

Lecture : Maximum Likelihood Estimate

Lecturer: Gang Liang

1 Motivating example

There is a biased coin with $P(\text{head}) = p$, and the parameter p is unknown to us. Now we toss this coin n times independently, and let X_1, X_2, \dots, X_n be the outcome of the experiments.

Question: why we use capital letters to denote the outcome of the experiment? **Answer:** the setup is general in the sense it works for all such experiments. The real observations are the realizations of these random variables:

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n,$$

with

$$X_i = \begin{cases} 1 & \text{if getting a head} \\ 0 & \text{if getting a tail} \end{cases}$$

So all X_i 's are independent Bernoulli random variables.

The probability of obtaining these observations is

$$P(X_1 = x_1, \dots, X_n = x_n; p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_i x_i} (1-p)^{n-\sum_i x_i}.$$

Note that p is usually unknown in practice.

The principle of maximum likelihood estimate (MLE) is one way to systematically estimate the unknown parameter, and it has a close analog with the philosophical principle “all that is real is rational; and all that is rational is real”. First, the observations are real. Now the question is what do we mean by “rational” in the probability framework? The answer is that a rational observation is the one with a large probability. Of course, events with small probabilities do happen, but it is always rational to bet on ones with large probabilities. So the principle of maximum likelihood estimate (MLE) is to find p which maximizes the above probability!

As an exercise, check that

$$\hat{p} = \sum_i x_i / n \text{ maximizes } P(X_1 = x_1, \dots, X_n; p).$$

2 General Maximum Likelihood Estimate

Let X be a random variable with pdf/pmf $f(x; \theta)$ where θ is the parameter to be estimated. We have many independent copies of random variables with the same distribution: X_1, X_2, \dots, X_n .

Definition: the **likelihood function** of the observations X_1, \dots, X_n is defined as

$$L(X_1, X_2, \dots, X_n; \theta) = \prod_{i=1}^n f(X_i; \theta).$$

MLE: The MLE of the problem is the parameter which maximizes the likelihood function, i.e.,

$$\hat{\theta}_{MLE} = \arg \max_{\theta} L(X_1, X_2, \dots, X_n; \theta),$$

where “arg max” is just a notation for taking the parameter which maximizes the given function.

3 How to obtain MLE?

1. Derive the **log-likelihood** function of the given problem:

$$l(\theta) = \log(L(X_1, X_2, \dots, X_n; \theta)).$$

The benefit of the log-likelihood function are (1) it turns multiplication into summation; (2) the logarithmic function is monotone so it does not change the maximum point of the likelihood function.

2. Maximize the log-likelihood function. Usually we obtain a likelihood equation by taking a derivative of the function wrt θ , then letting it be zero

$$l'(\theta) = 0.$$

Solving it to obtain \hat{p} .

3. Verify that the $\hat{\theta}$ obtained above is actually the maximal point of the log-likelihood function. Many people tend to ignore this step. It is not required for this class, but at least, you should keep it in mind.

4 Example: exponential distribution

Let X_1, \dots, X_n be independent exponential random variables with pdf

$$f(x; \lambda) = \lambda \exp(-\lambda x), \text{ for } x > 0.$$

Question: find the MLE of λ .

The likelihood function of the problem is

$$L(X_1, X_2, \dots, X_n; \lambda) = \prod_{i=1}^n \lambda \exp(-\lambda X_i),$$

so the log-likelihood function is

$$l(\lambda) = \log(L(X_1, X_2, \dots, X_n; \lambda)) = \sum_{i=1}^n (\log(\lambda) - \lambda X_i).$$

Taking the derivative of the function with respect to λ and letting it be zero:

$$l'(\lambda) = \sum_{i=1}^n \left(\frac{1}{\lambda} - X_i \right) = 0,$$

i.e.

$$\sum_{i=1}^n X_i = \frac{n}{\lambda}.$$

so we have

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^n X_i}.$$