Lecture 25: Method of Least Squares for Curve Fitting

(28-Spt-2012)

LECTURE 25
28-SPT-2012

APPROXIMATE POLYNOMIAL FITS

In the last class, we suggested that if you have data points (x_i, f_i) ; $i = 0, 1, 2, 3, \ldots, n$,

→ Then you can fit exact polynomial or approximate polynomial to the above data set.

→ Approximate polynomials need not pass through all the data points.

→ We can form approximate polynomials fits by obtaining the best combination (or coefficients) of the polynomial.

→ For the given data set one can approximately fit, a best

→ straight line

→ quadratic polynomial

– cubic polynomial, etc.

→ The method used to find the best values of coefficients can be many.

→ A particular method is the Least Squares Method

→ For a given value of x, say x_i, the observed value of function is f_i. However if we are approximating $f_i \approx y_i$ (a polynomial). Then the error or difference is $e_i = f_i - y_i$

→ Least squares method suggest that the sum of squares of the differences (or errors) should be minimum to ~~obtain~~ consider y as the best fit for the function f.

i	x_i	f_i	y
0	x_0	f_0	y_0
1	x_1	f_1	y_1
2	x_2	f_2	y_2
⋮	⋮		⋮
n	x_n	f_n	y_n

Sum of square of errors
$$S = (f_0 - y_0)^2 + (f_1 - y_1)^2 + \cdots$$
$$\cdots + (f_n - y_n)^2$$
$$= \sum_{i=0}^{n} (f_i - y_i)^2$$

Least Square Method for Straight Line Approximation

If we are approximating a straight line for finding relation between x and f, the best straight line is given as

$$y = A + Bx$$

such that sum of square of errors

$$S = \sum_{i=0}^{n} \left(f_i - A - B x_i\right)^2$$

Now S is $S(A, B)$

For S to be minimum, we need to have

$$\frac{\partial S}{\partial A} = 0 \quad \text{and} \quad \frac{\partial S}{\partial B} = 0$$

i.e.

$$\frac{\partial S}{\partial A} = \sum_{i=0}^{n} y_i - A\sum_{i=0}^{n} 1 - B\sum_{i=0}^{n} x_i$$

$$\frac{\partial S}{\partial A} = \sum_{i=0}^{n}\left[2\left(y_i - A - Bx_i\right)(-1)\right] = 0$$

i.e. $$\sum_{i=0}^{n} y_i - A\sum_{i=0}^{n}(1) - B\sum_{i=0}^{n} x_i = 0 \quad \to (1)$$

$$\frac{\partial S}{\partial B} = \sum_{i=0}^{n}\left[2\left(y_i - A - Bx_i\right)(-x_i)\right] = 0$$

i.e. $$\sum_{i=0}^{n}(-y_i x_i) + A\sum_{i=0}^{n} x_i + B\sum_{i=0}^{n} x_i^2 = 0 \quad \to (2)$$

Equation (1) and (2) are normal equations for straight line approximations. On solving (1) and (2) you get the coefficients A and B for the best straight line relating x and y.

→ If the number of data point $0, 1, 2, 3, \ldots, n = N$

Then normal equations

$$AN + B\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} y_i$$

$$A\sum_{i=0}^{n} x_i + B\sum_{i=0}^{n} x_i^2 = \sum_{i=0}^{n} y_i x_i$$

Example

If you have the data set. Fit the best straight line.

x	0	5	10	15
y	10	17	25	31

Let $y \approx y$ where $y = A + Bx$

Using normal equations for straight line, fit :

$N = 4$
$n = 3$

$\therefore \quad 4A + B \sum_{i=0}^{3} x_i = \sum_{i=0}^{3} y_i$

i.e. $\quad 4A + 30B = 83$

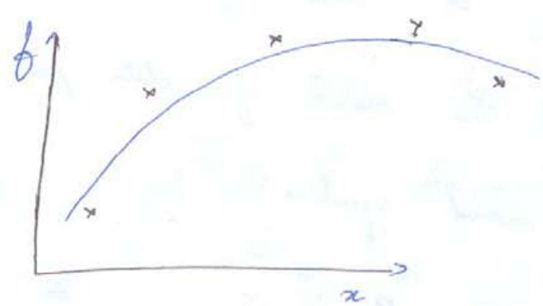
$\text{Illy} \quad 30A + 350B = 600$

$$\left[\begin{array}{cc|c} 4 & 30 & 83 \\ 3 & 35 & 60 \end{array}\right]$$

Solving: you get A and B

⇒ Similar approach by forming normal equations to find the coefficient of higher order polynomials can also be used for quadratic, cubic, and other higher-degree polynomials.

→ You need that many number of normal equations as are the coefficients.



The Multi-variate polynomial

The least squares method can be used to provide approximate fits for multi-variate problems.
As seen earlier if there are some functions

$z = f(x, y)$

Then the best straight line fit for the observations of z versus x and y ~~are~~ can be given as:

$$z = A + Bx + Cy$$

Again, we want the sum of squares of differences between z and ~~the~~ $\mathfrak{z}$ to be minimum

i.e. $$e_i = \mathfrak{z}_i - z_i$$

$$\text{Minimise } \sum e_i^2 = \sum_{i=0}^{n} (\mathfrak{z}_i - z_i)^2 = S$$

We need to have

$$\frac{\partial S}{\partial A} = 0, \quad \frac{\partial S}{\partial B} = 0, \quad \text{and} \quad \frac{\partial S}{\partial C} = 0$$

$$\frac{\partial S}{\partial A} = \sum_{i=0}^{n} (\mathfrak{z}_i - A - Bx_i - Cy_i)(-1) = 0$$

$$\frac{\partial S}{\partial B} = \sum_{i=0}^{n} (\mathfrak{z}_i - A - Bx_i - Cy_i)(-x_i) = 0$$

$$\frac{\partial S}{\partial C} = \sum_{i=0}^{n} (\mathfrak{z}_i - A - Bx_i - Cy_i)(-y_i) = 0$$

i.e.
$$\begin{cases} AN + B\sum_{i=0}^{n} x_i + C\sum_{i=0}^{n} y_i = \sum_{i=0}^{n} \mathfrak{z}_i \\ A\sum_{i=0}^{n} x_i + B\sum_{i=0}^{n} x_i^2 + C\sum_{i=0}^{n} x_i y_i = \sum_{i=0}^{n} x_i \mathfrak{z}_i \\ A\sum_{i=0}^{n} y_i + B\sum_{i=0}^{n} x_i y_i + C\sum_{i=0}^{n} y_i^2 = \sum_{i=0}^{n} y_i \mathfrak{z}_i \end{cases}$$

You need to solve these normal equations to get A, B & C.

$\Rightarrow$ The same procedure may be adopted for higher order polynomials as well.

Extension of Least Squares Method for fitting curves that may not be polynomials

As seen till now, least squares method is a curve fitting technique. We demonstrated the method for fitting straight lines (i.e. a first order polynomial) to relate given set of data points. You can also use higher order polynomials like - quadratic, cubic, etc. to fit the curve.

→ However, not all data sets can be related using polynomials. Other non-linear equations may give better approximations

e.g. like Power equation $y = Ax^B$

" like Exponential equation $y = Ae^{Bx}$

Other forms of non-linear equations that can relate the functional values with the independent variables may also be possible.

→ To fit such non-linear equations, you may again use Least Squares Method

Case - Power equation fit

In the given data set (x_i, f_i) if we want to approximate $f \approx y$, where y is a power equation

$$y = Ax^B$$

Using the properties of logarithms to find unknowns A and B

$$\ln y = \ln A + B \ln x$$

Now assign $\ln y = y'$ (Not first derivative)

IIIly $\ln A = A'$

$\ln x = x'$

$$\therefore \quad y' = A' + B x' \quad \rightarrow (2)$$

Now this equation is a straight line which can be evaluated by previous methods.

Case - Exponential Fit

If you are approximating $f \approx y$

where $y = A e^{Bx}$ (Exponential function).

then to obtain the best values of A and B

$$\ln y = \ln A + Bx$$

Again here, you can assign $\ln y = y'$

$\ln A = A'$

$$\therefore \quad y' = A' + Bx \quad \rightarrow (3)$$

So you can use linear fit methods on (3) to obtain the best ~~fit~~ exponential fit.

Case - A Non-linear fit of form $y = \frac{A}{1+Bx}$

In this case also we need to find A and B that are best values to relate y with f.

Using least squares method

$$e_i^2 = (\hat{y}_i - y_i)^2$$

ie $$S = \sum_{i=0}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=0}^{n} \left(\hat{y}_i - \frac{A}{1+Bx_i}\right)^2$$

If we assign $\frac{\partial S}{\partial A} = 0$ and $\frac{\partial S}{\partial B} = 0$

→ The expressions will be non-linear.

i.e. $$\frac{\partial S}{\partial A} = \sum_{i=0}^{n} 2\left(\hat{y}_i - \frac{A}{1+Bx_i}\right)\left(-\frac{1}{1+Bx_i}\right) = 0$$

Also $$\frac{\partial S}{\partial B} = \sum_{i=0}^{n} 2\left(\hat{y}_i - \frac{A}{1+Bx_i}\right)\frac{x_i}{(1+Bx_i)^2} = 0$$

This is a system of non-linear equation. You solve them to find A and B (recall Newton's method to solve system of non-linear equations).

QUIZ (26-SEPT-2012)

1) In the given data set

i	x	$\hat{y}$
0	0.00	12.0
1	5.00	15.0
2	10.00	17.0

Fit linear splines for the given data. How many linear spline equations are formed? Develop these linear spline equations

(20 MARKS)

2) If there are N data points from $i = 0, 1, 2, \ldots, n-1, n$. You may give an approximate fit using least squares method. If the approximate fit is $y_i = A + Bx_i + Cx_i^2$, how many normal equations are required for such curve fitting.