Maximum Likelihood Estimation

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

August 8, 2008

Mathematical modeling often follows a certain recipe. First, you design an objective function that specifies constraints on how you want the model to perform. This objective function is often a probability model (i.e. a likelihood function). Next you find the values of the model's parameters that maximize this function.

Maximum likelihood estimation (MLE) is perhaps best explained via an example. Suppose that you are given a set of data consisting of three data items, labeled x_1, x_2, and x_3. Also suppose that you have reason to believe that the data items are independent (i.e. the value of x_i does not depend on the values of x_j for $j \neq i$) and that they are identically distributed (i.e. they come from the same population). Moreover, let's assume that this distribution is a Normal distribution (also known as a Gaussian distribution) whose mean, denoted μ, is unknown but whose variance is believed to have a value of 1 (this is not necessary, but it will help keep this example simple). We would like to estimate the value of the mean, and we are seeking a maximum likelihood estimate of this value. An MLE of the mean is the value of the mean that results in a Normal distribution that is most likely to have produced the data.

For concreteness, let's suppose that $x_1 = \frac{1}{2}$, $x_2 = \frac{3}{2}$, and $x_3 = \frac{5}{2}$. Our likelihood function, denoted $L(\mu|x_1, x_2, x_3)$, is the same as the probability of the data given the mean, denoted $p(x_1, x_2, x_3|\mu)$, so long as we keep in mind the fact that we know the values of x_1, x_2, and x_3, but we don't know the value of μ:

$$L(\mu|x_1, x_2, x_3) = p(x_1, x_2, x_3|\mu). \tag{1}$$

Because the data items are independent, this means that

$$p(x_1, x_2, x_3|\mu) = p(x_1|\mu)\ p(x_2|\mu)\ p(x_3|\mu) \tag{2}$$

and so we can re-write the likelihood function as follows:

$$L(\mu|x_1, x_2, x_3) = p(x_1|\mu)\ p(x_2|\mu)\ p(x_3|\mu). \tag{3}$$

That is, to compute the likelihood function, we need to compute the probability $p(x_i|\mu)$ where the probability distribution is Normal. Recall the form of the Normal probability distribution (whose variance equals one):

$$p(x_i|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \mu)^2}. \tag{4}$$

Putting this all together we have:

$$L(\mu|x_1, x_2, x_3) = \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \mu)^2}. \tag{5}$$

Which value of μ maximizes the likelihood function? Let's try a few different values. First, let's try $\mu = \frac{1}{2}$. Then

$$p(x_1 = \frac{1}{2}|\mu = \frac{1}{2}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{1}{2}-\frac{1}{2})^2} = 0.399 \quad (6)$$

$$p(x_2 = \frac{3}{2}|\mu = \frac{1}{2}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{3}{2}-\frac{1}{2})^2} = 0.242 \quad (7)$$

$$p(x_3 = \frac{5}{2}|\mu = \frac{1}{2}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{5}{2}-\frac{1}{2})^2} = 0.054 \quad (8)$$

meaning that

$$L(\mu = \frac{1}{2}|x_1, x_2, x_3) = 0.399 \times 0.242 \times 0.054 = 0.005. \quad (9)$$

Next, let's try $\mu = \frac{3}{2}$. If we go through an analogous set of equations, we find that $L(\mu = \frac{3}{2}|x_1, x_2, x_3) = 0.023$. Lastly, let's try $\mu = \frac{5}{2}$. After going through the equations, we find that $L(\mu = \frac{5}{2}|x_1, x_2, x_3) = 0.005$. Between $\mu = \frac{1}{2}$, $\mu = \frac{3}{2}$, and $\mu = \frac{5}{2}$, which value gives a Normal distribution that is most likely to have produced the data? It is the value $\mu = \frac{3}{2}$.

In this example, we have assumed that we know the value of the variance, and we have obtained a point estimate of the mean. Note that maximum likelihood estimates can be found for any parameter, not just the mean. For instance, if we did not know the value of the variance, then we can obtain a point estimate for this parameter too. For the sake of completeness, let's go through a case of a Normal distribution in which both the mean and the variance are unknown and, thus, need to be estimated.

Again, suppose that you are given a set of data consisting of three data items: $x_1 = \frac{1}{2}$, $x_2 = \frac{3}{2}$, and $x_3 = \frac{5}{2}$. Which values of μ and σ^2 maximize the likelihood function? Let's try some different values. First, let's try $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{2}$. Then

$$p(x_1 = \frac{1}{2}|\mu = \frac{1}{2}, \sigma^2 = \frac{1}{2}) = \frac{1}{\sqrt{\pi}} e^{-(\frac{1}{2}-\frac{1}{2})^2} = 0.564 \quad (10)$$

$$p(x_2 = \frac{3}{2}|\mu = \frac{1}{2}, \sigma^2 = \frac{1}{2}) = \frac{1}{\sqrt{\pi}} e^{-(\frac{3}{2}-\frac{1}{2})^2} = 0.208 \quad (11)$$

$$p(x_3 = \frac{5}{2}|\mu = \frac{1}{2}, \sigma^2 = \frac{1}{2}) = \frac{1}{\sqrt{\pi}} e^{-(\frac{5}{2}-\frac{1}{2})^2} = 0.010 \quad (12)$$

meaning that

$$L(\mu = \frac{1}{2}, \sigma^2 = \frac{1}{2}|x_1, x_2, x_3) = 0.564 \times 0.208 \times 0.010 = 0.0012. \quad (13)$$

Table 1 lists the likelihood for several possible values of μ and σ^2. Of these possible values, $\mu = \frac{3}{2}$ and $\sigma^2 = \frac{1}{2}$ are the best point estimates.

Keep in mind that maximum likelihood estimation can be used to estimate parameter values for any parametric distribution, not just the Normal distribution. To illustrate this point, let's estimate

Mean	Variance	Likelihood
0.5	0.5	0.0012
0.5	1.0	0.0052
0.5	2.0	0.0064
1.5	0.5	0.0243
1.5	1.0	0.0234
1.5	2.0	0.0136
2.5	0.5	0.0012
2.5	1.0	0.0052
2.5	2.0	0.0064

Table 1: Several possible values for the mean and variance, along with their corresponding likelihood values.

the probability, denoted μ, that a flipped coin lands heads-up. Let $x = 1$ denote the event that a flipped coin lands heads-up, and $x = 0$ denote the event that a flipped coin lands tails-up. Our data set consists of three coin flips: heads-up ($x_1 = 1$), tails-up ($x_2 = 0$), and heads-up ($x_3 = 1$). Based on this data, we want a maximum likelihood estimate of μ.

Because the coin flips in the data set are statistically independent, we have

$$L(\mu|x_1, x_2, x_3) = p(x_1, x_2, x_3|\mu) = p(x_1|\mu) \times p(x_2|\mu) \times p(x_3|\mu). \tag{14}$$

Recall that binary events are characterized by a Bernoulli distribution:

$$p(x|\mu) = \mu^x \, (1-\mu)^{1-x}. \tag{15}$$

Let's try some different values for μ. First, let's try $\mu = \frac{1}{3}$. Then

$$p(x_1 = 1|\mu = \frac{1}{3}) = (\frac{1}{3})^1 \, (1 - \frac{1}{3})^0 = \frac{1}{3} \tag{16}$$

$$p(x_2 = 0|\mu = \frac{1}{3}) = (\frac{1}{3})^0 \, (1 - \frac{1}{3})^1 = \frac{2}{3} \tag{17}$$

$$p(x_3 = 1|\mu = \frac{1}{3}) = (\frac{1}{3})^1 \, (1 - \frac{1}{3})^0 = \frac{1}{3} \tag{18}$$

meaning that $L(\mu = \frac{1}{3}|x_1, x_2, x_3) = \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} = 0.0741$. Analogous equations for $\mu = 0.5$ yields a likelihood of 0.1250; $\mu = \frac{2}{3}$ produces a likelihood of 0.1481. Of these possible values, $\mu = \frac{2}{3}$ is the maximum likelihood estimate.