IIIT Vadodara
WINTER 2020-21
MA202 Numerical Techniques
LAB#9 Regression[1]

Regression establishes a function f that describes the trend (relationship) among a set of input x and output y data points, i.e., $y = f(x)$. It may further leads to system identification and more. Regression is a well-known approach for function approximation especially when the data points are perturbed or noisy, unlike polynomial interpolation. To start with, a simple approach is linear least-squares (LS) that minimizes sum-of-squared vertical distances, i.e., it assumes that the output data points y are corrupted.

As a reasonable means, we consider the least-squares (LS) approach to minimizing the sum of squared errors, where the error is described by the vertical distance to the curve from the data points. We will look over various types of fitting functions in this section.

Linear least-squares regression using straight line fit (polynomial approximation of first degree)

Given a set of M input/output data points' pairs $(x_1, y_1), (x_2, y_2), ..., (x_M, y_M)$, we would like to establish following linear relationship using linear least-squares regression,

$$\theta_1 x + \theta_0 = y \tag{1}$$

Where θ_1 is the slope and θ_0 is the intercept of the resultant straight line equation. Using the available data points pairs, we can write a set of linear simultaneous equations as follows:
$\theta_1 x_1 + \theta_0 = y_1$
$\theta_1 x_2 + \theta_0 = y_2$
$\theta_1 x_3 + \theta_0 = y_3$
.....
$\theta_1 x_M + \theta_0 = y_M$

The corresponding matrix form can be written as,

$$A\theta = y \; with \; A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ . & . \\ x_M & 1 \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_M \end{bmatrix} \tag{2}$$

Since it yields an overdetermined system of expression, i.e., more number of equations than unknowns (θ_0 and θ_1), we can solve the same in norm-2 sense, i.e., Euclidean norm, by writing the normal equations,

$$A\theta = y, A^T A\theta = A^T y = \theta^0 = \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \end{bmatrix} \quad = [A^T A]^{-1} A^T y \tag{3}$$

This is the normal equation which minimizes the objective function $J = \|e\|^2 = J = \|A\theta - y\|^2 = [A\theta - y]^T [A\theta - y]$

Polynomial Curve Fit: A Polynomial Function of Higher Degree

If there is no reason to limit the degree of fitting polynomial to one, then we may increase the degree of fitting polynomial to, say, N in expectation of decreasing the error. Still, we can

[1]submission deadline : 11^{th} April 11 PM

use minimizing the objective function as discussed earlier, but with different definitions of A and θ as

$$A = \begin{bmatrix} x_1^N & \cdot & x_1 & 1 \\ x_2^N & \cdot & x_2 & 1 \\ \cdot & \cdot & \cdot & \cdot \\ x_M^N & \cdot & x_M & 1 \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_N \\ \cdot \\ \theta_1 \\ \theta_0 \end{bmatrix} \tag{4}$$

Procedure for Linear Regression (Fitting $y = \theta_0 + \theta_1 x$) using Least Square Method

1. Form normal equations:

$\sum y = n\theta_0 + \theta_1 \sum x$
$\sum xy = \theta_0 \sum x + \theta_1 \sum x^2$

2. Solve normal equations as simulataneous equations for θ_0 and θ_1
3. Substitute the value of θ_0 and θ_1 in $y = \theta_0 + \theta_1 x$ which is required line of best fit.

Linear Regression Algorithm (Fitting $y = \theta_0 + \theta_1 x$)

1. Start
2. Read Number of Data (n)
3. For $i = 1$ to n:

Read X_i and Y_i
Next i

4. Initialize:

$sumX = 0$
$sumX2 = 0$
$sumY = 0$
$sumXY = 0$

5. Calculate Required Sum

For $i = 1$ to n:
$sumX = sumX + X_i$
$sumX2 = sumX2 + X_i * X_i$
$sumY = sumY + Y_i$
$sumXY = sumXY + X_i * Yi$
Next i

6. Calculate Required Constant θ_0 and b of $y = \theta_0 + \theta_1 x$:

$b = (n * sumXY - sumX * sumY)/(n * sumX2 - sumX * sumX)$
$\theta_0 = (sumY - \theta_1 * sumX)/n$

7. Display value of θ_0 and θ_1
8. Stop

Polynomial Curve Fit by LS (Least Squares).

Q. 1: Given these data points:
x: [-3 -2 -1 0 1 2 3]
y: [0.2774 0.8958 1.5651 3.4565 3.0601 4.8568 3.8982]
Fit these data into polynomials of degree 1, 3, 5, and 7. Plot the polynomials of different degree in independent subgraphs using MATLAB command subplot. Observe oscillation of the fitting curve between the data points with higher degree.

Q. 2: Given the following data:
x: [0.8 1.4 2.7 3.8 4.8 4.9]
y: [0.69 1.00 2.00 2.39 2.34 2.83]
Fit these data into polynomials of degree 1, 3, 5, and 7. Observe oscillation of the fitting curve between the data points with higher degree.