Maximum Likelihood Estimation for Fitting

Shakthi Visagan

October 2018

Consider the problem of estimating the mean of a sequence of N observations $\{\boldsymbol{y}_n\}$ each drawn from a Gaussian distribution with known mean μ and known variance $\sigma_{\boldsymbol{y}}^2$.

Here $\boldsymbol{y}$ is a **random variable**, which might be different notation than other texts. A random variable is not really a variable more so than it is a *function* that takes **events** from a **sample space** as an input to a number on the real line $\mathbb{R}$ as an output. For example, if tossing a coin was our context, then obtaining a Heads or a Tails would be our events, since the sample space, often denoted by Ω, is $\Omega = \{\text{Heads, Tails}\}$, since those would be the only possible events from the context of flipping a coin and taking the side that's facing up. Our random variable would then be constructed as $\boldsymbol{y}(\text{Heads}) = 1$ and $\boldsymbol{y}(\text{Tails}) = 0$.

What we can then do is ask questions like the following: what is the probability that we get Heads, or using the random variable, the probability that $\boldsymbol{y} = 1$ or that $\mathbb{P}(\boldsymbol{y} = 1)$? Likewise, we can ask what is the probability that we get Tails, or using the random variable, the probability that $\boldsymbol{y} = 0$ or that $\mathbb{P}(\boldsymbol{y} = 0)$? All random variables will be noted by a bold, lowercase, mathematical letter from the greek or roman alphabet.

One can then see that the random variable can then be used for another function where the input is the real number from the random variable, like in our example 1 or 0, and the output would be the probability of the random variable being that value, a real number between 0 and 1. This new function is called the **probability mass function** in the case where our random variable is discrete, or a **probability density function** where the random variable is continuous. Rolling a dice or flipping a coin would give a discrete random variable, but the heights of people or the lifetime of a cell would be continuous. We can define the probability mass function for our coin flipping problem as $f_{\boldsymbol{y}}(y)$, and in general probability mass functions and probability density functions will be noted as such.

Continuing with the example, let's assume we have a fair coin. We can then see that the probability mass function looks like the following: $f_{\boldsymbol{y}}(1) = 0.5$ and $f_{\boldsymbol{y}}(0) = 0.5$. Note that the probability mass function takes the random variable as inputs to give out probabilities. If our coin wasn't fair, we'd have $f_{\boldsymbol{y}}(1) = p_{\text{H}}$ and $f_{\boldsymbol{y}}(0) = 1 - p_{\text{H}}$ where p_{H} would be the probability of landing a Heads.

Sometimes the shapes of probability functions are shared among many phenomena, like the distribution of scores on a test, or the length of arms on humans, or errors made in measuring volumes in a graduated cylinder. Some of these distributions have names, and the examples in the previous sentence are all examples of the Gaussian distribution. The Gaussian distribution, as seen in the examples, is a continuous one since scores, lengths, and errors can all take continuous values, and since it is nearly impossible to find the probability of every single event happening from a continuous set, we can **generate** the distribution using its probability density function and some **parameters** that describe our data. A parameter for coin flipping could be the chance of landing Heads. In general, most probability density functions can then be described by their input random variable (or sometimes multiple random variables in the multivariate case), and the parameters for each random variable (or the sets of parameters in the multivariate case). We can then denote this as $f_{\boldsymbol{y}}(y; \theta_1, \theta_1, \ldots, \theta_m)$. A univariate Gaussian distribution distribution only takes the mean μ and the variance $\sigma_{\boldsymbol{y}}^2$ as parameters.

Assume that $\mu \in \mathbb{R}$. Compute the maximum-likelihood estimator $\hat{\boldsymbol{\mu}}$ for μ from the observations $\boldsymbol{y}_n$. Compute the variance of $\hat{\boldsymbol{\mu}}$. Is the estimator $\hat{\boldsymbol{\mu}}$ unbiased?

The symbol $\in$ means *belongs in* or *is from*. This is an example of a classic machine learning problem since we have data (our N observations of the random variable $\boldsymbol{y}$) and our unknown is the parameter that could describe the distribution of the data, in particular the mean of the Gaussian distribution. This is important because, once we have enough data, we can **fit** data to our model, back out these parameters that describe a certain distribution, and then use those parameters to generate data from the distribution. We will see later that our efforts in **estimating** parameters only get better as we get more data.

We first define $f_{\boldsymbol{y}}\left(y; \mu, \sigma_{\boldsymbol{y}}^2\right)$ as the following Gaussian distribution

$$f_{\boldsymbol{y}}\left(y; \mu, \sigma_{\boldsymbol{y}}^2\right) = \frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{y}}^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma_{\boldsymbol{y}}^2}\right)$$

Note how the probability density function takes in the random variable as an input but its output also depends on the parameters chosen. Getting more accurate parameters will approximate the underlying distribution of the data better.

Something that is confusing about this problem is that the parameters of the distribution that produced each of the individual observations, that is for the y_i in the sequence of our N observations, is known, μ and $\sigma_{\boldsymbol{y}}^2$. What we are trying to find is, given a sequence of these observations, can we build an estimator (a random variable) that will give us the mean of the sequence of the data, where each of the observations comes from a known Gaussian distribution.

We have a sequence of N independent and identically distributed observations so we then have the following **likelihood**

$$f_{\boldsymbol{y}}\left(y_1, y_2, \ldots, y_N; \mu, \sigma_{\boldsymbol{y}}^2\right) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{y}}^2}} \exp\left(\frac{-\left(y_n-\mu\right)^2}{2\sigma_{\boldsymbol{y}}^2}\right)$$

One way to think about the above is returning to the coin example. The data we're given is a sequence of N observations, or N random variable outputs. In our coin example, the sequence could be N coin tosses, which could look like $\{\text{H,T,T,H,T},\ldots,\text{H}\}$ or $\{1,0,0,1,0,\ldots,1\}$. Each coin toss is independent from the other, so if we wanted to know the probability of an entire sequence happening, it would just be the product of the probability of each coin toss, which in our case would be $\prod_{n=1}^{N} 0.5$ or equivalently 0.5^N.

Consider if we did *not* know the parameter of our coin, the chance it would land Heads, but we were still given the sequence of coin flips and were asked to get that parameter. We could call it p_{H} and see that the probability of a specific sequence of Heads or Tails happening would then be

$$p_{\text{H}} \times (1-p_{\text{H}}) \times (1-p_{\text{H}}) \times p_{\text{H}} \times (1-p_{\text{H}}) \times \cdots \times p_{\text{H}} = p_{\text{H}}^k (1-p_{\text{H}})^{(N-k)}$$

where k would be the number of Heads in our sequence and $N-k$ would be the number of Tails. Furthermore, since we know what our sequence is, we would know what k, N, and $N-k$ are. This function that we have constructed, that is, $\mathbb{P}(\{1,0,0,1,0,\ldots,1\}) = p_{\text{H}}^k(1-p_{\text{H}})^{(N-k)}$, is called the likelihood function since it takes the parameters as an input, and some information about the given data, to give a likelihood, the likelihood that this is our model, given our data.

We take the natural logarithm to find the log-likelihood, knowing that the logarithm of a product of numbers is the sum of the logarithm of the numbers.

$$\begin{aligned}
\ln\left(f_{\boldsymbol{y}}\left(y_1, y_2, \ldots, y_n; \mu, \sigma_{\boldsymbol{y}}^2\right)\right) &= \ln\left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{y}}^2}} \exp\left(\frac{-\left(y_n-\mu\right)^2}{2\sigma_{\boldsymbol{y}}^2}\right)\right) \\
&= \sum_{n=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{y}}^2}} \exp\left(\frac{-\left(y_n-\mu\right)^2}{2\sigma_{\boldsymbol{y}}^2}\right)\right) \\
&= \sum_{n=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{y}}^2}}\right) + \ln\left(\exp\left(\frac{-\left(y_n-\mu\right)^2}{2\sigma_{\boldsymbol{y}}^2}\right)\right) \\
&= \sum_{n=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{y}}^2}}\right) + \left(\frac{-\left(y_n-\mu\right)^2}{2\sigma_{\boldsymbol{y}}^2}\right)
\end{aligned}$$

If we continue with our coin flipping example, taking the logarithm of the likelihood function would give us $\ln(p_{\text{H}}^k(1-p_{\text{H}})^{(N-k)}) = k\ln(p_{\text{H}}) + (N-k)\ln(1-p_{\text{H}})$. Note how our product turned into a sum when using the logarithm. We take the logarithm or the natural logarithm because it is easier to work with sums rather than products. We also know that the logarithm is strictly increasing, so the maximum of the likelihood will also give you the maximum of the log of the likelihood. We care about the maximum of the likelihood because we want to know which parameters when used in our model maximize the probability of our data occurring. Mathematically this can be written as $\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \log(f_{\boldsymbol{y}}\left(y_1, y_2, \ldots, y_N; \theta\right))$. We already have most of this equation done. The rest isn't that bad either since we learned how to find the argument that maximizes a function in high school: take the derivative, and find what input takes the derivative to 0. Keep

in mind that we haven't done anything difficult so far, besides manipulate variables and perform some simple algebra.

We take the derivative of the log-likelihood with respect to μ and obtain the following

$$\begin{aligned}
\frac{\partial \ln\left(f_{\boldsymbol{y}}\left(y_1, y_2, \ldots, y_N; \mu, \sigma_{\boldsymbol{y}}^2\right)\right)}{\partial \mu} &= \frac{\partial}{\partial \mu}\left[\sum_{n=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{y}}^2}}\right) + \left(\frac{-\left(y_n-\mu\right)^2}{2\sigma_{\boldsymbol{y}}^2}\right)\right] \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mu}\left(\frac{-\left(y_n-\mu\right)^2}{2\sigma_{\boldsymbol{y}}^2}\right) \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mu}\left(\frac{-\mu^2+2y_n\mu-y_n^2}{2\sigma_{\boldsymbol{y}}^2}\right) \\
&= \sum_{n=1}^{N} \frac{-2\mu+2y_n}{2\sigma_{\boldsymbol{y}}^2} \\
&= \sum_{n=1}^{N} \frac{-\mu+y_n}{\sigma_{\boldsymbol{y}}^2}
\end{aligned}$$

We set this equal to 0 to arrive at the estimator $\hat{\boldsymbol{\mu}}$.

$$\begin{aligned}
\sum_{n=1}^{N} \frac{-\hat{\boldsymbol{\mu}}+y_N}{\sigma_{\boldsymbol{y}}^2} &= 0 \\
\sum_{n=1}^{N} -\hat{\boldsymbol{\mu}}+y_n &= 0 \\
\sum_{n=1}^{N} \hat{\boldsymbol{\mu}} &= \sum_{n=1}^{N} y_n \\
N\hat{\boldsymbol{\mu}} &= \sum_{n=1}^{N} y_n \\
\hat{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{n=1}^{N} y_n
\end{aligned}$$

In other words, the maximum likelihood estimator for the mean of the sequence is just the sample average, the average of all the data points.

The above may not be so astonishing, but is another reason the Gaussian distribution is so powerful. We ended up with fact that the random variable that best outputs the mean of a sequence of Gaussian observations is one that just takes the mean of the sequence given. Consider what happens when we do this to our coin flipping example: we have to take the derivative of the log-likelihood with respect to our parameter.

$$\begin{aligned}
\frac{\partial \ln\left(f_{\boldsymbol{y}}\left(y_1, y_2, \ldots, y_n; \hat{\boldsymbol{p}}_{\mathrm{H}}\right)\right)}{\partial p_{\mathrm{H}}} &= \frac{\partial}{\partial p_{\mathrm{H}}}\left[k\ln(p_{\mathrm{H}}) + (N-k)\ln(1-p_{\mathrm{H}})\right] \\
&= \frac{k}{p_{\mathrm{H}}} - \frac{N-k}{1-p_{\mathrm{H}}} \\
&= \frac{k-kp_{\mathrm{H}}-Np_{\mathrm{H}}+kp_{\mathrm{H}}}{p_{\mathrm{H}}(1-p_{\mathrm{H}})} \\
&= \frac{k-Np_{\mathrm{H}}}{p_{\mathrm{H}}(1-p_{\mathrm{H}})}
\end{aligned}$$

We then set this equal to 0 and solve for our parameter p_{H}:

$$\begin{aligned}
\frac{k - N\hat{\boldsymbol{p}}_{\mathrm{H}}}{p_{\mathrm{H}}(1 - \hat{\boldsymbol{p}}_{\mathrm{H}})} &= 0 \\
k - N\hat{\boldsymbol{p}}_{\mathrm{H}} &= 0 \\
N\hat{\boldsymbol{p}}_{\mathrm{H}} &= k \\
\hat{\boldsymbol{p}}_{\mathrm{H}} &= \frac{k}{N}
\end{aligned}$$

Let's unpack what this means. We know that k represented the number of Heads in our sequence, and N was the number of total flips. This is saying that the best estimator for the parameter of our coin flipping distribution, which was the chance the coin lands on Heads, is just the number of times the coin landed Heads divided by the total number of flips. This intuitively makes sense, since if we flipped a coin 10 times and got Heads 9 out of the 10 times, our best bet would be to say that the coin was loaded to land Heads $\frac{9}{10}$ or 90% of the time. However, imagine if our coin was truly fair and we just happened to get a serendipitous string of Heads those first 9 tosses. What we can do to better our estimator is to take more samples. Say we did 1000 coin flips and we produced 489 Heads. Our estimator would then say our coin is loaded to land on Heads almost half the time. Our estimate of the underlying distribution gets better when we have more data.

We then calculate the variance of this estimator, knowing that the variance of the sum of independent random variables is the sum of their variances.

$$\begin{aligned}
\mathrm{Var}\left(\hat{\boldsymbol{\mu}}\right) &= \mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^{N} y_n\right) \\
&= \frac{1}{N^2}\mathrm{Var}\left(\sum_{n=1}^{N} y_n\right) \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{Var}\left(y_n\right) \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\sigma_{\boldsymbol{y}}^2 \\
&= \frac{\sigma_{\boldsymbol{y}}^2 N}{N^2} \\
&= \frac{\sigma_{\boldsymbol{y}}^2}{N}
\end{aligned}$$

Taking the variance of our estimator also shows how obtaining more samples increases our estimate because as N grows larger, the variance of our estimator decreases, and the closer it gets to approximating the actual value.

This estimator is obviously unbiased.

$$\begin{aligned}\mathbb{E}\left[\hat{\boldsymbol{\mu}}\right] - \mu &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} y_n\right] - \mu \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[y_n\right] - \mu \\ &= \frac{1}{N}\sum_{n=1}^{N}\mu - \mu \\ &= \frac{N\mu}{N} - \mu = 0 \\ &= 0\end{aligned}$$

The bias of an estimator is just a good way to check how good your estimator is. Does the **expected value** of your estimator differ from the true value it's trying to estimate? The expected value is different from a sample average because the expected value depends on the underlying distribution of the data and not the actual data observed. Having no bias means your estimator captures both the underlying distribution and your sample data well.

Knowing this, let's calculate the bias of estimator in our coin flip example. However, recall that our estimator is a random variable; it is precisely its status as a random variable that let's us take the expectation of it. However our estimator currently looks like a scalar and this because in trying to simplify our explanation we skipped a step in thinking the number of Heads, k, is a scalar when it is actually a random variable since it always depends on the number of Heads that we do actually obtain. Since we denoted each head as when the random variable $y = 1$, we can rewrite our estimator $\hat{\boldsymbol{p}}_{\mathrm{H}}$ as a function of $\boldsymbol{y}$, and since any function of a random variable is another random variable, this will fix our small clerical error.

$$\hat{\boldsymbol{p}}_{\mathrm{H}} = \frac{k}{N} = \frac{\sum_{n=1}^{N} y_n}{N}$$

Recall that since we let $y_i = 1$ if the ith toss was a Heads, this lets us say that summing all the outputs of our coin toss will give us precisely the number of Heads, which is equivalent to k. Also keep in mind that the expectation of our coin toss will be just the value of our parameter because of the following: $\mathbb{E}[\boldsymbol{y}] = 1 \times p_{\mathrm{H}} + 0 \times (1 - p_{\mathrm{H}}) = p_{\mathrm{H}}$. We can now proceed with finding the bias.

$$\begin{aligned}\mathbb{E}\left[\hat{\boldsymbol{p}}_{\mathrm{H}}\right] - p_{\mathrm{H}} &= \mathbb{E}\left[\frac{\sum_{n=1}^{N} y_n}{N}\right] - p_{\mathrm{H}} \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[y_n\right] - p_{\mathrm{H}} \\ &= \frac{1}{N}\sum_{n=1}^{N} p_{\mathrm{H}} - p_{\mathrm{H}} \\ &= \frac{Np_{\mathrm{H}}}{N} - p_{\mathrm{H}} = 0 \\ &= 0\end{aligned}$$

Coincidentally, our coin flip paramter estimator is also unbiased.

■