# 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 $\theta$.
- A good intuitive rule might be to select the value of $\theta$ 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 $\theta$ 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 $\Omega$ is the parameter space or the set of all possible values of $\theta$). 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 $\theta$.

# Few Mathematics

## Necessary Condition

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

## Sufficient Condition

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

## Concavity Conditions

If g is continuous on $\Theta$ and $g^{''}(\theta) < 0$ for all $\theta$ that are interior points of $\Theta$, 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 $\theta$ is an interior point of $\Theta$ and a local maximum of g, then $\bigtriangledown g(\theta) = 0$. If $\theta$ is an interior point of $\Theta$ and a local maximum of g, then $\bigtriangledown^2 g()$ negative semi-definite matrix.

### necessary Condition

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

### Concavity Conditions

If g is continuous on $\theta$ and $\bigtriangledown^2 g(\theta)$ is a negative definite matrix for all $\theta$ that are interior points of $\Theta$, 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 $\mu$.
- Let $X_i^n$ be a random sample from $N(\mu, \sigma^2)$. Find out the maximum Likelihood estimate of $\mu$ and $\sigma^2$.
- Let $X_1, X_2, \cdots, X_n$ be a random sample from Weibull($\beta$, $\theta$). Find out the maximum likelihood estimate of $\beta$, $\theta$.

# Example - continued

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

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

Find out the maximum likelihood estimate of $\theta$.

# Example

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

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

Find out the maximum likelihood estimate of $\theta$.

# 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 $\theta$).

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

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

- 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 $\theta$.
- Consequently we can maximize $L(\theta; \theta^{(t)})$ wrt $\theta$ 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 $\theta$:

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

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** $\theta$ **to obtain** $\theta^{(t+1)}$**.**