Maximum Likelihood Estimate

Dr. Arabin Kumar Dey

August 11, 2014

Suppose $X_1, X_2, \cdots, X_n$ is a sample from $f(X, \theta)$. The joint p.d.f. of the sample $X_1, X_2, \cdots, X_n$ can be written as

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

- Key Idea : probability of observing the given sample if true value is θ.
- A good intuitive rule might be to select the value of θ which makes the sample most likely.

Definition: In continuous case

Let X_i^n be a sample from $f(x;\theta)$. The quantity $L(X_1,\cdots,X_n;\theta) = \prod_{i=1}^n f(X_i;\theta)$ which is regarded as function of θ given the observation $X_1, X_2, \cdots, X_n$ is called the likelihood of the sample.

Definition : Maximum Likelihood Estimate

Let X_i^n be a sample from $f(X_i;\theta)$, $\theta \in \Omega$ (where Ω is the parameter space or the set of all possible values of θ). Suppose, $\tilde{\theta} \in \Omega$ is such that $L(\tilde{\theta}) = Max_{\theta\in\Omega} L(\theta)$, then $\tilde{\theta}$ is said to be the maximum likelihood estimator (MLE) of θ.

Few Mathematics

Necessary Condition

If θ is an interior point of Θ and a local maximum of g, then $g^{'}(\theta) = 0$. If θ is an interior point of Θ and a local maximum of g, then $g^{''}(\theta) \leq 0$.

Sufficient Condition

If $g^{'}(\theta) = 0$, then we say θ is a stationary point of g. If $g^{'}(\theta) = 0$ and $g^{''}(\theta) < 0$, then θ is a local maximum of g.

Concavity Conditions

If g is continuous on Θ and $g^{''}(\theta) < 0$ for all θ that are interior points of Θ, then we say g is a strictly concave function. In this case, any stationary point of g is the unique global maximum of g.

necessary Condition

If θ is an interior point of Θ and a local maximum of g, then $\bigtriangledown g(\theta) = 0$. If θ is an interior point of Θ and a local maximum of g, then $\bigtriangledown^2 g()$ negative semi-definite matrix.

necessary Condition

If $\bigtriangledown g = 0$ then we say θ is a stationary point of g. If $\bigtriangledown g = 0$ and $\bigtriangledown^2 g(\theta)$ is a negative definite matrix, then θ is a local maximum of g.

Concavity Conditions

If g is continuous on θ and $\bigtriangledown^2 g(\theta)$ is a negative definite matrix for all θ that are interior points of Θ, then we say g is a strictly concave function.
In this case, any stationary point of g is the unique global maximum of g.

Examples

- Let X_i^n be a random sample from $N(\mu, 1)$. Find out the maximum likelihood estimate of μ.
- Let X_i^n be a random sample from $N(\mu, \sigma^2)$. Find out the maximum Likelihood estimate of μ and σ^2.
- Let $X_1, X_2, \cdots, X_n$ be a random sample from Weibull(β, θ). Find out the maximum likelihood estimate of β, θ.

Example - continued

Let $X_1, X_2, \cdots, X_n$ be a random sample from the p.d.f.

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq X \leq \theta \\ 0, & \text{o.w.} \end{cases}$$

Find out the maximum likelihood estimate of θ.

Example

Let $X_1, X_2, \cdots, X_n$ be a random sample from the p.d.f.

$$f(x;\theta) = \begin{cases} 1 & \theta \leq X \leq \theta + 1 \\ 0, & \text{o.w.} \end{cases}$$

Find out the maximum likelihood estimate of θ.

Discrete Case: Example 1

Let $X \sim b(n,p)$. One observation on X is available, and it is known that n is either 2 or 3 and p = $\frac{1}{2}$ or $\frac{1}{3}$. Our object is to find an estimate of the pair (n,p). The following table gives the probability that $X = x$ for each possible pair (n,p).

x	$(2,\frac{1}{2})$	$(2,\frac{1}{2})$	$(3,\ \frac{1}{2})$	$(3,\frac{1}{3})$	Maximum Probability
0	$\frac{1}{4}$	$\frac{4}{9}$	$\frac{1}{8}$	$\frac{8}{27}$	$\frac{4}{9}$
1	$\frac{1}{2}$	$\frac{4}{9}$	$\frac{3}{8}$	$\frac{12}{27}$	$\frac{1}{2}$
2	$\frac{1}{4}$	$\frac{1}{9}$	$\frac{3}{8}$	$\frac{6}{27}$	$\frac{3}{8}$
3	0	0	$\frac{1}{8}$	$\frac{1}{27}$	$\frac{1}{8}$

What is value of parameters which maximizes the probability of a particular observed value ?

$$(\hat{n}, \hat{p})(x) = \begin{cases} (2, \frac{1}{3}) & x = 0 \\ (2, \frac{1}{2}) & x = 1 \\ (3, \frac{1}{2}) & x = 2 \\ (3, \frac{1}{2}) & x = 3 \end{cases}$$

- The Expectation Maximization (EM) algorithm is one of the most widely used algorithms in statistics.
- The basic idea of EM is actually quite simple: when direct maximization of $p(X|\theta)$ is complicated we can augment the data X by introducing some hidden variable Z such that

$$p(X, Z|\theta)$$

can be computed easily (for example when you observe both X and Z it can be easily maximized with respect to θ).

- Suppose we have a guess of the parameter value $\theta^{(t)}$ and want to find θ such that $p(X|\theta) > p(X|\theta^{(t)})$.
- This can be done by considering the difference between observed-data log-likelihood

$$\Delta L = L(\theta) - L(\theta^{(t)}) = \log\left(\frac{p(X|\theta)}{p(X|\theta^{(t)})}\right).$$

- Now we introduce the hidden variable Z such that $p(X, Z|\theta)$ is easy to compute (usually in a product form so that $\log(p(X, Z|\theta))$ can be factorized).

We have

$$
\begin{aligned}
L(\theta) - L(\theta^{(t)}) &= \log \frac{\int p(x, z|\theta) dz}{p(x|\theta^{(t)})} \\
&= \log \left[\int \frac{p(z|\theta^{(t)}, x) p(x, z|\theta)}{p(z|\theta^{(t)}, x) p(x|\theta^{(t)})} \right] dz \\
&\geq \int \left[p(z|\theta^{(t)}, x) \log \frac{p(x, z|\theta)}{p(z|\theta^{(t)}, x) p(x|\theta^{(t)})} \right] dz \\
&= \Delta L(\theta; \theta^{(t)})
\end{aligned}
$$

where the last inequality is due to Jensen's inequality and the fact that $\log(\cdot)$ is concave.

- We have $L(\theta) \geq L(\theta^{(t)}) + \Delta L(\theta; \theta^{(t)})$, which says that $L(\theta^{(t)}) + \Delta L(\theta; \theta^{(t)})$ is a global lower bound of $L(\theta)$ for any θ.
- Consequently we can maximize $L(\theta; \theta^{(t)})$ wrt θ to obtain $\theta^{(t+1)}$, and as long as $\Delta L(\theta^{(t+1)}; \theta^{(t)}) \geq 0$.
- We have $L(\theta^{(t+1)}) \geq L(\theta^{(t)})$ (and verify that $\Delta L(\theta^{(t)}; \theta^{(t)}) = 0$).

Now back to the problem of maximizing $L(\theta; \theta^{(t)})$ wrt θ:

$$
\begin{aligned}
\theta^{(t+1)} &= argmax_{\theta} \Delta L(\theta; \theta^{(t)}) \\
&= argmax_{\theta} \int p(z|\theta^{(t)}, x) \log \left(\frac{p(x, z|\theta)}{p(z|\theta^{(t)}, x) p(x|\theta^{(t)})} \right) dz \\
&= argmax_{\theta} \int p(z|\theta^{(t)}, x) \log(p(x, z|\theta)) dz
\end{aligned}
$$

Define,

$$
\begin{aligned}
Q(\theta;\theta^{(t)}) &= \int p(z|\theta^{(t)}, x)\log(p(x,z|\theta))dz \\
&= E_{Z|\theta^{(t)},X}(\log(X,Z|\theta))
\end{aligned}
$$

- **E-step: compute $Q(\theta;\theta^{(t)})$, which is the expectation of complete-data log-likelihood $\log p(X,Z|\theta^{(t)})$ and the expectation is wrt $p(Z|\theta^{(t)},X)$.** $\square$
- **M-step: maximize $Q(\theta;\theta^{(t)})$ wrt θ to obtain $\theta^{(t+1)}$.**

Example

We now apply EM to fit a mixture of two normal distributions. Suppose we observe $x_1, \cdots, x_n$ from a mixture of normal distributions $p(x) = \lambda N(\mu_1, \sigma_1^2) + (1-\lambda)N(\mu_2, \sigma_2^2)$. So in our case the observed data is $x_1, \cdots, x_n$ and the $\theta = \lambda, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$. We introduce hidden variables $z_1, ..., z_n$ where $z_i = 1$ if x_i comes from the first mixture component and 1 otherwise.

$$
\begin{aligned}
& \log p(x_i, z_i|\theta) \\
= \ & \log[\lambda \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}}]^{z_i}[(1-\lambda)\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}}]^{(1-z_i)} \\
= \ & z_i \log[\lambda \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}}] + (1-z_i)\log[(1-\lambda)\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}}]
\end{aligned}
$$

$$
\begin{aligned}
Q(\theta;\theta^{(t)}) &= E_{Z|X,\theta^{(t)}}\left[\sum_{i=1}^{n} z_i \log\lambda + (1-z_i)\log(1-\lambda)\right] \\
&+ E_{Z|X,\theta^{(t)}}\left[\sum_{i=1}^{n} z_i \log\sigma_1 - (1-z_i)\log\sigma_2\right] \\
&+ E_{Z|X,\theta^{(t)}}\left[\sum_{i=1}^{n} -z_i\frac{(x_i-\mu_1)^2}{2\sigma_1^2} - (1-z_i)\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right] \\
&= \sum_{i=1}^{n}\left[-E_{Z|X,\theta^{(t)}}(z_i)\log\lambda - (1-E_{Z|X,\theta^{(t)}}(z_i))\log(1-\lambda)\right] \\
&+ \sum_{i=1}^{n}\left[-E_{Z|X,\theta^{(t)}}(z_i)\log(\sigma_1) - (1-E_{Z|X,\theta^{(t)}}(z_i))\log(\sigma_2)\right] \\
&+ \sum_{i=1}^{n} -E_{Z|X,\theta^{(t)}}(z_i)\frac{(x_i-\mu_1)^2}{2\sigma_1^2} - (1-E_{Z|X,\theta^{(t)}}(z_i))\frac{(x_i-\mu_2)^2}{2\sigma_2^2}
\end{aligned}
$$

Define, $m_i^1 = E(Z_i|X_i)$ and $m_i^2 = 1 - E(Z_i|X_i)$. **Very Very important Remark : expectation is taken over Z given the observation and the parameter (expectation over posterior distribution)** and we first work out the M-step assuming that we already know m_i^1 and m_i^2's (which depend on the value of $\theta^{(t)}$). By maximizing $Q(\theta;\theta^{(t)})$ w.r.t θ we have,

$$\lambda^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} m_i^1$$

,

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n m_i^j x_i}{\sum_{i=1}^n m_i^j} \quad (j = 1, 2)$$

,

$$\sigma_j^{(t+1)} = \frac{\sum_{i=1}^n m_i^j (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^n m_i^j} \quad (j = 1, 2)$$

Note that the M-step makes perfect sense if we split each x_i into two particles, the first comes from mixture component one with weight m_i^1 , etc. The quantity $m_i^1 = E_{Z|X,\theta^{(t)}}[Z_i]$ which is needed in the E-step can be computed as

$$
\begin{aligned}
E(Z_i) &= 1 \cdot P(Z_i = 1|\theta^{(t)}, x_1, x_2, \cdots, x_n) + 0 \cdot P(Z_i = 0|\theta^{(t)}, x_1, x_2, \cdots \\
&= \frac{p(x_i, z_i = 1|\theta^{(t)})}{p(x_i, z_i = 0|\theta^{(t)}) + p(x_i, z_i = 1|\theta^{(t)})} \\
&= \frac{\lambda^{(t)} N(x_i|\mu_1^{(t)}, (\sigma_1^{(t)})^2)}{\lambda^{(t)} N(x_i|\mu_1^{(t)}, (\sigma_1^{(t)})^2) + (1 - \lambda^{(t)}) N(x_i|\mu_2^{(t)}, (\sigma_2^{(t)})^2)}
\end{aligned}
$$