Chapter 7. Statistical Estimation

7.2: Maximum Likelihood Examples

Slides (Google Drive) | Alex Tsun | Video (YouTube)

We spend an entire section just doing examples because maximum likelihood is such a fundamental concept used everywhere (especially machine learning). I promise that the idea is simple: find θ that maximizes the likelihood of the data. The computation and notation can be confusing at first though.

7.2.1 MLE Example (Poisson)

Example(s)

Let's say $x_1, x_2, ..., x_n$ are iid samples from Poi(θ). (These values might look like $x_1 = 13, x_2 = 5, x_3 = 6$, etc...) What is the MLE of θ?

Solution Remember that we discussed that the sample mean might be a good estimate of θ. If we observed 20 events over 5 units of time, a good estimate for λ, the average number of events per unit of time, would be $\frac{20}{5} = 4$. This turns out to be the maximum likelihood estimate!
Let's follow the recipe provided in 7.1.

1. **Compute the likelihood and log-likelihood of data.** To do this, we take the following product of the Poisson PMFs at each sample x_i, over all the data points:

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i \mid \theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

Again, this is the probability of seeing x_1, then x_2, and so on. This function is pretty hard to differentiate, so to make it easier, let's compute the log-likelihood instead, using the following identities:

$$\log(ab) = \log(a) + \log(b) \qquad \log(a/b) = \log(a) - \log(b) \qquad \log(a^b) = b\log(a)$$

In most cases, we'll want to optimize the log-likelihood instead of the likelihood (since we don't want to use the product rule of calculus)!

$$\begin{aligned}
\ln L(\mathbf{x} \mid \theta) &= \ln\left(\prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}\right) && \text{[def of likelihood]} \\
&= \sum_{i=1}^{n} \ln\left[e^{-\theta} \frac{\theta^{x_i}}{x_i!}\right] && \text{[log of product is sum of logs]} \\
&= \sum_{i=1}^{n} [\ln(e^{-\theta}) + \ln(\theta^{x_i}) - \ln x_i!)] && \text{[log of product is sum of logs]} \\
&= \sum_{i=1}^{n} [-\theta + x_i \ln\theta - \ln x_i!)] && \text{[other log properties]}
\end{aligned}$$

2. **Take the partial derivative(s) with respect to θ and set to 0. Solve the equation(s).**

   Now we want to take the derivative of the log likelihood with respect to θ, so the derivative of $-\theta$ is just -1, and the derivative of $x_i \ln \theta$ is just $\frac{x_i}{\theta}$, because remember x_i is a constant with respect to θ.

$$\frac{\partial}{\partial \theta} \ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \left[-1 + \frac{x_i}{\theta} \right]$$

   And now we want to set the derivative equal to 0, and solve for θ, and $\hat{\theta}$ is actually the estimate that we solve for. We do some algebra, and get $\frac{1}{n} \sum_{i=1}^{n} x_i$, which is actually just the sample mean!

$$\sum_{i=1}^{n} \left[-1 + \frac{x_i}{\theta} \right] = 0 \rightarrow -n + \frac{1}{\hat{\theta}} \sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

3. **Optionally, verify $\hat{\theta}_{MLE}$ is indeed a (local) maximizer by checking that the second derivative at $\hat{\theta}_{MLE}$ is negative (if θ is a single parameter), or the Hessian (matrix of second partial derivatives) is negative semi-definite (if θ is a vector of parameters).**

   We want to take the second derivative also, because otherwise we don't know if this is a maximum or a minimum. We differentiate the first derivative $\sum_{i=1}^{n} [-1 + \frac{x_i}{\theta}]$ again with respect to θ, and we notice that because θ^2 is always positive, the negative of that is always negative, so the second derivative is always less than 0, so that means that it's concave down everywhere. This means that anywhere the derivative is zero is a global maximum, so we've successfully found the global maximum of our likelihood equation.

$$\frac{\partial^2}{\partial \theta^2} \ln L(x \mid \theta) = \sum_{i=1}^{n} \left[-\frac{x_i}{\theta^2} \right] < 0 \rightarrow \text{concave down everywhere}$$

□

7.2.2 MLE Example (Exponential)

Example(s)

Let's say $x_1, x_2, ..., x_n$ are iid samples from Exp(θ). (These values might look like $x_1 = 1.354, x_2 = 3.198, x_3 = 4.312$, etc...) What is the MLE of θ?

Solution Now that we've seen one example, we'll just follow the procedure given in the previous section.

1. **Compute the likelihood and log-likelihood of data.**

   Since we have a continuous distribution, our likelihood is the product of the PDFs:

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

   The log-likelihood is

$$\ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \ln \left(\theta e^{-\theta x_i} \right) = \sum_{i=1}^{n} \left[\ln(\theta) - \theta x_i \right]$$

2. **Take the partial derivative(s) with respect to θ and set to 0. Solve the equation(s).**

$$\frac{\partial}{\partial\theta}\ln L(\mathbf{x}\mid\theta)=\sum_{i=1}^{n}\left[\frac{1}{\theta}-x_i\right]$$

Now, we set the derivative to 0 and solve (here we replace θ with $\hat{\theta}$):

$$\sum_{i=1}^{n}\left[\frac{1}{\hat{\theta}}-x_i\right]=0\rightarrow\frac{n}{\hat{\theta}}-\sum_{i=1}^{n}x_i=0\rightarrow\hat{\theta}=\frac{n}{\sum_{i=1}^{n}x_i}$$

This is just the inverse of the sample mean! This makes sense because if the average waiting time was $1/2$ hours, then the average rate per unit of time λ should be $\frac{1}{1/2}=2$ per hour!

3. **Optionally, verify $\hat{\theta}_{MLE}$ is indeed a (local) maximizer by checking that the second derivative at $\hat{\theta}_{MLE}$ is negative (if θ is a single parameter), or the Hessian (matrix of second partial derivatives) is negative semi-definite (if θ is a vector of parameters).** The second derivative of the log-likelihood just requires us to take one more derivative:

$$\frac{\partial^2}{\partial\theta^2}\ln L(\mathbf{x}\mid\theta)=\sum_{i=1}^{n}\left[\frac{-1}{\theta^2}\right]<0$$

Since the second derivative is negative everywhere, the function is concave down, and any critical point is a global maximum!

□

7.2.3 MLE Example (Uniform)

Example(s)

Let's say $x_1, x_2, ..., x_n$ are iid samples from (continuous) Unif$(0,\theta)$. (These values might look like $x_1 = 2.325, x_2 = 1.1242, x_3 = 9.262$, etc...) What is the MLE of θ?

Solution It turns out our usual procedure won't work on this example, unfortunately. We'll explain why once we run into the problem!

To compute the likelihood, we first need the individual density functions. Recall

$$f_X(x\mid\theta)=\begin{cases}\frac{1}{\theta} & 0\le x\le\theta\\ 0 & \text{otherwise}\end{cases}$$

Let's actually define an indicator function for whether or not some boolean condition A is true or false:

$$I_A=\begin{cases}1 & A\text{ is true}\\ 0 & A\text{ is false}\end{cases}$$

This way, we can rewrite the uniform density in one line as ($1/\theta$ for $0\le x\le\theta$ and 0 otherwise):

$$f_X(x\mid\theta)=\frac{1}{\theta}I_{\{0\le x\le\theta\}}$$

First, we take the product over all data points of the density at that data point, and plug in the density of the uniform distribution. How do we simplify this? First of all, we notice that in every term in the product, there is still a $\frac{1}{\theta}$, so multiply it by itself n times and get $\frac{1}{\theta^n}$. How do we multiply indicators? If we want the product of 1's and 0's to be 1, they ALL have to be 1. So,

$$I_{\{0\leq x_1\leq\theta\}}\cdot I_{\{0\leq x_2\leq\theta\}}\cdots I_{\{0\leq x_n\leq\theta\}}=I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$$

and our likelihood is

$$L(\mathbf{x}\mid\theta)=\prod_{i=1}^{n}f_X(x_i\mid\theta)=\prod_{i=1}^{n}\frac{1}{\theta}I_{\{0\leq x_i\leq\theta\}}=\frac{1}{\theta^n}I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$$

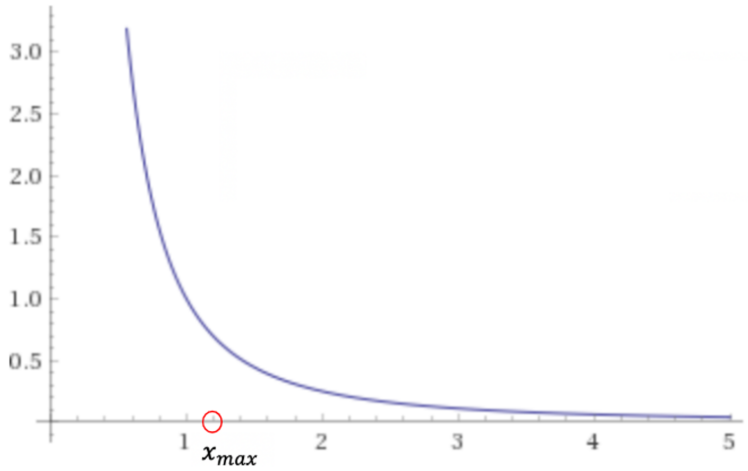
We could take the log-likelihood before differentiating, but this function isn't too bad-looking, so let's take the derivative of this. The $I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$ just says the function is $\frac{1}{\theta^n}$ when it the condition is true and 0 otherwise. So our derivative will just be the derivative of $\frac{1}{\theta^n}$ when that condition is true and 0 otherwise.

$$\frac{d}{d\theta}L(\mathbf{x}\mid\theta)=-\frac{n}{\theta^{n+1}}I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$$

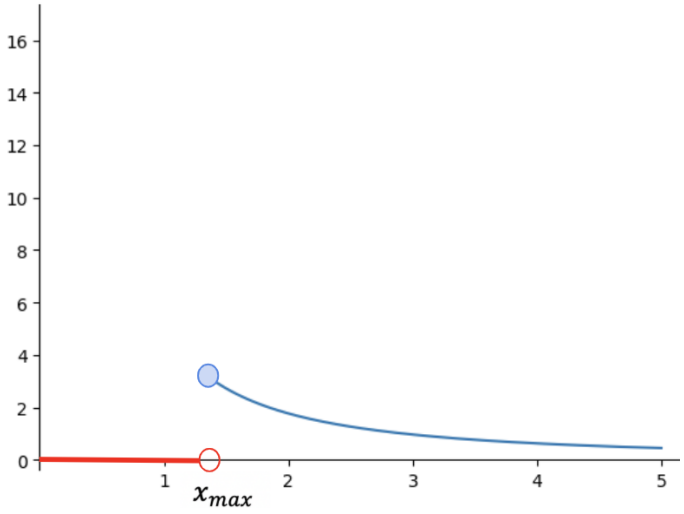
Now, let's set the derivative equal to 0 and solve for θ.

$$-\frac{n}{\theta^{n+1}}=0\rightarrow\theta=???$$

There seems to be no value of θ that solves this, what's going on? Let's plot the likelihood. First, we plot just $\left(\frac{1}{\theta}\right)^n$ (not quite the likelihood) where θ is on the x-axis:



Above is a graph of $\frac{1}{\theta^n}$, and so if we wanted to maximize this function, we should choose $\theta=0$. But remember that the likelihood, was $\frac{1}{\theta^n}I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$, which can also be written as $\frac{1}{\theta^n}I_{\{x_{\max}\leq\theta\}}$, because all the samples are $\leq\theta$ if and only if the maximum is. Below is the graph of the actual likelihood:



Notice that multiplying by the indicator function just kept the function as is when the condition was true, $x_{\max} \leq \theta$, but zeroed it out otherwise. So now we can see that our maximum likelihood estimator should be $\hat{\theta}_{MLE} = x_{\max} = \max\{x_1, x_2, \ldots, x_n\}$, since it achieves the highest value.

Why? Remember $x_1, \ldots, x_n \sim \text{Unif}(0, \theta)$, so θ has to be at least as large as the biggest x_i, because if it's not as large as the biggest x_i, then it would have been impossible for that uniform to produce that largest x_i. For example, if our samples were $x_1 = 2.53, x_2 = 8.55, x_3 = 4.12$, our θ had to be at least 8.55 (the maximum sample), because if it were 7 for example, then $\text{Unif}(0, 7)$ could not possibly generate the sample 8.55.

So our likelihood remember $\frac{1}{\theta^n}$ would have preferred as small a θ as possible to maximize it, but subject to $\theta \geq x_{\max}$. Therefore the "compromise" was reached by making them equal!

I'd like to point out this is a special case because the range of the uniform distribution depends on its parameter(s) a, b (the range of $\text{Unif}(a, b)$ is $[a, b]$). On the other hand, most of our distributions like Poisson or Exponential have the same range no matter what value the value of their parameters. For example, the range of $\text{Poi}(\lambda)$ is always $\{0, 1, 2, \ldots\}$ and the range of $\text{Exp}(\lambda)$ is always $[0, \infty)$, independent of λ.

Therefore, most MLE problems will be similar to the first two examples rather than this complicated one!

□