Libby MacKinnon
CSE 527 notes
Lecture 7, October 17, 2007
MLE and EM

Review of Maximum Likelihood Estimators

MLE is one of many approaches to parameter estimation. The likelihood of independent observations is expressed as a function of the unknown parameter. Then the value of the parameter that maximizes the likelihood of the observed data is solved for. This is typically done by taking the derivative of the likelihood function (or log of the likelihood function) with respect to the unknown parameter, setting that equal to zero, and solving for the parameter.

Example 1 - Coin Flips

Take n coin flips, $x_1,\ x_2, \ldots, x_n$. The number of heads and tails are, n_0 and n_1, respectively. θ is the probability of getting heads. That makes the probability of tails $1-\theta$. This example shows how to estimate θ using data from the n coin flips and maximum likelihood estimation. Since each flip is independent, we can write the likelihood of the n coin flips as

$$\begin{aligned} L(x_1,\ x_2, \ldots, x_n|\theta) &= (1-\theta)^{n_0}\theta^{n_1} \\ \log L(x_1,\ x_2, \ldots, x_n|\theta) &= n_0 \log(1-\theta) + n_1 \log\theta \\ \frac{\partial}{\partial\theta} \log L(x_1,\ x_2, \ldots, x_n|\theta) &= \frac{-n_0}{1-\theta} + \frac{n_1}{\theta} \end{aligned}$$

To find the max of the original likelihood function, we set the derivative equal to zero and solve for θ. We call this estimated parameter $\hat{\theta}$.

$$\begin{aligned} 0 &= \frac{-n_0}{1-\theta} + \frac{n_1}{\theta} \\ n_0\theta &= n_1 - n_1\theta \\ (n_0+n_1)\theta &= n_1 \\ \theta &= \frac{n_1}{(n_0+n_1)} \\ \hat{\theta} &= \frac{n_1}{n} \end{aligned}$$

It is also important to check that this is not a minimum of the likelihood function and that the function maximum is not actually on the boundaries. In this case, this is the true maximum. This makes intuitive sense. The θ (probability of getting heads) that maximizes the likelihood of seeing this particular data is the number times heads came up over the total number of flips.

Example 2 - Normal with Known Variance

Example 2 is a from a continuous distribution, specifically the normal distribution. This time the x_i's are assumed to be chosen from normally distributed data with $\sigma^2 = 1$ and μ unknown.

The estimate for μ, $\hat{\theta}$, is found similarly to example 1:

$$\begin{aligned} L(x_1,\ x_2, \ldots, x_n|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} \\ \log L(x_1,\ x_2, \ldots, x_n|\theta) &= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi - \frac{(x_i-\theta)^2}{2} \\ \frac{\partial}{\partial\theta} \log L(x_1,\ x_2, \ldots, x_n|\theta) &= \sum_{i=1}^{n} (x_i - \theta) \end{aligned}$$

To find the theta that maximizes the first equation, we set the derivative equal to zero and solve.

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \theta) &= 0 \\ \sum_{i=1}^{n} x_i - n\theta &= 0 \\ \hat{\theta} &= \sum_{i=1}^{n} x_i/n \\ \hat{\theta} &= \bar{x} \end{aligned}$$

Again, it must be verified that this isn't a minimum and that the likelihood isn't higher on the boundaries. In this case this is the maximum. This tells us that the estimation for μ that maximizes the likelihood of seeing this data is the sample mean.

Example 3 - Normal with Both Parameters Unknown

Consider the x_i's drawn from a normal distribution with both μ and σ^2 both unknown. This time we have to estimate both parameters. θ_1 is the estimate of the mean and θ_2 is the estimate of the variance. We can go through the same steps as before:

$$\begin{aligned} L(x_1,\ x_2, \ldots, x_n|\theta_1, \theta_2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} e^{-(x_i-\theta_1)^2/2\theta_2} \\ \log L(x_1,\ x_2, \ldots, x_n|\theta_1, \theta_2) &= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi\theta_2 - \frac{(x_i-\theta_1)^2}{2\theta_2} \\ \frac{\partial}{\partial\theta_1} \log L(x_1,\ x_2, \ldots, x_n|\theta_1, \theta_2) &= \sum_{i=1}^{n} \frac{(x_i - \theta_1)}{\theta_2} \end{aligned}$$

We can set the derivative with respect to θ_1 equal to zero and solve for θ_1. The result is the same as in example 2:

$$\begin{aligned}
\sum_{i=1}^{n} \frac{(x_i - \theta_1)}{\theta_2} &= 0 \\
\sum_{i=1}^{n} x_i - n\theta_1 &= 0 \\
\hat{\theta_1} &= \sum_{i=1}^{n} x_i/n \\
\hat{\theta_1} &= \bar{x}
\end{aligned}$$

This result is the same as in example 2. We can use this estimation to find the estimate for the variance, θ_2.

$$\begin{aligned}
L(x_1,\ x_2, \ldots, x_n|\theta_1, \theta_2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} e^{-(x_i-\theta_1)^2/2\theta_2} \\
\log L(x_1,\ x_2, \ldots, x_n|\theta_1, \theta_2) &= \sum_{i=1}^{n} -\frac{1}{2} \ln 2\pi\theta_2 - \frac{(x_i - \theta_1)^2}{2\theta_2} \\
\frac{\partial}{\partial\theta_2} \log L(x_1,\ x_2, \ldots, x_n|\theta_1, \theta_2) &= \sum_{i=1}^{n} -\frac{1}{2}\frac{2\pi}{2\pi\theta_2} + \frac{(x_i - \theta_1)^2}{2\theta_2^2}
\end{aligned}$$

Setting equal to zero and solving for θ_2 gives:

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

This is a *consistent* estimate of the population variance, i.e., in the limit as n grows it equals the population variance. However, this estimate is *biased*. The unbiased estimate is:

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

(Why does it happen? Think about $n = 2$. $\hat{\theta}_1$ is exactly in the middle of the two sample points, whereas the population mean is unlikely to be (probability 0 in fact). This does not introduce bias in the mean estimate (it's too large as often as it's too small), but does systematically underestimate the variance.) The moral of the story is that MLE is a good idea, but it does not always perfectly estimate the true population parameters.

Expectation Maximization

The expectation maximization algorithm is commonly used when there is *hidden data*. For example, suppose samples are being drawn from a *mixture* of two or more different distributions, but the specification of which distribution is being sampled at each point is hidden. If

the data is believed to be from several distributions, then there are many more parameters. Each distribution has a mean, each distribution has a variance, and there are new parameters that relate the distributions called mixing parameters.

A typical example of this type of problem is a sample of heights of people. The female and male heights sampled come from different distributions. If we had thought to record sex along with height, the data would have a simple form and we could separately estimate the parameters of the two groups. But if sex is hidden, there may be a cluster of data points around the average female height and a cluster of data points around the average male height, and the best we can hope to do is jointly estimate the parameters somehow. For this example, there are two means μ_1 and μ_2. There are two variances σ_1^2 and σ_2^2. The mixing parameters(proportion of males/females in the sample) are τ_1 and $\tau_2 = 1 - \tau_1$. The PDF's are $f(x|\mu_1, \sigma_1^2)$ and $f(x|\mu_2, \sigma_2^2)$. Then the likelihood function becomes:

$$L(x_1,\ x_2, \ldots, x_n|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2) = \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i|\mu_j, \sigma_j^2)$$

There is no closed form solution for finding the set of parameters (θ) which maximize L. The likelihood surface is complex. There are multiple maxima, which arise because the group labels can be switched to get the same solution. There are also local maxima, which arise from different grouping of the data.

These problems wouldn't arise if the hidden data was known. That is, if there was a formula for separating the data into the correct groups. These are known as the z_{ij}'s. $z_{ij} = 1$ if x_i is drawn from distribution f_j, and $z_{ij} = 0$ otherwise. There is somewhat of a chicken vs. egg problem with z_{ij} and θ. If the z_{ij} were known, then θ could be estimated, and if θ was known, z_{ij} could be estimated. Since we know neither, we iterate through two steps (E and M), alternately estimating z's and θ's.

Classification EM

A simple version is to rigidly split the data in half (or into n groups) based on the estimated z_{ij}—if the probability that $z_{ij} = 1$ is greater (less) than 1/2, then assume it is 1 (0, resp.). Then recalculate θ based on that partition of data. Then recalculate the z_{ij} based on the new θ. Then recalculate θ assuming the newest z_{ij}. And on and on.

Full EM

The x_i's are known. θ, the set of parameters, is unknown. The goal is to find the θ which maximizes the hidden data likelihood, $L(x_1,\ x_2, \ldots, x_n|\theta)$. This would be easy if the z_{ij}'s were known. Then we would have $L(x_1,\ x_2, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2}|\theta)$, the complete data likelihood. However, the z_{ij}'s are now known, so instead maximize the expected likelihood of the known data: $E(L(x_1,\ x_2, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2}|\theta))$, where expectation is over the z_{ij}'s, the hidden data.

The E-Step

Assume θ is known and fixed.
A: the event that x_i was drawn from f_1.

B: the event that x_i was drawn from f_2.
D: the observed datum x_i.
The expected value of z_{i1} is $P(A|D)$, the probability of x_i being from f_1 given the observed x_i. $P(A|D) = \frac{P(D|A)P(A)}{P(D)}$ (Bayes' Rule).
From the law of total probability,

$$P(D) = P(D|A)P(A) + P(D|B)P(B)$$

$$P(D) = f_1(x_i|\theta_1)\tau_1 + f_2(x_i|\theta_2)\tau_2$$

This is repeated for all x_i.

The likelihood function for each x_i is $L(x_i, z_{ij}|\theta) = \tau_1 f_1(x_i|\theta)$ if $z_{i1} = 1$, and $L(x_i, z_{ij}|\theta) = \tau_2 f_2(x_i|\theta)$ otherwise. To get rid of if statement in the function, exploiting the fact that the z_{ij} are 0/1 indicators, the likelihood function can be written as (for x_1):

$$L(x_1, z_{1j}|\theta) = z_{11}\tau_1 f_1(x_1|\theta) + z_{12}\tau_2 f_2(x_1|\theta)$$

Or:

$$L(x_1, z_{1j}|\theta) = (\tau_1 f_1(x_1|\theta))^{z_{11}}(\tau_2 f_2(x_1|\theta))^{z_{12}}$$

The later form is more convenient, since we're about to take a log.

The M-Step

For simplicity, σ_1 and σ_2 are assumed to be σ. Also, τ_1 and τ_2 are assumed to be $\tau = 0.5$. Then we have a likelihood function in terms of a vector x and a vector z.

$$\begin{aligned}
L(x, z|\theta) &= \prod_{i=1}^{n} \frac{\tau}{\sqrt{2\pi\sigma^2}} \exp(-\sum_{j=1}^{2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2}) \\
E(\log L(x, z|\theta)) &= E\left[\sum_{i=1}^{n}\left(\log\tau - \frac{1}{2}\log 2\pi\sigma^2 - \sum_{j=1}^{2} z_{ij}\frac{(x_i - \mu_j)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{n}\left(\log\tau - \frac{1}{2}\log 2\pi\sigma^2 - \sum_{j=1}^{2} E[z_{ij}]\frac{(x_i - \mu_j)^2}{2\sigma^2}\right)
\end{aligned}$$

θ is μ_j in this case. This parameter is found as before, using $E[z_{ij}]$ found in the E-Step. The result is $\mu_j = \sum_{i=1}^{n} E[z_{ij}]x_i / \sum_{i=1}^{n} E[z_{ij}]$. This is an average, weighted by the subpopulation probability.

EM Summary

EM is fundamentally an MLE problem. $E(z)$ is estimated given θ, the E-Step. Then θ is estimated by maximizing E(likelihood) given $E(z)$, the M-Step. Then the steps are repeated.

EM may converge to a local, not global, max. However, it is still an effective method and widely used.