Maximum Likelihood Estimation

Mustaf Ahmed

May 28, 2025

1 Maximizing log-likelihood

In statistical inference, Maximum Likelihood Estimation (MLE) is a method of producing a model that best accounts for the observed data by maximizing the likelihood of the observed data; among all competing models, we should select the one that best predicts the observed data.

Optimizing the likelihood function becomes easier when we instead maximize the log-likelihood. We'll justify this alternative optimization in what follows.

Theorem 1 (Equivalence of Likelihood and Log-Likelihood Maximizers)**.** *Let $L(\theta \mid \mathbf{x})$ be the likelihood function for parameters θ given data $\mathbf{x}$, assumed to be positive. Let $\log L(\theta \mid \mathbf{x})$ be the corresponding log-likelihood function. A parameter value $\hat{\theta}$ maximizes the likelihood function if and only if it also maximizes the log-likelihood function. Formally:*

$$\hat{\theta} = \arg\max_{\theta} L(\theta \mid \mathbf{x}) \iff \hat{\theta} = \arg\max_{\theta} \log L(\theta \mid \mathbf{x})$$

Proof. For the forward direction, assume that $\hat{\theta} = \arg\max_{\theta} L(\theta \mid \mathbf{x})$. By the definition of argument maximization, we have $L(\theta^{'} \mid \mathbf{x}) \leq L(\hat{\theta} \mid \mathbf{x})$. Since $\log(x)$ is strictly increasing, $\log(L(\theta^{'} \mid \mathbf{x})) \leq \log(L(\hat{\theta} \mid \mathbf{x}))$, so $\hat{\theta} = \arg\max_{\theta} \log L(\theta \mid \mathbf{x})$.

For the backward direction, assume that $\hat{\theta} = \arg\max_{\theta} \log L(\theta \mid \mathbf{x})$. By the definition of argument maximization, we have $\log(L(\theta^{'} \mid \mathbf{x})) \leq \log(L(\hat{\theta} \mid \mathbf{x}))$. Since e^x is strictly increasing, we have

$$e^{\log(L(\theta^{'}|\mathbf{x}))} \leq e^{\log(L(\hat{\theta}|\mathbf{x}))}$$

, and by simplifying we get

$$L(\theta^{'} \mid \mathbf{x}) \leq L(\hat{\theta} \mid \mathbf{x})$$

□

1.1 Example: MLE for a Normal Sample

Suppose $x_1, x_2, \dots x_n$ are iid with $x_i \sim \mathcal{N}(\mu, \sigma^2)$.

First, we will derive the likelihood function:

$$
\begin{aligned}
L(\theta \mid x) &= P(x_1, x_2, \dots, x_n \mid \mu, \sigma^2) \\
&= \prod_{i=1}^{n} P(x_i \mid \mu, \sigma^2) \quad \text{(due to independence)} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}
\end{aligned}
$$

We can instead maximize the log-likelihood by theorem 1, so

$$
\begin{aligned}
\log(L(\theta \mid x)) &= \log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right) \\
&= \log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\right) + \log\left(e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right) \\
&= n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}
$$

Maximizing the log-likelihood requires finding its partial derivatives:

$$
\begin{aligned}
\frac{\partial \log L}{\partial \mu} &= \frac{\partial}{\partial \mu}\left[n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= \frac{\partial}{\partial \mu}\left[n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right] - \frac{\partial}{\partial \mu}\left[\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= \frac{1}{\sigma^2}\left(\left(\sum_{i=1}^{n} x_i\right) - n\mu\right)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\frac{\partial \log L}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left[n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \\
&= \frac{\partial}{\partial \sigma}\left[n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right] - \frac{\partial}{\partial \sigma}\left[\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \\
&= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^3}
\end{aligned}
$$

We can set each partial derivative to 0 and solve, so with some algebra, the result is

$$
\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\hat{\sigma} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}
\end{aligned}
$$

The critical points $(\hat{\mu}, \hat{\sigma})$ represent a global maximum due to the second derivative test.