Maximum Likelihood Estimation

Mustafa Hajij

The Likelihood Function

Setup: We have a random variable Y with a known type but an unknown parameter θ (typically a vector if the distribution of Y admits more than one parameter).

Let us consider a sample {Y1, ..,Yn} of i.i.d. random variables with the same arbitrary distribution as Y given above.

The realization of {Y1, ..,Yn} (the data set..) is denoted {y1, .., yn}.

Problem: Estimate the parameter θ.

Setting the scene : Denote by $p_Y(y;\theta)$to the PDF of the random variable Y. Since we are trying to estimate θ, we put θ in the notation.

The joint PDF of Y1, ..,Yn is can be written as follows :

$$p_{Y_1 \ldots Y_n}(y_1, \ldots, y_n; \theta) = p_Y(y_1, ; \theta) \ldots p_Y(y_n, ; \theta)$$

Maximal likelihood estimate

$$L(\theta; y_1, \dots, y_n) := p_{Y_1 \dots Y_n}(y_1, \dots, y_n; \theta) = p_Y(y_1, ; \theta) \dots p_Y(y_n, ; \theta)$$

The likelihood function is a function of the unknown parameter θ.

Definition : the value of $\theta = \ \theta_{MLE}$ that maximizes the function L is called the maximal likelihood estimate.

Method to find θ :

Using methods we learned in calculus. When θ is a high dimensional vector, we usually rely on optimization techniques.

Log likelihood

$$\text{log likelihood} = \ln(\text{likelihood}) = \ln(p_Y(y_1, ; \theta) \ldots p_Y(y_n, ; \theta))$$

$$= \ln(\sum_{i=0}^{n} (p_Y(y_i, ; \theta))$$

It is usually easier to deal with last quantity over the likelihood

Maximizing the likelihood function is the same as maximizing the log likelihood (why ?)

Remark : in practice when trying to find θ, one usually uses an optimization algorithm.

Example 1

- Suppose $x_1, x_2, \ldots, x_n$ is a random sample from an exponential distribution with parameter λ. Because of independence, the likelihood function is a product of the individual pdf's:

$$\begin{aligned} f(x_1, \ldots, x_n; \lambda) &= (\lambda e^{-\lambda x_1}) \cdot \cdots \cdot (\lambda e^{-\lambda x_n}) \\ &= \lambda^n e^{-\lambda \Sigma x_i} \end{aligned}$$

- The natural logarithm of the likelihood function is

- $$\ln[f(x_1, \ldots, x_n; \lambda)] = n \ln(\lambda) - \lambda \Sigma x_i$$

Example 1

- Equating $(d/d\lambda)[\ln(\text{likelihood})]$ to zero results in

$$n/\lambda - \Sigma x_i = 0, \text{ or } \lambda = n/\Sigma x_i =$$

$$\hat{\lambda} = 1/\overline{X};$$

Homework : check second derivative is negative

Example 2

- Let $x_1, \dots, x_n$ be a random sample from a normal distribution. The likelihood function is

$$f(x_1, \dots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdot \dots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\Sigma(x_i-\mu)^2/(2\sigma^2)}$$

- so

$$\ln[f(x_1, \dots, x_n; \mu, \sigma^2)] = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(x_i - \mu)^2$$

Example 2

- To find the maximizing values of μ and σ^2, we must take the partial derivatives of $\ln(f)$ with respect to μ and σ^2, equate them to zero, and solve the resulting two equations.

- Omitting the details (homework)

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

- Note that the MLE of σ^2 is not the unbiased estimator.

Example 3

Find MLE for $f(x|\theta)=1/\theta$ for $0 \le xi \le \theta$ assuming that we are giving the data x1,….,x10.

Solution : we know that :

$$L(\theta)=\theta^{-10}$$

Take the derivative of the log Likelihood wrt θ:

$$(d/d\theta)[\ln(\text{likelihood})] = -10/\theta<0$$ So L is a decreasing function

We are trying to find the max of L(θ) while satisfying the condition $0 \le xi \le \theta$. This implies that

$$\theta_{MLE} = \max(x1, \dots, x10)$$

Example 4

Let $x_1, x_2, \ldots, x_n \in R$ be a random sample from a Poisson distribution. Find MLE of λ.

Example 5

Let $x_1, x_2, \dots, x_n \in R$ be random samples from the geometric distribution. Find MLE of p.

Nice properties of MLE

Fact1: MLE of i.i.d observation is consistent :

Let $\{Y_1, \cdots, Y_n\}$ be a sequence of i.i.d. observations where $Y_k \overset{iid}{\sim} f_\theta(y)$.

Then the MLE of θ is consistent.

Nice properties of MLE

Fact2: Invariance property of MLE

If $\hat{\theta}(\mathbf{x})$ is a maximum likelihood estimate for θ, then $g(\hat{\theta}(\mathbf{x}))$ is a maximum likelihood estimate for $g(\theta)$.

Example :

Let X denotes binomial random variable with parameter p. Lets find the MLE of the binomial parameter p.

Denote by x the total number of successes where xi is a single trial (0 or 1), then :

$$\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_1^n x_i}(1-p)^{\sum_1^n 1-x_i} = p^{x}(1-p)^{n-x}$$ Why ?

Consider the log likelihood :

$$\ln\left(nC_x\, p^x(1-p)^{n-x}\right) = \ln(nC_x) + x\ln(p) + (n-x)\ln(1-p)$$

Take the derivative and set to zero :

$$\frac{d}{dp}\ln(nC_x) + x\ln(p) + (n-x)\ln(1-p) = \frac{x}{p} - \frac{n-x}{1-p} = 0$$

Thus $\implies \frac{n}{x} = \frac{1}{p} \implies p = \frac{x}{n}$

Example :

Lets find the MLE for variance of X.

The variance of a binomial random variable X is given by

$V(X) = np(1 - p)$.

Because $V(X)$ is a function of the binomial parameter by the invariance property

the MLE of $V(X)$ *is*

.

$$\widehat{V(X)} = n(x/n)(1-x/n)$$

Remarks on the relation to Bayesian inference

Lets reconsider the equation :

$$L(\theta; y_1, \dots, y_n) \coloneqq p_{Y_1 \dots Y_n}(y_1, \dots, y_n; \theta)$$

Recall that in the MLE method we like to estimate the parameter θ_{MLE} such that

$$\theta_{MLE} = argmax_\theta L(\theta; y_1, \dots, y_n)$$

This means that we are relying on the data $(y_1, \dots, y_n)$ and the data alone to find the parameter final θ_{MLE}

Remarks on the relation to Bayesian inference

By Bayes rule we have the following identity

$$P(y_1, \ldots, y_n|\theta) = (L(\theta; data)P(\theta))/P(y_1, \ldots., y_n)$$

Posterior=likelihood*prior/data

Hence Posterior ~ likelihood*prior

Proportional to

Notice how the likelihood function shows up in the equation.

Now we may ask the following question, what does is the parameter θ that gives us the max value for the posterior ?

Such a parameter is called maximum a posteriori estimate and it coincides with $\theta_{MLE}$with the prior $P(\theta)$ is uniform.