Curve Fitting

Goal: Given a set of data (x_i, y_i) with $1 \leq i \leq n$, find a mathematical model, or function of $y(x)$, that would match as closely as possible, the given data.

Step 1: assume a model, or the format of the function $y(x)$, e.g., linear, exponential, log, …

Step 2: find the parameters of the function that minimize the squares of the errors/differences between the model prediction and the data.

1. Linear fitting

Step 1: the function is given by:

$$y(x) = ax + b$$

where a, b are constants to be determined.

Step 2: Find values of a, b

Approach: the sum of the squares of the differences between model and data is:

$$S = [y(x_1) - y_1]^2 + [y(x_2) - y_2]^2 + \cdots + [y(x_n) - y_n]^2 = \sum_{i=1}^{n} [y(x_i) - y_i]^2 = \sum_{i=1}^{n} [ax_i + b - y_i]^2$$

S is a function of a, b.

To minimize S,

$$\frac{\partial S}{\partial a} = 0$$

$$\frac{\partial S}{\partial a} = 2\sum_{i=1}^{n} [ax_i + b - y_i]\, x_i = 0$$

$$\frac{\partial S}{\partial b} = 2\sum_{i=1}^{n} [ax_i + b - y_i] \; = 0$$

After some calculation, you will get:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{x=i}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

The coefficient of determination, a measure of how well model fits the data, is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} [y_i - y(x_i)]^2}{\sum_{i=1}^{n} \left[y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right]^2}$$

2. General theory of curve fitting

Find a function of the format:

$$y(x) = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_m f_m(x) = \sum_{j=1}^{m} a_j f_j(x)$$

To fit the data of (x_i, y_i) with $1 \le i \le n$.

where $a_1, a_2, \ldots, a_m$ are parameters to be determined, $f_1(x), f_2(x), \ldots, f_m(x)$ are functions of x, e.g., 1, x, x^2, exp(x), sin(x), ….

$$S = [y(x_1) - y_1]^2 + [y(x_2) - y_2]^2 + \cdots + [y(x_n) - y_n]^2 = \sum_{i=1}^{n} [y(x_i) - y_i]^2 = \sum_{i=1}^{n} \left[\sum_{j=1}^{m} a_j f_j(x_i) - y_i \right]^2$$

Let $\frac{\partial S}{\partial a_1} = 0, \frac{\partial S}{\partial a_2} = 0, \ldots, \frac{\partial S}{\partial a_m} = 0$

We will get m equations, and we have m unknown of $(a_1, a_2, \ldots, a_m)$ -> determine a_j

$$\begin{aligned} S = [y(x_1) - y_1]^2 + [y(x_2) - y_2]^2 + \cdots + [y(x_n) - y_n]^2 \\ = [a_1 f_1(x_1) + a_2 f_2(x_1) + \cdots + a_m f_m(x_1) - y_1]^2 \\ + [a_1 f_1(x_2) + a_2 f_2(x_2) + \cdots + a_m f_m(x_2) - y_2]^2 \\ + \cdots \\ + [a_1 f_1(x_n) + a_2 f_2(x_n) + \cdots + a_m f_m(x_n) - y_n]^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial S}{\partial a_1} = \quad & 2[a_1 f_1(x_1) + a_2 f_2(x_1) + \cdots + a_m f_m(x_1) - y_1] f_1(x_1) \\ & + 2[a_1 f_1(x_2) + a_2 f_2(x_2) + \cdots + a_m f_m(x_2) - y_2] f_1(x_2) \\ & + \cdots \\ & + 2[a_1 f_1(x_n) + a_2 f_2(x_n) + \cdots + a_m f_m(x_n) - y_n] f_1(x_n) = 0 \end{aligned}$$

So,

$$\begin{aligned} & [a_1 f_1(x_1) + a_2 f_2(x_1) + \cdots + a_m f_m(x_1) - y_1] f_1(x_1) \\ & + [a_1 f_1(x_2) + a_2 f_2(x_2) + \cdots + a_m f_m(x_2) - y_2] f_1(x_2) \\ & + \cdots \\ & + [a_1 f_1(x_n) + a_2 f_2(x_n) + \cdots + a_m f_m(x_n) - y_n] f_1(x_n) = 0 \end{aligned}$$

So,

$$\begin{aligned} & f_1(x_1) f_1(x_1) a_1 + f_1(x_2) f_1(x_2) a_1 + \cdots + f_1(x_n) f_1(x_n) a_1 \\ & + f_1(x_1) f_2(x_1) a_2 + f_1(x_2) f_2(x_2) a_2 + \cdots + f_1(x_n) f_2(x_n) a_2 \\ & + \cdots \\ & + f_1(x_1) f_m(x_1) a_m + f_1(x_2) f_m(x_2) a_m + \cdots + f_1(x_n) f_m(x_n) a_m \\ & = y_1 f_1(x_1) + y_2 f_1(x_2) + \cdots + y_n f_1(x_n) \end{aligned}$$

So, if we let $\frac{\partial S}{\partial a_1} = 0$, we get an equation of:

$$\sum_{i=1}^{n} f_1(x_i) f_1(x_i)\, a_1 + \sum_{i=1}^{n} f_1(x_i) f_2(x_i) a_2 + \cdots + \sum_{i=1}^{n} f_1(x_i) f_m(x_i)\, a_m = \sum_{i=1}^{n} y_i f_1(x_i)$$