机器学习与数据挖掘

Machine Learning & Data Mining

权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

Preface

Preface

50 Best Jobs in America for 2019

Preface

50 Best Jobs in America for 2019

Best Jobs | 2019 | United States | Share

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Data Scientist	$108,000	4.3/5	6,510	View Jobs
#2	Nursing Manager	$83,000	4/5	13,931	View Jobs
#3	Marketing Manager	$82,000	4.2/5	7,395	View Jobs
#4	Occupational Therapist	$74,000	4/5	17,701	View Jobs
#5	Product Manager	$115,000	3.8/5	11,884	View Jobs
#6	Devops Engineer	$106,000	4.1/5	4,657	View Jobs
#7	Program Manager	$87,000	3.9/5	14,753	View Jobs
#8	Data Engineer	$100,000	3.9/5	4,739	View Jobs
#9	HR Manager	$85,000	4.2/5	3,908	View Jobs
#10	Software Engineer	$104,000	3.6/5	49,007	View Jobs

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

Preface

数据科学家(Data Scientist)?

Preface

While having a strong coding ability is important, data science isn't all about software engineering.

Source: https://medium.com/cracking-the-data-science-interview/the-10-statistical-techniques-data-scientists-need-to-master-1ef6dbd531f7

Preface

Data scientists live at the intersection of coding, statistics, and critical thinking.

Source: https://medium.com/cracking-the-data-science-interview/the-10-statistical-techniques-data-scientists-need-to-master-1ef6dbd531f7

Preface

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."

Source: https://medium.com/cracking-the-data-science-interview/the-10-statistical-techniques-data-scientists-need-to-master-1ef6dbd531f7

Preface

The 10 Statistical Techniques Data Scientists Need to Master

Source: https://medium.com/cracking-the-data-science-interview/the-10-statistical-techniques-data-scientists-need-to-master-1ef6dbd531f7

Preface

1 — Linear Regression

2 — Classification

3 — Resampling Methods

4 — Subset Selection

5 — Shrinkage

6 — Dimension Reduction

7 — Nonlinear Models

8 — Tree-Based Methods

9 — Support Vector Machines

10 — Unsupervised Learning

Source: https://medium.com/cracking-the-data-science-interview/the-10-statistical-techniques-data-scientists-need-to-master-1ef6dbd531f7

Lecture 7: Linear Model

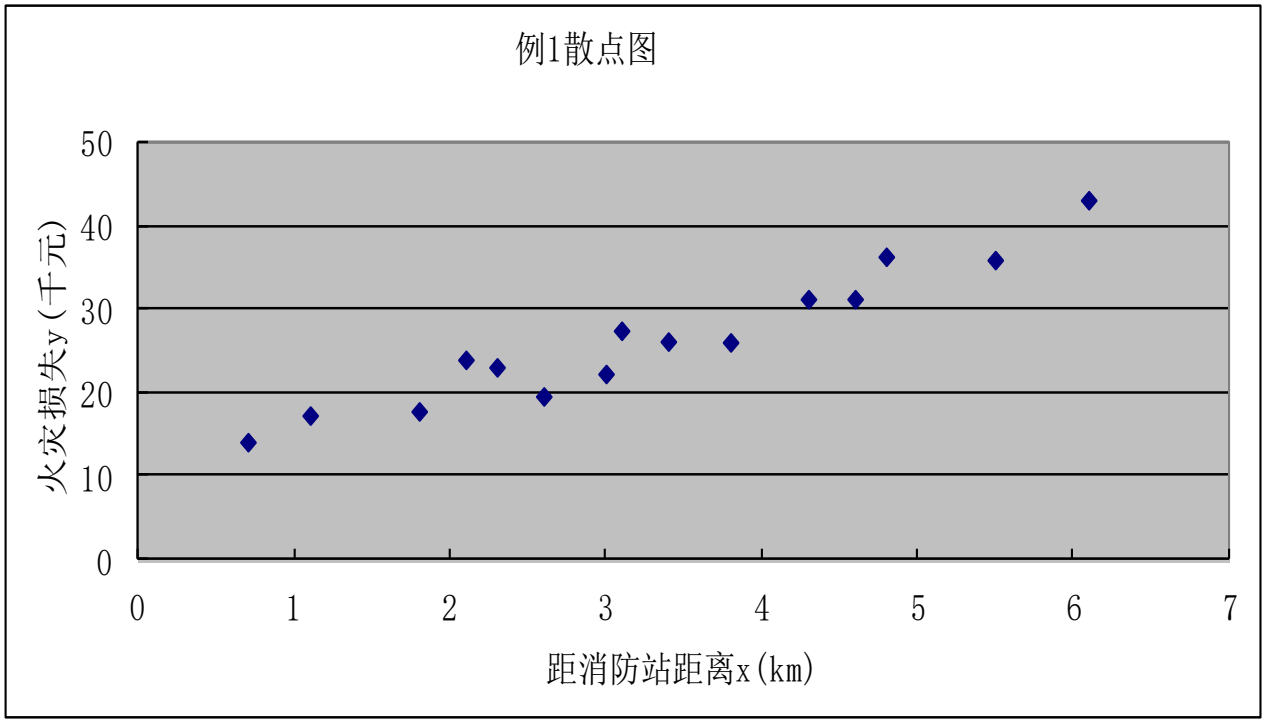
回归的概念

- **例1** 下表列出了15起火灾事故的损失及火灾发生地与最近的消防站的距离。

火灾损失表

距消防站离 x(km)	3．4	1．8	4．6	2．3	3．1	5．5	0．7	3．0
火灾损失 y(千元)	26.2	17.8	31.3	23.1	27.5	36.0	14.1	22.3
距消防站离 x(km)	2．6	4．3	2．1	1．1	6．1	4．8	3．8	
火灾损失 y(千元)	19.6	31.3	24.0	17.3	43.2	36.4	26.1	

回归的概念



结论：离消防站越近，发生火灾时损失越小

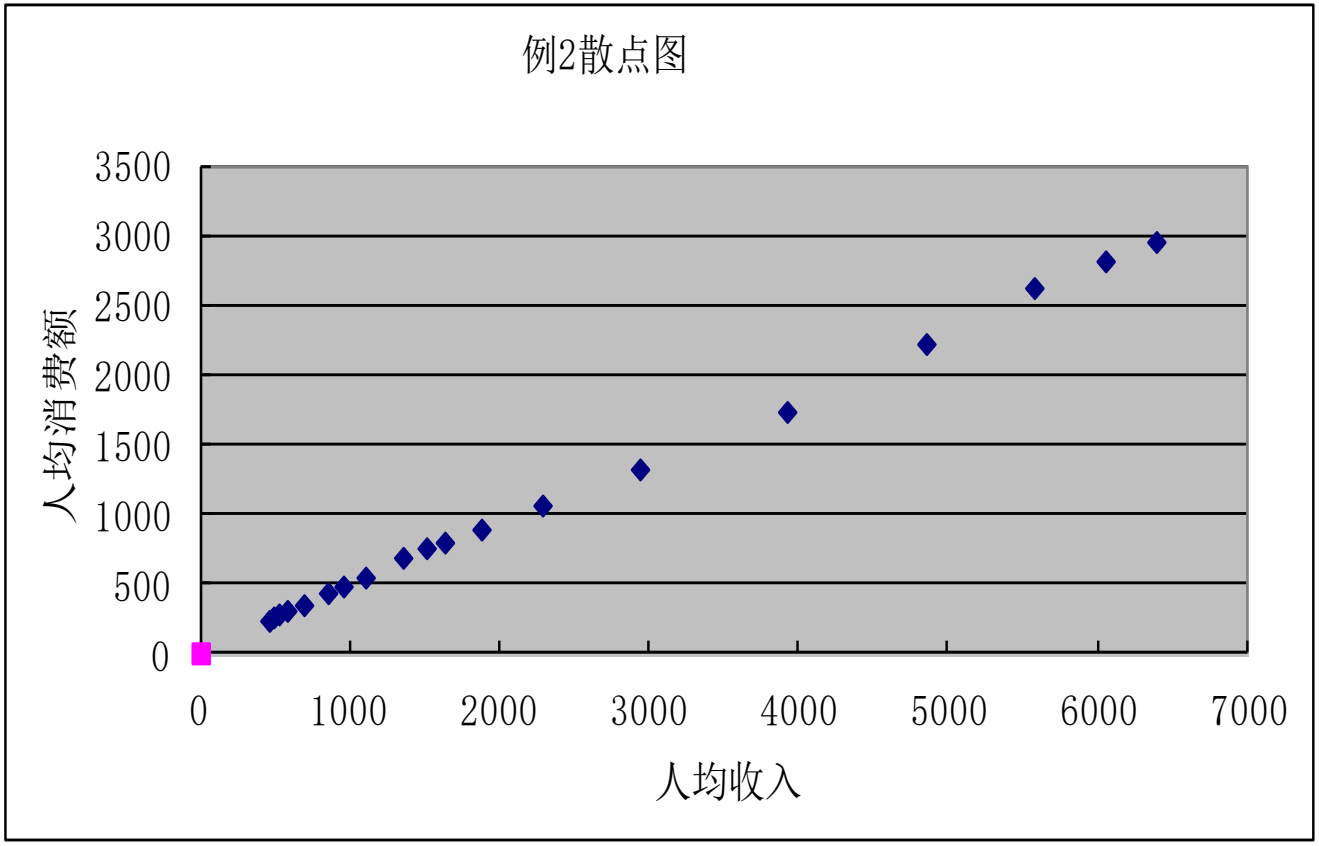
回归的概念

- **例2** 全国人均消费金额记作y(元)，人均国民收入记为x(元)

人均国民收入表

年份	人均国民收入（元）	人均消费金额（元）	年份	人均国民收入（元）	人均消费金额（元）
1980	460	234.75	1990	1634	797.08
1981	489	259.26	1991	1879	890.66
1982	525	280.58	1992	2287	1063.39
1983	580	305.97	1993	2939	1323.22
1984	692	347.15	1994	3923	1736.32
1985	853	433.53	1995	4854	2224.59
1986	956	481.36	1996	5576	2627.06
1987	1104	545.40	1997	6053	2819.36
1988	1355	687.51	1998	6392	2958.18
1989	1512	756.27			

回归的概念



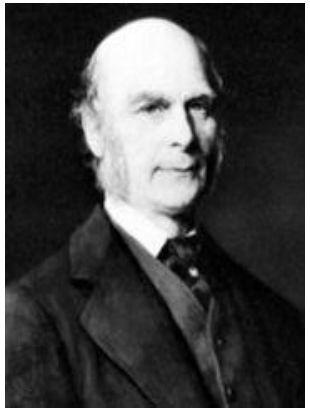
结论：人均消费和人均收入线性相关

回归的概念

问题：父母身高和子女身高之间的关系？

回归的概念

- 回归(*Regression*)这一概念最早由英国生物统计学家高尔顿和他的学生皮尔逊在研究父母亲和子女的身高遗传特性时提出
- 高个子父代的子代在成年之后的身高平均来说不是更高，而是稍矮于其父代水平，而矮个子父代的子代的平均身高不是更矮，而是稍高于其父代水平
- 高尔顿将这种趋向于种族稳定的现象称之“回归 ”

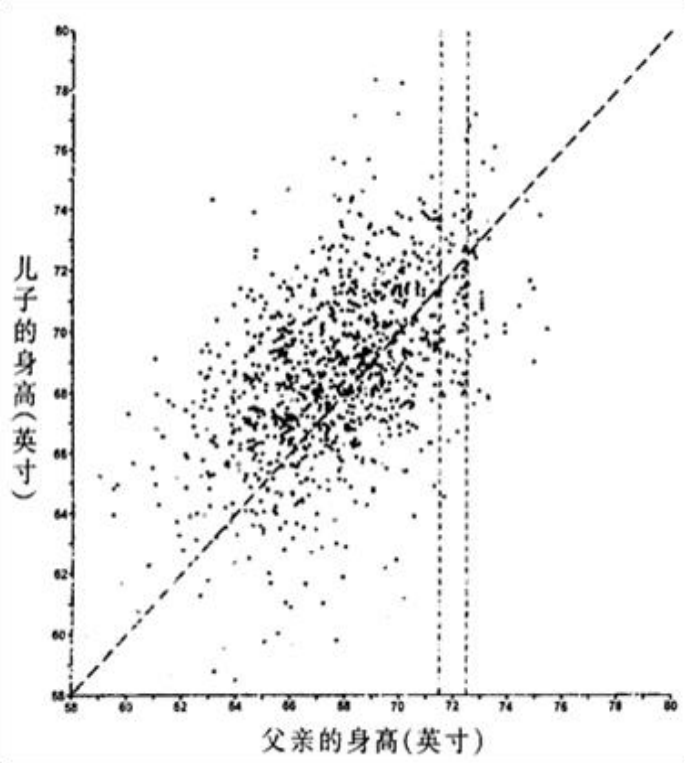


高尔顿

皮尔逊

回归的概念

高尔顿搜集了1078对父亲及其儿子的身高数据，他发现这些数据的散点图大致呈直线状态，也就是说总的趋势是父亲身高增加时，儿子的身高也倾向于增加

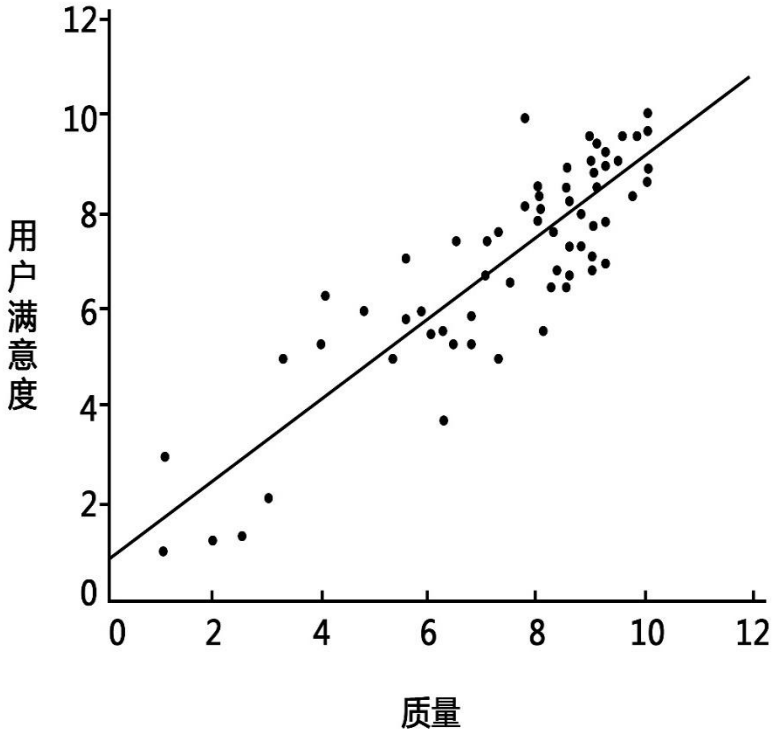


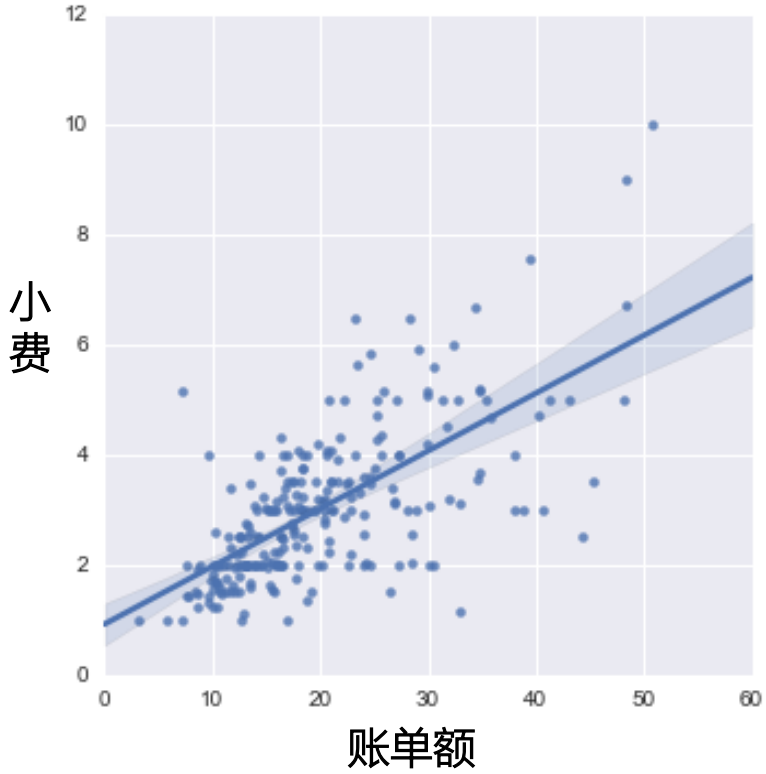
图中回归直线的线性拟合方程为（单位 英寸）：

$$y = 33.73 + 0.516\,x$$ （x为父亲身高，y为儿子身高）

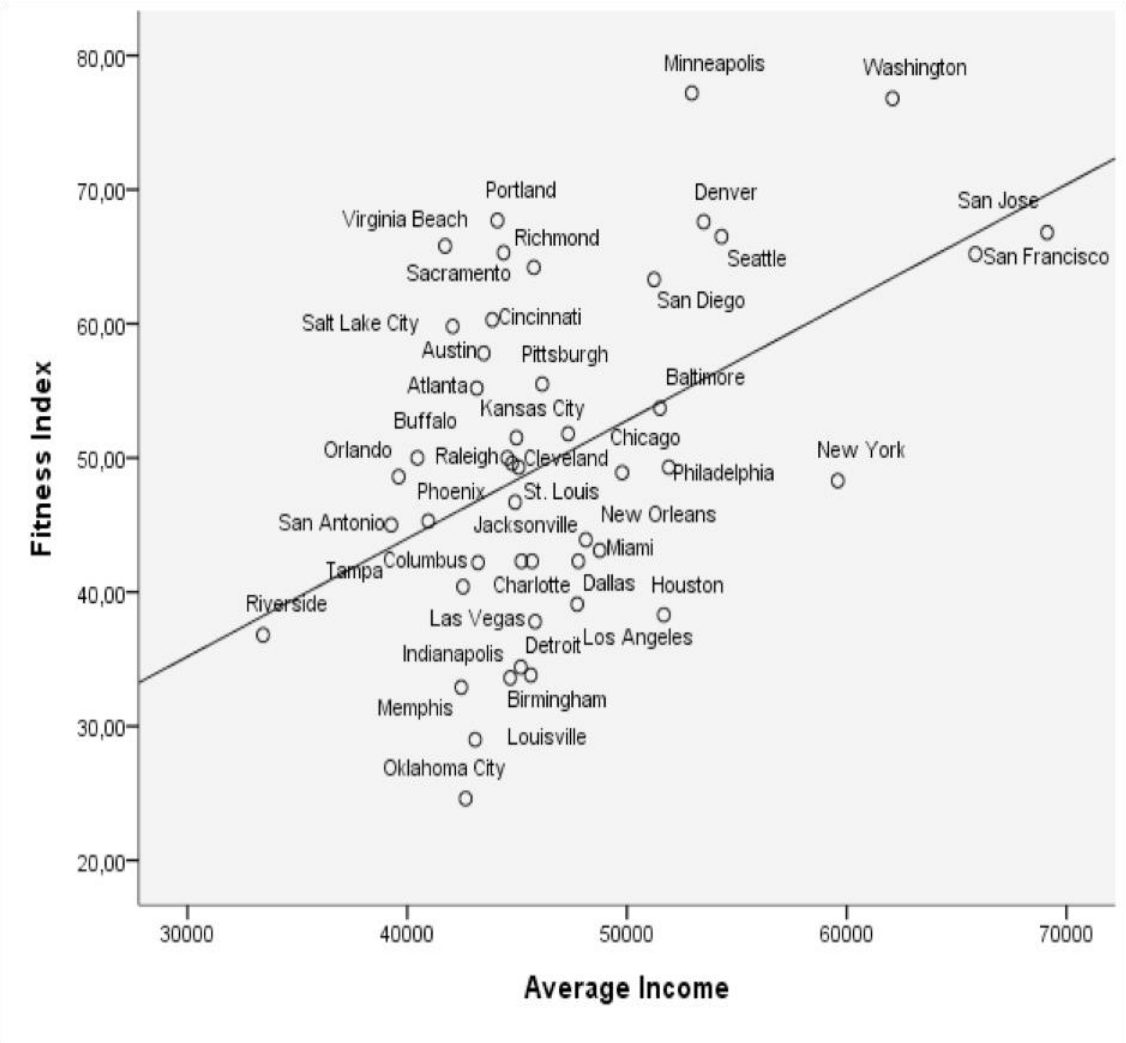
回归的概念：自变量

- 在一个回归模型中，我们需要关注或预测的变量叫做因变量（响应变量或结果变量），我们选取的用来解释因变量变化的变量叫做自变量（解释变量或预测变量）。

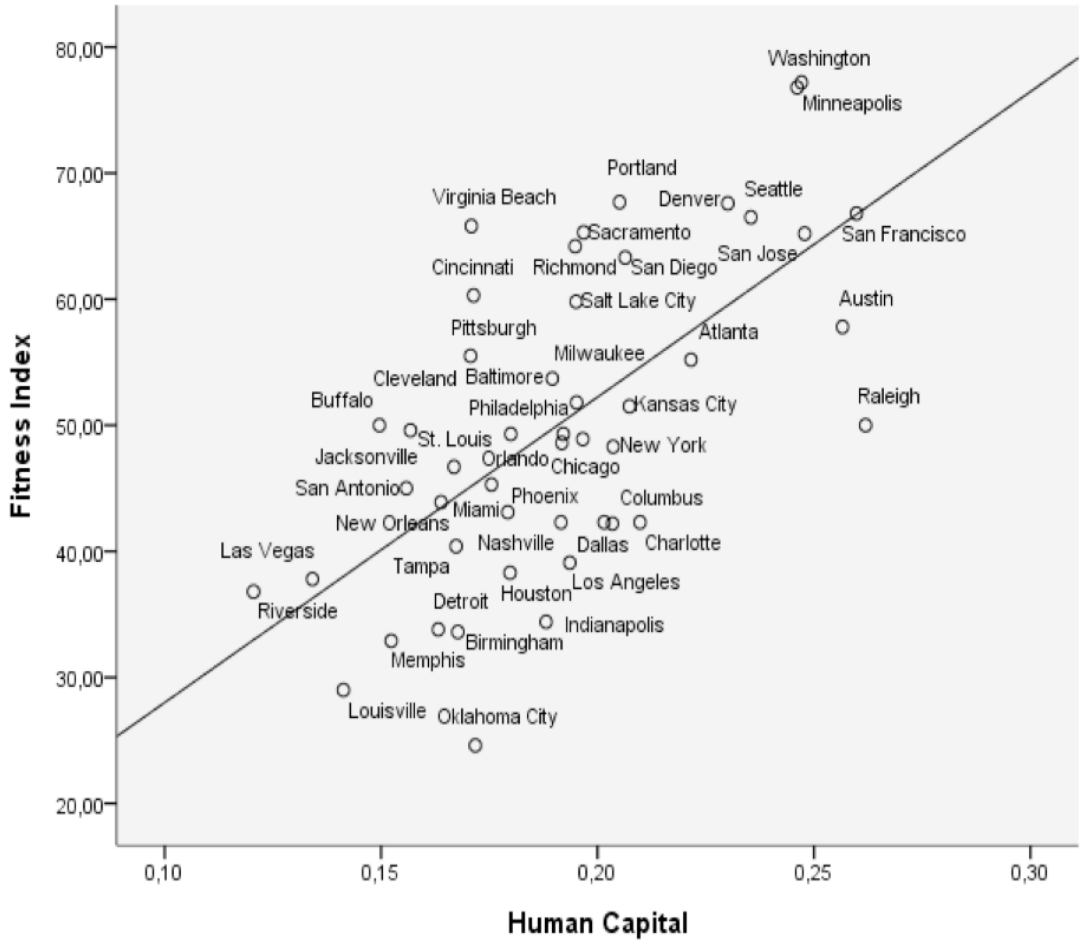




自变量和因变量：例一


Fitness Index
Average Income
80,00
70,00
60,00
50,00
40,00
30,00
20,00
30000
40000
50000
60000
70000
Minneapolis
Washington
Portland
Denver
San Jose
Virginia Beach
Richmond
Seattle
San Francisco
Sacramento
San Diego
Salt Lake City
Cincinnati
Austin
Pittsburgh
Baltimore
Atlanta
Kansas City
Buffalo
Chicago
Orlando
Raleigh
Cleveland
New York
Philadelphia
Phoenix
St. Louis
San Antonio
New Orleans
Jacksonville
Miami
Tampa
Columbus
Charlotte
Dallas
Houston
Riverside
Las Vegas
Detroit
Los Angeles
Indianapolis
Memphis
Birmingham
Louisville
Oklahoma City

自变量和因变量：例二


Fitness Index
Human Capital
80,00
70,00
60,00
50,00
40,00
30,00
20,00
0,10
0,15
0,20
0,25
0,30
Washington
Minneapolis
Portland
Virginia Beach
Denver
Seattle
Sacramento
San Francisco
San Jose
Cincinnati
Richmond
San Diego
Salt Lake City
Austin
Pittsburgh
Atlanta
Milwaukee
Cleveland
Baltimore
Buffalo
Philadelphia
Kansas City
Raleigh
St. Louis
New York
Jacksonville
Orlando
Chicago
San Antonio
Phoenix
Columbus
Miami
New Orleans
Nashville
Dallas
Charlotte
Las Vegas
Tampa
Los Angeles
Houston
Detroit
Riverside
Indianapolis
Birmingham
Memphis
Louisville
Oklahoma City

Lecture 7.1: 基本形式

基本形式

- 线性模型一般形式

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$ 是由属性描述的示例，其中 x_i 是 $\boldsymbol{x}$ 在第 i 个属性上的取值

- 向量形式 $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b$

其中 $\boldsymbol{w} = (w_1; w_2; \ldots; w_d)$

线性模型优点

- 形式简单、易于建模
- 可解释性
- 非线性模型的基础
- 一个例子
  - 综合考虑色泽、根蒂和敲声来判断西瓜好不好
  - 其中根蒂的系数最大，表明根蒂最要紧；而敲声的系数比色泽大，说明敲声比色泽更重要

$$f_{\text{好瓜}}(\boldsymbol{x}) = 0.2 \cdot x_{\text{色泽}} + 0.5 \cdot x_{\text{根蒂}} + 0.3 \cdot x_{\text{敲声}} + 1$$

线性回归

- 给定数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$

  其中 $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$ ， $y_i \in \mathbb{R}$

- 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

- 离散属性处理
  - 有“序”关系
    - 连续化为连续值
  - 无“序”关系
    - 有k个属性值，则转换为k维向量

线性回归

- 单一属性的线性回归目标

$$f(x) = wx_i + b \text{ 使得 } f(x_i) \simeq y_i$$

- 参数/模型估计：最小二乘法（least square method）

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

$$= \arg\min_{(w,b)} \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

线性回归 - 最小二乘法

- 最小化均方误差

$$E_{(w,b)} = \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

怎么估计参数？？？

线性回归 - 最小二乘法

- 最小化均方误差

$$E_{(w,b)} = \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

- 分别对 w 和 b 求导，可得

$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)\,x_i\right)$$

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right)$$

线性回归 - 最小二乘法

- 得到闭式（closed-form）解

$$w = \frac{\sum_{i=1}^{m} y_i \left(x_i - \bar{x}\right)}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2}$$

$$b = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - w x_i\right)$$

其中

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

多元线性回归

- 给定数据集

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$$

$$\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id}) \quad y_i \in \mathbb{R}$$

- 多元线性回归目标

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \text{ 使得} f(\boldsymbol{x}_i) \simeq y_i$$

多元线性回归

- 把 $\boldsymbol{w}$和b 吸收入向量形式$\hat{\boldsymbol{w}} = (\boldsymbol{w}; b)$, 数据集表示为

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

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

Lecture 7.2: 参数求解

多元线性回归 - 最小二乘法

- 最小二乘法（least square method）

$$\hat{\boldsymbol{w}}^* = \arg\min_{\hat{w}} \left(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}^{\mathrm{T}}\right)\left(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}\right)$$

令 $E_{\hat{\boldsymbol{w}}} = \left(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}\right)^{\mathrm{T}}\left(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}\right)$，对 $\hat{\boldsymbol{w}}$ 求导得到

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y}\right)$$

令上式为零可得 $\hat{\boldsymbol{w}}$ 最优解的闭式解

多元线性回归 - 满秩讨论

- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 是满秩矩阵或正定矩阵，则

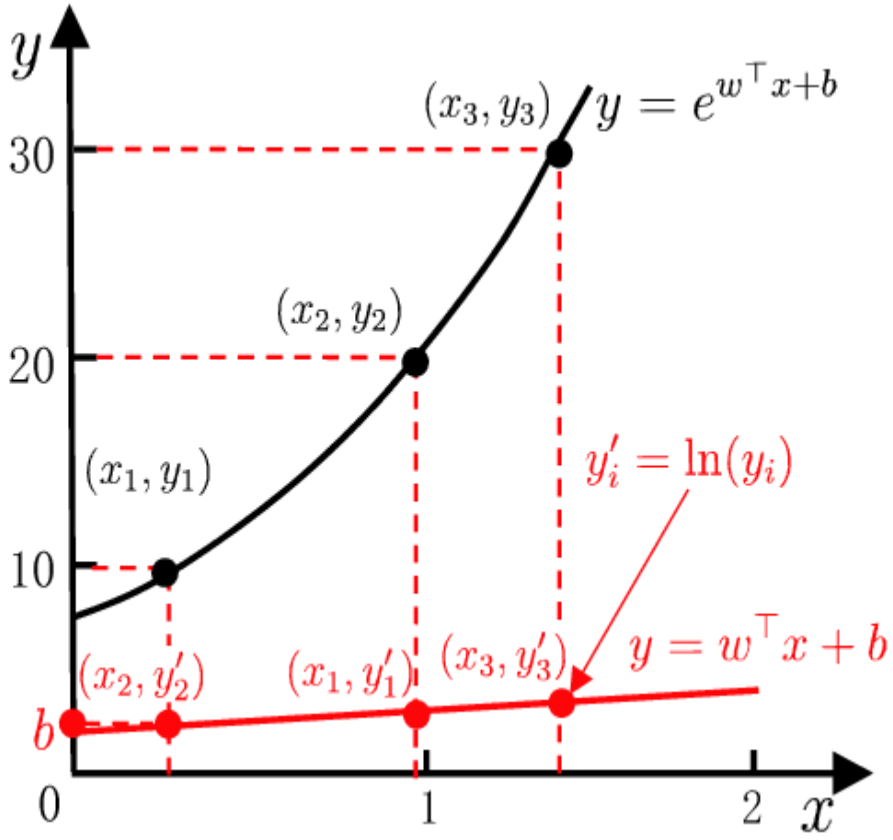
$$\hat{\boldsymbol{w}}^{*} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

其中$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$是 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的逆矩阵，线性回归模型为

$$f\left(\hat{\boldsymbol{x}}_i\right) = \hat{\boldsymbol{x}}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

对数线性回归

- 输出标记的对数为线性模型逼近的目标



$$\ln y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

$$\Updownarrow$$

$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

线性回归 - 广义线性模型

- 一般形式

$$y = g^{-1}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right)$$

- $g(\cdot)$ 称为联系函数（link function）
  - 单调可微函数

- 对数线性回归是 $g(\cdot) = \ln(\cdot)$ 时广义线性模型的特例

Thank you!

权小军 中山大学数据科学与计算机学院