Lecture 13: Maximum Likelihood

Sta 111

Colin Rundel

June 2, 2014

Likelihood

Last time, as part of discussing Bayesian inference we defined $P(\boldsymbol{X}|\boldsymbol{\theta})$ as the likelihood, which is the probability of the data, $\boldsymbol{X}$, given the model parameters $\boldsymbol{\theta}$.

In the case where the observations are iid

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

Our inference goal is still the same, given the observed data $(x_1, \ldots, x_n)$ we want to come up with an estimate for the parameters $\hat{\boldsymbol{\theta}}$

Maximum Likelihood

The maximum likelihood approach is straight forward, if we don't know $\boldsymbol{\theta}$ we might as well guess a value that maximizes the likelihood, or in other words pick the value of $\boldsymbol{\theta}$ that has the greatest probability of producing the observed data.

To do this we construct a likelihood function by considering the likelihood as a function of the parameter(s) given the data

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) = f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

The maximum likelihood estimate (estimator), $\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}$, is therefore given by

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}} = \arg\max_{\boldsymbol{\theta}} \; \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X})$$

Why use the MLE?

The maximum likelihood estimator has the following properties:

- *Consistency* - as the sample size tends to infinity the MLE tends to the 'true' value of the parameter(s)
- *Asymptotic normality* - as the sample size increases, the distribution of the MLE tends to the normal distribution
- *Efficiency* - as the sample size tends to infinity, there are not any other consistent estimator with a lower mean squared error

Example - Discrete Parameter Space

Assume that you have a box where you keep trick coins (biased towards heads or tails), the labels have fallen off three of them: one of which comes up heads 25% of the time, one which comes up head 75% of the time, and one that was fair (heads 50% of the time). If you pick one of the coins at random and flip it 100 times and you get 65 heads what coin do you think it is?

$$\theta \in \{0.25, 0.5, 0.75\}$$

$$\begin{aligned}
\mathcal{L}(\theta|X=65) = P(X=65|\theta) = \binom{100}{65}\theta^{65}(1-\theta)^{35} \\
\mathcal{L}(\theta=0.25|X=65) = \binom{100}{65}0.25^{65}(1-0.25)^{35} \approx 0 \\
\mathcal{L}(\theta=0.50|X=65) = \binom{100}{65}0.50^{65}(1-0.50)^{35} = 0.00086 \\
\mathcal{L}(\theta=0.75|X=65) = \binom{100}{65}0.75^{65}(1-0.75)^{35} = 0.00702
\end{aligned}$$

Example - Continuous Parameter Space

Consider the same situation but the labels has fallen off of only one coin and you have no idea how biased it might be. You once again flip the coin 100 times and get 65 heads what would MLE be for θ?

Maximum log-Likelihood

For the previous examples it was easy to write down the likelihood function and find its derivative. For many common likelihoods this can be difficult, or at the very least tediuous. Consider the case of n observations from a normal distribution with known variance σ^2 the likelihood function has the form

$$\mathcal{L}(\mu|\boldsymbol{X},\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

Such problems can be simplified by instead maximizing the log-likelihood function, $\ell(\boldsymbol{\theta}|\boldsymbol{X})$ which is equivalent as log is a monotone function.

$$\begin{aligned}
\ell(\boldsymbol{\theta}|\boldsymbol{X}) = \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) = \sum_{i=1}^{n} \log f(x_i,\boldsymbol{\theta}) \\
\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}} = \underset{\boldsymbol{\theta}}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) = \underset{\boldsymbol{\theta}}{\arg\max}\ \ell(\boldsymbol{\theta}|\boldsymbol{X})
\end{aligned}$$

Example - Normal with known Variance

Let $X_1,\ldots,X_n \overset{iid}{\sim} \mathcal{N}(\mu,\sigma^2)$ then the MLE of μ is

Example - Normal with known Mean

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ then the MLE of σ^2 is

Example - Poisson

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Pois}(\lambda)$ then the MLE of λ is

Example - Exponential

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$ then the MLE of λ is

Example - Uniform - Open vs Closed

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$ then the MLE of θ is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{L}(\lambda|\boldsymbol{X}) = \frac{1}{\theta^n}$$

Clearly, $\mathcal{L}(\lambda|\boldsymbol{X})$ is a decreasing function of θ, therefore to maximize $\mathcal{L}$ we need to minimize θ but we have additional constraints on θ from the likelihood: $0 \le x \le \theta$.

Therefore, $\hat{\theta}_{MLE} = \max(x_1, \ldots, x_n)$.

What would happen if we had defined the uniform on the open interval $(0, \theta)$?

Example - Uniform - Uniqueness

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(\theta, \theta+1)$ then the MLE of θ is

$$f(x|\theta) = \begin{cases} 1 & \text{if } \theta \le x \le \theta+1 \\ 0 & \text{otherwise} \end{cases}$$

Based on our previous result we maximizing $\mathcal{L}$ does not depend on θ but we still have to choose a θ such that $\theta \le x_i \le \theta+1$ for all x_i.

From the lower bound it is clear that $\theta \le \min(x_1, \ldots, x_n)$ and from the upper bound $\theta \ge \max(x_1, \ldots, x_n) - 1$. Obviously, there are many potential values that will satisfy these conditions.

Example - Gamma - Fixed λ

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(k, 1)$ then the MLE of k is

$$f(x|k, \lambda=1) = \frac{1}{\Gamma(k)} x^{k-1} e^{-x}$$

Example - Gamma - Known k

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(k, \lambda)$ then the MLE of k is

$$f(x|k, \lambda) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x}$$