Introduction to Maximum Likelihood Estimation

1. General formulation

A. Problem statement

Given data set $\mathbf{X} = \{\mathbf{x}_j\}$, $j \in \{1, 2, \ldots, n\}$ that is identically and independently distributed, and a parametric density function (pdf) $p(\mathbf{x}|\theta)$ with parameters θ, find θ^* such that:

$$p(\mathbf{X}|\theta^*) \geq p(\mathbf{X}|\theta),\ \forall\lambda. \tag{1}$$

That is, we want to find parameter values θ^* that maximizes the joint likelihood of all the data given the probability density function. For example, for the case of a one-dimensional normal density,

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \tag{2}$$

B. Maximum-likelihood estimation

Let us first write an expression for $p(\mathbf{X}|\theta)$, given the assumption that the data are identically and independently distributed:

$$p(\mathbf{X}|\theta) = \prod_{j=1}^{n} p(\mathbf{x}_j|\theta) \tag{3}$$

The product in equation (3) can be converted to a sum by taking the natural logarithm of both sides of the equation:

$$\ln p(\mathbf{X}|\theta) = \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\theta) \tag{4}$$

[Note: $\ln(ab) = \ln(a) + \ln(b)$.]

Typically, equation (4) is easier to maximize; note, however, that maximizing equation (4) with respect to θ also maximizes equation (3) since the logarithm function is monotonic and increasing. One approach to maximizing the log-likelihood of $\mathbf{X}$ with respect to θ is to take the gradient with respect to the parameters θ, setting the resulting set of equations equal to zero, and solving for the parameters θ:

$$\nabla_\theta \ln p(\mathbf{X}|\theta) = 0 \tag{5}$$

$$\nabla_\theta \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\theta) = 0 \tag{6}$$

$$\sum_{j=1}^{n} \nabla_\theta \ln p(\mathbf{x}_j|\theta) = 0 \tag{7}$$

Whether or not equation (7) is easy or difficult to solve for θ largely depends on the functional form of the likelihood function $p(\mathbf{x}|\theta)$. Note, however, that equation (7) is easy to solve for a large family of exponential probability density functions. In such cases, a closed-form solution exists for the maximum-likelihood parameter estimates. In other cases, equation (7) cannot be solved directly which will lead us to the development of an iterative algorithm for maximum-likelihood estimation known as Expectation-Maximization.

2. Maximum-likelihood estimation examples

A. Univariate Normal density

Problem statement: Given a one-dimensional set of identically and independently distributed data $\mathbf{X} = \{x_j\}$, $j \in \{1, 2, \ldots, n\}$, compute the maximum-likelihood estimates for the parameters θ of the Gaussian probability density function.

Solution: The parameters for a one-dimensional Gaussian are given by,

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \tag{8}$$

and the likelihood function is given by,

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \tag{9}$$

The log-likelihood of $\mathbf{X}$ given θ is given by,

$$\begin{aligned} \ln p(\mathbf{X}|\theta) &= \sum_{j=1}^{n} \ln p(x_j|\theta) \\ &= \sum_{j=1}^{n}\left[\frac{-(x_j-\mu)^2}{2\sigma^2} - \ln\sigma - \ln\sqrt{2\pi}\right] \end{aligned} \tag{10}$$

To maximize equation (10) with respect to μ and σ^2, we now compute the derivatives with respect to each, set the derivatives equal to zero, and solve for the two parameters:

$$\frac{\partial}{\partial\mu}\ln p(\mathbf{X}|\theta) = \sum_{j=1}^{n}\frac{(x_j-\mu^*)}{\sigma^2} = 0 \tag{11}$$

$$\sum_{j=1}^{n}(x_j-\mu^*) = 0 \tag{12}$$

$$\left(\sum_{j=1}^{n} x_j\right) - n\mu^* = 0 \tag{13}$$

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_j \tag{14}$$

Now, solving for σ^2:

$$\frac{\partial}{\partial\sigma}\ln p(\mathbf{X}|\theta) = \sum_{j=1}^{n}\left[\frac{(x_j-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\right] = 0 \tag{15}$$

$$\sum_{j=1}^{n}[(x_j-\mu)^2 - \sigma^2] = 0 \tag{16}$$

$$\left(\sum_{j=1}^{n}(x_j-\mu)^2\right) - n\sigma^2 = 0 \tag{17}$$

$$(\sigma^2)^* = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)^2 \tag{18}$$

Thus, the maximum-likelihood estimates for μ and σ^2 are given by,

$$\mu = \frac{1}{n}\sum_{i=1}^{n}x_j \tag{19}$$

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)^2 \tag{20}$$

B. Multivariate Normal density

The results in equations (19) and (20) generalize to the d-dimensional Normal density,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] \tag{21}$$

and d-dimensional data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$:

$$\mu = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j \tag{22}$$

$$\Sigma = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j-\mu)(\mathbf{x}_j-\mu)^T \tag{23}$$

C. Non-Gaussian density

<u>Problem statement</u>: Given a one-dimensional set of identically and independently distributed data $\mathbf{X} = \{x_j\}$, $j \in \{1, 2, \ldots, n\}$, compute the maximum-likelihood estimate for the parameter γ of the following probability density function:

$$p(x|\gamma) = \begin{cases}(\gamma+1)x^{\gamma} & 0 < x \le 1, \gamma > -1 \\ 0 & \text{elsewhere}\end{cases} \tag{24}$$

The likelihood function $p(x|\gamma)$ in equation (24) is plotted in Figure 1 below.

<u>Solution</u>: The log-likelihood of $\mathbf{X}$ given γ is given by,

$$\begin{aligned}\ln p(\mathbf{X}|\gamma) &= \sum_{j=1}^{n}\ln p(x_j|\gamma) \\ &= \sum_{j=1}^{n}[\ln(\gamma+1) + \gamma\ln(x_j)]\end{aligned} \tag{25}$$

To maximize equation (25) with respect to γ, we now compute the derivative with respect to γ, set the derivative equal to zero, and solve for the unknown parameter:

$$\frac{\partial}{\partial\gamma}\ln p(\mathbf{X}|\gamma) = \sum_{j=1}^{n}\left[\frac{1}{(\gamma^*+1)} + \ln(x_j)\right] = 0 \tag{26}$$

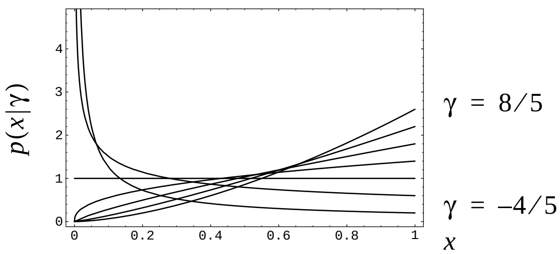

Figure 1

$$\frac{n}{(\gamma^* + 1)} + \sum_{j=1}^{n} \ln(x_j) = 0 \tag{27}$$

$$(\gamma^* + 1)\sum_{j=1}^{n} \ln(x_j) = -n \tag{28}$$

$$\gamma^* = -n/\left(\sum_{j=1}^{n} \ln(x_j)\right) - 1 \tag{29}$$