Basic Notions of Maximum Likelihood Estimation and Regression

Renato Lo Cigno

Simulation and Performance Evaluation 2018-19

The basic idea of MLE is simple

- Given I observed event B, what is the probability the event A occurred?
- Also: Given I have the sample $\{X_i\}$ what is the most likely population / process that generated it?
- MLE under certain hypotheses can be shown to be asymptotically optimum
- For small sample sets the estimation can be biased and give wrong results
- Unless there are some additional strong constraints MLE can be computationally very heavy
  - There are no "general" closed form solutions
  - If the state space of A is continuous, then we can in general only have an approximate solution

MLE is based on Bayes' Theorem

$$\mathbf{P}[B_j|A] = \frac{\mathbf{P}[A|B_j]\mathbf{P}[B_j]}{\mathbf{P}[A]} \quad \Leftrightarrow \quad \mathbf{P}[A] = \frac{\mathbf{P}[A|B_j]\mathbf{P}[B_j]}{\mathbf{P}[B_j|A]}$$

- MLE maximizes the a-posteriori probability of a conditional probability
- The maximization is done on some parameters of the conditioning events

Let $\{X_i;\ i = 1, 2, \ldots, n\}$ be a sample set and $\Theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$ be a set or vector of parameters to be estimated
Define a likelihood function $L(\Theta)$ as:

$$L(\Theta) = \mathbf{P}[X_1 = x1, X_2 = x_2, \ldots, X_n = x_n | \Theta]$$

if the population is described by a discrete PMF

or

$$L(\Theta) = f_X(x|\Theta)$$

if the population is described by a continuous pdf

Now the problem is trivial: find Θ such that $L(\Theta)$ is maximum

In math

$$\hat{\Theta} : \text{argmax}_{\Theta} L(\Theta)$$

- We need to know the joint probability of n random variables
- If they are not i.i.d. ... **game over!**
- If we know the sample set is i.i.d. then the likelihood functions reduce to

$$L(\Theta) = \prod_{i=1}^{n} \mathsf{P}[X_i = x_i | \Theta]$$

if the population is described by a discrete PMF, or

$$L(\Theta) = \prod_{i=1}^{n} f_{X_i}(x_i | \Theta)$$

if the population is described by a continuous pdf

In case of i.i.d. sets (& some other cases), as the likelihood function $L(\Theta)$ is described as a product it is custom to use logarithmic likelihood function $l(\Theta) = log[L(\Theta)]$ so that the maximization problem is described by a sum and not by a product

- $$l(\Theta) = \sum_{i=1}^{n} \mathbf{P}[X_i = x_i|\Theta]$$

  if the population is described by a discrete PMF, or

  $$l(\Theta) = \sum_{i=1}^{n} f_{X_i}(x_i|\Theta)$$

  if the population is described by a continuous pdf

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- Depending on Θ the problem can still be computationally very difficult (even in i.i.d. cases)
- Under some fairly general conditions of regularity of both the distributions and the Θ parameter set, then the optimization, in general an NP-complete problem, can be reduced to a set of k joint partial differential equations, where finding the zeros may be easy (?!?)

$$\frac{\delta L(\Theta)}{\delta \theta_i}; \quad i = 1, 2, \ldots, k$$

- Really the only case where MLE is simple and works without hassles is when θ_i are orthogonal and the partial differential equations either reduce to normal differential equations or we can in any case apply the gradient algorithm

- Really the only case where MLE is simple and works without hassles is when θ_i (the set of parameters) are orthogonal and the partial differential equations either reduce to normal differential equations

$$\frac{dL(\Theta)}{d\theta_i}; \quad i = 1, 2, \ldots, k$$

- or we can in any case apply the gradient algorithm (only one minimum exists)

- For instance if $\{X_i\}$ is drawn from a gamma distribution and θ_1 and θ_2 are the parameters λ and α of the distribution, then the set of 2 partial differential equations have no closed form solution and we have to resort to numerical methods (that's why you find the function in Matlab!!)

- For another totally "casual" example, if $\{X_i\}$ is drawn from a gamma distribution affected by random Gaussian noise samples Y_i distributed as $N(0, \sigma)$ and θ_1, θ_2 and θ_3 are the parameters λ, α, and σ of the two distributions, then we have to compute the distribution of

$$Z_i = X_i + Y_i$$

$$f_Z(z) = f_X(x) * f_Y(y)$$

where $*$ is the convolutional product so

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-z)^2}{2\sigma^2}}\, dx$$

and there is no solution to the MLE, unless we resort to (complex) numerical methods

MLE is instead simple when Θ is a partition of a probability space or a finite set of deterministic conditions. For example, it is the base for optimal detection in digital communications

- The key "problem" of digital transmission is finding the best strategy to decide what symbol $S_i(t)$ has been transmitted given we have received a symbol $R(t)$
- Find the maximum over j of

$$\mathbf{P}[S_j|R] = \frac{\mathbf{P}[R|S_j]\mathbf{P}[S_j]}{\mathbf{P}[R]}$$

- $R(t)$ can be modeled as $R(t) = S_i(t) + N(0, \sigma)$

$$\mathbf{P}[S_j|(S_j(t) + N(0,\sigma))] = \frac{\mathbf{P}[(S_j(t) + N(0,\sigma))|S_j]\mathbf{P}[S_j]}{\mathbf{P}[(S_j(t) + N(0,\sigma))]}$$

- Thus the MLE problem is reduced to a minimum distance problem

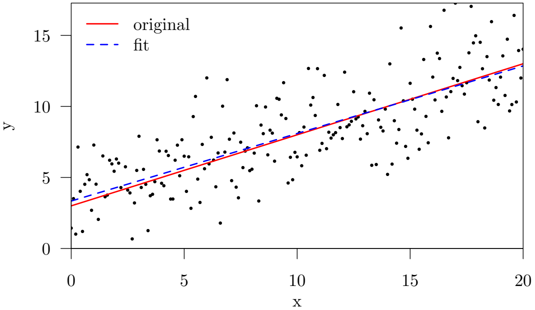
$$\min_j(||S_j - R||)$$

- More reasoning at the blackboard.

- Consider two joint RV X, Y and a dependence function $d(\cdot)$ such that $Y = d(X) + \epsilon$ where ϵ is a residual error
- Our problem is finding $d(\cdot)$ such that $d(X)$ is as close as possible to Y in some appropriate sense, e.g., minimizing a euclidean distance or a generic norm such as l_∞ or any proper measure
- Let $D = Y - d(X)$ be the random variable that measures the residual error done because we do not know $f_{X,Y}(x, y)$, and we approximate the dependence with the function $d(\cdot)$
- The most common measure of the difference is $E[D^2]$

- The function $d(x)$ that minimizes $E[D^2]$ is called the **Least-square regression curve**
- It is not difficult to show that this function is $d(x) = E[Y|x]$
- However the conditional distribution $f_{Y|x}(y|x)$ is normally very difficult to find
- It is common practice to limit the structure of $d(x)$ (e.g., to a polynomial function) to make the problem more tractable

A scatter diagram is nothing else than an (x, y) plot of the outcome of n random experiments on the pair X, Y



Scatter diagram with the linear regression of the points and the "true" linear relationship

- The simplest form of dependence is assuming that the function is linear: $d(x) = a + bx$
- Clearly this is a huge limitation to the dependence relationship, but in many cases it is useful and it can be treated easily
- In this case the problem of finding the optimal fitting curve reduces to minimize the following

$$G(a, b) = e[D^2] = E[(Y - d(X))^2] = E[(Y - a - bX)^2]$$

- Let $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$ be the mean and variance of X and Y respectively, and also $\rho = \dfrac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$
- Then expanding $G(a,b)$ yields

$$\begin{aligned} G(a,b) &= \sigma_y^2 + b^2\sigma_x^2 + (\mu_y - a)^2 + b^2\mu_x^2 - 2b\rho\sigma_x\sigma_y \\ &\quad -2b\mu_x(\mu_y - a) \\ &= \sigma_y^2 + b^2\sigma_x^2 + (\mu_y - a - b\mu_x)^2 - 2b\rho\sigma_x\sigma_y \end{aligned}$$

- To find the minimum of $G(a,b)$ we have to find the point where the partial derivatives with respect to a and b are zero

$$\frac{\delta G(a,b)}{\delta a} = -2(\mu_y - a - b\mu_x) = 0$$

$$\frac{\delta G(a,b)}{\delta b} = 2b\sigma_x^2 - 2\mu_x(\mu_y - a - b\mu_x) - 2\rho\sigma_x\sigma_y = 0$$

Solving the equations we find that the optimal values of a and b are

$$b = \rho\frac{\sigma_y}{\sigma_x}$$
$$a = \mu_y - b\mu_x$$

You normally find subroutines and function to perform a linear regression in any statistical tool

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- If the relationship is not linear, then finding the regression can be very difficult, even if the polynomial structure is given (it is not like the deterministic case of fitting)
- The exception is the exponential relation

$$Y = ae^{bX}$$

where we can simply take the logarithm and do a linear fitting of the logarithm