Maximum Likelihood and Least Squares

Maximum Likelihood. This is a method to select one of the possible models based on the experiments. The idea is very straightforward: you select the model for which the probability of the actually observed events is the largest possible.

Example of a practical situation. In many real-life situations, we know that the quantity y is determined by the quantities $\mathbf{x} = (x_1, \ldots, x_n)$ but we do not know the exact form of this dependence. In many cases, we know the typical shape of this dependence, but we do not know the values of the parameters of this dependence.

For example, we may know that the dependence is linear:

$$y = c_0 + c_1 \cdot x_1 + \ldots + c_n \cdot x_n,$$

but we do not know the values c_i. Another example: for radioactive decay, we may know that the amount y of remaining material depends on the time x_1 as

$$y = c_1 \cdot \exp(-c_2 \cdot x_1) + c_3 \cdot \exp(-c_4 \cdot x_1),$$

but we do not know the coefficients c_j.

In general, we know that the dependence has the form $y = f(\mathbf{x}, \mathbf{c})$ for some values $\mathbf{c} = (c_0, c_1, \ldots, c_m)$. To find these coefficients, we:

- measure both $\mathbf{x}$ and y in several situations $k = 1, \ldots, N$, and
- estimate the values c_i based on the results y_k and $\mathbf{x}_k = (x_{1,k}, \ldots, x_{n,k})$ of these measurements.

In many cases, the measurement errors of measuring y are much larger than the errors in measuring x_i. So, in the first approximation, we can ignore the measurement errors in measuring x_i's and assume that the measurement results $x_{i,k}$ are the actual values of the corresponding quantities. Under this assumption, we would expect the actual value of y in the k-th measurement to be equal to $f(\mathbf{x}_k, \mathbf{c})$ for the actual (to be determined) values of the coefficients c_j.

Of course, due to measurement errors in measuring y – which cannot be ignored – the measurement results y_k are, in general, different from the actual values $f(\mathbf{x}_k, \mathbf{c})$. By definition of the measurement errors Δy_k of each measurement is the difference between the measurement result y_k and the true value, i.e., $\Delta y_k = y_k - f(\mathbf{x}_k, \mathbf{c})$.

For normal distributions, Maximum Likelihood approach leads to Least Squares. In many practical situations, the measurement errors are characterized by a Gaussian (normal) distribution with 0 mean and known standard deviation σ. In this case, the probability of different values of the measurement error Δy_k are described by the formula

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\left(-\frac{(\Delta y_k)^2}{2\sigma^2}\right).$$

Measurement errors corresponding to different measurements are usually independent, so the probability p of having all observed measurement errors is equal to the product of all the above probabilities:

$$p=\prod_{k=1}^{N}\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\left(-\frac{(\Delta y_k)^2}{2\sigma^2}\right).$$

According the Maximum Likelihood idea, we need to select the coefficients $\mathbf{c}$ for which this probability is the largest.

The corresponding maximization problem can be simplified if we take into account that for every a, b, and c, we have $a^b\cdot a^c=a^{b+c}$. In particular, $e^b\cdot e^c=e^{b+c}$, i.e., $\exp(b)\cdot\exp(c)=\exp(b+c)$. Thus, the maximized probability can be rewritten as:

$$p=\text{const}\cdot\exp\left(-\sum_{k=1}^{N}\frac{(\Delta y_k)^2}{2\sigma^2}\right),$$

where we denoted

$$\text{const}=\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^N.$$

The relation between numbers – and thus, which alternative is the largest – does not change if we multiply all the values by the same constant. For example, the richest person in Mexico is the same whether we measure wealth in US dollars or in Mexican Pesos. Thus, to find the coefficients $\mathbf{c}$ for which the probability is the largest, it is sufficient to divide all the probability values by the constant and thus, select $\mathbf{c}$ for which the value

$$\exp\left(-\sum_{k=1}^{N}\frac{(\Delta y_k)^2}{2\sigma^2}\right)$$

is the largest.

The function $\exp(-z)$ is decreasing, so the largest value of this function corresponds to the smallest value of z. Thus, it is sufficient to find the values $\mathbf{c}$ for which the following sum is the smallest possible:

$$\sum_{k=1}^{N}\frac{(\Delta y_k)^2}{2\sigma^2}.$$

Similarly to what we did before, we can multiply all these values by the same constant $2\sigma^2$ and conclude that we need to find the values $\mathbf{c}$ for which the sum

$$\sum_{k=1}^{N}(\Delta y_k)^2$$

is the smallest possible. This is known as the *Least Squares* approach. Let us summaize it.

Least Squares approach. In the above problem, we need to find the coefficient $\mathbf{c}$ for which the following sum of the squares attains its least possible value:

$$\sum_{k=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{c}))^2.$$

An alternative explanation of the Least Squares approach. For the actual values $\mathbf{c}$, the only difference between the measurement results y_k nd the computed values $f(\mathbf{x}_k, \mathbf{c})$ is caused by the measurement errors. If we select a different set of coefficients $\mathbf{c}' \neq \mathbf{c}$, then the new difference $y_k - f(\mathbf{x}_k, \mathbf{c}')$ will, in general, increase – since this difference will also contain the inaccuracy of our new model $f(\mathbf{x}, \mathbf{c}'$. Thus, it makes sense to select the coefficients $\mathbf{c}$ for which, for every k, the difference Δy_k is the closest to 0, i.e., in other words, for which the multi-dimensional point $(\Delta y_1, \ldots, \Delta y_N)$ is the closest to the ideal point $(0, \ldots, 0)$.

According to Pythagoras Theorem, in general, the distance between the points $(a_1, \ldots, a_N)$ and $(b_1, \ldots, b_N)$ is equal to

$$\sqrt{(a_1 - b_1)^2 + \ldots + (a_N - b_N)^2}.$$

In particular, in our case, the distance is equal to

$$\sqrt{(\Delta y_1)^2 + \ldots + (\Delta y_N)^2}.$$

The square root is a strictly increasing function, so maximizing a square root of a number leads to the same choices as minimizing the number itself. Thus, to find the coefficients $\mathbf{c}$, we need to minimize the expression:

$$(\Delta y_1)^2 + \ldots + (\Delta y_N)^2,$$

i.e.,

$$(y_1 - f(\mathbf{x}_1, \mathbf{c}))^2 + \ldots + (y_N - f(\mathbf{x}_N, \mathbf{c}))^2.$$

This is exactly the Least Squares method.

The Least Squares approach: first example. Let us first consider a simple example in which we do not have any inputs, we just have several measurement result $y_1, \ldots, y_N$ that measure the same unknown value. This corresponds to

the model $f(\mathbf{x}, \mathbf{c}) = c_1$, where c_1 is the to-be-detemined actual value of the measured quantity.

In this case, the Least Squares approach means that we minimize the sum

$$(y_1 - c_1)^2 + \ldots + (y_N - c_1)^2.$$

To minimize this expression, let us differentiate it with respect to the unknown c_1 and equate the resulting derivative to 0. As a result, we get the following equation:

$$2(y_1 - c_1) \cdot (-1) + \ldots + 2(y_N - c_1) \cdot (-1) = 0.$$

If we divide both sides by -2 and open parentheses, we get the following equation:

$$y_1 - c_1 + \ldots + y_N - c_1 = 0.$$

As usual, we can move all the terms containing the unknown c_1 to the right-hand side and get

$$y_1 + \ldots + y_N = N \cdot c_1,$$

hence

$$c_1 = \overline{y},$$

where we denoted, by $\overline{y}$, the arithmetic average of all the measurement results:

$$\overline{y} = \frac{y_1 + \ldots + y_N}{N}.$$

The Least Squares approach: second example. Let us now consider a somewhat more complex case, when $n = 1$ and we are looking for th coefficients of a linear dependence $y = c_0 + c_1 \cdot x_1$. In this case, the Least Squares method means that we minimise the following sum:

$$(y_1 - (c_0 + c_1 \cdot x_{1,1}))^2 \ldots + (y_N - (c_0 + c_1 \cdot x_{1,N}))^2.$$

According to calculus, both the derivative with respect to c_0 and the derivative with respect to c_1 should be equal to 0.

Differentiating with respect to c_0 and equating the derivative to 0, we get:

$$2(y_1 - (c_0 + c_1 \cdot x_{1,1})) \cdot (-1) + \ldots + 2(y_N - (c_0 + c_1 \cdot x_{1,N})) \cdot (-1) = 0.$$

Dividing both sides by -2 and opening the parentheses, we get:

$$y_1 - c_0 - c_1 \cdot x_{1,1} + \ldots + y_N - c_0 - c_1 \cdot x_{1,N} = 0.$$

If we move all the terms containing the unknowns to the right-hand side, we get:

$$y_1 + \ldots + y_N = N \cdot c_0 + c_1 \cdot (x_{1,1} + \ldots + x_{1,N}).$$

If we divide both sided by N, we get the following equation:

$$\overline{y} = c_0 + c_1 \cdot \overline{x},$$

where we denoted

$$\overline{x} = \frac{x_{1,1} + \ldots + x_{1,N}}{N} \text{ and } \overline{y} = \frac{y_1 + \ldots + y_N}{N}.$$

Differentiating with respect to c_1 and equating the derivative to 0, we get:

$$2(y_1 - (c_0 + c_1 \cdot x_{1,1})) \cdot (-x_{1,1}) + \ldots + 2(y_N - (c_0 + c_1 \cdot x_{1,N})) \cdot (-x_{1,N}) = 0.$$

Dividing both sides by -2 and opening the parentheses, we get:

$$x_{1,1} \cdot y_1 - c_0 \cdot x_{1,1} - c_1 \cdot x_{1,1}^2 + \ldots + x_{1,N} \cdot y_N - c_0 \cdot x_{1,N} - c_1 \cdot x_{1,N}^2 = 0.$$

If we move all the terms containing the unknown to the right-hand side, we get:

$$x_{1,1} \cdot y_1 + \ldots + x_{1,N} \cdot y_N = N \cdot c_0 \cdot (x_{1,1} + \ldots + x_{1,N}) + c_1 \cdot \left(x_{1,1}^2 + \ldots + x_{1,N}^2\right).$$

If we divide both sided by N, we get the following equation:

$$\overline{x \cdot y} = c_0 \cdot \overline{x} + c_1 \cdot \overline{x^2},$$

where we denoted

$$\overline{x \cdot y} = \frac{x_{1,1} \cdot y_1 + \ldots + x_{1,N} \cdot y_N}{N} \text{ and } \overline{x^2} = \frac{x_{1,1}^2 + \ldots + x_{1,N}^2}{N}.$$

So, we have two equations for two unknowns c_0 and c_1:

$$\overline{y} = c_0 + c_1 \cdot \overline{x},$$
$$\overline{x \cdot y} = c_0 \cdot \overline{x} + c_1 \cdot \overline{x^2}.$$

To solve this systems of equations, let us multiply the first equation by $\overline{x}$ and subtract from the first equation. This was, the term proportional to c_0 disappear, and we have an equation with only one unknown c_1:

$$c_1 \cdot \left(\overline{x^2} - (\overline{x})^2\right) = \overline{x \cdot y} - \overline{x} \cdot \overline{y},$$

so

$$c_1 = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}.$$

Once we know c_1, we can find c_0 from the first of the two above equations as:

$$c_0 = \overline{y} - c_1 \cdot \overline{x}.$$

So, we arrive at the following algorithm for computing the coefficients c_0 and c_1:

- First, we compute the following averages:

$$\overline{x} = \frac{x_{1,1} + \ldots + x_{1,N}}{N}, \quad \overline{y} = \frac{y_1 + \ldots + y_N}{N},$$
$$\overline{x \cdot y} = \frac{x_{1,1} \cdot y_1 + \ldots + x_{1,N} \cdot y_N}{N}, \quad \overline{x^2} = \frac{x_{1,1}^2 + \ldots + x_{1,N}^2}{N}.$$

- Then, we compute

$$c_1 = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}, \text{ and}$$
$$c_0 = \overline{y} - c_1 \cdot \overline{x}.$$