第3章 机器学习的起点：线性回归

机器学习作为人工智能的研究核心，通过模拟或实现人类的学习行为以获取新的知识或技能，并且重新组织已有的知识结构来不断改善自身的性能。一般来说，机器学习分为监督学习、半监督学习、无监督学习和强化学习等。本章介绍的线性回归是一种监督学习方法，它是机器学习的基石，很多复杂的机器学习算法都一定程度构建在线性回归基础上。

机器学习中的两个常见的问题：回归任务和分类任务。那什么是回归任务和分类任务？在研究两个或两个以上变量之间的关系时，通常需要由一个或几个变量来预测另一个变量。换句话说，对于一组数据和其标记值$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，需要使用$x_i$对$y_i$进行预测。如果$y_i$是连续的，则称为回归；如果$y_i$是离散的，则称为分类。简单地说，在监督学习中(也就是有标签的数据中)，标记值为连续值时是回归任务，标记值为离散值时是分类任务。线性回归模型就是处理回归任务最基础的模型。

首先，本章结合实例从机器学习和统计学角度来介绍线性回归模型的建立方法；其次，具体介绍一元线性回归和多元线性回归的基本概念和一些回归模型参数的估计方法；最后，结合案例和习题加深对线性回归的理解。

3.1 线性回归模型建立

线性回归(linear regression)的目的是建立尽可能准确预测实值输出标记的线性模型。许多非线性模型也可以在线性模型的基础上通过引入层级结构或高维映射得到。总的来说，线性回归虽然形式简单，但蕴含着机器学习的一些基本思想。

首先，用一个简单的例子来介绍线性回归模型。表3.1所示的数据为玩具厂某工人制作玩偶的数量和成本之间的关系。仔细研究所记录的数据，似乎玩偶个数和成本是线性关系。这也很符合我们的预期，为了更加直观地了解数据，验证猜测，将数据可视化，即把上面的数据点表示在一个直角坐标系中，如图3.1所示。

表3.1 生产记录表

日　　期	玩偶个数	成　　本	第　几　天
4月1日	10	7.7	1
4月2日	10	9.87	2
4月3日	11	10.87	3
4月4日	12	12.18	4
4月5日	13	11.43	5
4月6日	14	13.36	6
4月7日	15	15.15	7
4月8日	16	16.73	8
4月9日	17	17.4	9
…	…	…	…

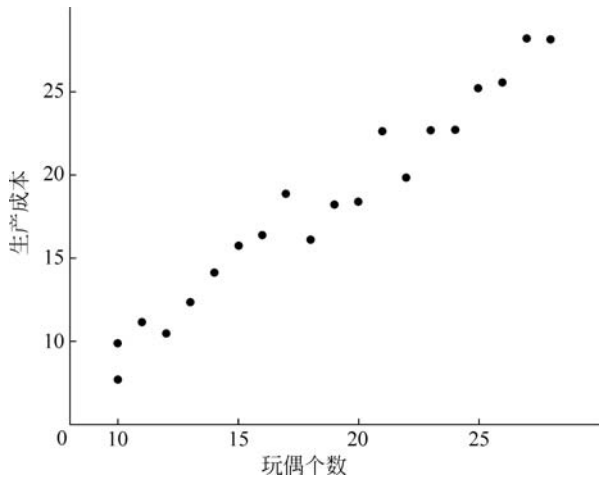

图 3.1　玩偶个数与生产成本图

由图 3.1 可以看出，生产成本和生产个数并不呈一条严格的直线，似乎是沿着某条直线上下随机地波动。其实上面展示的这些数据，是按照下面的数学公式产生的

$$y_i = x_i + \varepsilon_i \tag{3.1}$$

- x_i 是某一天制作的玩偶个数，y_i 是对应此天的生产成本。由图 3.1 可知，制作玩偶的平均成本是 1。
- 其中 $\{\varepsilon_i\}$ 是一个随机变量，它服从期望为 0，方差为 1 的正态分布。它表示在生产玩偶时，一些随机产生或随机节约的成本。比如对于制作失败的玩偶，ε_i 为正；又如制作过程中刚好发现有可用的旧布料，此时 ε_i 为负。
- ε_i 所代表的随机成本和制作玩偶的个数是相互独立的。

为了分析数据之间的关联性、趋势和统计特征等，需要采用适当的方法或模型对特定的数据进行分析。对于线性回归模型，一般可以从机器学习和统计学的两个角度进行分析。

3.1.1　机器学习角度

1. 确定场景类型

(1) 在现有的数据集中，有玩偶的生产个数(记为 x_i)和生产成本(记为 y_i)。其中，i 表示第 i 天的数据，x_i 表示第 i 天生产的玩偶数。我们的目的是通过生产个数的信息去预测生产成本。

(2) 因为需要被预测的成本是一个数量。它是一个连续变化的量，而并非表示类别的离散量，所以这是一个回归问题。

2. 定义损失函数(loss function)

(1) 搭建模型的目标是使得模型预测的生产成本和实际成本接近。

(2) 换用数学语言描述上述问题,已知实际成本为 y_i,假设模型预测的成本为 $\hat{y}_i$。定义如式(3.2)所示的损失函数,表示预测值和真实值的差。

$$\mathrm{LL}=\sum_i(y_i-\hat{y}_i)^2 \tag{3.2}$$

而搭建模型的目标就是使损失函数达到最小值。

(3) 式(3.2)并不是一个处处可导的函数,数学上处理起来比较麻烦。因此,重新定义一个数学上容易处理的损失函数,即真实值与预测值之间的欧几里得距离平方和,如式(3.3)所示。

$$L=\sum_i(y_i-\hat{y}_i)^2 \tag{3.3}$$

下面模型参数的估计就依赖于此损失函数。

3. 提取特征

(1) 经过检查,提供的数据中没有记错或者特别异常的情况。换句话说,数据可以直接使用。

(2) 在现有的数据中,只有一个表示个数的原始特征 $X=\{x_i\}$。在这个变量上面的加减乘除是有明确含义的。也就是说,变量本身的数学运算是有意义的,所以 X 可以直接在模型里面使用。

(3) 也可以对 X 做某种数学变换,得到一个新的特征。比如,对它做平方运算得到新的特征 X^2。这些新提取的特征也可以被应用到模型中。但对于此问题,我们先只用原始特征建模。如果效果不好,则再考虑提取新的特征。

4. 确定模型形式并估计参数

(1) 根据分析,x_i 和 y_i 之间是线性关系。因此,可以直接使用线性模型,不需要考虑非线性问题到线性问题的转化。

(2) 模型的定义如式(3.4)所示,其中,a 表示生产一个玩偶的变动成本,b 表示生产的固定成本

$$\hat{y}_i=ax_i+b \tag{3.4}$$

(3) 参数的估计值为使得损失函数达到最小值的情况,如式(3.5)所示。

$$(\hat{a},\hat{b})=\mathrm{argmin}_{a,b}\sum_i(y_i-ax_i-b)^2 \tag{3.5}$$

5. 评估模型效果

(1) 从预测的角度看,我们希望模型的预测成本越接近真实成本越好。所以,定义如式(3.6)的线性模型均方差,均方差越小,模型效果越好。

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{1}{n}L \tag{3.6}$$

(2) 从解释数据的角度来看，我们希望模型能最大限度地解释成本变化的原因。换句话说，未被模型解释的成本($y_i-\hat{y}_i$)占成本变化($y_i-\frac{1}{n}\sum y_i$)的比例越小越好。因此，定义模型的决定系数 R^2。决定系数越接近 1，模型的效果越好。

$$\begin{aligned}\bar{y}&=\frac{1}{n}\sum_{i=1}^{n}y_i\\ \mathrm{SS}_{\mathrm{tot}}&=\sum_i(y_i-\bar{y})^2\\ \mathrm{SS}_{\mathrm{res}}&=\sum_i(y_i-\hat{y}_i)^2\\ R^2&=1-\frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}\end{aligned} \tag{3.7}$$

3.1.2 统计学角度

1. 假设条件概率

(1) 根据描述，数据集中有两个变量：一个是自变量玩偶个数(记为 x_i)；另一个是因变量生产成本(记为 y_i)，其中，i 表示第 i 天的数据，比如 x_i 表示第 i 天生产的玩偶数。根据前面的分析，y_i 和 x_i 两者之间似乎是线性关系，但又带着一些随机波动。因此可以假设关系如下：

$$y_i=ax_i+b+\varepsilon_i \tag{3.8}$$

(2) 在式(3.8)中，a 和 b 是模型的参数，分别表示生产一个玩偶的变动成本和固定成本；而 ε_i 被称为噪声项，表示没被已有数据捕捉到的随机成本。它服从期望为 0、方差为 σ^2(σ^2 也是模型的参数)的正态分布，记为 $\varepsilon_i\sim N(0,\sigma^2)$。这里假设$\{\varepsilon_i\}$之间相互独立，而且$\{\varepsilon_i\}$和$\{x_i\}$之间也是相互独立的，这两点假设非常重要。

(3) 从左到右看式(3.8)，如果给定一组参数 a，b 以及噪声项的方差 σ^2。由于 x_i 表示玩偶个数，是一个确定的量。那么 y_i 就与 ε_i 一样是一个随机变量，服从期望为 ax_i+b，方差为 σ^2 的正态分布，即 $y_i\sim N(ax_i+b,\sigma^2)$。换句话说，提供的数据 y_i 只是 $N(ax_i+b,\sigma^2)$这个正态分布的一个观测值，而且$\{y_i\}$之间也是互相独立的。

(4) 把上面的第(3)点翻译成数学语言就是：y_i 在已知 a，b，x_i，σ 时的条件概率是 $N(ax_i+b,\sigma^2)$，如式(3.9)所示。

$$P(y_i\mid a,b,x_i,\sigma^2)\sim N(ax_i+b,\sigma^2) \tag{3.9}$$

2. 估计参数

(1) 根据上面的分析，$\{y_i\}$之间也是相互独立的。所以得到$\{y_i\}$出现的联合概率如

式(3.10)所示。此概率称为模型的似然函数(likelihood function),通常也记为 L。

$$P(Y \mid a,b,X,\sigma^2)=\prod P(y_i \mid a,b,x_i,\sigma^2) \tag{3.10}$$

$$\ln P(Y \mid a,b,X,\sigma^2)=-0.5n\ln(2\pi\sigma^2)-(1/2\sigma^2)\sum_i (y_i-ax_i-b)^2 \tag{3.11}$$

(2) 对于不同的模型参数,$\{y_i\}$出现的概率(即参数的似然函数)并不相同。这个概率当然是越大越好,所以使这个概率最大的参数将是参数估计的最佳选择。此方法也称为极大似然估计法(Maximum Likelihood Estimation,MLE)。根据式(3.10),参数(a,b)的估计值$(\hat{a},\hat{b})$如下:

$$(\hat{a},\hat{b})=\text{argmax}_{a,b}P(Y \mid a,b,X,\sigma^2)=\text{argmin}_{a,b}\sum_i (y_i-ax_i-b)^2 \tag{3.12}$$

(3) 同理,可以得到参数 σ^2 的估计值 $\hat{\sigma^2}$,如式(3.13)所示。

$$\begin{cases}\hat{\sigma^2}=\text{argmax}_{a,b}P(Y \mid a,b,X,\sigma^2)=\dfrac{\sum_{i=1}^{n}(y_i-\hat{y}_i)}{n} \\ \hat{y}_i=\hat{a}x_i+\hat{b}\end{cases} \tag{3.13}$$

3. 推导参数的分布

(1) 其实上面得到的参数估计值$(\hat{a},\hat{b},\hat{\sigma^2})$都是随机变量。具体的推导过程有些烦琐,限于篇幅,这里略去数学细节。仅以 $\hat{a}$ 为例,用一个不太严谨的数学推导来说明这个问题。假设数据集里只有两对数据(x_k,y_k),(x_l,y_l),且 $x_k \neq x_l$。这时可以通过解式(3.13)中的方程组来得到表达式。式(3.13)的右半部分表示 $\hat{a}$ 是一个随机变量,而且服从一个以参数真实值 a 为期望的正态分布。由$\begin{cases}y_k=\hat{a}x_k+\hat{b} \\ y_l=\hat{a}x_l+\hat{b}\end{cases}$可得

$$\hat{a}=\frac{y_k-y_l}{x_k-x_l}=\frac{a(x_k-x_l)+\varepsilon_k-\varepsilon_l}{x_k-x_l}=a+\frac{\varepsilon_k-\varepsilon_l}{x_k-x_l} \tag{3.14}$$

(2) 通过更加细致的数学运算可以得到

$$\begin{aligned}&\hat{b} \sim N(b,\sigma^2/n) \\ &\hat{a} \sim N\left(a,\sigma^2/\sum_i (x_i-\bar{x})^2\right) \\ &\hat{\sigma^2} \sim \chi^2_{n-2}\frac{\sigma^2}{n}\end{aligned} \tag{3.15}$$

(3) 既然参数估计值都是随机变量,那么我们更关心的是这些估计值所服从的概率分布,而不仅仅是根据式(3.12)和式(3.13)得到的数值。因为这些数值只是对应分布的一次观测值,它们并不总是等于真实参数,而是严重依赖于估计参数时所使用的数据。比如针对提供的数据集,只使用前 3 天数据估计出来的参数和使用 4～6 天数据估计出来的参数就不一样。

(4) 由式(3.15)可得，参数估计值的方差随着数据量的增大而减少。换句话说，数据量越大，模型估计的参数就越接近真实值。这也是大数据的价值之一：数据量越大，模型预测的效果就越好。

4. 假设检验与置信区间

(1) 参数的概率分布可以透露许多很有用的信息。比如在95%的情况下，参数 a 的真实值(生产一个玩偶的变动成本)会落在一个怎样的区间里？我们称此为参数 a 的95%置信区间。类似地，也可以计算模型预测结果的置信区间，即对于被预测对象，真实值的大致范围是怎样。这一点非常重要，因为模型几乎不可能准确地预测真实值。知道真实值的概率分布情况，能使我们更有信心地使用模型结果。

(2) 又比如在1%犯错的概率下，我们能不能拒绝参数的真实值其实等于0这个假设？在学术上它被称为参数的99%显著性假设检验。对于这个假设检验，更通俗一点的理解是：参数 b 的真实值等于0的概率是否小于1%？这可以帮助我们更好地理解数据之间的关系，比如生产玩偶时，固定成本(参数 b)是否真实存在？或者模型估计的固定成本 $\hat{b}$ 只是由于模型搭建得不准确而导致的“错误”结论？

上面由简单的例子入手，分别从机器学习和统计学两个角度分析数据之间的关联性、趋势和统计特征等。下面具体介绍一元线性回归和多元线性回归的基本概念和最小二乘法。

3.2 线性回归原理

线性回归通过对大量的观测数据进行处理，以得到比较符合事物内部规律的数学表达式。也就是说，寻找到数据与数据之间的规律，对结果进行预测。一元线性回归是研究由单个变量预测另一个与之有关的变量的回归分析。例如，如果已知广告费用和销售额之间的关系，当知道广告水平时，通过回归分析，可以预测销售额。

定义数据集 $D=\{(x_i,y_i)\}_{i=1}^{m}$，其中 $x_i\in\mathbf{R}$，$y_i\in\mathbf{R}$。对离散属性，若属性值间存在“序”(order)关系，可通过连续化将其转化为连续值，例如二值属性“身高”的取值“高”“矮”可转化为{1,0}，三值属性“高度”的取值“高”“中”“低”可转化为{1,0.5,0}。

线性回归试图学得

$$f(x_i)=wx_i+b,\quad \text{使得 } f(x_i)\simeq y_i \tag{3.16}$$

即通过衡量 $f(x)$ 与 y 之间的差别来确定 w 和 b。其中，均方误差是回归任务中最常用的性能度量，因此可试图让均方误差最小化，即

$$\begin{aligned}(w^*,b^*)&=\operatorname{argmin}_{w,b}\sum_{i=1}^{m}(f(x_i)-y_i)^2\\&=\operatorname{argmin}_{w,b}\sum_{i=1}^{m}(y_i-wx_i-b)^2\end{aligned} \tag{3.17}$$

均方误差对应了常用的欧几里得距离或简称“欧氏距离”(Euclidean distance)。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”(least square method)。在线性回归中,最小二乘法就是试图找到一条直线,使所有样本到直线上的欧氏距离之和最小。

求解 w 和 b 使 $E_{(w,b)}=\sum_{i=1}^{m}(y_i-wx_i-b)^2$ 最小化的过程,称为线性回归模型的最小二乘法“参数估计”(parameter estimation)。可将 $E_{(w,b)}$ 分别对 w 和 b 求导,得到

$$\frac{\partial E_{(w,b)}}{\partial w}=2\Big(w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}(y_i-b)x_i\Big) \tag{3.18}$$

$$\frac{\partial E_{(w,b)}}{\partial b}=2\Big(mb-\sum_{i=1}^{m}(y_i-wx_i)\Big) \tag{3.19}$$

然后令式(3.18)和式(3.19)为零,可得到 w 和 b 最优解的闭式(closed-form)解

$$w=\frac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}\big(\sum_{i=1}^{m}x_i\big)^2}=\frac{\sum_{i=1}^{m}x_iy_i-m\bar{y}}{\sum_{i=1}^{m}x_i^2-m\bar{x}^2} \tag{3.20}$$

$$b=\frac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)=\bar{y}-w\bar{x} \tag{3.21}$$

其中,$\bar{x}=\frac{1}{m}\sum_{i=1}^{m}x_i$ 为 x 的均值,$\bar{y}=\frac{1}{m}\sum_{i=1}^{m}y_i$ 为 y 的均值。

如果由两个或两个以上变量预测另一个与之有关的变量,则称为多元线性回归。此时,定义数据集 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{m}$,其中,$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{id})$,$y_i\in\mathbf{R}$,即样本由 d 个属性描述。此时我们试图学得

$$f(\boldsymbol{x}_i)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b,\quad \text{使得 } f(\boldsymbol{x}_i)\approx y_i \tag{3.22}$$

类似地,可利用最小二乘法来对 $\boldsymbol{w}$ 和 b 进行估计。为便于讨论,把 $\boldsymbol{w}$ 和 b 吸收入向量形式 $\hat{\boldsymbol{w}}=(w;b)$,相应地,把数据集 D 表示为一个 $m\times(d+1)$ 大小的矩阵 $\boldsymbol{X}$,其中每行对应于一个示例,该行前 d 个元素对应于示例的 d 个属性值,最后一个元素恒置为 1,即

$$\boldsymbol{X}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1d}&1\\x_{21}&x_{11}&\cdots&x_{21}&1\\\vdots&\vdots&\ddots&\vdots&\vdots\\x_{m1}&x_{m1}&\cdots&x_{m1}&1\end{pmatrix}=\begin{pmatrix}\boldsymbol{x}_1^{\mathrm{T}}&1\\\boldsymbol{x}_2^{\mathrm{T}}&1\\\vdots&\vdots\\\boldsymbol{x}_m^{\mathrm{T}}&1\end{pmatrix} \tag{3.23}$$

再把标记也写成向量形式 $\boldsymbol{y}=(y_1,y_2,\cdots,y_m)$,则类似于式(3.17),有

$$\hat{\boldsymbol{w}}^{*}=\operatorname{argmin}_{\hat{\boldsymbol{w}}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{w}}) \tag{3.24}$$

令 $E_{\hat{\boldsymbol{w}}}=(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{w}})$,对 $\hat{\boldsymbol{w}}$ 求导可得

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial\hat{\boldsymbol{w}}}=2\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\hat{\boldsymbol{w}}-\boldsymbol{y}) \tag{3.25}$$

令上式为零可得 $\hat{\boldsymbol{w}}$ 最优解的闭式解,但由于涉及矩阵逆的计算,比单变量情形要复杂一

些。下面做一个简单的讨论。

当 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 为满秩矩阵(full-rank matrix)或正定矩阵(positive definite matrix)时，令式(3.25)为零，可得

$$\hat{\boldsymbol{w}}^{*}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{3.26}$$

其中$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 是矩阵$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$的逆矩阵。令$\hat{\boldsymbol{x}}_i=(\boldsymbol{x}_i,1)$，则最终学得的多元线性回归模型为

$$f(\hat{\boldsymbol{x}}_i)=\hat{\boldsymbol{x}}_i^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{3.27}$$

现实任务中，如果$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 不是满秩矩阵，常见的做法是引入正则化(regularization)项。

以上就是最小二乘法的数学原理，“二乘”表示取平方，“最小”表示损失函数最小。可以看出，有了损失函数，机器学习的过程被转化成对损失函数求最优解过程，即求一个最优化问题。

例 3.1 高三·一班学生每周用于数学学习的时间 x(单位：小时)与数学成绩 y(单位：分)之间有如表 3.2 所示的对应关系。

表 3.2 学习时间与学习成绩关系表

学习时间 x	24	15	23	19	16	11	20	16	17	13
学习成绩 y	92	79	97	89	64	47	83	68	71	59

如果 y 与 x 之间具有线性相关关系，求回归线性方程。

解：从表 3.2 中可以看出：同样是每周用 16 小时学数学，一位同学的成绩是 64 分，另一位却是 68 分，这反映了 y 和 x 只有相关关系，没有函数关系。列出表 3.3。

表 3.3 学习时间与学习成绩关系表

i	1	2	3	4	5	6	7	8	9	10
x_i	24	15	23	19	16	11	20	16	17	13
y_i	92	79	97	89	64	47	83	68	71	59

经计算可得$\bar{x}=\frac{1}{10}\sum_{i=1}^{10}x_i=17.4$，$\bar{y}=\frac{1}{10}\sum_{i=1}^{10}y_i=74.9$，$\sum_{i=1}^{10}x_i^2=3182$，$\sum_{i=1}^{10}y_i^2=58375$，$\sum_{i=1}^{10}x_iy_i=13578$。设回归直线方程 $\hat{y}=wx+b$，代入式(3.20)和式(3.21)可得

$$w=\frac{\sum_{i=1}^{10}x_iy_i-10\bar{y}}{\sum_{i=1}^{10}x_i^2-10\bar{x}^2}=\frac{545.4}{154.4}\approx 3.53 \tag{3.28}$$

$$b=\bar{y}-w\bar{x}=74.9-3.53\times 17.4\approx 13.5 \tag{3.29}$$

因此所求的回归直线方程是 $\hat{y}=3.53x+13.5$。

例 3.2 某公司的企业管理费主要取决于两种重点产品的产量，其管理费用与产量的关系表如表 3.4 所示，试估计企业管理费线性回归模型。

表 3.4　管理费用与产量关系表

年　份	企业管理费 Y(千元)	甲产品产量 X_1(万吨)	甲产品产量 X_2(万吨)
1	3	3	5
2	1	1	4
3	8	5	6
4	3	2	4
5	5	4	6

解：令 $\boldsymbol{Y}=\begin{bmatrix}3\\1\\8\\3\\5\end{bmatrix}$，$\boldsymbol{X}=\begin{bmatrix}1&3&5\\1&1&4\\1&5&6\\1&2&4\\1&4&6\end{bmatrix}$

求解可得

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}=\begin{bmatrix}5&15&25\\15&55&81\\25&81&129\end{bmatrix},\quad \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\begin{bmatrix}20\\76\\109\end{bmatrix},\quad (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}=\begin{bmatrix}26.7&4.5&-8.0\\4.5&1.0&-1.5\\-8.0&-1.5&2.5\end{bmatrix}$$

结合式(3.26)，可得

$$\hat{\boldsymbol{w}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\begin{bmatrix}26.7&4.5&-8.0\\4.5&1.0&-1.5\\-8.0&-1.5&2.5\end{bmatrix}\begin{bmatrix}20\\76\\109\end{bmatrix}=\begin{bmatrix}4\\2.5\\-1.5\end{bmatrix}$$

因此，回归模型为 $\hat{y}=4+2.5x_1-1.5x_2$。

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。它是统计学中最基础的数学模型，但却蕴含着机器学习中一些重要的基本思想。更为重要的是，线性模型的易解释性使得它在物理学、经济学、商学等领域中占据了难以取代的地位。线性回归的一种最直观解法是最小二乘法，其损失函数是误差的平方，具有最小值点，可以通过解矩阵方程求得这个最小值。简单来说，线性回归就是选择一条线性函数来很好地拟合已知数据并预测未知数据。

线性回归的优点是：建模速度快，不需要很复杂的计算，在数据量大的情况下依然运行速度很快；可以根据系数给出每个变量的理解和解释。建立出来的线性模型可以直观地表达自变量和因变量之间的关系。缺点是不能很好地拟合非线性数据，因此在使用时需要先判断变量之间是否是线性关系。因此，对于不能用线性回归很好进行拟合的数据，可以使用机器学习的其他算法进行预测，后面将深入介绍这些内容。

习题

1. 统计学角度分析时，式(3.8)中随机误差项 ε_i 的意义是什么？

2. 一台机器可以按不同的速度运转，且通常其生产的物件会有一定的次品比例。每

小时生产次品物件的多少会随机器运转速度的变化而不同，实验结果如表 3.5 所示。

表 3.5　机器运转速度与每小时生产次品数关系表

速度	8	12	14	16
每小时生产次品数	5	8	9	11

（1）求机器速度影响每小时生产次品物件数的线性回归方程；

（2）若实际生产中所允许的每小时最大次品数不超过 10，那么，机器的速度不超过多少转/秒？

3. 设有模型 $y=b_0+b_1x_1+b_2x_2+u$，试在下列条件下：

（1）$b_1+b_2=1$；

（2）$b_1=b_2$。

分别求出 b_1 和 b_2 的最小二乘估计量。

参考文献

[1] Allwein E L, Schapire R E, Singer Y. Reducing multiclass to binary: A unifying approach for margin classifiers[J]. *Journal of Machine Learning Research*, 2000(1): 113-141.

[2] Boyd S, Vandenberghe L. Convex Optimization[M]. Cambridge, UK: Cambridge University Press, 2004.

[3] Chawla N V, Bowyer K W, Hall L O, et al., SMOTE: Synthetic minority over-sampling technique [J]. *Journal of Artificial Intelligence Research*, 2002(16): 321-357.

[4] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[5] Freedman D A, Pisami R, Purves R. 统计学[M]. 北京：中国统计出版社，1997.

[6] 李航. 统计学习办法[M]. 北京：清华大学出版社，2012.

[7] 唐亘. 精通数据科学：从线性回归到深度学习[M]. 北京：人民邮电出版社，2018.