Maximum likelihood estimators and least squares

November 11, 2010

1 Maximum likelihood estimators

A maximum likelihood estimate for some hidden parameter λ (or parameters, plural) of some probability distribution is a number $\hat{\lambda}$ computed from an i.i.d. sample $X_1, ..., X_n$ from the given distribution that maximizes something called the "likelihood function". Suppose that the distribution in question is governed by a pdf $f(x; \lambda_1, ..., \lambda_k)$, where the λ_i's are all hidden parameters. The likelihood function associated to the sample is just

$$L(X_1, ..., X_n) \ = \ \prod_{i=1}^{n} f(X_i; \lambda_1, ..., \lambda_k).$$

For example, if the distribution is $N(\mu, \sigma^2)$, then

$$L(X_1, ..., X_n; \hat{\mu}, \hat{\sigma}^2) \ = \ \frac{1}{(2\pi)^{n/2}\hat{\sigma}^n} \exp\left(-\frac{1}{2\hat{\sigma}^2} \left((X_1 - \hat{\mu})^2 + \cdots + (X_n - \hat{\mu})^2 \right) \right). \tag{1}$$

Note that I am using $\hat{\mu}$ and $\hat{\sigma}^2$ to indicate that these are variable (and also to set up the language of estimators).

Why should one expect a maximum likelihood esimate for some parameter to be a "good estimate"? Well, what the likelihood function is measuring is how likely $(X_1, ..., X_n)$ is to have come from the distribution assuming particular values for the hidden parameters; the more likely this is, the closer one would think that those particular choices for hidden parameters are to the true values. Let's see two examples:

Example 1. Suppose that $X_1, ..., X_n$ are generated from a normal distribution having hidden mean μ and variance σ^2. Compute a MLE for μ from the sample.

Solution. As we said above, the likelihood function in this case is given by (1). It is obvious that to maximize L as a function of $\hat{\mu}$ and $\hat{\sigma}^2$ we must minimize

$$\sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

as a function of $\hat{\mu}$. Upon taking a derivative with respect to $\hat{\mu}$ and setting it to 0, we find that

$$\hat{\mu} \ = \ \frac{X_1 + \cdots + X_n}{n} \ = \ \overline{X},$$

the sample mean. So, the sample mean is the MLE for μ in this case.

Example 2. Now we give an example where calculus does not so easily apply: Suppose that $X_1, ..., X_n$ are random samples from a distribution that is uniform on $[0, N]$, where N is now the hidden parameter. We wish to produce a maximum likelihood estimate for N. In this case, the likelihood function is

$$L(X_1, ..., X_n; \hat{N}) \ = \ \begin{cases} 0, & \text{if any } X_i \text{ outside } [0, \hat{N}]; \\ (1/\hat{N})^n, & \text{if all } X_i \in [0, \hat{N}]. \end{cases}$$

Clearly, to maximize L, given $X_1, ..., X_n$, we should choose $\hat{N}$ to be $\max(X_1, ..., X_n)$; and note that we got this MLE without using calculus.

Now it turns out that $\max(X_1, ..., X_n)$ is actually a *biased* estimator for N; and so, *maximum likelihood estimates need not be unbiased.* Let us see that this is so: We must compute the expected value for $\hat{N}$, and for this we will need its pdf. Naturally, we start by finding the cdf for $\hat{N}$. We have that

$$\begin{aligned} \mathbb{P}(\hat{N} \ \leq \ x) \ &= \ \mathbb{P}(X_1 \leq x, \ X_2 \leq x, \ ..., \ X_n \leq x) \\ &= \ \mathbb{P}(X_1 \leq x) \cdots \mathbb{P}(X_n \leq x) \\ &= \ \mathbb{P}(X_1 \leq x)^n \\ &= \ (x/N)^n. \end{aligned}$$

Taking a derivative, then, we find that if $f(x)$ is the pdf for $\hat{N}$, then

$$f(x) \ = \ \begin{cases} 0, & \text{if } x < 0 \text{ or } x > N; \\ n(x/N)^{n-1}, & \text{if } x \in [0, N]. \end{cases}$$

So,

$$\mathbb{E}(\hat{N}) \ = \ \int_0^N nx^n/N^{n-1}dx \ = \ nN/(n+1),$$

which is *not* equal to N, meaning that $\hat{N}$ is a *biased* estimator for N.

2 Least squares

Suppose that you are presented with a sequence of data points (X_1, Y_1), ..., (X_n, Y_n), and you are asked to find the "best fit" line passing through those points. Well, of course, in order to answer this you need to know precisely how to tell whether one line is "fitter" than another. A common measure of fitness is the square-error, given as following: Suppose $y = \lambda_1 x + \lambda_2$ is your candidate line. Then, the error associated with this line is

$$E \ := \ \sum_{i=1}^n (Y_i - \lambda_1 X_i - \lambda_2)^2.$$

In other words, it is the sum of the square distance between the y-value at the data points $X = X_1, X_2, ..., X_n$ and the y-value for the line at those data points.

Why use the sum of square errors? Well, first of all, the fact that we compute squares means that all the terms in the sum are non-negative and error at a given point $X = X_i$ is the same if the point (X_i, Y_i) is t units above the line $y = \lambda_1 x + \lambda_2$ as it is if it is t units below the line. Secondly, squaring is a "smooth operation"; and so, we can easily compute derivatives of E – in other words, using sum of square errors allows us to use calculus. And finally, at the end of this note we will relate the sum of square error to MLE's.

Minimizing E over all choices for (λ_1, λ_2) results in what is called the "least squares approximation". Let us see how to compute it: Well, basically we take a partial of E with respect to λ_1 and λ_2 and then set those equal to 0; so, we have the equations

$$\begin{aligned} 0 \ &= \ \frac{\partial E}{\partial \lambda_1} \ = \ \sum_{i=1}^n 2(Y_i - \lambda_1 X_i - \lambda_2)(-X_i) \\ 0 \ &= \ \frac{\partial E}{\partial \lambda_2} \ = \ \sum_{i=1}^n 2(Y_i - \lambda_1 X_i - \lambda_2)(-1). \end{aligned}$$

Upon rearranging these equations, collecting coefficients of λ_1 and λ_2, we find they are equivalent to:

$$\begin{aligned} \lambda_1 \left(\sum_{i=1}^n X_i^2 \right) + \lambda_2 \left(\sum_{i=1}^n X_i \right) &= \sum_{i=1}^n X_i Y_i \\ \lambda_1 \left(\sum_{i=1}^n X_i \right) + \lambda_2 \cdot n &= \sum_{i=1}^n Y_i. \end{aligned}$$

Now suppose that instead of $y = \lambda_1 x + \lambda_2$ we have a sequence of differentiable functions $f_1(x), ..., f_k(x)$ and that we seek paramters $\lambda_1, ..., \lambda_k$ so that

$$y = \lambda_1 f_1(x) + \cdots + \lambda_k f_k(x)$$

is a best-fit curve to a set of data points $(X_1, Y_1), ..., (X_k, Y_k)$. Then, it turns out that the above "least squares" approach will also work for this problem.

3 MLE's again, and least squares

In this section we consider a different sort of problem related to "best fit lines". Suppose that we know *a priori* that the data points (X_i, Y_i) fit a straight line, except that there is a little error involved. That is to say, suppose that $X_1, ..., X_n$ are fixed and that we think of $Y_1, ..., Y_n$ as being random variables satisfying

$$Y_i = \lambda_1 X_i + \lambda_2 + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma^2),$$

where all the ε_i's are assumed to be independent.

This sort of situation is quite common in the sciences, particularly physics: Imagine that the position of a particle as a function of time satisfies $P = \lambda_1 T + \lambda_2$; however, in tracking the particle, there is some uncertainty as to its exact position, and this uncertainty is roughly normal with mean 0 and variance σ^2. Then, if we let P' be the *observed* position of the particle, we have that $P' = \lambda_1 T + \lambda_2 + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$.

Now we find a MLE estimate for λ_1, λ_2. Our likelihood function is given by (we assume $X_1, ..., X_n$ are fixed)

$$L(Y_1, ..., Y_n; \lambda_1, \lambda_2, \sigma^2) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left(\frac{-1}{2\sigma^2} \left(\sum_{i=1}^n (Y_i - \lambda_1 X_i - \lambda_2)^2 \right) \right).$$

Clearly, for any fixed $\sigma > 0$, maximizing L is equivalent to minimizing

$$E \ := \ \sum_{i=1}^{n} (Y_i - \lambda_1 X_i - \lambda_2)^2.$$

So we see that the least squares estimate we saw before is really equivalent to producing a maximum likelihood estimate for λ_1 and λ_2 for variables X and Y that are linearly related up to some Gaussian noise $N(0, \sigma^2)$. The significance of this is that it makes the least-squares method of linear curve fitting a little more natural – it's not as artificial as it might have seemed at first: What made it seem artificial, at first, was the fact that there are many, many different error functions that we could have written down that measure how well the line $y = \lambda_1 x + \lambda_2$ fits the given data. And what we have shown is that the "sum of square errors" error function happens to have a privileged position among them.