WIKIPEDIA

最大似然估计

维基百科，自由的百科全书

在统计学中，**最大似然估计**，也称为**最大概似估计**，是用来估计一个概率模型的参数的一种方法。

目录

预备知识

下边的讨论要求读者熟悉概率论中的基本定义，如概率分布、概率密度函数、随机变量、数学期望等。同时，还要求读者熟悉连续实函数的基本技巧，比如使用微分来求一个函数的极值（即极大值或极小值）。

最大似然估计的原理

给定一个概率分布D，已知其概率密度函数（连续分布）或概率质量函数（离散分布）为f_D，以及一个分布参数θ，我们可以从这个分布中抽出一个具有n个值的采样$X_1, X_2, \ldots, X_n$，利用f_D计算出其似然函数：

$$\mathrm{lik}(\theta \mid x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n).$$

若D是离散分布，f_θ即是在参数为θ时观测到这一采样的概率。若其是连续分布，f_θ则为$X_1, X_2, \ldots, X_n$联合分布的概率密度函数在观测值处的取值。一旦我们获得$X_1, X_2, \ldots, X_n$，我们就能求得一个关于θ的估计。最大似然估计会寻找关于θ的最可能的值（即，在所有可能的θ取值中，寻找一个值使这个采样的“可能性”最大化）。从数学上来说，我们可以在θ的所有可能取值中寻找一个值使得似然函数取到最大值。这个使可能性最大的$\hat{\theta}$值即称为θ的**最大似然估计**。由定义，最大似然估计是样本的函数。

注意

- 这裡的似然函数是指$x_1, x_2, \ldots, x_n$不变时，关于θ的一个函数。
- 最大似然估计不一定存在，也不一定唯一。

例子

离散分布，离散有限参数空间

考虑一个抛硬币的例子。假设这个硬币正面跟反面轻重不同。我们把这个硬币抛80次（即，我们获取一个采样$x_1 = \mathrm{H}, x_2 = \mathrm{T}, \ldots, x_{80} = \mathrm{T}$并把正面的次数记下来，正面记为H，反面记为T）。并把抛出一个正面的概率记为p，抛出一个反面的概率记为$1-p$（因此，这裡的p即相当于上边的θ）。假设我们抛出了49个正面，31个反面，即49次H，31次T。假设这个硬币是我们从一个装了三个硬币的盒子里头取出的。这三个硬币抛出正面的概率分别为$p = 1/3$, $p = 1/2$, $p = 2/3$.这些硬币没有标记，所以我们无法知道哪个是哪个。使用**最大似然估计**，基于**二项分布**中的**概率质量函数**公式，通过这些试验数据（即采样数据），我们可以计算出哪个硬币的可能性最大。这个似然函数取以下三个值中的一个：

$$\begin{aligned}
\mathbb{P}(\mathrm{H}{=}49, \mathrm{T}{=}31 \mid p = 1/3) &= \binom{80}{49}(1/3)^{49}(1-1/3)^{31} \approx 0.000 \\
\mathbb{P}(\mathrm{H}{=}49, \mathrm{T}{=}31 \mid p = 1/2) &= \binom{80}{49}(1/2)^{49}(1-1/2)^{31} \approx 0.012 \\
\mathbb{P}(\mathrm{H}{=}49, \mathrm{T}{=}31 \mid p = 2/3) &= \binom{80}{49}(2/3)^{49}(1-2/3)^{31} \approx 0.054
\end{aligned}$$

我们可以看到当$\hat{p} = 2/3$时，似然函数取得最大值。这就是p的最大似然估计。

离散分布，连续参数空间

现在假设例子1中的盒子中有无数个硬币，对于$0 \le p \le 1$中的任何一个p， 都有一个抛出正面概率为p的硬币对应，我们来求其似然函数的最大值：

$$\mathrm{lik}(\theta) = f_D(\mathrm{H}{=}49, \mathrm{T}{=}80\text{-}49 \mid p) = \binom{80}{49} p^{49}(1-p)^{31}$$

其中$0 \le p \le 1$. 我们可以使用微分法来求最值。方程两边同时对p取微分，并使其为零。

$$\begin{aligned}
0 &= \frac{d}{dp}\left(\binom{80}{49} p^{49}(1-p)^{31}\right) \\
&\propto 49p^{48}(1-p)^{31} - 31p^{49}(1-p)^{30} \\
&= p^{48}(1-p)^{30}\left[49(1-p) - 31p\right]
\end{aligned}$$

其解为$p = 0$, $p = 1$ ，以及$p = 49/80$.使可能性最大的解显然是$p = 49/80$ （因为$p = 0$和$p = 1$这两个解会使可能性为零）。因此我们说**最大似然估计值**为$\hat{p} = 49/80$.

这个结果很容易一般化。只需要用一个字母t代替49用以表达伯努利试验中的被观察数据（即样本）的“成功”次数，用另一个字母n代表伯努利试验的次数即可。使用完全同样的方法即可以得到**最大似然估计值**:

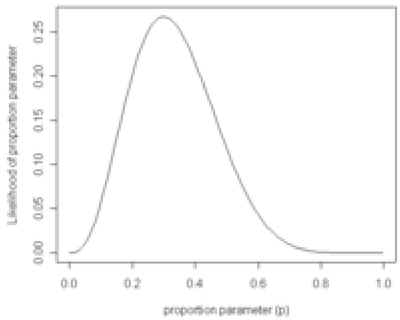
$$\hat{p} = \frac{t}{n}$$

对于任何成功次数为t，试验总数为n的伯努利试验。

连续分布，连续参数空间

最常见的连续概率分布是正态分布，其概率密度函数如下：

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



在不同比例参数值下一个二项式过程的可能性曲线t = 3, n = 10；其最大似然估计值发生在其众数并在曲线的最大值处。

现在有n个正态随机变量的采样点，要求的是一个这样的正态分布，这些采样点分布到这个正态分布可能性最大（也就是概率密度积最大，每个点更靠近中心点），其n个正态随机变量的采样的对应密度函数（假设其独立并服从同一分布）为：

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

或：

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right),$$

这个分布有两个参数：μ, σ^2.有人可能会担心两个参数与上边的讨论的例子不同，上边的例子都只是在一个参数上对可能性进行最大化。实际上，在两个参数上的求最大值的方法也差不多：只需要分别把可能性$\mathrm{lik}(\mu, \sigma) = f(x_1, , \ldots, x_n \mid \mu, \sigma^2)$在两个参数上最大化即可。当然这比一个参数麻烦一些，但是一点也不复杂。使用上边例子同样的符号，我们有$\theta = (\mu, \sigma^2)$.

最大化一个似然函数同最大化它的自然对数是等价的。因为自然对数log是一个连续且在似然函数的值域内严格递增的上凸函数。[注意：可能性函数（似然函数）的自然对数跟信息熵以及Fisher信息联系紧密。]求对数通常能够一定程度上简化运算，比如在这个例子中可以看到：

$$\begin{aligned}
0 &= \frac{\partial}{\partial\mu} \log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}}\right) \\
&= \frac{\partial}{\partial\mu}\left(\log\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right) \\
&= 0 - \frac{-2n(\bar{x}-\mu)}{2\sigma^2}
\end{aligned}$$

这个方程的解是$\widehat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i/n$.这的确是这个函数的最大值，因为它是$\mu$里头惟一的一阶导数等于零的点并且二阶导数严格小于零。

同理，我们对σ求导，并使其为零。

$$\begin{aligned}
0 &= \frac{\partial}{\partial\sigma} \log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}}\right) \\
&= \frac{\partial}{\partial\sigma}\left(\frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{\sigma^3}
\end{aligned}$$

这个方程的解是$\widehat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \widehat{\mu})^2/n$.

因此，其关于$\theta = (\mu, \sigma^2)$的*最大似然估计*为：

$$\hat{\theta} = (\widehat{\mu}, \widehat{\sigma}^2) = (\bar{x}, \sum_{i=1}^{n}(x_i - \bar{x})^2/n).$$

性质

泛函不变性（Functional invariance）

如果$\hat{\theta}$是θ的一个最大似然估计，那么$\alpha = g(\theta)$的最大似然估计是$\widehat{\alpha} = g(\hat{\theta})$.函数$g$无需是一个一一映射。请参见George Casella与Roger L. Berger所著的*Statistical Inference*定理Theorem 7.2.10的证明。（中国大陆出版的大部分教材上也可以找到这个证明。）

渐近线行为

最大似然估计函数在采样样本总数趋于无穷的时候达到最小方差（其证明可见于Cramer-Rao lower bound）。当最大似然估计非偏时，等价的，在极限的情况下我们可以称其有最小的均方差。 对于独立的观察来说，最大似然估计函数经常趋于正态分布。

偏差

最大似然估计的偏差是非常重要的。考虑这样一个例子，标有*1*到*n*的*n*张票放在一个盒子中。从盒子中随机抽取票。如果*n*是未知的话，那么*n*的最大似然估计值就是抽出的票上标有的*n*，尽管其期望值的只有$(n+1)/2$.为了估计出最高的*n*值，我们能确定的只能是*n*值不小于抽出来的票上的值。

历史

最大似然估计最早是由羅納德·費雪在1912年至1922年间推荐、分析并大范围推广的。[1]（虽然以前高斯、拉普拉斯、T. N. Thiele和F. Y. 埃奇沃思也使用过）。[2] 许多作者都提供了最大似然估计发展的回顾。[3]

大部分的最大似然估计理论都在贝叶斯统计中第一次得到发展，并被后来的作者简化。[1]

参见

- 均方差是衡量一个估计函数的好坏的一个量。
- 关于Rao-Blackwell定理（Rao-Blackwell theorem）的文章里头讨论到如何利用Rao-Blackwellisation过程寻找最佳非偏估计（即使均方差最小）的方法。而最大似然估计通常是一个好的起点。
- 读者可能会对最大似然估计（如果存在）总是一个关于参数的充分统计（sufficient statistic）的函数感兴趣。
- 最大似然估计跟一般化矩方法（generalized method of moments）有关。

参考文献

1. Pfanzagl (1994)
2. Edgeworth (September 1908) and Edgeworth (December 1908)
3. Savage (1976), Pratt (1976), Stigler (1978, 1986, 1999), Hald (1998, 1999), and Aldrich (1997)

外部链接

- 关于最大似然估计的历史的一篇论文，作者John Aldrich