Lecture 22: Maximum Likelihood Estimation

Anup Rao

February 22, 2018

We discuss a method to reconstruct distributions from their samples.

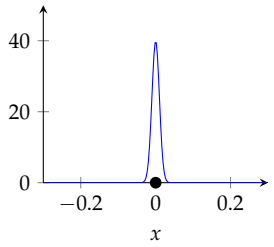
Throughout this course we have worked with the idea of an underlying probability space that controls how samples are generated. Sometimes, we know the parameters of the probability space—we know the pdf or cdf. But often, we are working with a distribution that we do not completely understand, and we want to reconstruct its parameters using data.

For example, suppose you work at a ride-sharing service. All the drivers that use the service get ratings from customers. How can you compare two different drivers? You could use just the average of the ratings, but this might be too coarse a measure. Clearly a driver whose every rating is 3 stars is very different than a driver who has 50% 1's and 50% 5's. The first driver seems to be much more consistent than the second driver, even though both have an average rating of 3.

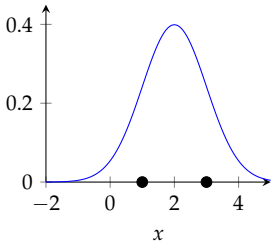
One approach for the above problem might be to assume that ratings for each driver are normally distributed (and then rounded to the nearest integer). Given that the normal distribution is so frequently observed in practice, this might (or might not) be a reasonable assumption. If this was true, it looks like both drivers have normals that have mean 3, but the seond driver has a higher variance than the first driver. Can we actually come up with an estimate for the variance given the data?

Another setting where this might be useful is in trying to make predictions. Suppose you are running google and you want to understand the probability that you get a certain number of search requests in between 10 am and 11 am next Saturday. One approach would be to assume that the number of requests is distributed according to a Poisson distribution, and then try to estimate the parameter λ for the Poisson based on data about the number of requests for all of the last 100 Saturdays.

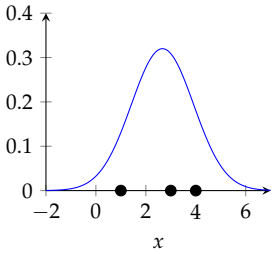
Maximum likelihood estimation is a method to reconstruct the parameters of the underlying distribution from the data. The intuition is to pick the parameters for the distribution that maximize the likelihood that the given data would have shown up.



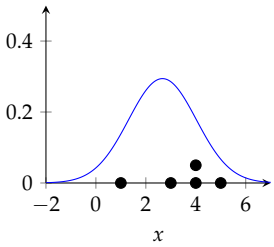
(a) $\mu = 0$, $\sigma = 1/100$.



(b) $\mu = 2$, $\sigma = 1$.



(c) $\mu = 8/3$, $\sigma = 1.2469$.



(d) $\mu = 17/5$, $\sigma = 1.3564$.

Figure 1: Some examples showing attempts to fit Normals to data.

Example: Estimating the bias of a coin

Suppose you toss a coin 5 times, and see 3 heads. What is the best estimate for the bias of the coin?

If the bias is p, then the probability of seeing k heads is exactly $\binom{5}{3}p^3(1-p)^2$. So, we define the likelihood of 3 given our estimate θ to be:

$$L(3|\theta) = \binom{5}{3}\theta^3(1-\theta)^2.$$

Our goal is to find the θ that maximizes the likelihood, so we take the derivative:

$$\frac{d}{d\theta}L(k|\theta) = \binom{5}{3} \cdot (3\theta^2(1-\theta)^2 - 2\theta^3(1-\theta)).$$

The derivative is 0 exactly when

$$\begin{aligned}
&(3\theta^2(1-\theta)^2 - 2\theta^3(1-\theta)) = 0 \\
\Rightarrow &3(1-\theta) = 2\theta \\
\Rightarrow &\theta = 3/5.
\end{aligned}$$

To check that this is in fact a maximum, we take the second derivative

$$\begin{aligned}
&\frac{d^2}{d\theta^2}L(k|\theta) \\
&= \binom{5}{3} \cdot (6\theta(1-\theta)^2 - 6\theta^2(1-\theta) \\
&- 6\theta^2(1-\theta) + 2\theta^3 \\
&= \binom{n}{k}\theta(6(1-\theta)^2 - 12\theta(1-\theta) + 2\theta^2) \\
&= \binom{n}{k}\theta(14\theta^2 - 24\theta + 6).
\end{aligned}$$

When $\theta = 3/5$, we have $(14\theta^2 - 24\theta + 6) = -3.36 \leq 0$, so this is actually a global maximum.

Example: Estimating the mean of a normal

Suppose you are given data points $x_1, x_2, \ldots, x_n$ and you know that they all came from the same normal, and that normal has variance 1. Can we estimate the mean using maximum likelihood?

Recall that the pdf of the normal is:

$$\frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2}},$$

since $\sigma = 1$ by assumption.

We define the *likelihood* of $x_1, \ldots, x_n$ to be a function of our estimate θ for the mean to be:

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x_i-\theta)^2}{2}},$$

and now we want to find the the θ that maximizes the likelihood. To solve this problem, we use calculus. First note that

$$\ln L(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{n} -\frac{1}{2} \cdot \ln(2\pi) - \frac{(x_i-\theta)^2}{2}.$$

To maximize this, we look for a point at which the derivative with respect to θ is 0.

$$\frac{d}{d\theta} \ln L(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{n} (x_i - \theta) = (\sum_{i=1}^{n} x_i) - n\theta.$$

This quantity is 0 when $\theta = (1/n) \cdot \sum_{i=1}^{n} x_i$, as you might expect. This setting of θ does in fact give the maximum, since you see that for all lower values of θ the derivative is positive, and for all higher values, the derivative is negative—the likelihood is increasing until you hit the average of the samples—and then it decreases.

The systematic way to verify that this is indeed the maximum is to consider the second derivative:

$$\frac{d^2}{d\theta^2} \ln L(x_1, \ldots, x_n|\theta) = -n.$$

Since the second derivative is negative, the point we have found must be a local maximum, and it is in fact a global maximum since it is the only point where the derivative is 0.

Example: Estimating both the mean and the standard deviation of a normal

Suppose the setup is exactly like before—we have n samples $x_1, \ldots, x_n$, and we want to estimate both the mean and the standard deviation of the underlying normal.

This time, there are two parameters, so the likelihood should be defined:

$$L(x_1, \ldots, x_n|\theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \cdot e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}.$$

As before, we have:

$$\ln L(x_1, \ldots, x_n|\theta_1, \theta_2) = \sum_{i=1}^{n} -\frac{1}{2} \ln(2\pi\theta_2) - \frac{(x_i-\theta_1)^2}{2\theta_2}.$$

The partial derivative with respect to θ_1 is

$$\frac{\partial}{\partial\theta_1} \ln L(x_1, \ldots, x_n|\theta_1, \theta_2) = \sum_{i=1}^{n} -\frac{(x_i-\theta_1)}{\theta_2},$$

and setting this to 0, we get

$$\begin{aligned}\sum_{i=1}^{n} -\frac{(x_i - \theta_1)}{\theta_2} &= 0\\ \Rightarrow \sum_{i=1}^{n}(x_i - \theta_1) &= 0\\ \Rightarrow \theta_1 &= (1/n) \cdot \sum_{i=1}^{n} x_i,\end{aligned}$$

exactly as before. To solve for θ_2, we take the partial derivative with respect to θ_2:

$$\frac{\partial}{\partial \theta_2} \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = \sum_{i=1}^{n} -\frac{1}{2} \cdot \frac{2\pi}{2\pi\theta_2} + \frac{(x_i - \theta_1)^2}{2\theta_2} = \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2 - \theta_2}{2\theta_2}.$$

Setting this to be 0, we see

$$\begin{aligned}\sum_{i=1}^{n} \frac{(x_i - \theta_1)^2 - \theta_2}{2\theta_2} &= 0\\ \Rightarrow \theta_2 &= (1/n) \cdot \sum_{i=1}^{n}(x_i - \theta_1)^2.\end{aligned}$$

In words, the maximum likelihood estimator for the variance is the average observed variance.

Our estimator for the mean of the distribution was $\frac{x_1 + x_2 + \ldots + x_n}{n}$. This is a *unbiased estimator*, in the sense that if the underlying distribution had mean μ, then the expected value of our own estimate is also μ.

It turns out that our estimate for the variance is not unbiased. One can show that the estimator

$$\theta_2 = (1/(n-1)) \cdot \sum_{i=1}^{n}(x_i - \theta_1)^2$$

is an unbiased estimator.