数据科学基础

从认知诊断到知识追踪

汇报人：童世炜

时间：2019.10.24

复习

从人工建模到数据智能

$$f(x)$$

人工设计

自主学习

复习

从数据到规律

- 针对一类问题的特征，如何设计合适的学习算法
- 如何使用数据来有效地获取函数参数具体值

$$D \mapsto f$$

数据　　　　规律

- 数据类型多样，包括各种格式和形态的数据
- 数据量大，但存在噪声，有效数据少
- 数据处理要求高实时性

- 如何验证所得规律
- 如何保证所得规律的正确性

01

背景

Background

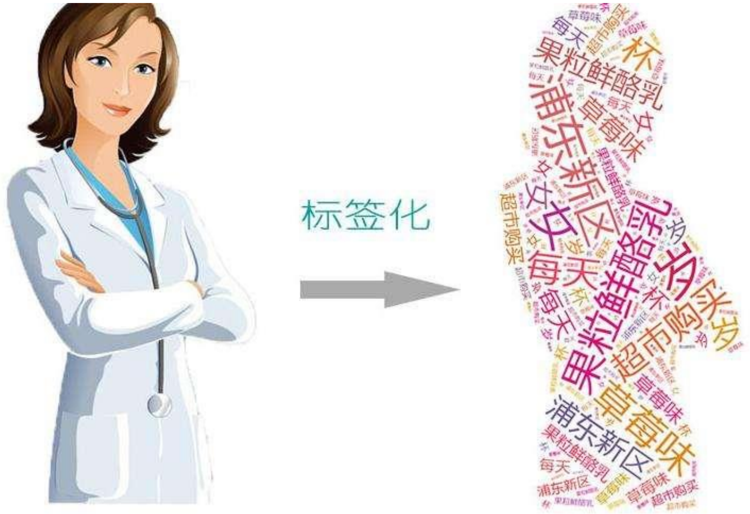
什么是认知诊断和知识追踪
它们有什么应用

PART ONE

用户画像

User Profiling

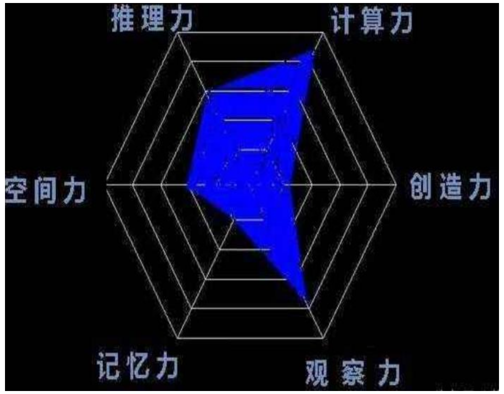
- 勾画目标用户，是连接用户需求和设计方向的有效工具
- 用户的每一个特定信息都被抽象成标签，用来描述用户的行为和偏好，从而提供有针对性的服务。

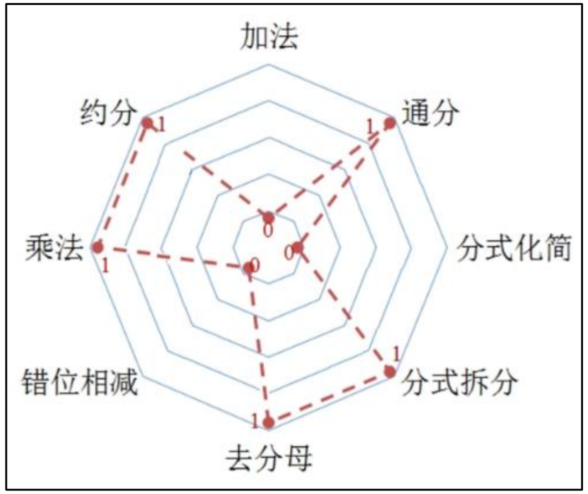


应用

Application

- 涉及游戏，运动，智慧教育等领域
- 一个更具体的应用：诊断/描述参与者对特定技能/概念的熟练程度
- 教育领域：认知诊断
  - 诊断/描述学生对特定问题/概念/知识点的熟练程度（知识状态）





认知诊断

Cognitive Diagnosis

- 输入
  - 学生的练习交互矩阵 R （响应矩阵）
    - R_{ij} 表示学生 i 在习题 j 上的得分
  - 习题 - 概念/知识点关联矩阵（Q 矩阵）
    - $Q_{jk}=1$ 表示习题 j 考查了概念/知识点 k
- 输出
  - 学生在每个概念/知识点/习题上的熟练度（知识状态）
    - 取值范围 $[0,1]$

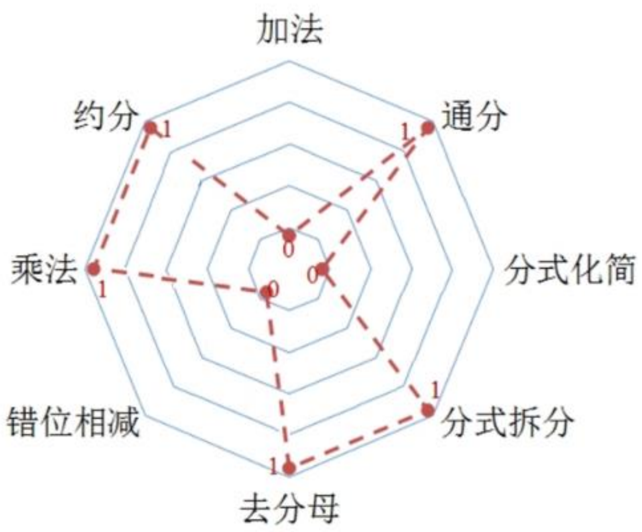
响应矩阵 R

1	0	1	2	3	4
0	1	0	0	5	3
0	1	0	1	6	5

Q矩阵

	一次函数	函数求导	线性规划
试题1	1	0	1
试题2	1	1	0
试题3	0	1	0
试题4	0	0	1
……			

认知诊断



传统的认知诊断模型

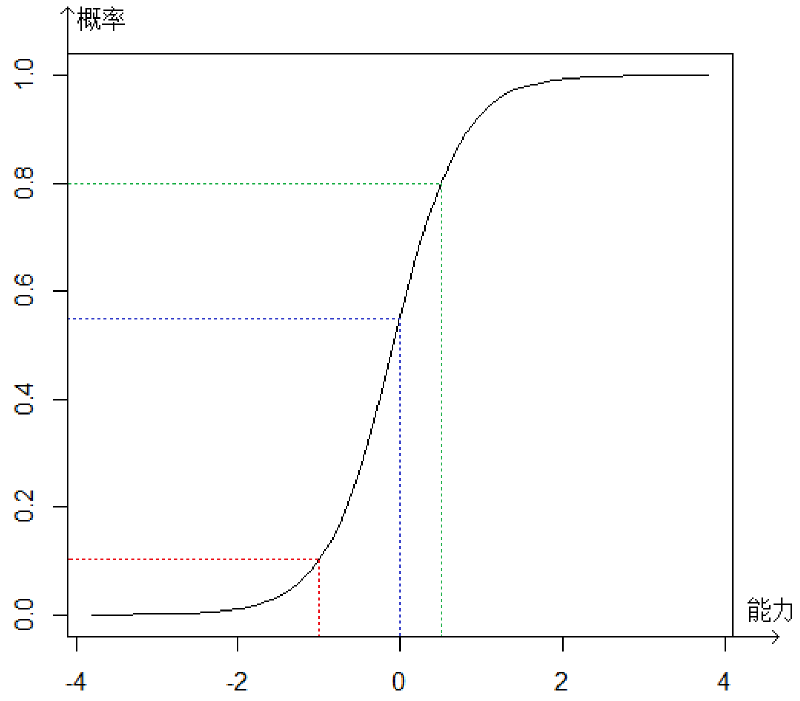
Traditional CDM

- Item Response Theory (IRT)

  - $P(R_{ij}=1|\theta_{ij},a_j,b_j,c_j)=c_j+\frac{1-c_j}{1+\exp(-1.7\cdot a_j(\theta_i-b_j))}$

    - θ_{ij} 指的是学生 i 对技能 j 的掌握度

    - a_j,b_j,c_j 对应试题的区分度，难度，猜测度

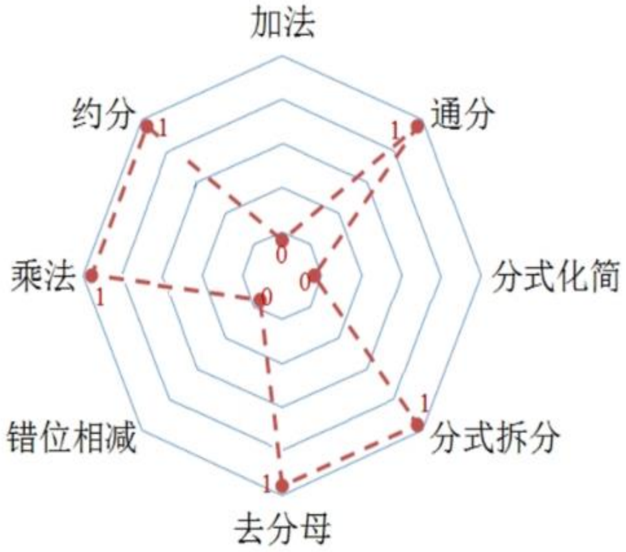


- DINA

  - $P(R_{ij}=1|\boldsymbol{\theta}_i)=g_j^{\eta_{ij}}(1-s_j)^{1-\eta_{ij}}$

    - $\eta_{ij}=\prod_K \theta_{ik}$

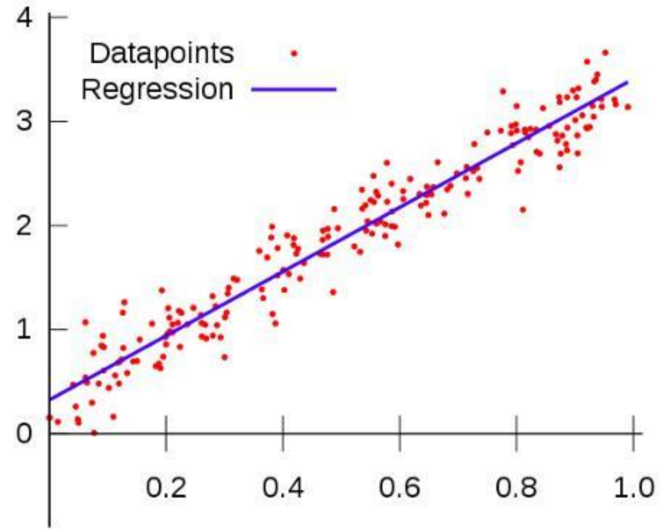
    - g_j 是猜测率，s_j 是失误率



知识课堂

线性回归

基本形式



- 线性模型一般形式
  - $f(x) = w_1 x_1 + w_2 x_x + \cdots + w_d x_d + d$
  - $\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$ 是由属性描述的示例（sample），其中 x_i 是 $\boldsymbol{x}$ 在第 i 个属性上的取值
    - 例如 IRT 中的能力、区分度、难度
- 向量形式
  - $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$
  - 其中 $\boldsymbol{w} = (w_1; w_2; \ldots; w_d)$

线性回归

Linear Regression

- 给定数据集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\dots,(\boldsymbol{x}_m,y_m)\}$
  - 其中 $\boldsymbol{x}_i=(x_{i1};x_{i2};\dots;x_{id}),y_i\in R$
- 线性回归（linear regression）目的
  - 学得一个线性模型以尽可能准确地预测实值输出标记
  - Eg. IRT 学得一个线性模型以尽可能准确地预测学生答对某答题的概率/某道题的得分
- 非数值/离散属性处理
  - 有“序”关系
    - 连续化为连续值
    - Eg. 难度（简单、中等、难）→ (简单, 0), (中等，1), (难, 2)
  - 无“序”关系
    - 有 k 个属性值，则转换为 k 维向量
    - Eg. 课程类别（生物、地理、数学）→（生物，001），（地理，010），（数学，100）
      - 这种编码方式被称为“独热码”（one-hot encoding）
      - 思考题：如果某道题、某个知识点涉及交叉学科，有多个课程类别，如何表示？

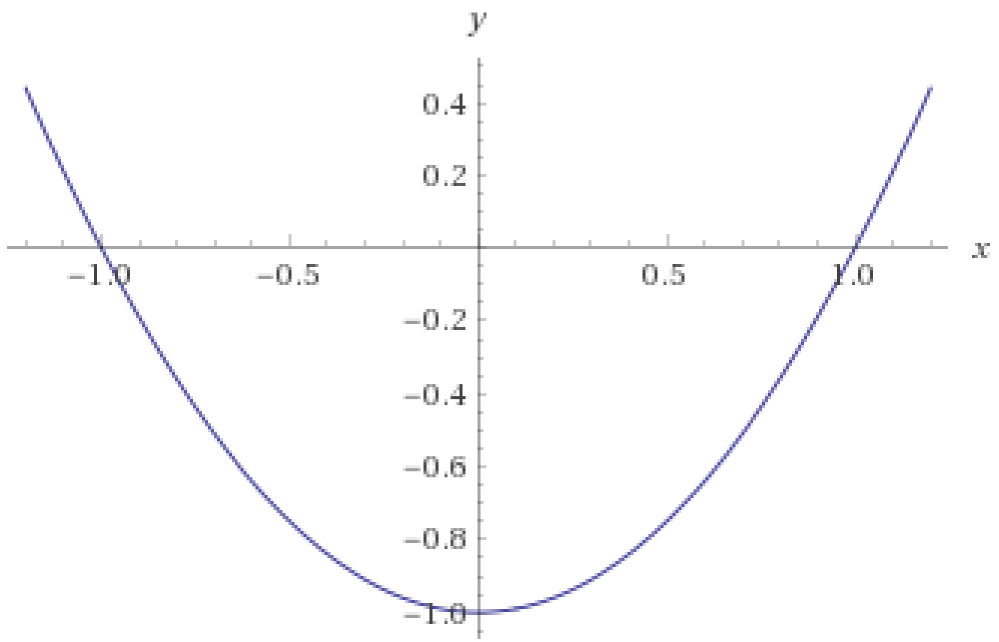
线性回归求解

- 单一属性的线性回归目标
  - $f(x) = wx_i + b$ 使得 $f(x_i) \approx y_i$
  - 使得预测值逼近于真实值
- 求解参数
  - 目标：求得 w, b
  - $(w^*, b^*) = \underset{(w,b)}{argmin} \sum_{i=1}^{m} (f(x_i) - y_i)^2 = \underset{(w,b)}{argmin} \sum_{i=1}^{m} (y_i - wx_i - b)^2$

最小二乘法

Least Square Method

- 最小化均方误差
  - $E_{(w,b)} = \sum_{i=1}^{m}(y_i - wx_i - b)^2$
- 极值点
  - 导数为 0 处为函数极值点
  - 二次函数的极值点即为最值点
- 分别对 w 和 b 求导
  - $\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)x_i\right)$
  - $\frac{\partial E_{(w,b)}}{\partial b} = 2(mb - \sum_{i=1}^{m}(y_i - wx_i))$



最小二乘法

- 令导数为 0，得解
- $\frac{\partial E_{(w,b)}}{\partial b} = 2(mb - \sum_{i=1}^{m}(y_i - wx_i)) = 0$
  - $b = \frac{\sum_{i=1}^{m}(y_i - wx_i)}{m}$
- $\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)x_i\right) = 0$
  - 代入 b 得 $2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}\left(y_i - \frac{\sum_{j=1}^{m}(y_j - wx_j)}{m}\right)x_i\right) = 0$
  - 即 $2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i-1}^{m} y_i x_i + \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m} y_j x_i - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m} w x_j x_i\right) = 0$
  - $w = \frac{\sum_{i-1}^{m} y_i x_i - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m} y_j x_i}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m} w x_j x_i} = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2}$
    - $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$

多元线性回归求解

- 给定数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_m, y_m)\}$
  - 其中 $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \dots; x_{id}), y_i \in R$
- 单一属性的线性回归目标
  - $f(x) = wx_i + b$ 使得 $f(x_i) \approx y_i$
- 多元属性的线性回归目标
  - $f(\boldsymbol{x}_i) = \boldsymbol{w}^T \boldsymbol{x}_i + b$ 使得 $f(\boldsymbol{x}_i) \approx y_i$

带偏置的数据形式

- 令 $\hat{\boldsymbol{w}} = (\boldsymbol{w}; b)$
- 改写数据集表示形式

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1 \end{pmatrix}$$

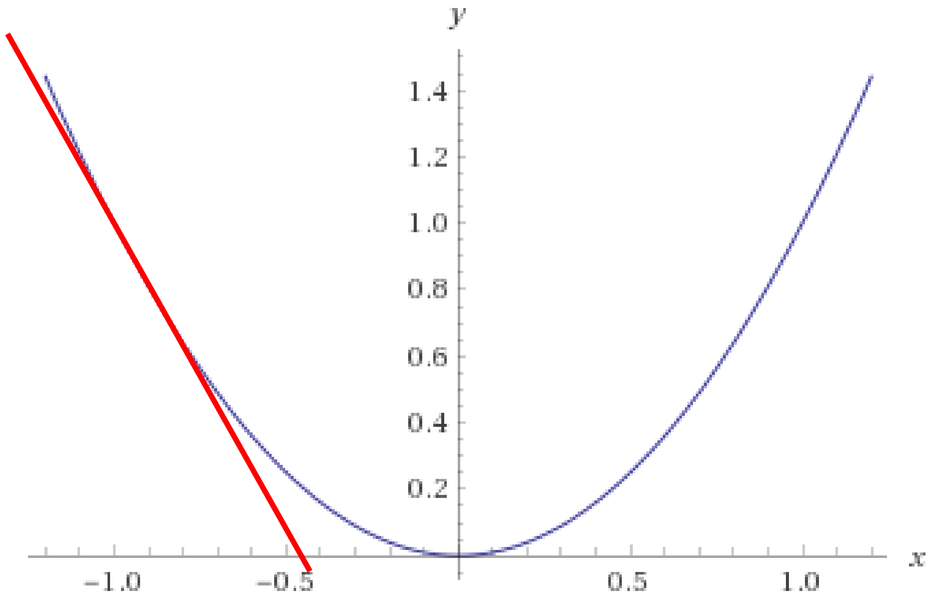
$$\boldsymbol{y} = (y_1; y_2; \ldots; y_m)$$

多元回归的最小二乘法

- 单一属性的最小二乘法
  - $(w^*, b^*) = \underset{(w,b)}{argmin} \sum_{i=1}^{m}(f(x_i) - y_i)^2 = \underset{(w,b)}{argmin} \sum_{i=1}^{m}(y_i - wx_i - b)^2$
- 多元属性的最小二乘法
  - $\hat{\boldsymbol{w}}^* = \underset{\hat{\boldsymbol{w}}}{argmin}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})$
  - $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})$
  - 对 $\hat{\boldsymbol{w}}$ 求导，并令导数为 0
    - $\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\boldsymbol{X}^T(\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})$

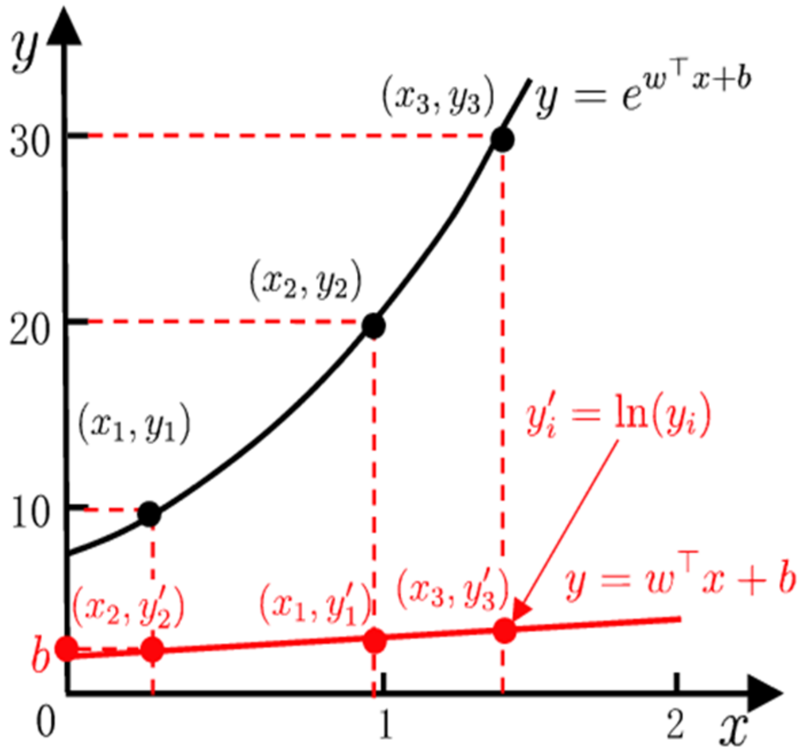
线性回归的求解方法

- 矩阵的逆与伪逆
  - $2\boldsymbol{X}^T(\boldsymbol{X}\widehat{\boldsymbol{w}}-\boldsymbol{y})=\boldsymbol{0}$ 等价于 $2\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{w}}=2\boldsymbol{X}^T\boldsymbol{y}\rightarrow\widehat{\boldsymbol{w}}^*=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
    - 条件：$\boldsymbol{X}^T\boldsymbol{X}$ 为满秩或正定矩阵 → 可求逆矩阵
    - 不满足情况下：伪逆法
- 梯度下降法与随机梯度下降法
  - 梯度的反方向是函数值减小最快的方向
    - Eg. $f(x)=x^2, f'(x)=2x,$
    - $-f'(-1)=2,$ 沿正轴指向 0
  - 梯度下降法
    - $\widehat{w}\leftarrow\widehat{w}-\frac{\partial E_{\widehat{w}}}{\partial\widehat{w}}\cdot\boldsymbol{\Delta}_{step}$
  - 随机梯度下降
    - 每次参数更新时，采用小批量方式替代全部样本
    - $\frac{\partial E_{\widehat{w}}^B}{\partial\widehat{w}}=2\boldsymbol{X}_B^T(\boldsymbol{X}_B\widehat{\boldsymbol{w}}-\boldsymbol{y}_B)$
    - 优点：计算量小 $m\rightarrow BatchSize$，速度快



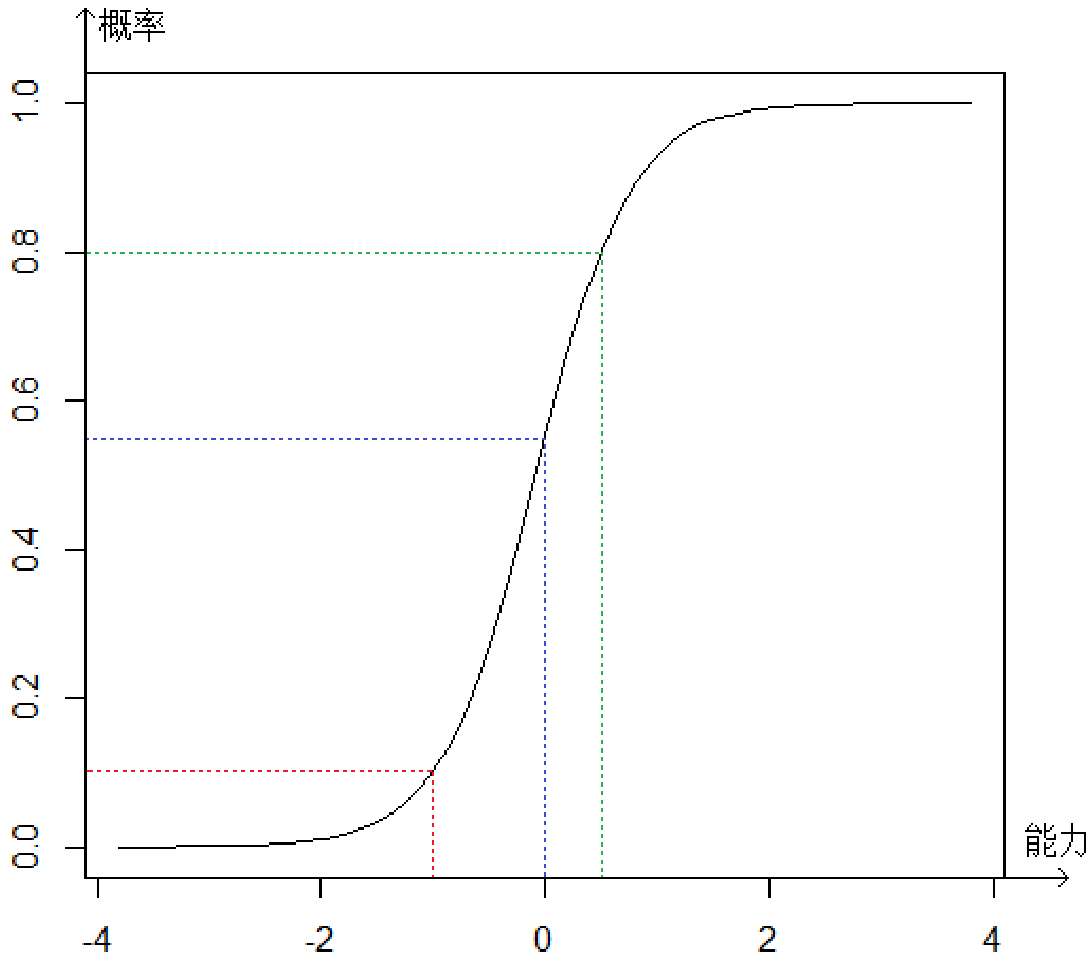
广义线性模型

- 一般线性回归仅能对纯线性关系建模
  - 现实生活中很多关系并不是线性的
- 广义线性模型
  - 一般形式 $y = g^{-1}(\boldsymbol{w}^T\boldsymbol{x} + b)$
    - 等价于 $g(y) = \boldsymbol{w}^T\boldsymbol{x} + b$
  - $g(\cdot)$ 称为联系函数，单调可微
  - 对数线性回归
    - $\ln y = \boldsymbol{w}^T\boldsymbol{x} + b$



从广义线性模型到二分类

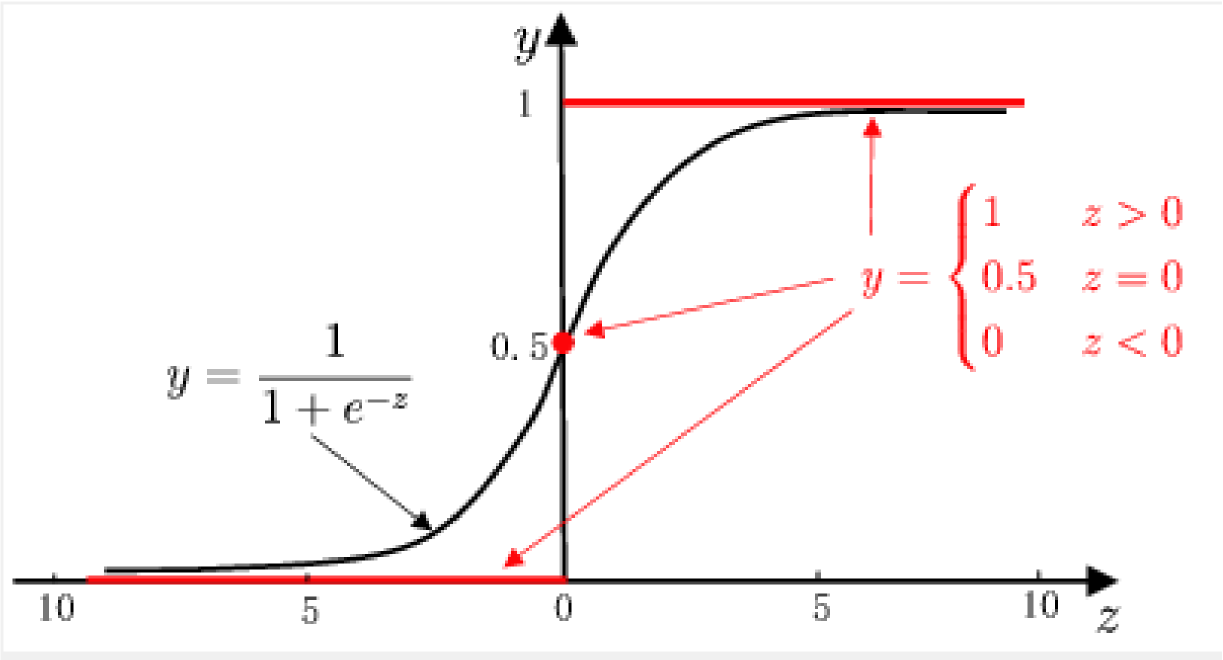
- 认知诊断
  - 答对或答错（0-1 问题）
  - 如何将线性模型改造成二分类模型
- 广义线性模型
  - $y = g^{-1}(\boldsymbol{w}^T\boldsymbol{x} + b)$
- 二分类模型
  - $z = \boldsymbol{w}^T\boldsymbol{x} + b$
  - $y = \sigma(z)$



逻辑斯蒂回归

Logistics Regression

- 单位阶跃函数 vs 对数几率函数
- 单位阶跃函数
  - $$\sigma(z) = \begin{cases} 0, & z < 0 \\ 0.5, & z = 0 \\ 1, & z > 0 \end{cases}$$
  - 缺陷：不连续
- 对数几率函数 (logistics function)
  - $\sigma(z) = \frac{1}{1+e^{-z}}$
  - 单调可微，任意阶可导

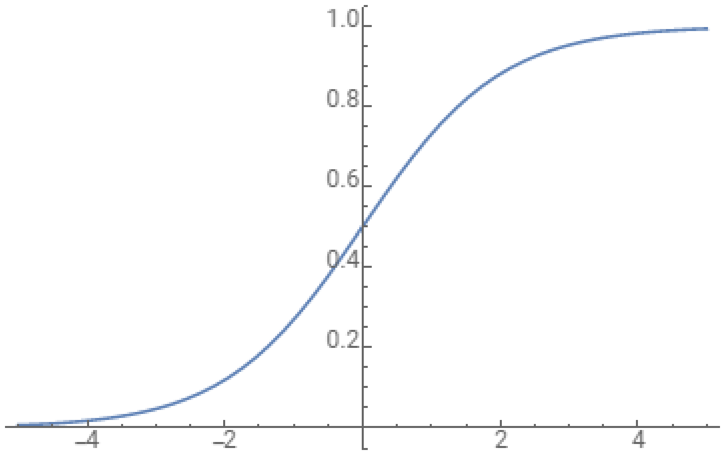


单位阶跃函数与对数几率函数

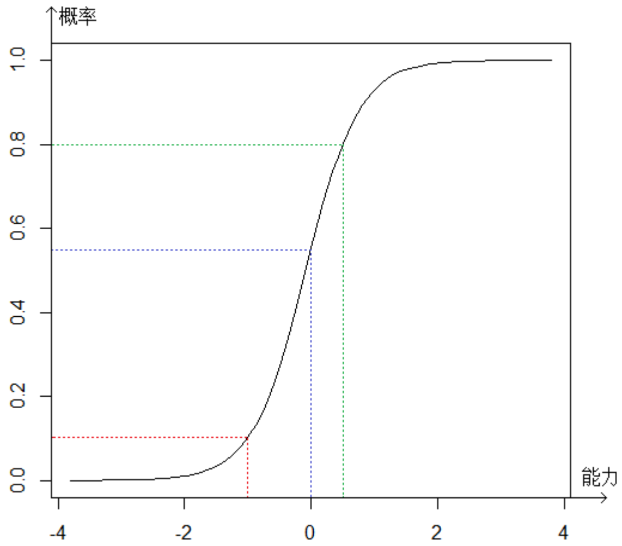
逻辑斯蒂回归

对数几率回归

- 逻辑斯蒂回归
  - $y = \sigma(z)$
  - $\sigma(z) = \frac{1}{1+e^{-z}}$
  - $z(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$
  - $y = \frac{1}{1+e^{-(w^T x+b)}} = \frac{e^{w^T x+b}}{1+e^{w^T x+b}}$
- IRT
  - $P(R_{ij} = 1|\theta_{ij}, a_j, b_j, c_j) = c_j + \frac{1-c_j}{1+\exp\left(-1.7 \cdot a_j(\theta_i-b_j)\right)}$



对数几率函数



IRT

逻辑斯蒂回归求解

对数几率回归

- 正例：$y = \frac{e^{w^T x+b}}{1+e^{w^T x+b}}$

- 负例： $1 - y = \frac{1}{1+e^{w^T x+b}}$

- 对数几率

  - $\ln \frac{P(y=1|x)}{P(y=0|x)} = \ln \frac{y}{1-y} = w^T x + b \rightarrow g(y) = \boldsymbol{w}^T \boldsymbol{x} + b$

  - 对数几率表示 $\boldsymbol{x}$ 为正例的相对可能性

逻辑斯蒂回归求解

对数几率回归

- 正例：$y = \frac{e^{w^T x+b}}{1+e^{w^T x+b}}$
- 负例： $1 - y = \frac{1}{1+e^{w^T x+b}}$
- 对数几率
  - $\ln \frac{P(y=1|x)}{P(y=0|x)} = \ln \frac{y}{1-y} = \boldsymbol{w}^T \boldsymbol{x} + b \rightarrow \boxed{g(y) = \boldsymbol{w}^T \boldsymbol{x} + b}$

    广义线性模型
  - 对数几率表示 $\boldsymbol{x}$ 为正例的相对可能性

极大似然估计

Maximum Likelihood

- 正例：$y = \frac{e^{w^T x+b}}{1+e^{w^T x+b}}$
- 负例： $1 - y = \frac{1}{1+e^{w^T x+b}}$
- 对数几率
  - $\ln \frac{P(y=1|\boldsymbol{x})}{P(y=0|\boldsymbol{x})} = \ln \frac{y}{1-y} = \boldsymbol{w}^T \boldsymbol{x} + b$
    - $P(y = 1|\boldsymbol{x}) = \frac{e^{w^T x+b}}{1+e^{w^T x+b}}$
    - $P(y = 0|\boldsymbol{x}) = \frac{1}{1+e^{w^T x+b}}$
  - 对数几率表示 $\boldsymbol{x}$ 为正例的相对可能性

极大似然估计

Maximum Likelihood

- 数据集
  - $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$
- 求解目标
  - 最大化样本属于其真实标记的概率
    - 如果样本标记是 1，那么 $P(y=1|\boldsymbol{x}) = \frac{e^{\boldsymbol{w}^T\boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^T\boldsymbol{x}+b}}$ 越大越好
    - 反之，则 $P(y=0|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{w}^T\boldsymbol{x}+b}}$ 越大越好
  - 最大化对数似然函数
    - $l(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i|\boldsymbol{x}_i; \boldsymbol{w}_i, b)$

极大似然估计

Maximum Likelihood

- 数据集
  - $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$
- 求解目标
  - 最大化样本属于其真实标记的概率
    - 如果样本标记是 1，那么 $P(y=1|\boldsymbol{x}) = \frac{e^{\boldsymbol{w}^T x+b}}{1+e^{\boldsymbol{w}^T x+b}}$ 越大越好
    - 反之，则 $P(y=0|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{w}^T x+b}}$ 越大越好
  - 最大化对数似然函数
    - $l(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln \boxed{p(y_i|\boldsymbol{x}_i; \boldsymbol{w}_i, b)}$ 似然项

极大似然估计

- 令 $\beta=(\boldsymbol{w};b),\hat{\boldsymbol{x}}=(\boldsymbol{x};1)$, 则 $\boldsymbol{w}^T\boldsymbol{x}+b$ 可简写为 $\beta^T\hat{\boldsymbol{x}}$
- 记
  - $p_0(\hat{\boldsymbol{x}};\beta)=P(y=0|\hat{\boldsymbol{x}};\beta)=\frac{\mathbf{1}}{1+e^{\beta^T\hat{\boldsymbol{x}}}}$
  - $p_1(\hat{\boldsymbol{x}};\beta)=P(y=1|\hat{\boldsymbol{x}};\beta)=\frac{e^{\beta^T\hat{\boldsymbol{x}}}}{1+e^{\beta^T\hat{\boldsymbol{x}}}}$
- 则原似然项 $p(y_i|\boldsymbol{x_i};\boldsymbol{w}_i,b)$ 变为
  - $p(y_i|\boldsymbol{x_i};\boldsymbol{w}_i,b)=y_ip_1(\hat{\boldsymbol{x}};\beta)+(1-y_i)p_0(\hat{\boldsymbol{x}};\beta)$
    - 当 $y_i=1$ 时，$p(y_i|\boldsymbol{x_i};\boldsymbol{w}_i,b)=p_1(\hat{\boldsymbol{x}};\beta)$
    - 当 $y_i=0$ 时，$p(y_i|\boldsymbol{x_i};\boldsymbol{w}_i,b)=p_0(\hat{\boldsymbol{x}};\beta)$
- 最大化对数似然函数 $l(\boldsymbol{w},b)=\sum_{i=1}^{m}\ln p(y_i|\boldsymbol{x_i};\boldsymbol{w}_i,b)$ 等价于最小化
  - $l(\beta)=\sum_{i=1}^{m}\left(-y_i\beta^T\hat{\boldsymbol{x}}_i+\ln\left(1+e^{\beta^T\hat{\boldsymbol{x}}_i}\right)\right)$

线性模型优点

- 形式简单、易于建模
  - 单一属性线性回归
  - 多元属性线性回归
  - 广义线性模型
    - 对数线性回归
    - 逻辑斯蒂（logistics）回归
      - 可用于二分类
  - 线性模型的求解
    - 最小二乘法，极大似然估计
    - 优化方法：矩阵求逆、梯度下降与随机梯度下降
- 可解释性
  - IRT
- 非线性模型的基础
  - 引入层级结构或高维映射