

# 回归

YangXiao

2018 年 3 月 24 日

## 1 引理：向量的导数

$\mathbf{A}$  为  $m \times n$  的矩阵,  $\mathbf{x}$  为  $n \times 1$  的列向量,

记  $\vec{y} = \mathbf{A} \cdot \vec{x}$

思考:  $\frac{\partial \vec{y}}{\partial \vec{x}} = ?$

向量导数的推导

令

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$
$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$
$$\mathbf{A} \cdot \vec{x} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 \cdots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 \cdots + a_{nn}x_n \end{bmatrix}$$

从而,

$$\frac{\partial \vec{y}}{\partial \vec{x}} = \frac{\partial \mathbf{A} \cdot \vec{x}}{\partial \vec{x}} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} = \mathbf{A}^T$$

结论与直接推广

向量偏导公式:

$$\frac{\partial \mathbf{A} \cdot \vec{x}}{\partial \vec{x}} = \mathbf{A}^T$$

$$\frac{\partial \mathbf{A} \cdot \vec{x}}{\partial \vec{x}^T} = \mathbf{A}$$

$$\frac{\partial (\vec{x}^T \cdot \mathbf{A})}{\partial \vec{x}} = \mathbf{A}$$

引理: 标量对向量求导

$\mathbf{A}$  为  $n \times n$  的矩阵,  $\mathbf{x}$  为  $n \times 1$  的列向量,

记  $y = \vec{x}^T \cdot \mathbf{A} \cdot \vec{x}$

同理可得:

$$\frac{\partial y}{\partial \vec{x}} = \frac{\vec{x}^T \cdot \mathbf{A} \cdot \vec{x}}{\partial \vec{x}} = (\mathbf{A} + \mathbf{A}^T) \cdot \vec{x}$$

$$\text{若 } \mathbf{A} = \mathbf{A}^T, \quad \frac{\vec{x}^T \cdot \mathbf{A} \cdot \vec{x}}{\partial \vec{x}} = 2\mathbf{A}\vec{x}$$

## 2 线性回归

$y = ax + b$  - 单变量

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

$$h_{\theta}(x) = \sum_{i=0}^n \theta_i x_i = \theta^T \mathbf{x} \quad (1)$$

使用极大似然估计解释最小二乘

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)} \quad (2)$$

误差  $\epsilon^{(i)} (1 \leq i \leq m)$  是独立同分布的, 服从均值为 0, 方差为某定值的  $\sigma^2$  的高斯分布

原因: 中心极限定理。众多因素的独立影响的综合反应, 往往近似服从正态分布。

似然函数

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right) \quad (3)$$

$$p(y^{(i)}|x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \quad (4)$$

$$\mathcal{L}(\theta) = \prod_{i=1}^m p(y^{(i)}|x^{(i)}; \theta) \quad (5)$$

$$= \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \quad (6)$$

在得到似然函数后，使似然函数值最大，即假定给定  $x$  和  $\theta$  后，使得  $y$  出现的可能性最大，事实上  $\theta$  未知，4 表示在  $i$  那个点处给定  $x$  和  $\theta$   $y^{(i)}$  出现的概率，所以连乘所有的点就可以得到所有概率，然后求最大。我们可以对似然函数  $L$  求导，使其等于 0，得到极值点。

为什么在求导等于 0 可以保证似然函数值最大？

因为，指数族函数都是凹函数，凹函数极值点便是最大值点。且为最大值点，我们平时看到的各种分布：高斯分布，泊松分布等可以写成指数族函数形式。下面即开始研究高斯对数与最小二乘的关系：

$$\mathcal{L}(\theta) = \log L(\theta) \quad (7)$$

$$= \log \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma^2}\right) \quad (8)$$

$$= \sum_{i=1}^m \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma^2}\right) \quad (9)$$

$$= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 \quad (10)$$

去掉无关常数项，所以目标函数变换为：

$$\mathcal{J}(\theta) = \frac{1}{2} \sum_{i=1}^m (h(\theta)(x^{(i)}) - y^{(i)})^2 \quad (11)$$

假设  $\mathbf{X}$  为  $m$  行， $n$  列的矩阵。 $m$  表示有  $m$  个样本， $n$  表示每个样本有多少个参数（特征）。

目标函数11 转变为如下：

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) \quad (12)$$

对上式12 求导等于 0，过程如下：

$$\nabla_{\boldsymbol{\theta}}\mathcal{J}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\left(\frac{1}{2}(\boldsymbol{\theta}^T\mathbf{X}^T - \mathbf{y}^T)(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})\right) \quad (13)$$

$$= \nabla_{\boldsymbol{\theta}}\left(\frac{1}{2}\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\theta} + \mathbf{y}^T\mathbf{y}\right) \quad (14)$$

$$= \frac{1}{2}(2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y} - (\mathbf{y}^T\mathbf{X})^T) \quad (15)$$

$$= \mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y} \quad (16)$$

令上式等于 0，可得  $\boldsymbol{\theta}$  解析式：

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (17)$$

若  $\mathbf{X}^T\mathbf{X}$  不可逆，则需要加入  $\lambda$  扰动，即正则，如下：

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad (18)$$

梯度下降算法最小化目标函数

- 初始化  $\boldsymbol{\theta}$ (随机初始化)
- 迭代，新的  $\theta$  能够使得  $J(\boldsymbol{\theta})$  更小
- 如果  $J(\boldsymbol{\theta})$  能继续减少，继续迭代
- $\boldsymbol{\theta}_j = \boldsymbol{\theta}_j - \alpha \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}$
- 其中， $\alpha$  称为学习率/步长

$$\alpha \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_j} = \alpha \frac{\partial}{\partial \theta_j} \frac{1}{2} (h_{\theta}(x) - y)^2 \quad (19)$$

$$= 2 \cdot \frac{1}{2} (h_{\theta}(x) - y) \cdot \frac{\partial}{\partial \theta_j} (h_{\theta}(x) - y) \quad (20)$$

$$= (h_{\theta}(x) - y) \cdot \frac{\partial}{\partial \theta_j} \left( \sum_{i=0}^n \theta_i x_i - y \right) \quad (21)$$

$$= (h_{\theta}(x) - y) \mathbf{x}_j \quad (22)$$

其中  $x_j$  表示某一个向量的第  $j$  维，是标量  
批处理梯度下降算法

$$\theta_j = \theta_j - \alpha \sum_{i=0}^m (h_{\theta}(x^i) - y^{(i)}) x_j^{(i)} \quad (23)$$

其中  $x_j^{(i)}$  表示第  $i$  个向量的第  $j$  维，没更新一个  $\theta$  需要遍历所有的样本，当样本很大时，显然需要的操作太多，下面介绍随机梯度下降

随机梯度下降——SGD

Loop

for  $i=1$  to  $m$ :

$$\theta_j = \theta_j - \alpha (h_{\theta}(x^i) - y^{(i)}) x_j^{(i)} \quad (24)$$

### 3 logistic 回归

**logistic 分布**

设  $X$  是连续随机变量，若  $X$  满足 logistic 分布，则  $X$  具有下列分布函数：

$$F(x) = P(X \leq x) = \frac{1}{1 + \exp^{-(x-\mu)/\gamma}} \quad (25)$$

其中  $\mu$  为位置参数， $\gamma > 0$  为形状参数

**logistic 函数**

$$g(z) = \frac{1}{1 + \exp^{-z}} h_{\theta}(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x}) = \frac{1}{1 + \exp^{-\boldsymbol{\theta}^T \mathbf{x}}} \quad (26)$$

对  $g(z)$  求导得：

$$g'(z) = g(z) (1 - g(z)) \quad (27)$$

假定：

$$P(y = 1 \mid \mathbf{x}; \boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\mathbf{x}) \quad (28)$$

$$P(y = 0 \mid \mathbf{x}; \boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(\mathbf{x}) \quad (29)$$

给定  $\mathbf{x}$  和  $\boldsymbol{\theta}$  分布律为：

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = (h_{\boldsymbol{\theta}}(\mathbf{x}))^y (1 - h_{\boldsymbol{\theta}}(\mathbf{x}))^{1-y} \quad (30)$$

似然函数:

$$\mathcal{L} = p(\vec{y}|\mathbf{X}; \boldsymbol{\theta}) \quad (31)$$

$$= \prod_{i=1}^m p(y^{(i)}|x^{(i)}; \boldsymbol{\theta}) \quad (32)$$

$$= \prod_{i=1}^m p(h_{\boldsymbol{\theta}}(x^{(i)}))^{y^{(i)}} (1 - h_{\boldsymbol{\theta}}(x^{(i)}))^{1-y^{(i)}} \quad (33)$$

对数似然函数:

$$\mathcal{L} = \log L(\boldsymbol{\theta}) = \sum_{i=1}^m y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log (1 - h(x^{(i)})) \quad (34)$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_j} = \left( y \frac{1}{g(\boldsymbol{\theta}^T \mathbf{X})} + (1 - y) \frac{-1}{1 - g(\boldsymbol{\theta}^T \mathbf{X})} \right) \frac{\partial g(\boldsymbol{\theta}^T \mathbf{X})}{\partial \theta_j} \quad (35)$$

$$= \left( y \frac{1}{g(\boldsymbol{\theta}^T \mathbf{X})} + (1 - y) \frac{-1}{1 - g(\boldsymbol{\theta}^T \mathbf{X})} \right) g(\boldsymbol{\theta}^T \mathbf{X})(1 - g(\boldsymbol{\theta}^T \mathbf{X})) \frac{\partial \boldsymbol{\theta}^T \mathbf{X}}{\partial \theta_j} \quad (36)$$

$$= (y(1 - g(\boldsymbol{\theta}^T \mathbf{X})) + (1 - y)g(\boldsymbol{\theta}^T \mathbf{X})) x_j \quad (37)$$

$$= (y - (\boldsymbol{\theta}^T \mathbf{X})) x_j \quad (38)$$

$$= (y - h_{\boldsymbol{\theta}}(\mathbf{X})) x_j \quad (39)$$

logistic 回归参数学习规则:

批处理梯度下降算法

$$\theta_j = \theta_j - \alpha \sum_{i=0}^m (y^{(i)} - h_{\boldsymbol{\theta}}(x^{(i)})) x_j^{(i)} \quad (40)$$

其中  $x_j^{(i)}$  表示第  $i$  个向量的第  $j$  维, 没更新一个  $\boldsymbol{\theta}$  需要遍历所有的样本, 当样本很大时, 显然需要的操作太多, 下面介绍随机梯度下降

**随机梯度下降——SGD**

Loop

for  $i=1$  to  $m$ :

$$\theta_j = \theta_j - \alpha (y^{(i)} - h_{\boldsymbol{\theta}}(x^{(i)})) x_j^{(i)} \quad (41)$$

对比23,40和24,41。可知线性回归和 logistic 回归有相同形式。

### 对数线性模型

一个事件的几率 odds，是指该事件发生的概率与该事件不发生的概率的比值。

对数几率：logit 函数

$$P(y = 1 \mid \mathbf{x}; \boldsymbol{\theta}) = h_{\theta}(\mathbf{x})$$

$$P(y = 0 \mid \mathbf{x}; \boldsymbol{\theta}) = 1 - h_{\theta}(\mathbf{x})$$

$$\text{logit}(p) = \log \frac{p}{1-p} \quad (42)$$

$$= \log \frac{h_{\theta}(\mathbf{x})}{1 - h_{\theta}(\mathbf{x})} \quad (43)$$

$$= \log \left( \frac{\frac{1}{1 + \exp(-w^T \mathbf{x})}}{\frac{\exp(-w^T \mathbf{x})}{1 + \exp(-w^T \mathbf{x})}} \right) \quad (44)$$

$$= w^T \mathbf{x} \quad (45)$$

所以 logistic 回归模型是，对输入  $\mathbf{x}$  的数线性模型  $w^T \mathbf{x}$ ，由26可知，当线性模型  $w^T \mathbf{x}$  越趋近无穷大，其概率值越接近 1，越接近负无穷，其概率值越趋近 0。

## 4 最大熵模型