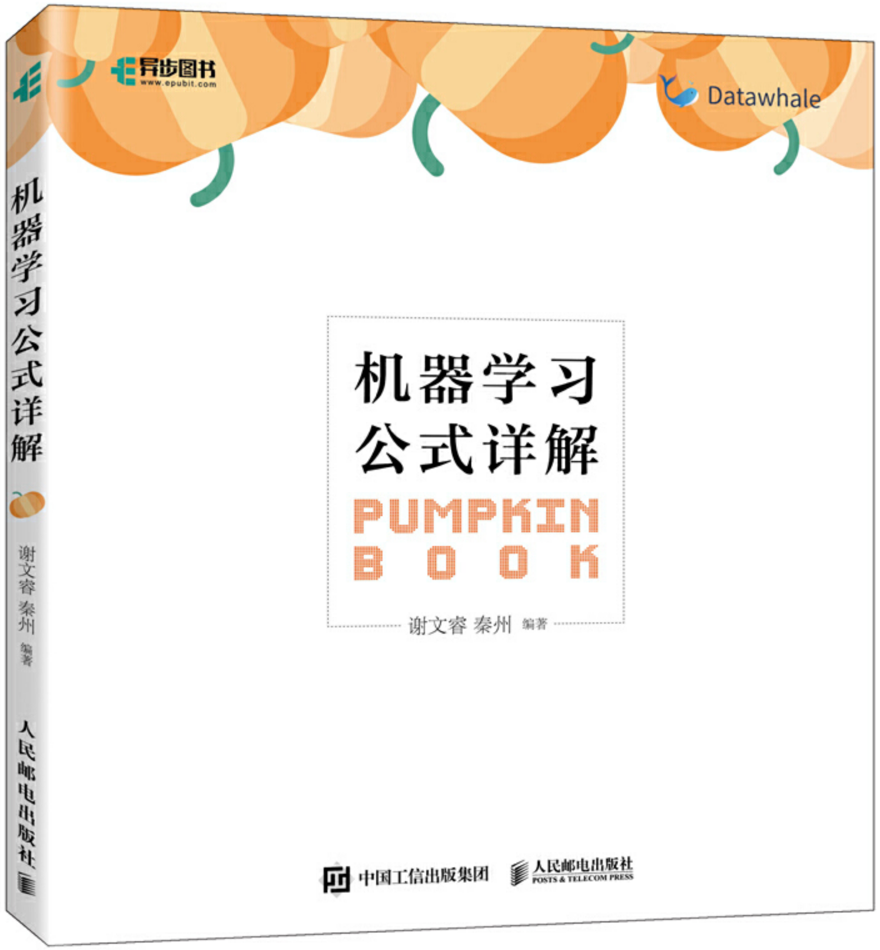
Datawhale 异步社区



《机器学习公式详解》（南瓜书）

第3章 一元线性回归

本节主讲：谢文睿

本节大纲

西瓜书对应章节：3.1、3.2

1. 算法原理
2. 线性回归的最小二乘估计和极大似然估计
3. 求解w和b

算法原理

举一个通过【发际线的高度】预测【计算机水平】的例子

算法原理

仅通过【发际线高度】预测【计算机水平】：

$$f(x) = w_1 x_1 + b$$

+二值离散特征【颜值】（好看：1，不好看：0）

$$f(x) = w_1 x_1 + w_2 x_2 + b$$

+有序的多值离散特征【饭量】（小：1，中：2，大：3）

$$f(x) = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$$

+无序的多值离散特征【肤色】（黄：[1,0,0]，黑：[0,1,0]，白：[0,0,1]）

$$f(x) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6 + b$$

最小二乘估计

Datawhale 异步社区

基于均方误差最小化来进行模型求解的方法称为“最小二乘法”

$$
\begin{aligned}
E_{(w,b)} &= \sum_{i=1}^{m}\left(y_i - f(x_i)\right)^2 \\
&= \sum_{i=1}^{m}\left(y_i - (wx_i + b)\right)^2 \\
&= \sum_{i=1}^{m}\left(y_i - wx_i - b\right)^2
\end{aligned}
$$

此即为公式3.4 $\underset{(w,b)}{\arg\min}$（解释一下）后面的部分

极大似然估计

用途：估计概率分布的参数值

方法：对于离散型（连续型）随机变量X，假设其概率质量函数为$P(x;\theta)$（概率密度函数为$p(x;\theta)$），其中θ为待估计的参数值（可以有多个）。现有$x_1, x_2, x_3, ..., x_n$是来自X的n个独立同分布的样本，它们的联合概率为

$$L(\theta) = \prod_{i=1}^{n} P(x_i;\theta)$$

其中$x_1, x_2, x_3, ..., x_n$是已知量，θ是未知量，因此以上概率是一个关于θ的函数，称$L(\theta)$为样本的似然函数。极大似然估计的直观想法：使得观测样本出现概率最大的分布就是待求分布，也即使得联合概率（似然函数）$L(\theta)$取到最大值的θ^*即为θ的估计值。

极大似然估计

例题：现有一批观测样本$x_1, x_2, x_3, ..., x_n$，假设其服从某个正态分布$X \sim N(\mu, \sigma^2)$，其中μ, σ为待估计的参数值，请用极大似然估计法估计μ, σ

【解】：第一步：写出随机变量X的概率密度函数

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

第二步：写出似然函数

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} p(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

第三步：求出使得$L(\mu, \sigma^2)$取到最大值的μ, σ

极大似然估计

由于对数函数ln是单调递增函数，所以$\ln L(\mu,\sigma^2)$和$L(\mu,\sigma^2)$拥有相同的最大值点，而且利用对数函数的性质可以化简$L(\mu,\sigma^2)$中的连乘项，因此通常会用$\ln L(\mu,\sigma^2)$代替$L(\mu,\sigma^2)$来求μ,σ，加了对数函数符号的似然函数$\ln L(\mu,\sigma^2)$称为对数似然函数。

$$
\begin{aligned}
\ln L(\mu,\sigma^2) &= \ln\left[\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{n}\ln\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

推荐视频：张宇考研数学——《概率论与数理统计》

推荐教材：陈希孺.《概率论与数理统计》、盛骤 等.《概率论与数理统计》

极大似然估计

对于线性回归来说，也可以假设其为以下模型

$$y = wx + b + \epsilon$$

其中ϵ为不受控制的随机误差，通常假设其服从均值为0的正态分布$\epsilon \sim N(0, \sigma^2)$（高斯提出的，也可以用中心极限定理解释），所以$\epsilon$的概率密度函数为

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

若将ϵ用$y-(wx+b)$等价替换可得

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-(wx+b))^2}{2\sigma^2}\right)$$

极大似然估计

上式显然可以看作$y \sim N(wx+b, \sigma^2)$，下面便可以用极大似然估计来估计w和b的值，似然函数为

$$L(w,b)=\prod_{i=1}^{m} p(y_i)=\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-(wx_i+b))^2}{2\sigma^2}\right)$$

$$\begin{aligned}\ln L(w,b)&=\sum_{i=1}^{m}\ln\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-wx_i-b)^2}{2\sigma^2}\right)\\&=\sum_{i=1}^{m}\ln\frac{1}{\sqrt{2\pi}\sigma}+\sum_{i=1}^{m}\ln\exp\left(-\frac{(y_i-wx_i-b)^2}{2\sigma^2}\right)\end{aligned}$$

极大似然估计

$$\ln L(w,b)=m\ln\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-wx_i-b)^2$$

其中m,σ均为常数，所以最大化$\ln L(w,b)$等价于最小化$\sum_{i=1}^{m}(y_i-wx_i-b)^2$，也即

$$(w^*,b^*)=\underset{(w,b)}{\arg\max}\ln L(\boldsymbol{w},b)=\underset{(w,b)}{\arg\min}\sum_{i=1}^{m}(y_i-wx_i-b)^2$$

此即为公式3.4，等价于最小二乘估计。

拓展阅读：靳志辉.《正态分布的前世今生》

求解w和b

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

求解w和b其本质上是一个多元函数求最值（点）的问题，更具体点是凸函数求最值的问题。

推导思路：

1. 证明$E_{(w,b)} = \sum_{i=1}^{m} (y_i - wx_i - b)^2$是关于$w$和$b$的凸函数
2. 用凸函数求最值的思路求解出w和b

求解w和b

凸集：设集合$D \subset \mathbb{R}^n$，如果对任意的$\boldsymbol{x}, \boldsymbol{y} \in D$与任意的$\alpha \in [0,1]$，有

$$\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in D$$

则称集合D是凸集。凸集的几何意义是：若两个点属于此集合，则这两点连线上的任意一点均属于此集合（此处应该有图）。常见的凸集有空集$\varnothing$，n维欧式空间$\mathbb{R}^n$

凸函数：设D是非空凸集，f是定义在D上的函数，如果对任意的$\boldsymbol{x}^1, \boldsymbol{x}^2 \in D, \alpha \in (0,1)$，均有

$$f\left(\alpha \boldsymbol{x}^1 + (1-\alpha)\boldsymbol{x}^2\right) \leqslant \alpha f(\boldsymbol{x}^1) + (1-\alpha) f(\boldsymbol{x}^2)$$

则称f为D上的凸函数（此处也应该有图）。

推荐教材：王燕军.《最优化基础理论与方法》

求解w和b

梯度（多元函数的一阶导数）：设n元函数$f(\boldsymbol{x})$对自变量$\boldsymbol{x}=(x_1,x_2,\ldots,x_n)^{\mathrm{T}}$的各分量$x_i$的偏导数$\frac{\partial f(\boldsymbol{x})}{\partial x_i}(i=1,\ldots,n)$都存在，则称函数$f(\boldsymbol{x})$在$\boldsymbol{x}$处一阶可导，并称向量

$$\nabla f(\boldsymbol{x})=\begin{bmatrix}\frac{\partial f(\boldsymbol{x})}{\partial x_1}\\ \frac{\partial f(\boldsymbol{x})}{\partial x_2}\\ \vdots\\ \frac{\partial f(\boldsymbol{x})}{\partial x_n}\end{bmatrix}$$

为函数$f(\boldsymbol{x})$在$\boldsymbol{x}$处的一阶导数或梯度（说一下为啥是列向量）。

求解w和b

Hessian（海塞）矩阵（多元函数的二阶导数）：设n元函数$f(\boldsymbol{x})$对自变量$\boldsymbol{x}=(x_1,x_2,\ldots,x_n)^{\mathrm{T}}$的各分量$x_i$的二阶偏导数$\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} \quad (i=1,2,\ldots,n;j=1,2,\ldots,n)$都存在，则称函数$f(\boldsymbol{x})$在$\boldsymbol{x}$处二阶可导，并称矩阵

$$\nabla^2 f(\boldsymbol{x})=\begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n^2} \end{bmatrix}$$

为函数$f(\boldsymbol{x})$在$\boldsymbol{x}$处的二阶导数或Hessian（海塞）矩阵。

求解w和b

定理：设$D \subset \mathbb{R}^n$是非空开凸集，$f: D \subset \mathbb{R}^n \rightarrow \mathbb{R}$，且$f(\boldsymbol{x})$在$D$上二阶连续可微，如果$f(\boldsymbol{x})$的Hessian（海塞）矩阵在$D$上是半正定的，则$f(\boldsymbol{x})$是$D$上的凸函数。(类比一元函数判断凹凸性)

因此，只需证明$E_{(w,b)} = \sum_{i=1}^{m}(y_i - wx_i - b)^2$的Hessian（海塞）矩阵

$$\nabla^2 E_{(w,b)} = \begin{bmatrix} \frac{\partial^2 E_{(w,b)}}{\partial w^2} & \frac{\partial^2 E_{(w,b)}}{\partial w \partial b} \\ \frac{\partial^2 E_{(w,b)}}{\partial b \partial w} & \frac{\partial^2 E_{(w,b)}}{\partial b^2} \end{bmatrix}$$

是半正定的，那么$E_{(w,b)}$就是关于w和b的凸函数。

求解w和b

$$
\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial w} &= \frac{\partial}{\partial w}\left[\sum_{i=1}^{m}\left(y_i - wx_i - b\right)^2\right] \\
&= \sum_{i=1}^{m}\frac{\partial}{\partial w}\left(y_i - wx_i - b\right)^2 \\
&= \sum_{i=1}^{m} 2\cdot\left(y_i - wx_i - b\right)\cdot\left(-x_i\right) \\
&= 2\left(w\sum_{i=1}^{m}x_i^2 - \sum_{i=1}^{m}\left(y_i - b\right)x_i\right)
\end{aligned}
$$

此即为公式3.5

求解w和b

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

求解w和b

$$
\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial w \partial b} &= \frac{\partial}{\partial b}\left(\frac{\partial E_{(w,b)}}{\partial w}\right) \\
&= \frac{\partial}{\partial b}\left[2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}\left(y_i - b\right)x_i\right)\right] \\
&= \frac{\partial}{\partial b}\left[-2\sum_{i=1}^{m}\left(y_i - b\right)x_i\right] \\
&= \frac{\partial}{\partial b}\left(-2\sum_{i=1}^{m} y_i x_i + 2\sum_{i=1}^{m} b x_i\right) \\
&= 2\sum_{i=1}^{m} x_i
\end{aligned}
$$

求解w和b

$$
\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial b} &= \frac{\partial}{\partial b}\left[\sum_{i=1}^{m}\left(y_i - wx_i - b\right)^2\right] \\
&= \sum_{i=1}^{m}\frac{\partial}{\partial b}\left(y_i - wx_i - b\right)^2 \\
&= \sum_{i=1}^{m} 2\cdot\left(y_i - wx_i - b\right)\cdot(-1) \\
&= 2\left(mb - \sum_{i=1}^{m}\left(y_i - wx_i\right)\right) \text{此即为公式3.6}
\end{aligned}
$$

求解w和b

$$
\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial b \partial w} &= \frac{\partial}{\partial w}\left(\frac{\partial E_{(w,b)}}{\partial b}\right) \\
&= \frac{\partial}{\partial w}\left[2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right)\right] \\
&= \frac{\partial}{\partial w}\left(2\sum_{i=1}^{m} wx_i\right) \\
&= 2\sum_{i=1}^{m} x_i
\end{aligned}
$$

求解w和b

Datawhale 异步社区

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

求解w和b

$$\nabla^2 E_{(w,b)} = \begin{bmatrix} \frac{\partial^2 E_{(w,b)}}{\partial w^2} & \frac{\partial^2 E_{(w,b)}}{\partial w \partial b} \\ \frac{\partial^2 E_{(w,b)}}{\partial b \partial w} & \frac{\partial^2 E_{(w,b)}}{\partial b^2} \end{bmatrix} = \begin{bmatrix} 2\sum_{i=1}^m x_i^2 & 2\sum_{i=1}^m x_i \\ 2\sum_{i=1}^m x_i & 2m \end{bmatrix}$$

半正定矩阵的判定定理之一：若实对称矩阵的所有顺序主子式均为非负，则该矩阵为半正定矩阵。

$$\left| 2\sum_{i=1}^m x_i^2 \right| > 0$$

$$\begin{vmatrix} 2\sum_{i=1}^m x_i^2 & 2\sum_{i=1}^m x_i \\ 2\sum_{i=1}^m x_i & 2m \end{vmatrix} = 2\sum_{i=1}^m x_i^2 \cdot 2m - 2\sum_{i=1}^m x_i \cdot 2\sum_{i=1}^m x_i$$

$$= 4m\sum_{i=1}^m x_i^2 - 4\left(\sum_{i=1}^m x_i\right)^2$$

求解w和b

$$4m\sum_{i=1}^{m}x_i^2-4\left(\sum_{i=1}^{m}x_i\right)^2=4m\sum_{i=1}^{m}x_i^2-4\cdot m\cdot\frac{1}{m}\cdot\left(\sum_{i=1}^{m}x_i\right)^2$$

$$=4m\sum_{i=1}^{m}x_i^2-4m\cdot\bar{x}\cdot\sum_{i=1}^{m}x_i=4m\left(\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}x_i\bar{x}\right)=4m\sum_{i=1}^{m}(x_i^2-x_i\bar{x})$$

由于$\sum_{i=1}^{m}x_i\bar{x}=\bar{x}\sum_{i=1}^{m}x_i=\bar{x}\cdot m\cdot\frac{1}{m}\cdot\sum_{i=1}^{m}x_i=m\bar{x}^2=\sum_{i=1}^{m}\bar{x}^2$

$$=4m\sum_{i=1}^{m}(x_i^2-x_i\bar{x}-x_i\bar{x}+x_i\bar{x})=4m\sum_{i=1}^{m}(x_i^2-x_i\bar{x}-x_i\bar{x}+\bar{x}^2)=4m\sum_{i=1}^{m}(x_i-\bar{x})^2$$

所以$4m\sum_{i=1}^{m}(x_i-\bar{x})^2\geqslant 0$，Hessian（海塞）矩阵$\nabla^2E_{(w,b)}$的所有顺序主子式均非负，该矩阵为半正定矩阵，进而$E_{(w,b)}$是关于$w$和$b$的凸函数得证。

求解w和b

凸充分性定理：若$f:\mathbb{R}^n\rightarrow\mathbb{R}$是凸函数，且$f(\boldsymbol{x})$一阶连续可微，则$\boldsymbol{x}^*$是全局解的充分必要条件是$\nabla f(\boldsymbol{x}^*)=\mathbf{0}$

所以，$\nabla E_{(w,b)}=\mathbf{0}$的点即为最小值点，也即

$$\nabla E_{(w,b)}=\begin{bmatrix}\frac{\partial E_{(w,b)}}{\partial w}\\ \frac{\partial E_{(w,b)}}{\partial b}\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$$

求解w和b

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right) = 0$$

$$mb - \sum_{i=1}^{m}(y_i - wx_i) = 0$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i) \text{公式3.8}$$

为了便于后续求解w，在此对b进行化简

$$b = \frac{1}{m}\sum_{i=1}^{m}y_i - w \cdot \frac{1}{m}\sum_{i=1}^{m}x_i = \bar{y} - w\bar{x}$$

求解w和b

$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}\left(y_i - b\right)x_i\right) = 0$$

$$w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)x_i = 0$$

$$w\sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} y_i x_i - \sum_{i=1}^{m} b x_i$$

把$b = \bar{y} - w\bar{x}$代入上式可得

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

求解w和b

$$w\sum_{i=1}^{m}x_i^2=\sum_{i=1}^{m}y_ix_i-\bar{y}\sum_{i=1}^{m}x_i+w\bar{x}\sum_{i=1}^{m}x_i$$

$$w\sum_{i=1}^{m}x_i^2-w\bar{x}\sum_{i=1}^{m}x_i=\sum_{i=1}^{m}y_ix_i-\bar{y}\sum_{i=1}^{m}x_i$$

$$w(\sum_{i=1}^{m}x_i^2-\bar{x}\sum_{i=1}^{m}x_i)=\sum_{i=1}^{m}y_ix_i-\bar{y}\sum_{i=1}^{m}x_i$$

求解w和b

$$w=\frac{\sum_{i=1}^{m} y_i x_i-\bar{y}\sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} x_i^2-\bar{x}\sum_{i=1}^{m} x_i}$$

其中

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

$$\bar{x}\sum_{i=1}^{m} x_i=\frac{1}{m}\sum_{i=1}^{m} x_i\sum_{i=1}^{m} x_i=\frac{1}{m}(\sum_{i=1}^{m} x_i)^2$$

所以

$$w=\frac{\sum_{i=1}^{m} y_i x_i-\bar{x}\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} x_i^2-\frac{1}{m}(\sum_{i=1}^{m} x_i)^2}=\frac{\sum_{i=1}^{m} y_i(x_i-\bar{x})}{\sum_{i=1}^{m} x_i^2-\frac{1}{m}(\sum_{i=1}^{m} x_i)^2} \quad \text{公式3.7}$$

求解w和b

w的向量化参见《机器学习公式详解》（南瓜书）式（3.7）的解析~

BTW

Datawhale 异步社区

机器学习三要素：

1. 模型：根据具体问题，确定假设空间
2. 策略：根据评价标准，确定选取最优模型的策略（通常会产出一个“损失函数”）
3. 算法：求解损失函数，确定最优模型

预告

下一节：多元线性回归

西瓜书对应章节：3.2

结束语

欢迎加入【南瓜书读者交流群】，我们将在群里进行答疑、勘误、本次直播回放、本次直播PPT发放、下次直播通知等最新资源发放和活动通知。

加入步骤：

1. 关注公众号【Datawhale】，发送【南瓜书】三个字获取机器人“小豚”的微信二维码
2. 添加“小豚”为微信好友，然后对“小豚”发送【南瓜书】三个字即可自动邀请进群