Math 541: Statistical Theory II

Maximum Likelihood Estimation

Lecturer: Songfeng Zheng

1 Maximum Likelihood Estimation

Maximum likelihood is a relatively simple method of constructing an estimator for an unknown parameter θ. It was introduced by R. A. Fisher, a great English mathematical statistician, in 1912. Maximum likelihood estimation (MLE) can be applied in most problems, it has a strong intuitive appeal, and often yields a reasonable estimator of θ. Furthermore, if the sample is large, the method will yield an excellent estimator of θ. For these reasons, the method of maximum likelihood is probably the most widely used method of estimation in statistics.

Suppose that the random variables $X_1, \cdots, X_n$ form a random sample from a distribution $f(x|\theta)$; if X is continuous random variable, $f(x|\theta)$ is pdf, if X is discrete random variable, $f(x|\theta)$ is point mass function. We use the given symbol — to represent that the distribution also depends on a parameter θ, where θ could be a real-valued unknown parameter or a vector of parameters. For every observed random sample $x_1, \cdots, x_n$, we define

$$f(x_1, \cdots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta) \tag{1}$$

If $f(x|\theta)$ is pdf, $f(x_1, \cdots, x_n|\theta)$ is the joint density function; if $f(x|\theta)$ is pmf, $f(x_1, \cdots, x_n|\theta)$ is the joint probability. Now we call $f(x_1, \cdots, x_n|\theta)$ as the *likelihood function*. As we can see, the likelihood function depends on the unknown parameter θ, and it is always denoted as $L(\theta)$.

Suppose, for the moment, that the observed random sample $x_1, \cdots, x_n$ came from a discrete distribution. If an estimate of θ must be selected, we would certainly not consider any value of θ for which it would have been impossible to obtain the data $x_1, \cdots, x_n$ that was actually observed. Furthermore, suppose that the probability $f(x_1, \cdots, x_n|\theta)$ of obtaining the actual observed data $x_1, \cdots, x_n$ is very high when θ has a particular value, say $\theta = \theta_0$, and is very small for every other value of θ. Then we would naturally estimate the value of θ to be θ_0. When the sample comes from a continuous distribution, it would again be natural to try to find a value of θ for which the probability density $f(x_1, \cdots, x_n|\theta)$ is large, and to use this value as an estimate of θ. For any given observed data $x_1, \cdots, x_n$, we are led by this reasoning to consider a value of θ for which the likelihood function $L(\theta)$ is a maximum and to use this value as an estimate of θ.

The meaning of maximum likelihood is as follows. We choose the parameter that makes the likelihood of having the obtained data at hand maximum. With discrete distributions, the likelihood is the same as the probability. We choose the parameter for the density that maximizes the probability of the data coming from it.

Theoretically, if we had no actual data, maximizing the likelihood function will give us a function of n random variables $X_1, \cdots, X_n$, which we shall call "maximum likelihood estimate" $\hat{\theta}$. When there are actual data, the estimate takes a particular numerical value, which will be the maximum likelihood estimator.

MLE requires us to maximum the likelihood function $L(\theta)$ with respect to the unknown parameter θ. From Eqn. 1, $L(\theta)$ is defined as a product of n terms, which is not easy to be maximized. Maximizing $L(\theta)$ is equivalent to maximizing $\log L(\theta)$ because log is a monotonic increasing function. We define $\log L(\theta)$ as *log likelihood function*, we denote it as $l(\theta)$, i.e.

$$l(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i|\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

Maximizing $l(\theta)$ with respect to θ will give us the MLE estimation.

2 Examples

Example 1: Suppose that X is a discrete random variable with the following probability mass function: where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ.

Solution: Since the sample is (3,0,2,1,3,2,1,0,2,1), the likelihood is

$$\begin{aligned} L(\theta) = \quad & P(X=3)P(X=0)P(X=2)P(X=1)P(X=3) \\ \times \quad & P(X=2)P(X=1)P(X=0)P(X=2)P(X=1) \end{aligned} \tag{2}$$

Substituting from the probability distribution given above, we have

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

Clearly, the likelihood function $L(\theta)$ is not easy to maximize.

Let us look at the log likelihood function

$$\begin{aligned} l(\theta) &= \log L(\theta) = \sum_{i=1}^{n} \log P(X_i|\theta) \\ &= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) + 2\left(\log\frac{1}{3} + \log(1-\theta)\right) \\ &= C + 5\log\theta + 5\log(1-\theta) \end{aligned}$$

where C is a constant which does not depend on θ. It can be seen that the log likelihood function is easier to maximize compared to the likelihood function.

Let the derivative of $l(\theta)$ with respect to θ be zero:

$$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$$

and the solution gives us the MLE, which is $\hat{\theta} = 0.5$. We remember that the method of moment estimation is $\hat{\theta} = 5/12$, which is different from MLE.

Example 2: Suppose $X_1, X_2, \cdots, X_n$ are i.i.d. random variables with density function $f(x|\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|x|}{\sigma}\right)$, please find the maximum likelihood estimate of σ.

Solution: The log-likelihood function is

$$l(\sigma) = \sum_{i=1}^{n}\left[-\log 2 - \log\sigma - \frac{|X_i|}{\sigma}\right]$$

Let the derivative with respect to θ be zero:

$$l'(\sigma) = \sum_{i=1}^{n}\left[-\frac{1}{\sigma} + \frac{|X_i|}{\sigma^2}\right] = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}|X_i|}{\sigma^2} = 0$$

and this gives us the MLE for σ as

$$\hat{\sigma} = \frac{\sum_{i=1}^{n}|X_i|}{n}$$

Again this is different from the method of moment estimation which is

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}X_i^2}{2n}}$$

Example 3: Use the method of moment to estimate the parameters μ and σ for the normal density

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

based on a random sample $X_1, \cdots, X_n$.

Solution: In this example, we have two unknown parameters, μ and σ, therefore the parameter $\theta = (\mu, \sigma)$ is a vector. We first write out the log likelihood function as

$$l(\mu, \sigma) = \sum_{i=1}^{n} \left[-\log \sigma - \frac{1}{2}\log 2\pi - \frac{1}{2\sigma^2}(X_i - \mu)^2 \right] = -n \log \sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Setting the partial derivative to be 0, we have

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0$$

$$\frac{\partial l(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

Solving these equations will give us the MLE for μ and σ:

$$\hat{\mu} = \overline{X} \quad \text{and} \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

This time the MLE is the same as the result of method of moment.

From these examples, we can see that the maximum likelihood result may or may not be the same as the result of method of moment.

Example 4: The Pareto distribution has been used in economics as a model for a density function with a slowly decaying tail:

$$f(x|x_0, \theta) = \theta x_0^{\theta} x^{-\theta-1}, \quad x \geq x_0, \quad \theta > 1$$

Assume that $x_0 > 0$ is given and that $X_1, X_2, \cdots, X_n$ is an i.i.d. sample. Find the MLE of θ.

Solution: The log-likelihood function is

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \log f(X_i|\theta) = \sum_{i=1}^{n} (\log \theta + \theta \log x_0 - (\theta + 1)\log X_i) \\ &= n \log \theta + n\theta \log x_0 - (\theta + 1)\sum_{i=1}^{n} \log X_i \end{aligned}$$

Let the derivative with respect to θ be zero:

$$\frac{dl(\theta)}{d\theta} = \frac{n}{\theta} + n \log x_0 - \sum_{i=1}^{n} \log X_i = 0$$

Solving the equation yields the MLE of θ:

$$\hat{\theta}_{MLE} = \frac{1}{\overline{\log X} - \log x_0}$$

Example 5: Suppose that $X_1, \cdots, X_n$ form a random sample from a uniform distribution on the interval $(0, \theta)$, where of the parameter $\theta > 0$ but is unknown. Please find MLE of θ.

Solution: The pdf of each observation has the following form:

$$f(x|\theta) = \begin{cases} 1/\theta, & \text{for} \quad 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Therefore, the likelihood function has the form

$$L(\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{for} \quad 0 \leq x_i \leq \theta \quad (i = 1, \cdots, n) \\ 0, & \text{otherwise} \end{cases}$$

It can be seen that the MLE of θ must be a value of θ for which $\theta \geq x_i$ for $i = 1, \cdots, n$ and which maximizes $1/\theta^n$ among all such values. Since $1/\theta^n$ is a decreasing function of θ, the estimate will be the smallest possible value of θ such that $\theta \geq x_i$ for $i = 1, \cdots, n$. This value is $\theta = \max(x_1, \cdots, x_n)$, it follows that the MLE of θ is $\hat{\theta} = \max(X_1, \cdots, X_n)$.

It should be remarked that in this example, the MLE $\hat{\theta}$ does not seem to be a suitable estimator of θ. We know that $\max(X_1, \cdots, X_n) < \theta$ with probability 1, and therefore $\hat{\theta}$ surely underestimates the value of θ.

Example 6: Suppose again that $X_1, \cdots, X_n$ form a random sample from a uniform distribution on the interval $(0, \theta)$, where of the parameter $\theta > 0$ but is unknown. However, suppose now we write the density function as

$$f(x|\theta) = \begin{cases} 1/\theta, & \text{for} \quad 0 < x < \theta \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

We will prove that in this case, the MLE for θ does not exist.

Proof: The only difference between Eqn. 3 and Eqn. 4 is that the value of the pdf at the two endpints 0 and θ has been changed by replacing the weak inequalities in Eqn. 3 with strict inequalities in Eqn. 4. Either equation could be used as the pdf of the uniform distribution.

However, if Eqn. 4 is used as the pdf, then an MLE of θ will be a value of θ for which $\theta > x_i$ for $i = 1, \cdots, n$ and which maximizes $1/\theta^n$ among all such values. It should be noted that the possible values of θ no longer include the value $\theta = \max(x_1, \cdots, x_n)$, since θ must be *strictly* greater than each observed value x_i for $i = 1, \cdots, n$. Since θ can be chosen arbitrarily close to the value $\max(x_1, \cdots, x_n)$ but cannot be chosen equal to this value, it follows that the MLE of θ does not exist in this case.

Example 5 and 6 illustrate one shortcoming of the concept of an MLE. We know that it is irrelevant whether the pdf of the uniform distribution is chosen to be equal to $1/\theta$ over the open interval $0 < x < \theta$ or over the closed interval $0 \leq x \leq \theta$. Now, however, we see that the existence of an MLE depends on this typically irrelevant and unimportant choice. This difficulty is easily avoided in Example 5 by using the pdf given by Eqn. 3 rather than that given by Eqn. 4. In many other problems, as well, in which there is a difficulty of this type in regard to the existence of an MLE, the difficulty can be avoided simply by choosing one particular appropriate version of the pdf to represent the given distribution.

Example 7: Suppose that $X_1, \cdots, X_n$ form a random sample from a uniform distribution on the interval $(\theta, \theta + 1)$, where the value of the parameter θ is unknown $(-\infty < \theta < \infty)$. Clearly, the density function is

$$f(x|\theta) = \begin{cases} 1, & \text{for} \quad \theta \leq x \leq \theta + 1 \\ 0, & \text{otherwise} \end{cases}$$

We will see that the MLE for θ is not unique.

Proof: In this example, the likelihood function is

$$L(\theta) = \begin{cases} 1, & \text{for} \quad \theta \leq x_i \leq \theta + 1 \quad (i = 1, \cdots, n) \\ 0, & \text{otherwise} \end{cases}$$

The condition that $\theta \leq x_i$ for $i = 1, \cdots, n$ is equivalent to the condition that $\theta \leq \min(x_1, \cdots, x_n)$. Similarly, the condition that $x_i \leq \theta + 1$ for $i = 1, \cdots, n$ is equivalent to the condition that $\theta \geq \max(x_1, \cdots, x_n) - 1$. Therefore, we can rewrite the the likelihood function as

$$L(\theta) = \begin{cases} 1, & \text{for} \quad \max(x_1, \cdots, x_n) - 1 \leq \theta \leq \min(x_1, \cdots, x_n) \\ 0, & \text{otherwise} \end{cases}$$

Thus, we can select any value in the interval $[\max(x_1, \cdots, x_n) - 1, \min(x_1, \cdots, x_n)]$ as the MLE for θ. Therefore, the MLE is not uniquely specified in this example.

3 Exercises

Exercise 1: Let $X_1, \cdots, X_n$ be an i.i.d. sample from a Poisson distribution with parameter λ, i.e.,

$$P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Please find the MLE of the parameter λ.

Exercise 2: Let $X_1, \cdots, X_n$ be an i.i.d. sample from an exponential distribution with the density function

$$f(x|\beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \text{ with} \quad 0 \leq x < \infty.$$

Please find the MLE of the parameter β.

Exercise 3: Gamma distribution has a density function as

$$f(x|\alpha,\lambda)=\frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x},\ \text{with}\quad 0\le x<\infty.$$

Suppose the parameter α is known, please find the MLE of λ based on an i.i.d. sample $X_1,\cdots,X_n$.

Exercise 4: Suppose that $X_1,\cdots,X_n$ form a random sample from a distribution for which the pdf $f(x|\theta)$ is as follows:

$$f(x|\theta)=\begin{cases}\theta x^{\theta-1}, & \text{for}\quad 0<x<1\\ 0, & \text{for}\quad x\le 0\end{cases}$$

Also suppose that the value of θ is unknown ($\theta>0$). Find the MLE of θ.

Exercise 5: Suppose that $X_1,\cdots,X_n$ form a random sample from a distribution for which the pdf $f(x|\theta)$ is as follows:

$$f(x|\theta)=\frac{1}{2}e^{-|x-\theta|}\quad \text{for}\quad -\infty<x<\infty$$

Also suppose that the value of θ is unknown ($-\infty<\theta<\infty$). Find the MLE of θ.

Exercise 6: Suppose that $X_1,\cdots,X_n$ form a random sample from a distribution for which the pdf $f(x|\theta)$ is as follows:

$$f(x|\theta)=\begin{cases}e^{\theta-x}, & \text{for}\quad x>\theta\\ 0, & \text{for}\quad x\le\theta\end{cases}$$

Also suppose that the value of θ is unknown ($-\infty<\theta<\infty$). a) Show that the MLE of θ does not exist. b) Determine another version of the pdf of this same distribution for which the MLE of θ will exist, and find this estimate.

Exercise 7: Suppose that $X_1,\cdots,X_n$ form a random sample from a uniform distribution on the interval (θ_1,θ_2), with the pdf as follows:

$$f(x|\theta)=\begin{cases}\frac{1}{\theta_2-\theta_1}, & \text{for}\quad \theta_1\le x\le\theta_2\\ 0, & \text{otherwise}\end{cases}$$

Also suppose that the values of θ_1 and θ_2 are unknown ($-\infty<\theta_1<\theta_2<\infty$). Find the MLE's of θ_1 and θ_2.