Maximum Likelihood Estimate

Consider a model parametrized by a vector θ, and let $X = (x_1, \dots x_N)$ be observed data samples from the model. Then the function $p(X|\theta)$ is called the *likelihood function* if viewed as a function of the parameter vector θ. It shows how probable the observed data X is for different values of θ. Note that the likelihood is not a probability distribution over θ (its integral with respect to θ may not be equal to one).

For example, if we consider the set X of independently drawn samples from the normal distribution with unknown parameters, then $\theta = (\mu, \sigma^2)$, and the likelihood function is

$$p(X|\theta) = \mathcal{N}(X|\mu, \sigma) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right)$$

treated as a function of μ and σ.

The Maximum Likelihood Estimate (MLE) for parameter is the value of θ which maximizes the likelihood. It is a very common way in statistics to estimate the unknown parameters for the model after observing the data.

Continuing the example above, let us find the MLE for parameter μ. As we assumed that samples are drawn independently from the model, the likelihood takes the form of a *product* of individual likelihood functions for each sample $p(x_i|\theta)$. When finding the MLE it is often convenient to find the maximum of the function $\log p(X|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta)$ (which in its turn, takes the form of the *sum* of individual log likelihood functions) instead of directly optimizing $p(X|\theta)$:

$$\log p(X|\theta) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

To maximize this expression with respect to μ, we set the partial derivative with respect to μ to zero and obtain:

$$\frac{\partial}{\partial\mu}p(X|\mu, \sigma) = -\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \mu) = 0$$

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Let us also consider multidimensional case for this problem: now each x_i is a d-dimensional vector drawn from the multivariate normal distribution with parameters mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Similarly, the log likelihood for this model takes the form:

$$\log p(X|\mu, \Sigma) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log\det\Sigma - \sum_{i=1}^{N}\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) =$$

$$= -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T\Sigma^{-1}x_i - \mu^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu\right) =$$

$$\Big[\text{use the fact that } \Sigma \text{ is symmetric, thus } \Sigma^{-1} \text{ is also symmetric which leads to:}$$

$$\mu^T\Sigma^{-1}x_i = \left(\mu^T\Sigma^{-1}x_i\right)^T = x_i^T\left(\Sigma^{-1}\right)^T\mu = x_i^T\Sigma^{-1}\mu\Big]$$

$$= -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{N}\left(x_i^T\Sigma^{-1}x_i - 2x_i^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu\right)$$

Now to obtain the MLE for μ, we need to compute the derivative of this expression with respect to vector μ and set it to zero. We will use the following vector differentiation rules:

$$\frac{\partial}{\partial y}(a^Ty) = a \quad \text{for } y \in \mathbb{R}^d, a \in \mathbb{R}^d$$

$$\frac{\partial}{\partial y}(y^TAy) = 2Ay \quad \text{for } y \in \mathbb{R}^d \text{ and symmetric matrix } A \in \mathbb{R}^{d\times d}$$

Applying them to the log likelihood expression, we get:

$$\frac{\partial}{\partial \mu}\log p(X|\mu, \Sigma) = -\frac{1}{2}\sum_{i=1}^{N}\left(-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu\right) = \sum_{i=1}^{N}\Sigma^{-1}(x_i-\mu) = 0$$

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N}x_i$$