# Introduction to Maximum Likelihood Estimation

A slecture by ECE student Wen Yi

## 1. Introduction

In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters.

In maximum likelihood estimation, we search over all possible sets of parameter values for a specified model to find the set of values for which the observed sample was most likely. That is, we find the set of parameter values that, given a model, were most likely to have given us the data that we have in hand.

## 2. Basic method

Suppose there is a sample $x_1, x_2, \dots, x_N$ of n independent and identically distributed observations from a distribution with an unknown probability density function $f_0$. We can say that the function $f_0$ belongs to a certain family of distributions $\{f(x \mid \theta), \theta \in \Theta\}$, where θ is a vector of parameters for this family, so that so that $f_0 = f(x|\theta_0)$. The value $\theta_0$ is unknown and is referred to as the true value of the parameter. So, using MLE, we want to find an estimator which would be as close to the true value $\theta_0$ as possible.

To use the method of maximum likelihood, one first specifies the joint density function for all observations. For an independent and identically distributed sample, this joint density function is

$$f(x_1, x_2, \dots, x_N|\theta) = f(x_1|\theta)f(x_2|\theta)f(x_3|\theta) \times \dots \times f(x_N|\theta)$$

As each sample $x_i$ is independent with each other, the likelihood of $\theta$ with the observation of samples $x_1, x_2, \dots, x_n$ can be defined as:

$$\mathrm{L}(\theta; x_1, x_2, \dots, x_N) = f(x_1, x_2, \dots, x_N|\theta) = \prod_{i=1}^{N} f(x_i \mid \theta)$$

In practice, it's more convenient to take ln for the both sides, called log-likelihhod. Then the formula becomes:

$$ln\mathrm{L}(\theta; x_1, x_2, \dots, x_N) = \sum_{i=1}^{N} lnf(x_i \mid \theta)$$

Then, for a fixed set of samples, to maximize the likelihood of $\theta$, we should choose the data that satisfied:

$$\{\hat{\theta}_{MLE}\} = \left\{\underset{\theta\in\Theta}{argmax}\; ln\mathrm{L}(\theta; x_1, x_2, \dots, x_N)\right\} = \left\{\underset{\theta\in\Theta}{argmax} \sum_{i=1}^{N} lnf(x_i|\theta)\right\}$$

To find the maximum of $ln\mathrm{L}(\theta; x_1, x_2, \dots, x_N)$, we take the derivative of $\theta$ on it and find the$\theta$ value that make the derivation equals to 0.

$$\frac{d}{d\theta} ln\mathrm{L}(\theta; x_1, x_2, \dots, x_N) = 0$$

To check our result we should garentee that the second derivative of $\theta$ on $ln\mathrm{L}(\theta; x_1, x_2, \dots, x_n)$ is negative.

$$\frac{d^2}{d^2\theta} ln\mathrm{L}(\theta; x_1, x_2, \dots, x_N) < 0$$

## 3. Practice considerations

3.1 Log-likelihood

Just as mentioned above, to make life a little easier, we can work with the natural log of likelihoods rather than the likelihoods themselves. The main reason for this is, computational rather than theoretical. If you multiply lots of very small numbers together (say all less than 0.0001) then you will very quickly end up with a number that is too small to be represented by any calculator or computer as different from zero. This situation will often occur in calculating likelihoods, when we are often multiplying the probabilities of lots of rare but independent events together to calculate the joint probability.

With log-likelihoods, we simply add them together rather than multiply them (log-likelihoods will always be negative, and will just get larger (more negative) rather than approaching 0).

So, log-likelihoods are conceptually no different to normal likelihoods. When we optimize the log-likelihood, with respect to the model parameters, we also optimize the likelihood with respect to the same parameters, for there is a one-to-one (monotonic) relationship between numbers and their logs.

3.2 Removing the constant

For example the likelihood function for the binomial distribution is:

$$L_{binomial} = \binom{n}{k} p^k (1-p)^{n-k}$$

In the context of MLE, we noted that the values representing the data will be fixed: these are *n* and *k*. In this case, the binomial 'co-efficient' depends only upon these constants. Because it does not depend on the value of the parameter *p* we can essentially ignore this first term. This is because any value for *p* which maximizes the above quantity will also maximize

$$p^k (1-p)^{n-k}$$

This means that the likelihood will have no meaningful scale in and of itself. This is not usually important, however, for as we shall see, we are generally interested not in the absolute value of the likelihood but rather in the *ratio* between two likelihoods - in the context of a likelihood ratio test.

We may often want to ignore the parts of the likelihood that do not depend upon the parameters in order to reduce the computational intensity of some problems. Even in the simple case of a binomial distribution, if the number of trials becomes very large, the calculation of the factorials can become infeasible.

3.3 Numerical MLE

Sometimes we cannot write an equation that can be differentiated to find the MLE parameter estimates. This is especially likely if the model is complex and involves many parameters and/or complex probability functions. (e.g. the normal mixture probability distribution)

In this scenario, it is also typically not feasible to evaluate the likelihood at all points, or even a reasonable number of points. In the parameter space of the problem in the coin toss example, the parameter space was only one-dimensional (i.e. only one parameter) and ranged between 0 and 1. Nonetheless, because p can theoretically take any value between 0 and 1, the MLE will always be an approximation (albeit an incredibly accurate one) if we just evaluate the likelihood for a finite number of parameter values. For example, we chose to evaluate the likelihood at steps of 0.02. But we could have chosen steps of 0.01, of 0.001, of 0.000000001, etc. In theory and practice, one has to set a minimum tolerance by which you are happy for your estimates to be out. This is why computers are essential for these types of problems: they can tabulate lots and lots of values very quickly and therefore achieve a much finer resolution.

## 4. Some basic examples

4.1 Poisson Distribution

For Poisson distribution the expression of probability is:

$$\mathrm{P}(\mathrm{X}=x)=\frac{\lambda^{x}e^{-\lambda}}{x!}$$

Let $X_1, X_2, \dots, X_N$ be the Independent and identically distributed (iid) Poisson random variables. Then, we will have a joint frequency function that is the product of marginal frequency functions. The log likelihood of Poisson distribution thus should be:

$$ln\mathrm{L}(\lambda; X_1, X_2, \dots, X_N)=\sum_{i=1}^{N} lnf(X_i|\lambda)=\sum_{i=1}^{N} ln\frac{\lambda^{X_i}e^{-\lambda}}{X_i!}=\sum_{i=1}^{N}(X_i ln\lambda-\lambda-\ln X_i!)$$

$$=ln\lambda\sum_{i=1}^{N}X_i-N\lambda-\sum_{i=1}^{N}\ln X_i!$$

Take the derivative of $\lambda$ on it and find the$\lambda$ value that make the derivation equals to 0.

$$\frac{d}{d\lambda}ln\mathrm{L}(\lambda; X_1, X_2, \dots, X_N)=0$$

$$\frac{d}{d\lambda}\left(ln\lambda\sum_{i=1}^{N}X_i-N\lambda-\sum_{i=1}^{N}\ln X_i!\right)=0$$

$$\frac{\sum_{i=1}^{N}X_i}{\lambda}-N=0$$

$$\lambda=\frac{\sum_{i=1}^{N}X_i}{N}$$

Thus, the ML estimation for Poisson distribution should be:

$$\hat{\lambda}=\bar{X}$$

4.2 Exponential distribution

For exponential distribution the expression of probability is:

$$\mathrm{P}(\mathrm{X}=x)=\begin{cases}\lambda e^{-\lambda x} & (x\geq 0)\\ 0 & (x<0)\end{cases}$$

Let $X_1, X_2, \dots, X_N$ be the Independent and identically distributed (iid) exponential random variables. As $\mathrm{P}(\mathrm{X}=x)=0$ when x<0, no samples can sit in x<0 region. Thus, for all $X_1, X_2, \dots, X_N$, we can only focus on the $x\geq 0$ part. Then, we will have a joint frequency function that is the product of marginal frequency functions. The log likelihood of exponential distribution thus should be:

$$ln\mathrm{L}(\lambda; X_1, X_2, \dots, X_N)=\sum_{i=1}^{N} lnf(X_i|\lambda)=\sum_{i=1}^{N} ln\lambda e^{-\lambda X_i}=\sum_{i=1}^{N}(ln\lambda-\lambda X_i)=\mathrm{N}ln\lambda-\lambda\sum_{i=1}^{N} X_i$$

Take the derivative of $\lambda$ on it and find the$\lambda$ value that make the derivation equals to 0.

$$\frac{d}{d\lambda}ln\mathrm{L}(\lambda; X_1, X_2, \dots, X_N)=0$$

$$\frac{d}{d\lambda}\left(\mathrm{N}ln\lambda-\lambda\sum_{i=1}^{N} X_i\right)=0$$

$$\frac{N}{\lambda}-\sum_{i=1}^{N} X_i=0$$

$$\lambda=\frac{N}{\sum_{i=1}^{N} X_i}$$

Thus, the ML estimation for exponential distribution should be:

$$\hat{\lambda}=\frac{1}{\overline{X}}$$

4.3 Gaussian distribution

For Gaussian distribution the expression of probability is:

$$\mathrm{P}(\mathrm{X}=x)=\frac{1}{(2\pi)^{\frac{1}{2}}}\cdot\frac{1}{|\Sigma|^{\frac{1}{2}}}\cdot\exp\left\{-\frac{(x-\mu)^2}{2\Sigma}\right\}$$

Let $X_1, X_2, \dots, X_N$ be the Independent and identically distributed (iid) Gaussian random variables. Then, we will have a joint frequency function that is the product of marginal frequency functions. The log likelihood of Gaussian distribution thus should be:

$$ln\mathrm{L}(\mu,\Sigma; X_1, X_2, \dots, X_N)=\sum_{i=1}^{N} lnf(X_i|\mu,\Sigma)=\sum_{i=1}^{N} ln\frac{1}{(2\pi)^{\frac{1}{2}}}\cdot\frac{1}{|\Sigma|^{\frac{1}{2}}}\cdot\exp\left\{-\frac{(X_i-\mu)^2}{2\Sigma}\right\}$$

$$=\frac{\mathrm{N}}{2}\ln\frac{1}{2\pi}+\frac{\mathrm{N}}{2}\ln\frac{1}{\Sigma}-\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{2\Sigma}$$

Take the derivative of $\mu,\Sigma$ on it and find the $\mu,\Sigma$ value that make the derivation equals to 0.

$$\frac{d}{d\mu}ln\mathrm{L}(\mu,\Sigma; X_1, X_2, \dots, X_N)=\frac{d}{d\mu}-\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{2\Sigma}=-\frac{\sum_{i=1}^{N}(X_i-\mu)}{\Sigma}=0$$

$$\sum_{i=1}^{N}(X_i-\mu)=0$$

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$\frac{d}{d\Sigma} ln\mathrm{L}(\mu, \Sigma; X_1, X_2, \dots, X_N) = \frac{d}{d\Sigma}\frac{\mathrm{N}}{2}\ln\frac{1}{\Sigma} - \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{2\Sigma} = -\frac{\mathrm{N}}{2\Sigma} + \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{2\Sigma^2} = 0$$

$$\Sigma = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Thus, the ML estimation for Gaussian distribution should be:

$$\hat{\mu} = \bar{X}$$

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

## 5. Some advanced examples

5.1 Expression of Estimated Parameters

The above estimation all base on the assumption that the distribution to be estimated follows the distribution of a single function, but how about the estimation of the mixture of functions?

To simplify the problem, we only talk about Gaussian Mixture Model (GMM) here. Using the same method, it's easy to extend it to other kind of mixture model and the mixture between different models.

To start with, we should know that if we set the number of Gaussian function to be used in the GMM estimation flexible, we will find out that the number of Gaussian function will never reach a best solution, as adding more Gaussian functions into the estimation will subsequently improve the accuracy anyway. As calculating how many Gaussian function is include in GMM is a clustering problem. We assume to know the number of Gaussian function in GMM as k here.

As this distribution is a mixture of Gaussian, the expression of probability is:

$$p(\mathrm{X} = x) = \sum_{j=1}^{k} g_j(x)\alpha_j$$

$\alpha_j$ is the weight of Gaussian function $g_j(x)$.

$$g_j(x) = \frac{1}{(2\pi)^{\frac{1}{2}}} \cdot \frac{1}{|\Sigma_j|^{\frac{1}{2}}} \cdot \exp\left\{-\frac{(x - \mu_j)^2}{2\Sigma_j}\right\}$$

Thus, the parameters to be estimated are:

$$\theta = (\mu_1, \mu_2, \dots, \mu_k, \Sigma_1, \Sigma_2, \dots, \Sigma_k, \alpha_1, \alpha_2, \dots, \alpha_k)$$

Let $X_1, X_2, \dots, X_N$ be the Independent and identically distributed (iid) Gaussian Mixture Model (GMM) random variables.

Following Bayes rule, the responsibility that a mixture component takes for explaining an observation $X_i$ is:

$$\mathrm{P}(\mathrm{j}|X_i, \theta) = \frac{p(X_i|j, \theta)P(j|\theta)}{p(X_i|\theta)} = \frac{g_j(X_i)\alpha_j}{\sum_{l=1}^{k} g_l(X_i)\alpha_l}$$

Then, we will have a joint frequency function that is the product of marginal frequency functions. The log likelihood of Gaussian Mixture Model distribution thus should be:

$$lnL(\theta; X_1, X_2, \dots, X_N) = \sum_{i=1}^{N} lnp(X_i|\theta)$$

Take the derivative of $\mu_j, \Sigma_j$ on it and find the $\mu_j, \Sigma_j$ value that make the derivation equals to 0.

$$\frac{d}{d\mu_j} lnL(\theta, \Sigma; X_1, X_2, \dots, X_N) = \frac{d}{d\mu_j} \sum_{i=1}^{N} lnp(X_i|\theta) = \sum_{i=1}^{N} \frac{d}{d\mu_j} lnp(X_i|\theta)$$

$$= \sum_{i=1}^{N} \frac{1}{p(X_i|\theta)} \frac{d}{d\mu_j} p(X_i|\theta) = \sum_{i=1}^{N} \frac{1}{p(X_i|\theta)} \frac{d}{d\mu_j} \sum_{j=1}^{k} g_j(X_i)\alpha_j$$

$$= \sum_{i=1}^{N} \frac{1}{p(X_i|\theta)} \frac{d}{d\mu_j} \left(g_j(X_i)\alpha_j\right)$$

$$= \sum_{i=1}^{N} \frac{1}{p(X_i|\theta)} \frac{d}{d\mu_j} \left( \frac{1}{(2\pi)^{\frac{1}{2}}} \cdot \frac{1}{\left|\Sigma_j\right|^{\frac{1}{2}}} \cdot \exp\left\{ -\frac{\left(X_i - \mu_j\right)^2}{2\Sigma_j} \right\} \alpha_j \right)$$

$$= \sum_{i=1}^{N} \frac{1}{p(X_i|\theta)} \frac{1}{(2\pi)^{\frac{1}{2}}} \cdot \frac{1}{\left|\Sigma_j\right|^{\frac{1}{2}}} \cdot \exp\left\{ -\frac{\left(X_i - \mu_j\right)^2}{2\Sigma_j} \right\} \alpha_j \left( -\frac{2\left(X_i - \mu_j\right)}{2\Sigma_j} \right)$$

$$= \sum_{i=1}^{N} \frac{g_j(X_i)\alpha_j}{p(X_i|\theta)} \left( -\frac{2\left(X_i - \mu_j\right)}{2\Sigma_j} \right) = -\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i, \theta) \frac{X_i - \mu_j}{\Sigma_j}$$

$$\frac{d}{d\mu_j} lnL(\theta, \Sigma; X_1, X_2, \dots, X_N) = 0$$

$$\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i, \theta) \frac{X_i - \mu_j}{\Sigma_j} = 0$$

$$\mu_j = \frac{\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i, \theta) X_i}{\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i, \theta)}$$

$$\frac{d}{d\Sigma_j} ln\mathrm{L}(\theta,\Sigma;X_1,X_2,\dots,X_N) = \sum_{i=1}^{N} \frac{1}{p(X_i|\theta)} \frac{d}{d\Sigma_j}\left(g_j(X_i)\alpha_j\right)$$

$$= \sum_{i=1}^{N} \frac{\alpha_j}{p(X_i|\theta)} \cdot \frac{1}{(2\pi)^{\frac{1}{2}}} \cdot \frac{d}{d\Sigma_j}\left(\frac{1}{\left|\Sigma_j\right|^{\frac{1}{2}}} \cdot \exp\left\{-\frac{\left(X_i-\mu_j\right)^2}{2\Sigma_j}\right\}\right)$$

$$= \sum_{i=1}^{N} \frac{\alpha_j}{p(X_i|\theta)}$$

$$\cdot \frac{1}{(2\pi)^{\frac{1}{2}}}\left(-\frac{1}{2\left|\Sigma_j\right|^{\frac{3}{2}}}\exp\left\{-\frac{\left(X_i-\mu_j\right)^2}{2\Sigma_j}\right\} + \frac{1}{\left|\Sigma_j\right|^{\frac{1}{2}}} \cdot \exp\left\{-\frac{\left(X_i-\mu_j\right)^2}{2\Sigma_j}\right\} \cdot \frac{\left(X_i-\mu_j\right)^2}{2{\Sigma_j}^2}\right)$$

$$= \sum_{i=1}^{N} \frac{\alpha_j}{p(X_i|\theta)} \cdot \frac{1}{2(2\pi)^{\frac{1}{2}}\left|\Sigma_j\right|^{\frac{3}{2}}}\exp\left\{-\frac{\left(X_i-\mu_j\right)^2}{2\Sigma_j}\right\}\left(-1+\frac{\left(X_i-\mu_j\right)^2}{\Sigma_j}\right)$$

$$= \sum_{i=1}^{N} \frac{g_j(X_i)\alpha_j}{p(X_i|\theta)} \cdot \frac{1}{2\Sigma_j}\left(-1+\frac{\left(X_i-\mu_j\right)^2}{\Sigma_j}\right)$$

$$= \sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta) \cdot \frac{1}{2\Sigma_j}\left(-1+\frac{\left(X_i-\mu_j\right)^2}{\Sigma_j}\right)$$

$$\frac{d}{d\Sigma_j} ln\mathrm{L}(\theta,\Sigma;X_1,X_2,\dots,X_N) = 0$$

$$\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta) \cdot \frac{1}{2\Sigma_j}\left(-1+\frac{\left(X_i-\mu_j\right)^2}{\Sigma_j}\right) = 0$$

$$\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta)\left(\left(X_i-\mu_j\right)^2 - \Sigma_j\right) = 0$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta)\left(X_i-\mu_j\right)^2}{\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta)}$$

The $\alpha_j$ is subject to $\sum_{j=1}^{k} \alpha_j = 1$. Basic optimization theories show that $\alpha_j$ $is\ optimized\ by$:

$$\alpha_j = \frac{\sum_{i=1}^{N} p(\mathrm{j}|X_i,\theta)}{N}$$

Thus, the ML estimation for Gaussian Mixture Model distribution should be:

$$\widehat{\mu_j} = \frac{\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta)X_i}{\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta)};\ \widehat{\Sigma_j} = \frac{\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta)\left(X_i-\widehat{\mu_j}\right)^2}{\sum_{i=1}^{N} \mathrm{P}(\mathrm{j}|X_i,\theta)};\ \alpha_j = \frac{\sum_{i=1}^{N} p(\mathrm{j}|X_i,\theta)}{N}$$

## 5.2 Practical Implementation

Now we can observe that, as the Gaussian Mixture Model with K Gaussian functions have 3K parameters, to find the best vector of parameters set, $\theta$, is to find the optimized parameters in 3K dimension space. As the Gaussian Mixture Model include more Gaussian functions, the complexity of computing the best $\theta$ will go incrediblily high. Also, we can see that all the expressions of $\mu$, $\Sigma$ and $\alpha$ include themselves directly or indirectly, it's implossible to get the value of the parameters within one time calculation.

Now it's time to introduce a method for finding maximum likelihood with large number of latent variables (parameters), Expectation–maximization (EM) algorithm.

In statistics, an expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables (the parameters). The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

In short words, to get the best $\theta$ for our maximum likelihood, firstly, for the expectation step, we should evaluate the weight of each cluster with the current parameters. Then, for the maximization step, we re-estimate parameters using the existing weight.

By repeating these calculation process for several times, the parameters will approach the value for the maximum likelihood.

## 6. References


http://www.cscu.cornell.edu/news/statnews/stnews50.pdf
http://en.wikipedia.org/wiki/Maximum_likelihood
http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_1.html
http://eniac.cs.qc.cuny.edu/andrew/gcml-11/lecture10c.pptx
http://statweb.stanford.edu/~susan/courses/s200/lectures/lect11.pdf