
监督学习 Supervised Learning

- 用已知答案的数据去训练 ➔ 监督学习



监督学习 Supervised Learning

- 用已知答案的数据去训练 ➔ 监督学习



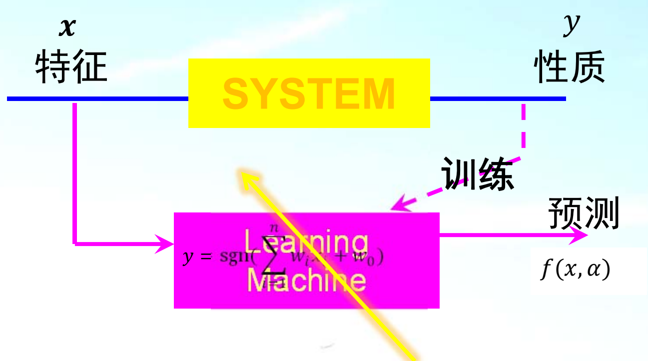
比如：

- x 身高体重→ y性别
- x 图像特征→ y性别
- x 图像像素→ y图像物体
- x 司机视觉→ y方向盘角度
- x 音频信号→ y语音内容
- x 基因表达→ y疾病类型
- x 肺CT影像→ y新冠病毒
- x 近期疫情→ y未来走向
- …

监督学习 Supervised Learning

- 用已知答案的数据去训练 ➔ 监督学习

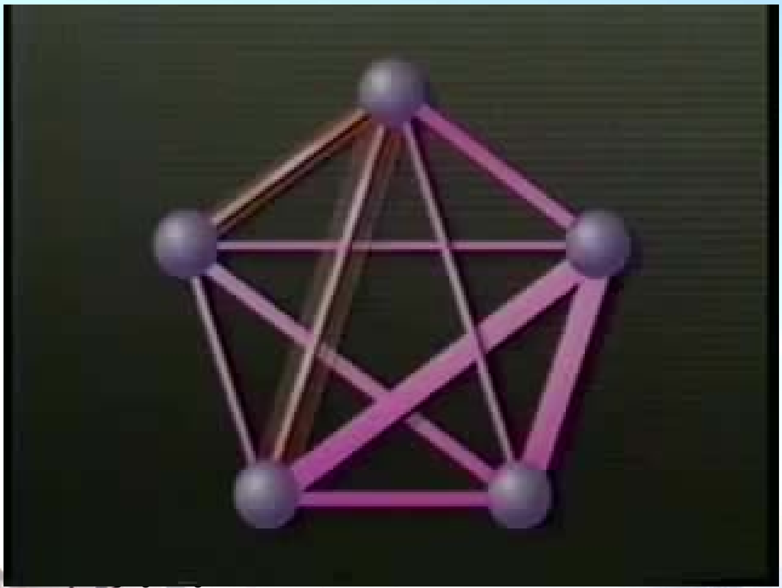
若y是离散类
就是模式识别
→监督模式识别



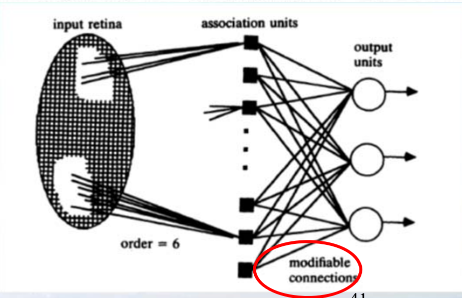
比如：

- x 身高体重→ y性别
- x 图像特征→ y性别
- x 图像像素→ y图像物体
- x 司机视觉→ y方向盘角度
- x 音频信号→ y语音内容
- x 基因表达→ y疾病类型
- x 肺CT影像→ y新冠病毒
- x 近期疫情→ y未来走向
- …

感知器是怎么学习的？



$$y = \mathrm{sgn}\left(\sum_{i=1}^{n} w_i x_i + w_0\right)$$



M. Olazaran, A sociological study of the official history of the perceptrons controversy, *Social Studies of Science*, 1996

41

单选题 1分

设置

休息4分钟，回到座位后请答题

A 已回座位

B 还没有

提交

Xuegong Zhang

42

用什么学习算法？

- 线性判别函数的齐次简化

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \boldsymbol{\alpha}^T\mathbf{y}$$

$\boldsymbol{y} = \begin{bmatrix}1\\ \mathbf{x}\end{bmatrix}$ 增广的特征向量

$\boldsymbol{\alpha} = \begin{bmatrix}w_0\\ \boldsymbol{w}\end{bmatrix}$ 增广的权向量

用什么学习算法？

- 线性判别函数的齐次简化

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \boldsymbol{\alpha}^T\mathbf{y}$$

$\boldsymbol{y} = \begin{bmatrix}1\\ \mathbf{x}\end{bmatrix}$ 增广的特征向量

$\boldsymbol{\alpha} = \begin{bmatrix}w_0\\ \boldsymbol{w}\end{bmatrix}$ 增广的权向量

为进一步简化推导，把样本向量再进行如下规范化：

$$\mathbf{y}_i' = \begin{cases}\mathbf{y}_i, & \text{if } \mathbf{y}_i \in \omega_1, \\ -\mathbf{y}_i, & \text{if } \mathbf{y}_i \in \omega_2,\end{cases} \quad i = 1, \cdots, N$$

$\boldsymbol{y}_i'$：规范化增广样本向量，仍记作$\boldsymbol{y}_i$

于是，对正确分类样本i：$\boldsymbol{\alpha}^T\mathbf{y}_i > 0$

用什么学习算法？

- 线性判别函数的齐次简化

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \boldsymbol{\alpha}^T\mathbf{y}$$

$\boldsymbol{y} = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}$ 增广的特征向量

$\boldsymbol{\alpha} = \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix}$ 增广的权向量

为进一步简化推导，把样本向量再进行如下规范化：

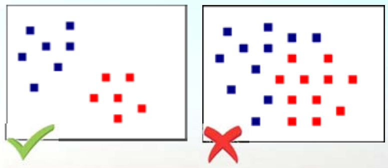
$$\mathbf{y}_i' = \begin{cases} \mathbf{y}_i, & \text{if } \mathbf{y}_i \in \omega_1, \\ -\mathbf{y}_i, & \text{if } \mathbf{y}_i \in \omega_2, \end{cases} \quad i = 1, \cdots, N$$

$\boldsymbol{y}_i'$：规范化增广样本向量，仍记作$\boldsymbol{y}_i$

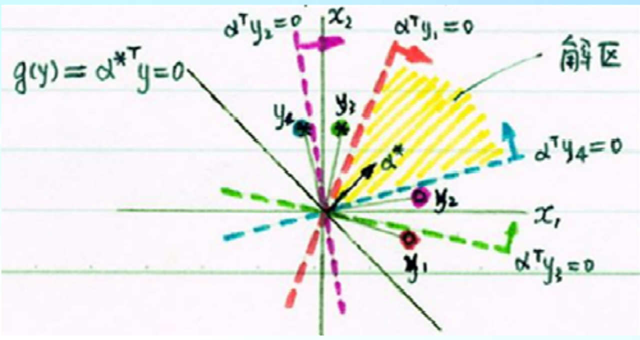
于是，对正确分类样本i：$\boldsymbol{\alpha}^T\mathbf{y}_i > 0$

- 线性可分性：

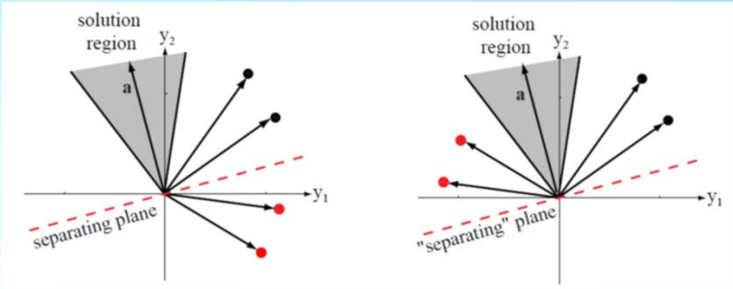
$\exists\boldsymbol{\alpha},\quad \boldsymbol{\alpha}^T\mathbf{y}_i > 0, i = 1, \cdots, N$



- 解向量$\boldsymbol{\alpha}^*$：满足$\boldsymbol{\alpha}^T\boldsymbol{y}_i > 0, i = 1, \cdots, N$
- 解区：权值空间中所有解向量组成的区域

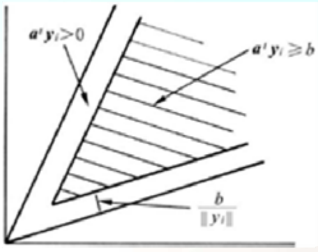


- 对每个样本$\boldsymbol{y}_i$，$\boldsymbol{\alpha}^T\boldsymbol{y}_i = 0$为权空间中的一个超平面，解区只可能在超平面的正侧
- 所有样本对应的超平面的正侧的交集就是解区



Duda et al, *Pattern Classification*

- 引入余量$\boldsymbol{\alpha}^T \boldsymbol{y}_i \geq b > 0$



感知器准则函数

$$J_P(\boldsymbol{\alpha}) = \sum_{\mathbf{y}_j \in \mathcal{Y}^k} (-\boldsymbol{\alpha}^T \mathbf{y}_j)$$

$\mathcal{Y}^k$: 在第k步被$\boldsymbol{\alpha}$ 错分的样本集合

- 感知器算法（Rosenblatt,1957）：

$$J_P(\boldsymbol{\alpha}^*) = \min J_P(\boldsymbol{\alpha}) = 0$$

感知器准则函数

$$J_P(\boldsymbol{\alpha}) = \sum_{\mathbf{y}_j \in \mathcal{Y}^k} (-\boldsymbol{\alpha}^T \mathbf{y}_j)$$

$\mathcal{Y}^k$: 在第k步被$\boldsymbol{\alpha}$ 错分的样本集合

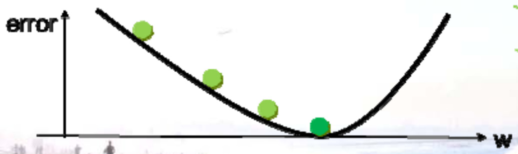
- 感知器算法(Rosenblatt,1957)：

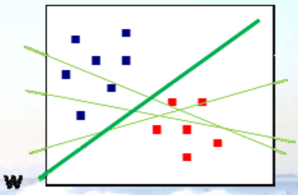
$$J_P(\boldsymbol{\alpha}^*) = \min J_P(\boldsymbol{\alpha}) = 0$$

- 用梯度下降法(Gradient descent)迭代求解

$$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) - \rho_k \nabla J$$

$$\nabla J = \frac{\partial J_P(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{\mathbf{y}_j \in \mathrm{Y}^k} (-\mathbf{y}_j), \qquad \therefore\ \boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \rho_k \sum_{\mathbf{y}_j \in \mathrm{Y}^k} \mathbf{y}_j$$





单样本修正：

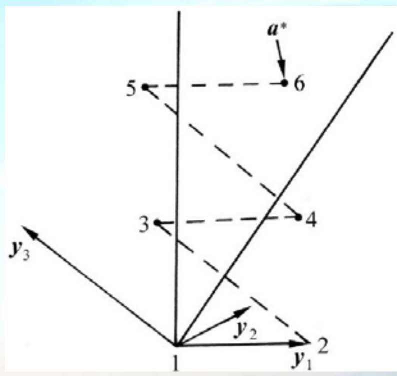
- 固定增量法：

$$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \rho_k \sum_{\mathbf{y}_j \in \mathrm{Y}^k} \mathbf{y}_j$$

① 初值任意

② 对样本y_j，若$\alpha(k)^T y_j \le 0$（或b），则$\alpha(k+1) = \alpha(k) + y_j$

③ 对所有样本重复(2)，直到$J_P = 0$



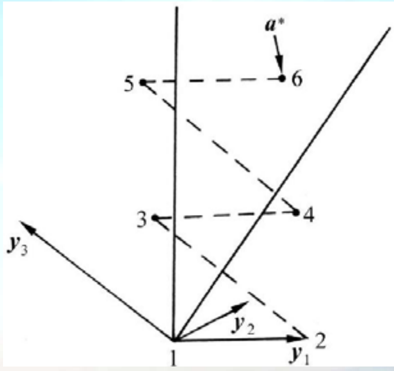
单样本修正：

- 固定增量法：
  ① 初值任意
  ② 对样本y_j，若$\alpha(k)^T y_j \le 0$（或b），则$\alpha(k+1) = \alpha(k) + y_j$
  ③ 对所有样本重复(2)，直到$J_P = 0$

$$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \rho_k \sum_{\mathbf{y}_j \in Y^k} \mathbf{y}_j$$

- 变增量法，如绝对修正法

$$\rho_k = \frac{|\alpha(k)^T y_j|}{\|y_j\|^2}$$



收敛性：

- 对线性可分样本集，
  经过有限次修正后一定可以找到一个解

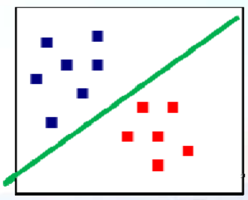
H

讨论

- 感知器有什么问题？

雨课堂随机点名，
视频会议中语音回答

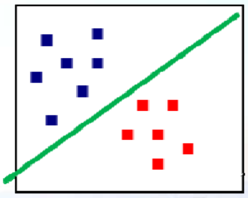
$$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \rho_k \sum_{\mathbf{y}_j \in Y^k} \mathbf{y}_j$$

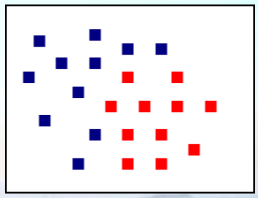


讨论

- 感知器有什么问题？
  - 样本线性不可分呢？
  - 线性可分时多解？
  - 多类呢？
- 有问题怎么办？

$$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \rho_k \sum_{\mathbf{y}_j \in Y^k} \mathbf{y}_j$$

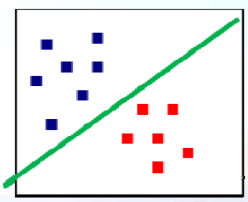


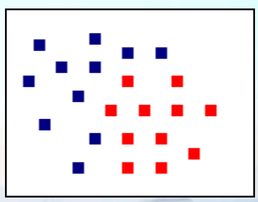


讨论

- 感知器有什么问题？
  - 样本线性不可分呢？
  - 线性可分时多解？
  - 多类呢？
- 有问题怎么办？
  - 容忍错误，使错误尽量小
    - 比如强制收敛、MSE
  - 寻求非线性方法
    - 比如神经网络、SVM
  - 寻求“最优分类器”
    - 比如支持向量机SVM
  - 多类分类方法
  - 用两类分类器完成多类分类

$$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \rho_k \sum_{\mathbf{y}_j \in Y^k} \mathbf{y}_j$$





2.3 线性回归
Linear Regression

线性分类器

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

预测 分类器 样本

$$\mathcal{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_m] = \begin{bmatrix} x_{11} & \cdots & x_{m1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix}$$

特征

样本

$$y = \mathrm{sgn}(\sum_{i=1}^{n} w_i x_i + w_0) = \mathrm{sgn}(\mathbf{w}^T\mathbf{x} + w_0), \qquad y = \begin{cases} +1 & \Rightarrow \mathbf{x} \in \omega_1 \\ -1 & \Rightarrow \mathbf{x} \in \omega_2 \end{cases}$$

线性回归

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

预测 回归 样本

$$x = [\mathbf{x}_1, \cdots, \mathbf{x}_m] = \begin{bmatrix} x_{11} & \cdots & x_{m1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix}$$

特征

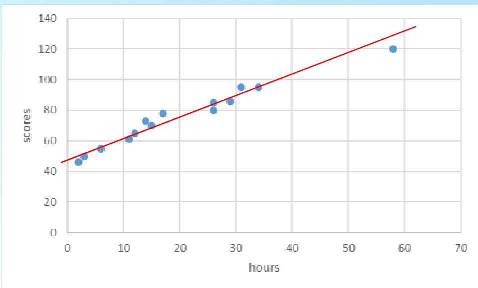
样本

$$y = \sum_{i=1}^{n} w_i x_i + w_0 = \mathbf{w}^T \mathbf{x} + w_0$$

Xuegong Zhang 57

简单线性回归

Student id	Final score	Study Hours per Week
1	50	3
2	95	34
3	78	17
4	55	6
5	65	12
6	70	15
7	80	26
8	86	29
9	73	14
10	120	58
11	46	2
12	95	31
13	85	26
14	61	11



Simple Linear Regression

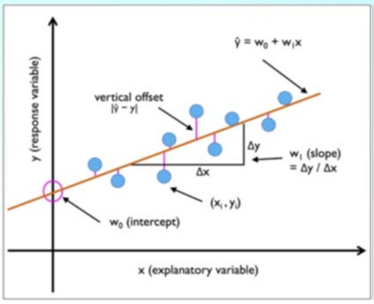
$$y = w_0 + w_1 x$$

Xuegong Zhang 58

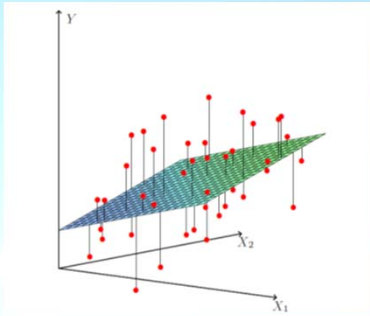
多元线性回归

Simple Linear Regression

$$y = w_0 + w_1 x$$



Raschka & Mirjalili, *Python Machine Learning*



Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*

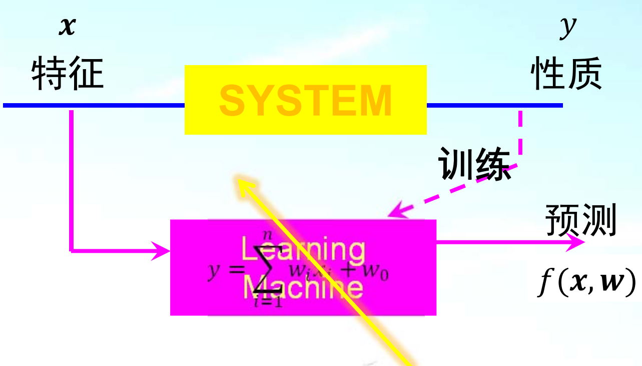
Multiple Linear Regression

$$y = w_0 + w_1 x + \cdots + w_d x_d = \sum_{i=0}^{d} w_i x_i = \boldsymbol{w}^T \boldsymbol{x}$$

Xuegong Zhang 59

线性回归版的机器学习

这里的y不再求符号



$$y = \sum_{i=1}^{n} w_i x_i + w_0 = \mathbf{w}^T \mathbf{x} + w_0$$

Xuegong Zhang 60

机器学习的基本要素

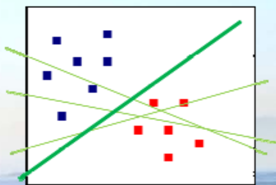
- 怎样造一个学习机器？
  - 它需要老师
    - → 我们设计它（特征和模型）
  - 它需要训练/学习材料
    - → 训练数据
  - 我们需要为它树立学习的目标
    - → 目标函数、学习准则
  - 我们需要告诉它怎样学
    - → 学习/训练算法

机器学习的基本要素：感知器版

- 怎样造一个学习机器？
  - 它需要老师
    - → 我们设计它（特征和模型）：线性判别函数
  - 它需要训练/学习材料
    - → 训练数据：知道分类标签的数据
  - 我们需要为它树立学习的目标
    - → 目标函数、学习准则：错误样本的累计
  - 我们需要告诉它怎样学
    - → 学习/训练算法：梯度下降

D

雨课堂随机点名，
视频会议中语音回答



清華大學

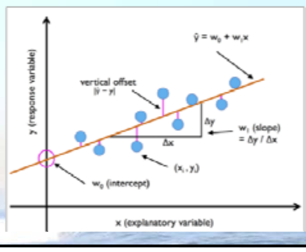
机器学习的基本要素：感知器版

- 怎样造一个学习机器？
  - 它需要老师　　计算结果y的符号
    - → 我们设计它（特征和模型）　$y=\mathrm{sgn}(\sum_{i=1}^{d} w_i x_i + w_0)$
  - 它需要训练/学习材料
    - → 训练数据　$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\},\ \boldsymbol{x}_j \in R^{d+1}, y_j \in \{-1,1\}$
  - 我们需要为它树立学习的目标
    - → 目标函数、学习准则　$\min J_P(\boldsymbol{\alpha}) = \sum_{\mathbf{y}_j \in \mathcal{Y}^k} (-\boldsymbol{\alpha}^T \mathbf{y}_j)$
  - 我们需要告诉它怎样学
    - → 学习/训练算法　$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) - \rho_k \nabla J = \boldsymbol{\alpha}(k) + \rho_k \sum_{\mathbf{y}_j \in \mathrm{Y}^k} \mathbf{y}_j$

清華大學

机器学习的基本要素：线性回归版

- 怎样造一个学习机器？
  - 它需要老师　　直接用y，连续版本
    - → 我们设计它（特征和模型）　$f(\boldsymbol{x}) = \sum_{i=0}^{d} w_i x_i = \boldsymbol{w}^T \boldsymbol{x}$
  - 它需要训练/学习材料
    - → 训练数据　$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\},\ \boldsymbol{x}_j \in R^{d+1}, y_j \in R$
  - 我们需要为它树立学习的目标
    - → 目标函数、学习准则　均方误差
  - 我们需要告诉它怎样学
    - → 学习/训练算法　？　最小二乘法



机器学习的基本要素：线性回归版

- 怎样造一个学习机器？
  - 它需要老师
    → 我们设计它（特征和模型）并训练它 $f(\boldsymbol{x}) = \sum_{i=0}^{d} w_i x_i = \boldsymbol{w}^T\boldsymbol{x}$
  - 它需要训练/学习材料
    → 训练数据 $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)\}$, $\boldsymbol{x}_j \in R^{d+1}, y_j \in R$
  - 我们需要为它树立学习的目标
    → 目标函数、学习准则 $\min E = \frac{1}{N}\sum_{j=1}^{N}\left(f(\boldsymbol{x}_j) - y_j\right)^2$
  - 我们需要告诉它怎样学
    → 学习/训练算法 ？

Xuegong Zhang 65

线性回归算法

$$\min_{\boldsymbol{w}} E(\boldsymbol{w}) = \frac{1}{N}\sum_{j=1}^{N}\left(f(\boldsymbol{x}_j) - y_j\right)^2 = \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = \frac{1}{N}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

其中 $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix}$, $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$.

Xuegong Zhang 66

线性回归算法

$$\min_{\boldsymbol{w}} E(\boldsymbol{w}) = \frac{1}{N}\sum_{j=1}^{N}\left(f(\boldsymbol{x}_j) - y_j\right)^2 = \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = \frac{1}{N}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

其中 $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$.

解：

令 $\nabla E(\boldsymbol{w}) = \frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{2}{N}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = 0,$

有 $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$.

若$(\boldsymbol{X}^T\boldsymbol{X})$可逆，则 $\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

其中，$\boldsymbol{X}^+ = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$也称作伪逆。

$$\Big[[\quad]\,[\quad]\Big]^{-1}[\quad]$$

dim: $(d+1)\times N$ $N\times(d+1)$ $(d+1)\times N$

Xuegong Zhang 67

Linear regression algorithm:

1: Construct the matrix X and the vector $\mathbf{y}$ from the data set $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$, where each $\mathbf{x}$ includes the $x_0 = 1$ bias coordinate, as follows

$$\mathrm{X} = \underbrace{\begin{bmatrix} —\mathbf{x}_1^{\mathrm{T}}— \\ —\mathbf{x}_2^{\mathrm{T}}— \\ \vdots \\ —\mathbf{x}_N^{\mathrm{T}}— \end{bmatrix}}_{\text{input data matrix}}, \qquad \mathbf{y} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\text{target vector}}.$$

2: Compute the pseudo-inverse $\mathrm{X}^\dagger$ of the matrix X. If $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible,

$$\mathrm{X}^\dagger = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}.$$

3: Return $\mathbf{w}_{\mathrm{lin}} = \mathrm{X}^\dagger\mathbf{y}$.

Abu-Mostafa, Magdon-Ismail, Lin, *Learning from Data*

---- Ordinary least squares (OLS) algorithm 最小二乘法

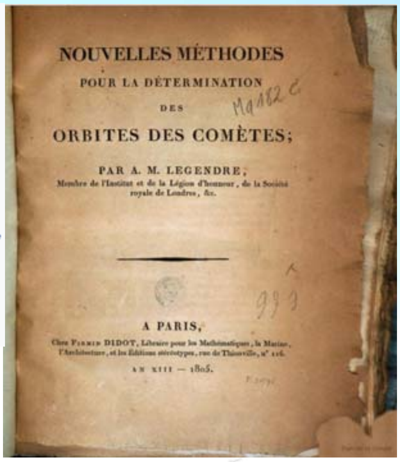
Xuegong Zhang 68

A.M. Legendre. Nouvelles méthodes pour la détermination des orbites des comètes, Firmin Didot, Paris, 1805. “Sur la Méthode des moindres quarrés” appears as an appendix.

Adrien-Marie Legendre

1820 watercolor caricature of Adrien-Marie Legendre by French artist Julien-Leopold Boilly (see portrait debacle), the only existing portrait known[1]

Born	18 September 1752 Paris, France
Died	10 January 1833 (aged 80) Paris, France



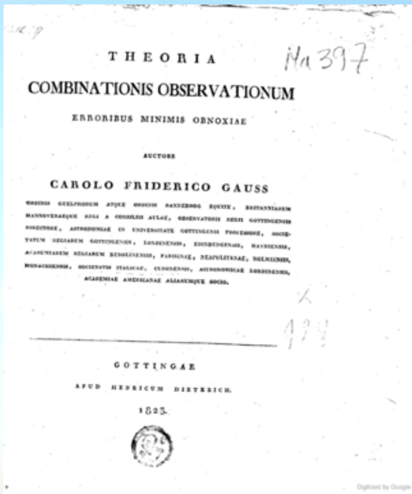
NOUVELLES MÉTHODES
POUR LA DÉTERMINATION
DES
ORBITES DES COMÈTES;
PAR A. M. LEGENDRE,
Membre de l'Institut et de la Légion d'honneur, de la Société royale de Londres, &c.
A PARIS,
Chez Firmin Didot, Libraire pour les Mathématiques, la Marine, l'Architecture, et les Éditions stéréotypes, rue de Thionville, n° 116.
AN XIII — 1805.

C.F. Gauss. Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum. (1809)

Carl Friedrich Gauss

Carl Friedrich Gauß (1777–1855), painted by Christian Albrecht Jensen

Born	Johann Carl Friedrich Gauss 30 April 1777 Brunswick, Principality of Brunswick-Wolfenbüttel
Died	23 February 1855 (aged 77) Göttingen, Kingdom of Hanover, German Confederation



THEORIA
COMBINATIONIS OBSERVATIONUM
ERRORIBUS MINIMIS OBNOXIAE
AUCTORE
CAROLO FRIDERICO GAUSS
GOTTINGAE
APUD HENRICUM DIETERICH.
1823.

\- Wikipedia

- 完美！
- 不过，这算机器学习吗？

请在雨课堂中弹幕回答

- 这解析解也算是“机器学习”？

- 完美！
- 不过，这算机器学习吗？

- 这解析解也算是“机器学习”？
  - 最小二乘法最早由Legendre在1805和Gauss在1809发明，那时远远没有机器学习的概念
  - 如FLD一样，既然算法能从数据中算出可用于预测的规律，通过解析公式算出来人们也倾向于算它是“机器学习”
  - 线性回归也可以迭代求解，也可以做成感知器那样学习机器（见下节课）

算法本身有没有问题？

请在雨课堂中弹幕提问

Linear regression algorithm:

1: Construct the matrix X and the vector $\mathbf{y}$ from the data set $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$, where each $\mathbf{x}$ includes the $x_0 = 1$ bias coordinate, as follows

$$\mathrm{X} = \underbrace{\begin{bmatrix} -\mathbf{x}_1^{\mathrm{T}}- \\ -\mathbf{x}_2^{\mathrm{T}}- \\ \vdots \\ -\mathbf{x}_N^{\mathrm{T}}- \end{bmatrix}}_{\text{input data matrix}}, \qquad \mathbf{y} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\text{target vector}}.$$

2: Compute the pseudo-inverse $\mathrm{X}^\dagger$ of the matrix X. If $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible,

$$\mathrm{X}^\dagger = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}.$$

3: Return $\mathbf{w}_{\text{lin}} = \mathrm{X}^\dagger \mathbf{y}$.

算法本身有没有问题？

- 若$(\boldsymbol{X}^T\boldsymbol{X})$可逆，则 $\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$；那如果不可逆呢？
- 可逆（非奇异，非退化，满秩）

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} & & \end{bmatrix}_{N\times(d+1)}$$

- $(\boldsymbol{X}^T\boldsymbol{X})$可逆：$\boldsymbol{X}$ 列满秩：特征间线性独立
  - 当$N \gg d+1$时通常成立
- 当特征不是线性独立时，仍然可以计算伪逆，但解不唯一
- 解决方案：
  – 通过特征选择或变换去除冗余
  – 通过引入其他准则对解加以限制(如SVD或正则化)

作业

- 第一章　该题不计分
  1. 调研与分析一些系统中的模式识别和机器学习问题
     （注意是分析问题而不是调研具体方法）
- 第二章
  1. 线性判别函数中基本几何关系
  2. Fisher线性判别的证明
  3. 【选做】感知器算法的收敛性证明
  4. 【选做】线性回归与皮尔森相关系数
- 截止日期：2月24日


清华大学
下课，下周见！
Xuegong Zhang
75