第3-1章 点估计
矩估计和极大似然估计

《统计推断》 第7章

感谢清华大学自动化系江瑞教授提供PPT

内容

- 矩估计
- 极大似然估计 (MLE)

Point estimator

Point estimator

A **point estimator** is any function $W(X_1,\ldots,X_n)$ of a sample; that is, any statistic is a point estimator.

Estimator: a function of the sample, a random variable.

$$W(X_1,\ldots,X_n)$$

Estimate: the realized value of an estimator, a number.

$$W(x_1,\ldots,x_n)$$

Making predictions

- What is the purpose of doing point estimation?
  - Estimate the parameters associated with a parametric distribution so that we can get full knowledge of the population
  - With the parameters estimated, we can calculate the value of the probability density (mass) for future values of the observation
  - In machine learning, we say **density estimation**
- How to calculate the probability density for new observations?

$$p(x^{\text{new}} \mid \theta^{\star}, \mathbf{x}) = p(x^{\text{new}} \mid \theta^{\star})$$

$$p(\mathbf{x} \mid \theta) \Rightarrow \theta^{\star} = \arg\max p(\mathbf{x} \mid \theta) \Rightarrow p(x^{\text{new}} \mid \theta^{\star}, \mathbf{x}) = p(x^{\text{new}} \mid \theta^{\star})$$

Method of moments

Let $X_1, \ldots, X_n$ be a sample from a population with k parameters $f(x \mid \theta_1, \ldots, \theta_k)$. Define

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^1, \qquad \mu'_1 = \mathrm{E}X^1;$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \qquad \mu'_2 = \mathrm{E}X^2;$$

$$\cdots$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad \mu'_k = \mathrm{E}X^k.$$

Solve the system of equations for $(\theta_1, \ldots, \theta_k)$, in terms of $(m_1, \ldots, m_k)$

$$m_1 = \mu'_1(\theta_1, \ldots, \theta_k);$$

$$m_2 = \mu'_2(\theta_1, \ldots, \theta_k);$$

$$\cdots$$

$$m_k = \mu'_k(\theta_1, \ldots, \theta_k).$$

Maximum likelihood estimate

Because a larger likelihood implies a bigger plausibility that a parameter is the true one. It is reasonable to choose the parameter $\theta^\star$ that can maximize the likelihood function $L(\theta \mid \mathbf{x})$ as our best guess of θ.

In other words,

$$\theta^\star = \arg\max_{\theta \in \Theta} L(\theta \mid \mathbf{x}).$$

Equivalently,

$$\theta^\star = \arg\max_{\theta \in \Theta} \log L(\theta \mid \mathbf{x}).$$

Obviously,

$$L(\theta^\star \mid \mathbf{x}) \geq L(\theta \mid \mathbf{x}), \text{ for any } \theta \in \Theta.$$

$\theta^\star$ is called the *maximum likelihood estimate* (MLE) of θ.

计算问题

- 全局极大值还是局部极大值(优化问题)
  – 可微驻点(导数的零点)，边界点，不可微点
  – 导数为零，二阶导数<0.

- 数值敏感性问题：如果样本的微小变化引起估计的巨大变化，显然这样的估计是不可取的

Bernoulli MLE

- 设 $X_1, X_2, \cdots, X_n \sim Bernoulli(p)$. 则对数似然为

$$l(p|x) = \log L(p|x) = \log \left[\prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} \right]$$
$$= y \log p + (n-y) \log(1-p)$$

$$y = \sum_i x_i$$

$$\frac{dl(p|x)}{dp} = \frac{y}{p} - \frac{n-y}{1-p} = 0$$

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

频率代替概率

二项分布试验次数的MLE

- 设 $X_1, X_2, \cdots, X_n \sim B(k,p)$. 这里p已知但k未知. 似然函数如下

$$L(k|x,p) = \prod_{i=1}^{n} C_k^{x_i} p^{x_i} (1-p)^{k-x_i}$$

- 对似然函数求导很难，考虑比值

$$\frac{L(k|x,p)}{L(k-1|x,p)} = \frac{[k(1-p)]^n}{\prod_{i=1}^{n}(k-x_i)}$$

二项分布试验次数的MLE

- 于是极大值条件

$$\begin{cases} [k(1-p)]^n \geq \prod_{i=1}^{n}(k-x_i) \\ [(k+1)(1-p)]^n < \prod_{i=1}^{n}(k+1-x_i) \end{cases}$$

- 两边除以k^n,令z=1/k, 上式等价于求解方程

$$(1-p)^n = \prod_{i=1}^{n}(1-x_i z), 0 < z \leq 1/\max x_i$$

二项分布试验次数的MLE

- 令

$$F(z)=\prod_{i=1}^{n}(1-x_i z), f(z)=\log F(z)$$

$$f'(z)=-\sum_{i=1}^{n}\frac{x_i}{1-x_i z}<0$$

- 因此F(z)严格单调下降, 而

$$F(0)=1>(1-p)^n, F(1/\max\{x_i\})=0<(1-p)^n$$

- 于是由中值定理方程存在零点。

正态均值MLE

- 设 $X_1, X_2, \cdots, X_n \sim N(\theta, 1)$, 其似然函数

$$L(\theta|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\theta)^2} = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}$$

- 对数似然

$$l(\theta|x) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2$$

- 求导并令导数为零得最大似然估计

$$\hat{\theta} = \frac{\sum_i x_i}{n}, \quad \frac{d^2 l(\theta|x)}{d\theta^2}\Big|_{\theta=\hat{\theta}} = -1 < 0$$

正态分布MLE

- 设 $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$, 其似然函数

$$L(\mu, \sigma^2|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2} = \frac{1}{2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2}$$

- 对数似然

$$l(\mu, \sigma^2|x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2$$

正态分布MLE

- 求偏导并令其为零

$$\begin{cases} \dfrac{\partial \log l(\mu, \sigma^2|x)}{\partial \mu} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i - \mu) = 0 \\ \dfrac{\partial \log l(\mu, \sigma^2|x)}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases}$$

- 解得

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

全局最大值的验证

方式一：Profile likelihood

- 对任意 σ^2，由 $\sum(x_i-\theta)^2 > \sum(x_i-\overline{x})^2, \theta \neq \overline{x}$

$$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum(x_i-\overline{x})^2} \geq \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2}$$

- 转化为一维问题

全局最大值的验证

二元微积分最大值条件验证

- 偏导数为0
- 至少有一个二阶偏导数为负
- 二阶偏导数的Jacobi行列式为正

$$\frac{\partial^2 l(\mu,\sigma^2)}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 l(\mu,\sigma^2)}{\partial \mu \partial \sigma^2} = \frac{1}{\sigma^4}\sum(x_i-\mu)$$

$$\frac{\partial^2 l(\mu,\sigma^2)}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum(x_i-\mu)^2$$

全局最大值的验证

- Jacobi行列式

$$
\left|\begin{matrix} \frac{\partial^2 l(\mu,\sigma^2)}{\partial\mu^2} & \frac{\partial^2 l(\mu,\sigma^2)}{\partial\mu\partial\sigma^2} \\ \frac{\partial^2 l(\mu,\sigma^2)}{\partial\mu\partial\sigma^2} & \frac{\partial^2 l(\mu,\sigma^2)}{\partial(\sigma^2)^2} \end{matrix}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2}
$$

$$
= -\frac{n}{\hat{\sigma}^2}\left(\frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}n\hat{\sigma}^2\right)
$$

$$
= \frac{n^2}{2\hat{\sigma}^6} > 0
$$

均匀分布MLE

- 设 $X_1, X_2, \cdots, X_n \sim U\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, 似然函数

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} I_{\{x_i \in [\theta - \frac{1}{2}, \theta + \frac{1}{2}]\}} \\
&= I_{\{\theta - \frac{1}{2} \le x_{(1)} \le \cdots \le x_{(n)} \le \theta + \frac{1}{2}\}} \\
&= I_{\{x_{(n)} - \frac{1}{2} \le \theta \le x_{(1)} + \frac{1}{2}\}}
\end{aligned}
$$

- 即 θ 可取区间 $\left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$中的任何值，因此MLE不唯一。

参数函数的MLE

- 定理：若 $\hat{\theta}$ 是 θ 的极大似然估计，则对 θ 的任何函数 $\tau(\theta)$，$\tau(\hat{\theta})$ 是 $\tau(\theta)$ 的极大似然估计.
- 诱导似然函数

$$L^*(\eta|x) = \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|x)$$

- 证明：

$$\begin{aligned} L^*(\hat{\eta}|x) &= \sup_{\eta} L^*(\eta|x) \\ &= \sup_{\eta} \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|x) \\ &= \sup_{\theta} L(\theta|x) = L(\hat{\theta}) \end{aligned}$$

$$\begin{aligned} L(\hat{\theta}|x) &= \sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta})\}} L(\theta|x) \\ &= L^*(\tau(\hat{\theta})|x) \end{aligned}$$

$$\Rightarrow \quad \hat{\eta} = \tau(\hat{\theta})$$