

## 1 极大似然估计基础

给定一个随机变量  $X$  的样本集  $T = \{X_1, \dots, X_n\}$ , 随机变量  $X$  满足参数为  $\theta$  的概率分布  $P(X|\theta)$ 。如果所有样本对于  $\theta$  条件独立, 那么样本集的联合概率分布可以写成:

$$P(X_1, \dots, X_n|\theta) = \prod_{i=1}^n P(X_i|\theta) \quad (1)$$

考虑到样本集的概率与  $\theta$  有关, 不同的  $\theta$  值会导致不同的概率, 由于采样到了样本集  $T$ , 我们可以认为  $\theta$  等于使样本集概率最大的  $\theta^*$  的概率最大。极大似然估计法就是基于这样一个思想来选取这样的  $\theta^*$  作为  $\theta$  的估计值, 使被采样出的样本出现的可能性最大。

极大似然估计法常用于估计模型中的未知参数。假设我们要求某一个模型中一个未知的参数  $\theta$ , 我们可以定义出以特定参数集为条件观察给定事件的概率, 即我们从实际中观察到了一系列的结果, 并推出这些结果出现的概率。那么我们选择使出现概率最大的参数值作为参数  $\theta$  的估计值, 可以写成

$$\hat{\theta} = \arg \max_{\theta} P(X_1, \dots, X_n|\theta) \quad (2)$$

函数  $L(\theta) = P(X_1, \dots, X_n|\theta)$  称之为似然函数。样本之间往往满足独立同分布的条件, 为了方便求导求最值, 常使用对数似然函数  $\log L(\theta) = \sum_{i=1}^n \log(P(X_i|\theta))$ , 它与似然函数是一致的, 而且更加便于计算。极大似然估计法为其他参数估计方法提供了一个标准。

## 2 离散型随机变量的极大似然估计

对于离散型随机变量, 可以使用极大似然分布来估计它概率分布的参数。本节以二项分布为例子, 来估算事件发生的概率  $p$ 。随机变量  $X \sim B(n, p)$ , 进行  $n$  次独立随机试验得到样本集  $T = \{x_1, \dots, x_n\}$ 。  $x_i = 1$  表示事件发生,  $x_i = 0$  表示事件没发生, 那么第  $i$  次试验结果出现的概率, 也就是质量密度函数可以表示为

$$P(x_i|p) = p^{x_i} (1-p)^{1-x_i}, x_i \in \{0, 1\} \quad (3)$$

令

$$x = \sum_{i=1}^n x_i \quad (4)$$

可以写出似然函数:

$$L(p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^x (1-p)^{n-x} \quad (5)$$

根据极大似然估计法, 接下来要最大化这个似然函数, 可以转化为最大化对数似然函数, 再转化为最小化负对数似然函数

$$\begin{aligned} \max L(p) &= \max p^x (1-p)^{n-x} \\ \Rightarrow \max l(p) &= \max \log L(p) = \max \log[p^x (1-p)^{n-x}] \\ \Rightarrow \min -l(p) &= \min -\log[p^x (1-p)^{n-x}] \end{aligned}$$

化简负对数似然函数:

$$\begin{aligned} -l(p) &= -\log(p^x) - \log((1-p)^{n-x}) \\ &= -x \log(p) - (n-x) \log(1-p) \end{aligned} \quad (6)$$

那么估计值  $\hat{p}$  可以写成

$$\hat{p} = \arg \min_p (-l(p)) \quad (7)$$

为求最值, 对负对数似然函数求导并让导数等于 0

$$\begin{aligned} \frac{d(-l(p))}{dp} &= -\frac{x}{p} + \frac{n-x}{1-p} = 0 \\ 0 &= \frac{-x(1-p) + p(n-x)}{p(1-p)} \\ 0 &= -x + px + pn - px \end{aligned} \quad (8)$$

最后我们可以得到估计值

$$\hat{p} = \frac{x}{n} \quad (9)$$

结果表明, 这个问题中的估计值就是实验中事件发生的频率。

### 3 连续型随机变量的极大似然估计

对于连续型随机变量，需使用概率密度函数来计算似然函数。以高斯分布为例子，随机变量  $X \sim N(\mu, \sigma^2)$ ，其概率密度函数为

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (10)$$

常用的联合概率分布只有多元高斯分布， $P$  元高斯分布的概率密度函数为

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{P/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (11)$$

其中  $\mu$  为各个随机变量均值的向量， $\Sigma$  为随机变量间的协方差矩阵。正态概率密度函数在均值处取到最大值，函数的等高线形如椭圆，如图1所示。

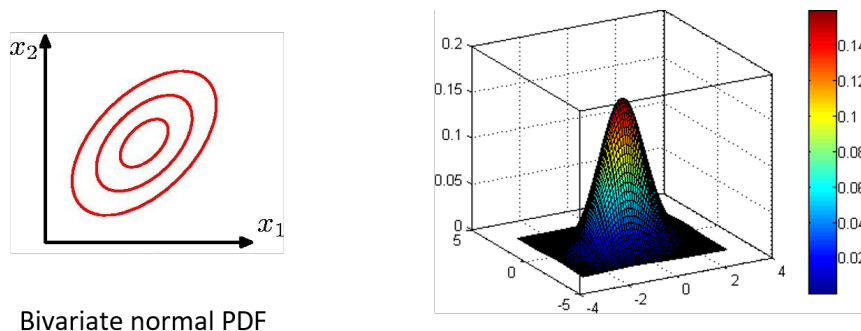


Figure 1: 二元高斯分布示意图

例 3.1. 二元高斯分布的概率密度函数为

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (12)$$

那么其中  $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ ,  $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ ,  $\rho$  为两变量之间的相关系数。二元高斯分布的曲面图，等高线图和散点图分别如图2，图3，图4所示。

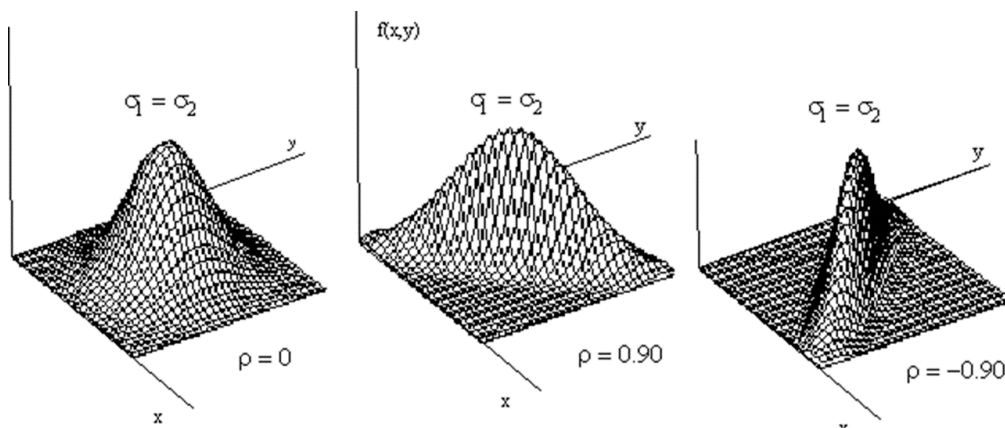


Figure 2: 二元高斯分布的曲面图

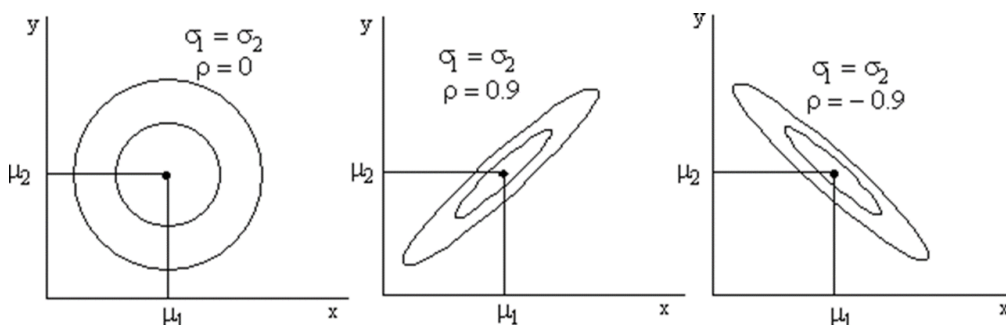


Figure 3: 二元高斯分布的等高线图

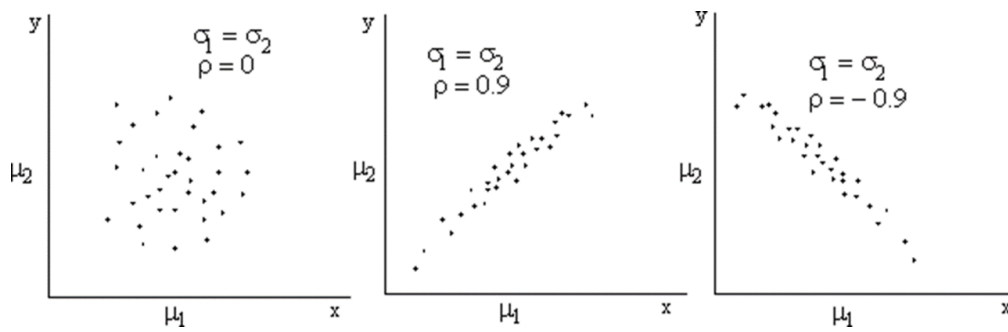


Figure 4: 二元高斯分布的散点图

对于一维的高斯分布，我们可以通过极大似然估计得到均值与方差的估计值

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

同样的，对于多元高斯分布，使用极大似然估计得到的均值和协方差矩阵为

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

#### 4 极大似然估计与线性回归中的标准方程法的联系

对于标准方程法而言，当从模型总体随机抽取  $n$  组样本观测值后，最合理的参数估计量应该使得模型能最好地拟合样本数据，也就是估计值和观测值之差的平方和最小。而对于极大似然估计，当从模型总体随机抽取  $n$  组样本观测值后，最合理的参数估计量应该使得从模型中抽取该  $n$  组样本观测值的概率最大。显然，这是从不同原理出发的两种参数估计法。

在极大似然估计中，通过选择参数，使已知数据在某种意义下最有可能出现，而某种意义通常指似然函数最大，而似然函数又往往指数据的概率分布函数。与标准方程法不同的是，极大似然估计需要已知这个概率分布函数，这在实践中是很困难的。一般假设其满足正态分布函数的特性，在这种情况下，极大似然估计与标准方程法相同。

**证明.** 已知拟合函数为  $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}, i = 1, 2, \dots, n$ ，我们假设误差服从正态分布，即  $\varepsilon \sim N(0, \sigma^2)$ ，因此有：

$$\begin{aligned} \therefore \varepsilon &= (\varepsilon^{(1)}, \varepsilon^{(2)}, \dots, \varepsilon^{(n)}) \\ \therefore P(\varepsilon^{(i)}) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) \\ \therefore P(y^{(i)}|x^{(i)}; \theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \end{aligned}$$

故而对于  $y$  来说， $P(y|x; \theta) = \prod_{i=1}^n P(y^{(i)}|x^{(i)}; \theta)$ ,

$$P(y|x; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \quad (13)$$

写出极大似然函数：

$$\begin{aligned} L(\theta) &= \log P(y|x; \theta) \\ &= -\frac{n}{2} \log 2\pi - n \log \sigma - \sum_{i=1}^n \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \\ \max_{\theta} L(\theta) &\Leftrightarrow \min_{\theta} \sum_{i=1}^n \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \Leftrightarrow \min_{\theta} \|y - X\theta\|^2 \end{aligned} \quad (14)$$

#### 附录 A：期望和方差的性质

最后我们简单介绍一下期望与方差常用的性质

## 简单相关系数

简单相关系数又被称为相关系数和线性相关系数，一般用于度量两个变量之间的线性关系，其表达式为：

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

其中， $Cov(X, Y)$  表示  $X$  和  $Y$  的协方差， $\sigma_x$  和  $\sigma_y$  表示  $X$  和  $Y$  的方差。相关系数的取值范围为：

$$-1 \leq \rho(X, Y) \leq 1$$

## 数学期望

数学期望又称均值或者简称期望，它反映了随机变量平均取值的大小，其满足性质：

$$E(X + Y) = E(X) + E(Y)$$

$$E(aX) = aE(X)$$

如果  $X$  和  $Y$  这两个变量是独立的，那么则有：

$$E(XY) = E(X)E(Y)$$

## 方差

方差在概率论中用于度量随机变量和其数学期望之间的偏离程度，其满足性质：

$$V(aX + b) = a^2 V(X)$$

如果  $X$  和  $Y$  这两个变量是独立的，那么则有：

$$V(X + Y) = V(X) + V(Y)$$

## 其他性质

- 当  $X = x$  时， $Y$  的条件期望为：

$$E(Y|X = x) = \int y * p(y|x) dy$$

- 总的期望可以用条件期望来表达：

$$E(Y) = E[E(Y|X)] = \int E(Y|X = x) p_x(x) dx$$

- 总的方差公式为：

$$V(Y) = V[E(Y|X)] + E[V(Y|X)]$$

证明.

$$\begin{aligned} V(Y) &= E(Y - E(Y))^2 \\ &= E(Y^2) - E(Y)^2 \\ &= E[E(Y^2|X)] - [E[E(Y|X)]]^2 \\ &= E[V(Y|X) + [E(Y|X)]^2] - [E[E(Y|X)]]^2 \\ &= E[V(Y|X)] + (E[E(Y|X)^2] - E[E(Y|X)]^2) \\ &= E[V(Y|X)] + V[E(Y|X)] \end{aligned}$$

## 引用

[1] Law of Total Variance [https://en.wikipedia.org/wiki/Law\\_of\\_total\\_variance](https://en.wikipedia.org/wiki/Law_of_total_variance)