Derivation of MLE for Binomial Distribution

Formalizing Maximum Likelihood Estimation

Let us formally verify the result obtained in the previous coin toss example. Namely, given multiple coin tosses resulting in n_T tails and n_H heads, the MLE estimate of the probability that a coin comes up heads is:

$$\mathrm{P}(H) = \frac{n_H}{n_H+n_T}$$

You can use MLE to estimate the parameters of a probability distribution, given the data.

To maximize the probability, take the derivative of the likelihood function and equate it with 0.

Calculating derivatives is cumbersome when a function is a product of many terms. However, you can apply the log function and maximize the log-likelihood instead.

This MLE estimate of the probability aligns with our empirical intuition: if you see 6 heads out of 10 coin tosses in an experiment, the likelihood of the experiment just completed would be maximum if the coin had a probability of 0.6 of showing heads. Therefore, MLE is doing exactly this: **given observations, what is the best estimate of the parameters of the experiment that maximize the likelihood of the observations?**

Let's derive the MLE result $\mathrm{P}(H) = \frac{n_H}{n_H+n_T}$ from first principles. First, define the parameters of the experiment. In the coin toss example, the experiment's observations only depend on 1 parameter: the "bias" of the coin towards showing heads (equivalently we can derive the same result by working with bias towards showing tails). For example, an "unbiased" coin has equal probability of showing heads and tails. A coin with bias 0.6 has a probability 0.6 of showing heads, and so on. Let's assume the probability of obtaining heads is $\mathrm{P}(H) = \theta$.

Further, assume you have observed the data D (sequence of heads and tails). With MLE, you would like to infer the true value of θ from the observed data D.

For any given probability θ, you can compute the likelihood of the observed data $\mathrm{P}(D|\theta)$. The MLE principle estimates the probability of heads θ so that it maximizes the likelihood of the observed data. The resulting estimate is referred to as the maximum likelihood estimate (denoted by $\hat{\theta}$) of θ:

$$\hat{\theta} = \arg\max_\theta \mathrm{P}(D|\theta)$$

Coin Toss Example

Coming back to our coin toss example, you are interested in estimating $\mathrm{P}(H) = \theta$ using MLE. You know that the result of coin tosses follows the binomial distribution, so you can model the following for some parameter θ:

$$\mathrm{P}(D|\theta) = \binom{n_H+n_T}{n_H}\theta^{n_H}(1-\theta)^{n_T}$$

For example, if you see 6 heads out of 10 coin tosses, you get different values of $\mathrm{P}(D|\theta)$ depending on the value of θ, i.e., the bias of the coin towards showing heads:

θ	$\mathrm{P}(D\|\theta)$
0.5	$\binom{10}{6}\cdot 0.5^6\cdot(1-0.5)^4 = \binom{10}{6}\cdot 0.5^6\cdot 0.5^4 \approx 0.205$
0.6	$\binom{10}{6}\cdot 0.6^6\cdot(1-0.6)^4 = \binom{10}{6}\cdot 0.6^6\cdot 0.4^4 \approx 0.251$
0.75	$\binom{10}{6}\cdot 0.75^6\cdot(1-0.75)^4 = \binom{10}{6}\cdot 0.75^6\cdot 0.25^4 \approx 0.146$
1	$\binom{10}{6}\cdot 1^6\cdot(1-1)^4 = \binom{10}{6}\cdot 1\cdot 0 = 0$

You are interested in finding the maximizer of $\mathrm{P}(D|\theta)$. To do so, you are going to take the derivative of the likelihood with respect to the parameter, equate the derivative expression with 0, and solve for the parameter. Since the likelihood $\mathrm{P}(D|\theta)$ is a product of many terms, the derivative expression will be long and cumbersome. Instead, you can find the MLE estimate with a trick: apply the $\log()$ function to the likelihood and maximize the log likelihood:

$$\hat{\theta} = \arg\max_\theta \mathrm{P}(D|\theta) = \arg\max_\theta \log \mathrm{P}(D|\theta)$$

Since the log function is monotonically increasing, the maximizer of $\mathrm{P}(D|\theta)$ is also the maximizer for $\log \mathrm{P}(D|\theta)$. The $\log()$ also turns any product into sums, so you obtain:

$$\log \mathrm{P}(D|\theta) = \log\binom{n_H+n_T}{n_H} + n_H \log\theta + n_T \log(1-\theta)$$

Now we'll compute the derivative of the log likelihood with respect to θ and equate it with zero (note that term $\binom{n_H+n_T}{n_H}$ is constant with respect to θ and drops out):

$$\begin{aligned}
&n_H\frac{\partial \log\theta}{\partial\theta} - n_T\frac{\partial\log(1-\theta)}{\partial\theta} = 0\\
\implies &\frac{n_H}{\theta} = \frac{n_T}{1-\theta}\\
\implies &n_H - n_H\theta = n_T\theta\\
\implies &\theta = \frac{n_H}{n_H+n_T}
\end{aligned}$$

This final form aligns with the intuition that:

$$\mathrm{P}(H) = \frac{n_H}{n_H+n_T}$$

If you plot the likelihood $\mathrm{P}(D|\theta) = \binom{n_H+n_T}{n_H}\theta^{n_H}(1-\theta)^{n_T}$ vs θ when 6 heads were seen out of 10 coin tosses, you will see that the likelihood is maximized for $\hat{\theta} = 0.6$, where the value of the likelihood is 0.251 – exactly what we calculated before.