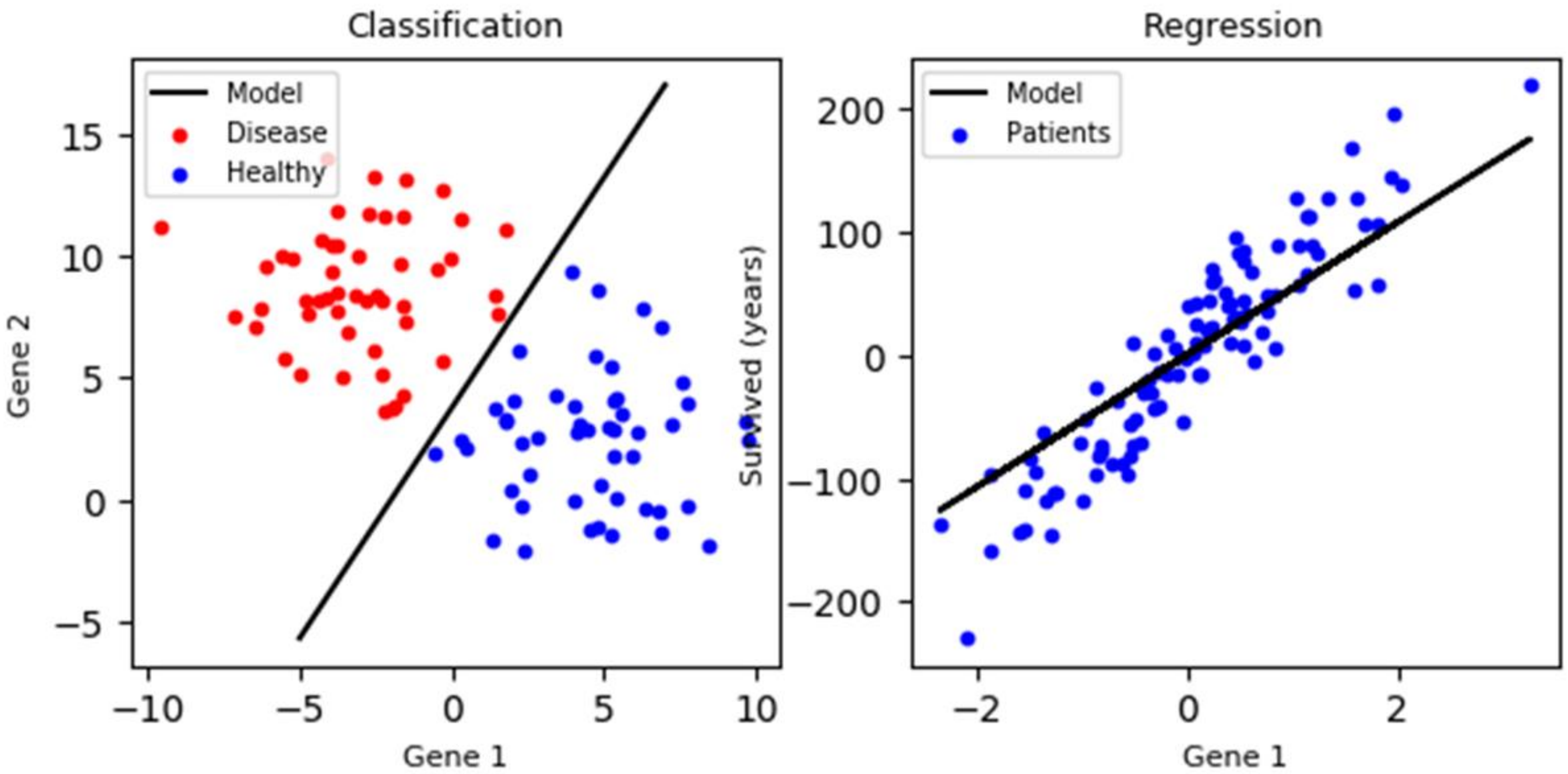
中国科学院南京分院
Nanjing Branch of Chinese Academy of Sciences

人工智能原理与算法
2. 线性回归模型

夏睿

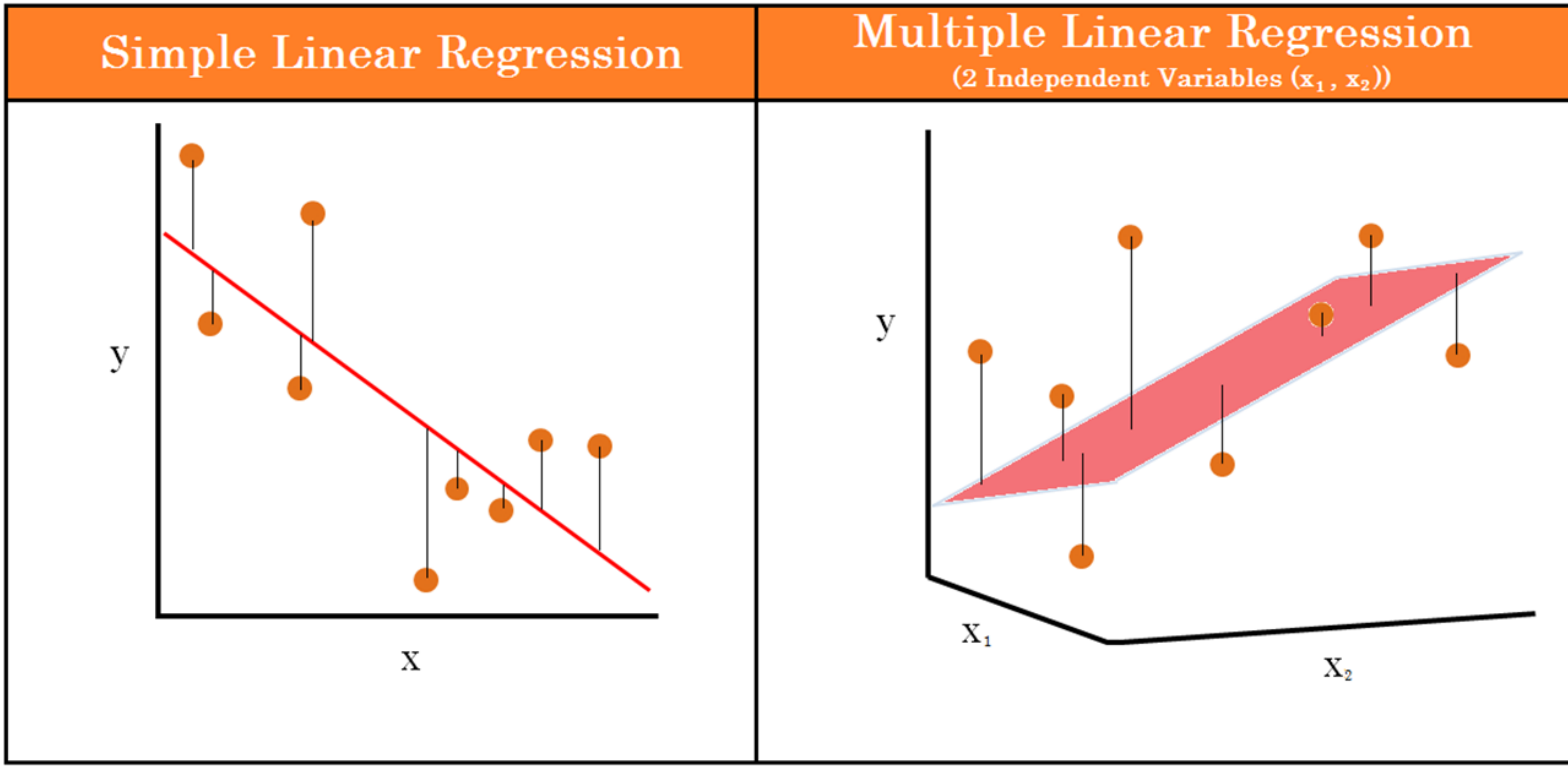
2023.2.22

回归 vs. 分类



Classification
Model
Disease
Healthy
Gene 2
Gene 1
15
10
5
0
−5
−10
−5
0
5
10
Regression
Model
Patients
Survived (years)
Gene 1
200
100
0
−100
−200
−2
0
2

线性回归


Simple Linear Regression
Multiple Linear Regression
(2 Independent Variables (x1 , x2))
y
x
y
X1
X2

输入、输出、关系

- 训练集

Living area (feet^2)	#bedrooms	Price (1000\$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

一个训练样本 $(x^{(k)}, y^{(k)})$，其中 k表示样本索引

输入：特征向量 $\boldsymbol{x} = [x_1, x_2]^{\mathrm{T}}$　　**输出：标量**y

- 输入与输出的关系

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

线性回归模型

- 模型假设

模型参数

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{M} \theta_i x_i + \theta_0 = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}$$

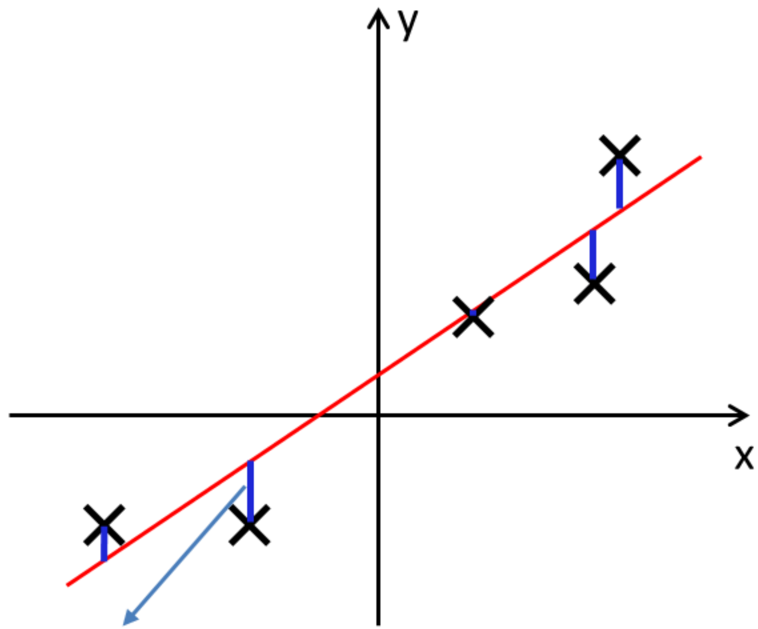
其中$\boldsymbol{\theta} = [\theta_0, \theta_1, \cdots, \theta_M]^{\mathrm{T}}, \boldsymbol{x} = [1, x_1, \cdots, x_M]^{\mathrm{T}}$

- 学习准则（损失函数）

均方误差（Mean Squared Error, MSE）

$$L_{lr}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{k=1}^{N}\left(h\left(\boldsymbol{x}^{(k)}\right) - y^{(k)}\right)^2$$

$$= \frac{1}{N}\sum_{k=1}^{N}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}^{(k)} - y^{(k)}\right)^2$$



误差 = 预测值 − 真实值

$$\boldsymbol{\theta}^* = \arg_{\boldsymbol{\theta}} \min L_{lr}(\boldsymbol{\theta})$$

最小均方误差（Minimum Mean Squared Error）
也称为最小二乘法（Least Square, LS）

LS问题求解

- 引入矩阵表示

$$\boldsymbol{X}=\begin{bmatrix}-\left(\boldsymbol{x}^{(1)}\right)^{\mathrm{T}}-\\-\left(\boldsymbol{x}^{(2)}\right)^{\mathrm{T}}-\\\vdots\\-\left(\boldsymbol{x}^{(N)}\right)^{\mathrm{T}}-\end{bmatrix}\qquad\qquad \boldsymbol{y}=\begin{bmatrix}y^{(1)}\\y^{(2)}\\\vdots\\y^{(N)}\end{bmatrix}$$

- 从而得到

$$\boldsymbol{X\theta}-\boldsymbol{y}=\begin{bmatrix}\left(\boldsymbol{x}^{(1)}\right)^{\mathrm{T}}\boldsymbol{\theta}\\\vdots\\\left(\boldsymbol{x}^{(N)}\right)^{\mathrm{T}}\boldsymbol{\theta}\end{bmatrix}-\begin{bmatrix}y^{(1)}\\\vdots\\y^{(N)}\end{bmatrix}=\begin{bmatrix}h\left(\boldsymbol{x}^{(1)}\right)-y^{(1)}\\\vdots\\h\left(\boldsymbol{x}^{(N)}\right)-y^{(N)}\end{bmatrix}$$

- 矩阵形式的损失函数

$$L_{lr}(\boldsymbol{\theta})=\frac{1}{N}\sum_{k=1}^{N}\left(h\left(\boldsymbol{x}^{(k)}\right)-y^{(k)}\right)^2=\frac{1}{N}(\boldsymbol{X\theta}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{X\theta}-\boldsymbol{y})$$

LS解析解

- 基于矩阵形式求导

$$\nabla_{\boldsymbol{\theta}} L_{lr}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \frac{1}{N} (\boldsymbol{X\theta} - \boldsymbol{y})^{\mathrm{T}} (\boldsymbol{X\theta} - \boldsymbol{y})$$

$$= \frac{1}{N} \nabla_{\boldsymbol{\theta}} \left(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X\theta} - \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}} \boldsymbol{X\theta} + \boldsymbol{y}^{\mathrm{T}} \boldsymbol{y} \right)$$

$$= \frac{2}{N} (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X\theta} - \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y})$$

$$\frac{\partial(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A x})}{\partial(\boldsymbol{x})} = (\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}) \boldsymbol{x}$$

$$\frac{\partial(\boldsymbol{b}^{\mathrm{T}} \boldsymbol{A x})}{\partial(\boldsymbol{x})} = \boldsymbol{A}^{\mathrm{T}} b$$

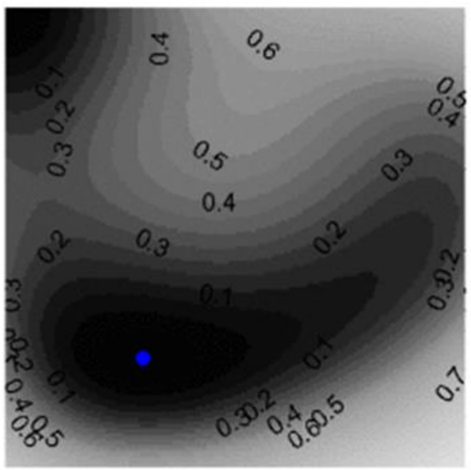
矩阵求导性质：https://en.wikipedia.org/wiki/Matrix_calculus

- 求导置零获得解析解

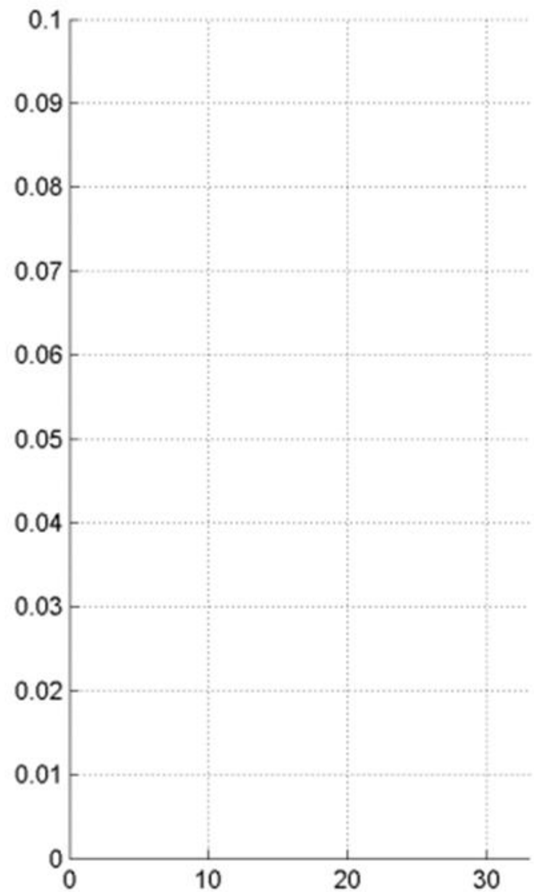
$$\boldsymbol{\theta}^* = \boxed{(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y}$$

逆矩阵常常难以计算，甚至不可逆！

梯度下降算法示意图



0.4
0.6
0.1
0.2
0.3
0.5
0.4
0.5
0.4
0.3
0.2
0.3
0.2
0.1
0.2
0.3
0.2
0.3
0.2
0.1
0.1
0.4
0.5
0.6
0.3
0.2
0.4
0.5
0.6
0.7



0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30

梯度下降算法

- 梯度下降是求函数$L(\boldsymbol{\theta})$最小值的一阶迭代优化算法

- 主要思想：
  - 梯度反方向是函数值下降最快的方向

- 优化过程：
  - 从初始位置开始（即初始参数$\boldsymbol{\theta}^{(0)}$）
  - 在当前位置$\boldsymbol{\theta}^{(t)}$，重复直到收敛
    - 计算当前位置梯度：$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$
    - 沿梯度反方向移动到下一个位置：$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha \cdot \nabla_{\boldsymbol{\theta}} L(\theta)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$，其中$\alpha$是学习率
    - $t = t + 1$

线性回归的梯度下降

- 梯度

$$
\begin{aligned}
\frac{dL_{lr}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} &= \frac{1}{N}\frac{d}{d\boldsymbol{\theta}}\sum_{k=1}^{N}\left(h\left(\boldsymbol{x}^{(k)}\right)-y^{(k)}\right)^2 \\
&= \frac{2}{N}\sum_{k=1}^{N}\left(h\left(\boldsymbol{x}^{(k)}\right)-y^{(k)}\right)\cdot\frac{d}{d\boldsymbol{\theta}}\left(h\left(\boldsymbol{x}^{(k)}\right)-y^{(k)}\right) \\
&= \frac{2}{N}\sum_{k=1}^{N}\left(h\left(\boldsymbol{x}^{(k)}\right)-y^{(k)}\right)\frac{d}{d\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}^{(k)}\right) \\
&= \frac{2}{N}\sum_{k=1}^{N}\boxed{\left(h\left(\boldsymbol{x}^{(k)}\right)-y^{(k)}\right)\boldsymbol{x}^{(k)}}
\end{aligned}
$$

误差 · 输入

- 梯度下降(GD) 优化

$$
\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha\frac{d}{d\boldsymbol{\theta}}L_{lr}(\boldsymbol{\theta}) = \boldsymbol{\theta} - \alpha\sum_{k=1}^{N}\left(h\left(\boldsymbol{x}^{(k)}\right)-y^{(k)}\right)\boldsymbol{x}^{(k)}
$$

从概率角度解读线性回归模型

- 假设一个具有高斯噪声的线性模型

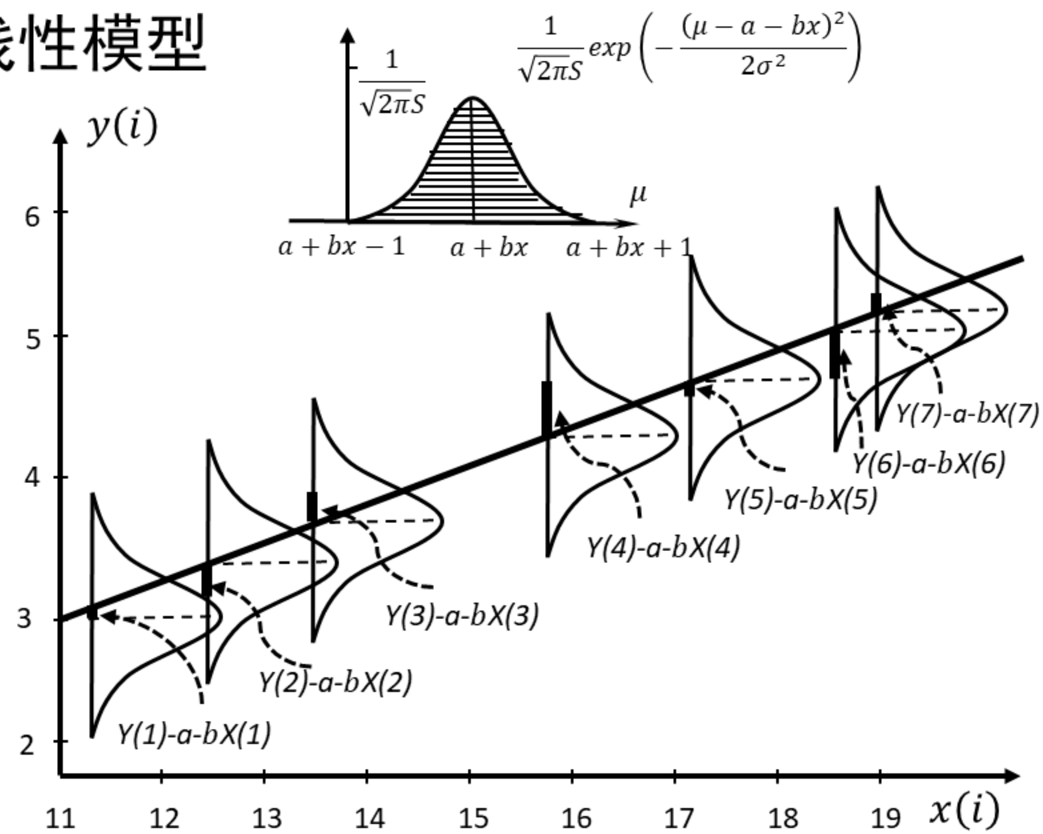
假设 $E\sim\mathcal{N}(0,\sigma^2)$

$$p(e;\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{e^2}{2\sigma^2}\right)$$

$\boldsymbol{x}$ 为一个确定性取值

$$y=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}+e\sim\mathcal{N}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x},\sigma^2\right)$$

$$p(y|\boldsymbol{x};\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})^2}{2\sigma^2}\right)$$



对于给定的取值$\boldsymbol{x}$，y服从$\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}$为期望、 σ^2为方差的高斯分布

最大似然估计(MLE)

- 似然函数

$$L(\boldsymbol{\theta}) = \prod_{k=1}^{N} p\left(y^{(k)} \middle| \boldsymbol{x}^{(k)}; \boldsymbol{\theta}\right) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(k)} - \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(k)}\right)^2}{2\sigma^2}\right)$$

- 对数似然

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log L(\boldsymbol{\theta}) = \log \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(k)} - \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(k)}\right)^2}{2\sigma^2}\right) \\ &= \sum_{k=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(k)} - \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(k)}\right)^2}{2\sigma^2}\right) \\ &= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{k=1}^{N} \left(y^{(k)} - \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(k)}\right)^2 \end{aligned}$$

- 最大化对数似然函数

MLE = LS

$$\max l(\boldsymbol{\theta}) \Leftrightarrow \min \frac{1}{N} \sum_{k=1}^{N} \left(y^{(k)} - \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(k)}\right)^2$$

回归任务的性能评估

- Mean Squared Error (MSE)

$$\mathrm{MSE}_{te} = \frac{1}{N_{te}}\sum_{k=1}^{N_{te}}\left(h\left(\boldsymbol{x}_{te}^{(k)}\right) - y_{te}^{(k)}\right)^2 = \frac{1}{N_{te}}\sum_{k=1}^{N_{te}}\left(\boldsymbol{\theta}^{*\mathrm{T}}\boldsymbol{x}_{te}^{(k)} - y_{te}^{(k)}\right)^2$$

- Mean Absolute Error (MAE)

$$\mathrm{MAE}_{te} = \frac{1}{N_{te}}\sum_{k=1}^{N_{te}}\left|h\left(\boldsymbol{x}_{te}^{(k)}\right) - y_{te}^{(k)}\right| = \frac{1}{N_{te}}\sum_{k=1}^{N_{te}}\left|\boldsymbol{\theta}^{*\mathrm{T}}\boldsymbol{x}_{te}^{(k)} - y_{te}^{(k)}\right|$$

- R Squared (R^2)

$$R^2{}_{te} = 1 - \frac{\mathrm{SS}_{residual}}{\mathrm{SS}_{total}} = 1 - \frac{\sum_{k=1}^{N_{te}}\left(y_{te}^{(k)} - h\left(\boldsymbol{x}_{te}^{(k)}\right)\right)^2}{\sum_{k=1}^{N_{te}}\left(y_{te}^{(k)} - \bar{y}_{te}\right)^2}$$

作业#1：南京房价预测

- 给定南京平均房价历史数据

  Year x = [2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013]

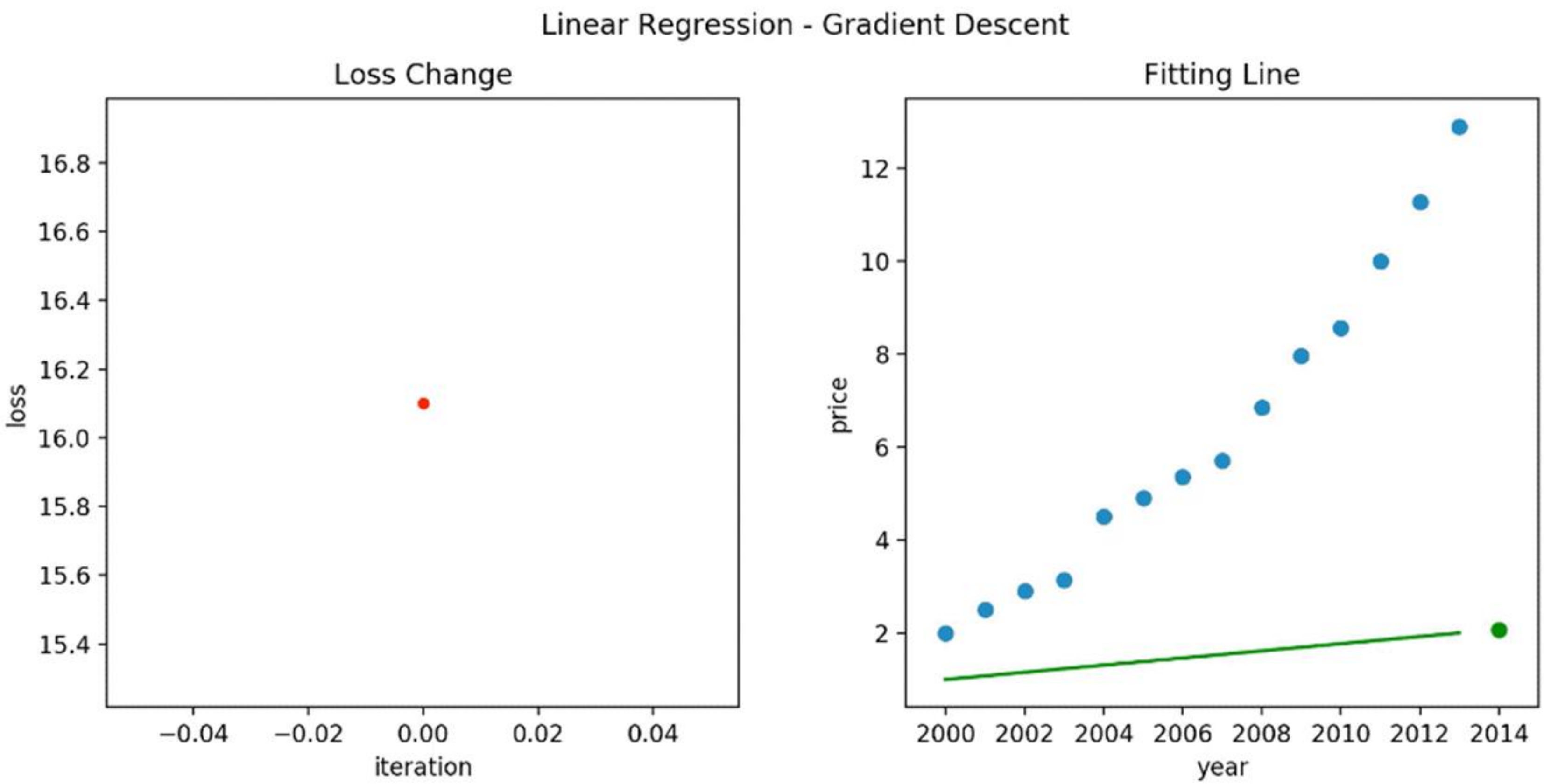
  Price y = [2.000, 2.500, 2.900, 3.147, 4.515, 4.903, 5.365, 5.704, 6.853, 7.971, 8.561, 10.000, 11.280, 12.900]

  http://www.nustm.cn/member/rxia/ml/data/Price.zip

- 假设：房价和年度呈线性关系

- 任务：基于下列两种方法编程实现线性回归：1) 解析解；2) 梯度下降法，获得x和y的关系，并预测2014年的南京房价

Demo


Linear Regression - Gradient Descent
Loss Change
16.8
16.6
16.4
16.2
16.0
15.8
15.6
15.4
loss
−0.04
−0.02
0.00
0.02
0.04
iteration
Fitting Line
12
10
8
6
4
2
price
2000
2002
2004
2006
2008
2010
2012
2014
year

本讲结束 欢迎提问