Fitting Data Using Linear and Nonlinear Models

Boyu Yue

ybyleo@126.com

3/20/25

Introduction

In this assignment, I will use three linear fitting methods to fit a given set of two-dimensional data, including **least squares method (LSM)**, **gradient descent method (GD)** and **Newton's method**, and compare their fitting results. In addition, I will try to use nonlinear models to find better fitting results.

Linear Models

1. Least Square Method (LSM)

LSM is a widely used mathematical technique for fitting a model to a set of observed data points by minimizing the sum of the squared differences between the observed values and the values predicted by the model. The core idea is to find the model parameters that result in the smallest possible error, where the error is defined as the difference between the actual data points and the corresponding points on the fitted model.

Methodology

Suppose we have n data points (x_1, y_1), (x_2, y_2), ... , (x_n, y_n), where x_i represents the independent variables and y_i is the observed dependent variable. For a linear model, we aim to fit:

$$y_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \epsilon_i$$

Where $\theta = [\theta_0, \theta_1, \cdots, \theta_p]$is the parameter vector. In matrix notation, the model becomes:

$$\mathbf{y} = \mathbf{X}\theta + \epsilon$$

The risidual vector is:

$$\epsilon = \mathbf{y} - \mathbf{X}\theta$$

The least squares method minimizes the sum of squared residuals, expressed as the squared Euclidean norm of the residual vector:

$$S(\theta) = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta)$$

Expanding:

$$S(\theta) = \mathbf{y}^T\mathbf{y} - 2\theta^T\mathbf{X}^T\mathbf{y} + \theta^T\mathbf{X}^T\mathbf{X}\theta$$

To find the optimal θ, we take the derivative of $S(\theta)$ with respect to θ and set it to zero.

$$\frac{\partial S}{\partial \theta} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\theta = 0$$

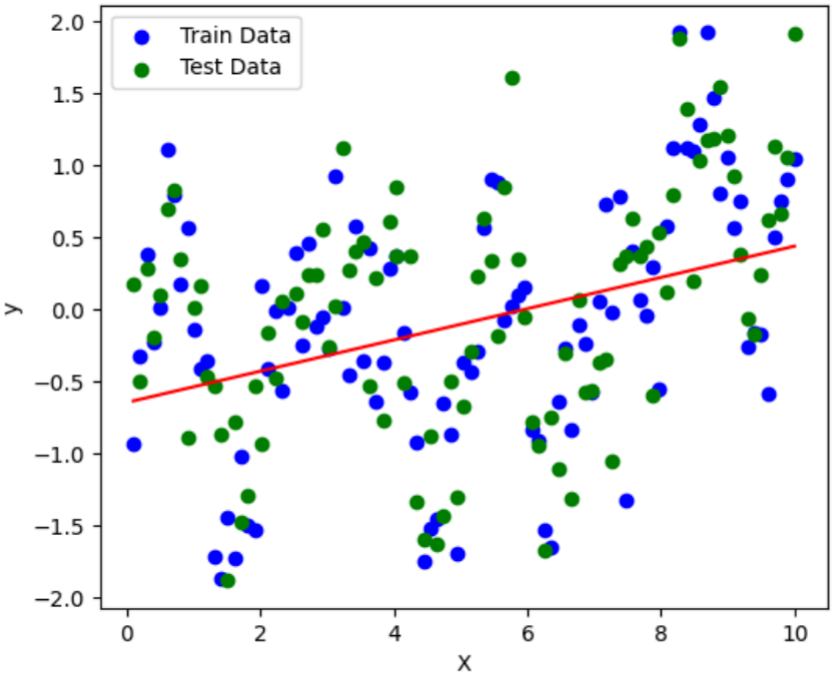
$$\mathbf{X}^T\mathbf{X}\theta = \mathbf{X}^T\mathbf{y}$$

This is the matrix form of the normal equations. Assuming $\mathbf{X}^T\mathbf{X}$ is invertible, we solve for $\hat{\theta}$:

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Experimental Studies

Here are the results of fitting the given data using LSM:



The analytical expression of the fitted line is:

$$y = 0.1083x - 0.6445$$

The mean squared error (MSE) of the traning data and the test data are:

$$MSE_{train} = 0.6195$$

$$MSE_{test} = 0.5990$$

2. Gradient Descent Method (GD)

GD is an iterative optimization technique used to fit a model to data by minimizing a loss function. GD approximates the optimal model parameters by iteratively adjusting them in the direction that reduces the error most rapidly, guided by the gradient of the loss function.

Methodology

The loss function commonly used in regression is the mean squared error (MSE), proportional to the sum of squared residuals. Consider the same set of data mentioned above, we have:

$$L(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i, \theta))^2 = \frac{1}{2n}(\mathbf{y} - \mathbf{X}\theta)^T(\mathbf{y} - \mathbf{X}\theta)$$

Gradient descent minimizes $L(\theta)$ by iteratively updating θ in the direction opposite to the gradient $\nabla L(\theta)$, which indicates the steepest increase in the loss. The gradient is the vector of partial derivatives:

$$\nabla L(\theta) = \frac{\partial L}{\partial \theta} = -\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\theta)$$

For a linear model:

$$L(\theta) = \frac{1}{2n}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\theta + \theta^T\mathbf{X}^T\mathbf{X}\theta)$$

$$\frac{\partial L}{\partial \theta} = \frac{1}{2n}(-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\theta) = \frac{1}{n}\mathbf{X}^T(\mathbf{X}\theta - \mathbf{y})$$

The parameters are updated iteratively using a learning rate η, which controls the step size:

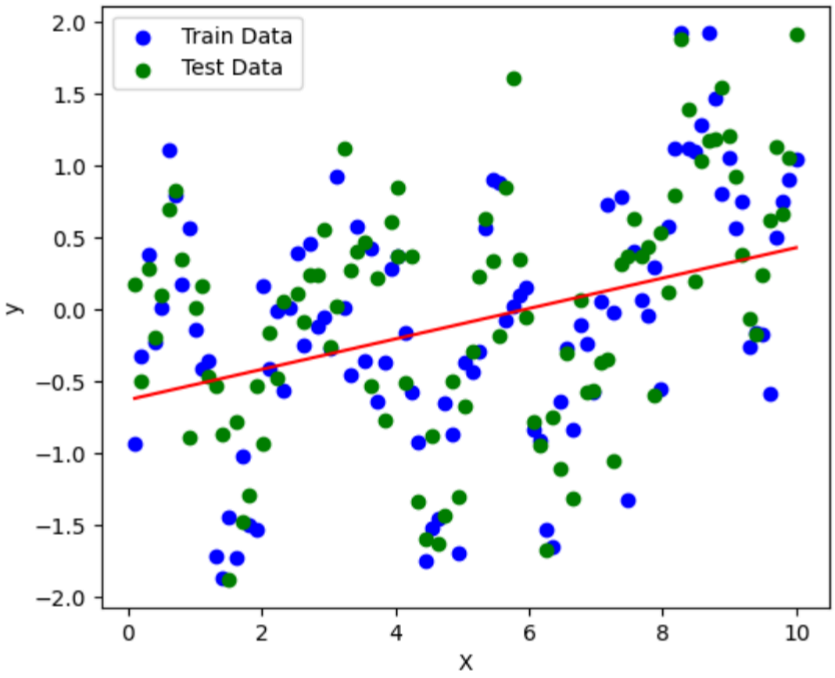
$$\theta^{(t+1)} = \theta^{(t)} - \eta\nabla L(\theta^{(t)})$$

Where $\theta^{(t)}$ is the parameter vector at iteration t. Substituting the gradient:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\theta^{(t)})$$

Experimental Studies

Here are the results of fitting the given data using GD:



The analytical expression of the fitted line is:

$$y = 0.1057x - 0.6271$$

The mean squared error (MSE) of the traning data and the test data are:

$$MSE_{train} = 0.6197$$

$$MSE_{test} = 0.5982$$

3. Newton's Method

Newton’s method is an iterative optimization technique used to fit a model to data by minimizing a loss function. Newton’s method incorporates both the gradient (first derivative) and the Hessian (second derivative) of the loss function, enabling faster convergence near the optimum by approximating the function locally as a quadratic.

Methodology

Newton's method requires the gradient (first derivative) and Hessian (second derivative) of the loss. Consider the same set of data and the loss function mentioned above, we have the gradient and the Hessian:

$$\nabla L(\theta) = \frac{\partial L}{\partial \theta} = -\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\theta)$$

$$\mathbf{H} = \frac{\partial^2 L}{\partial \theta \partial \theta^T} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$$

Newton's method approximates $L(\theta)$ around the current estimate $\theta^{(t)}$ with a second-order Taylor expansion and finds the minimum of this quadratic approximation.

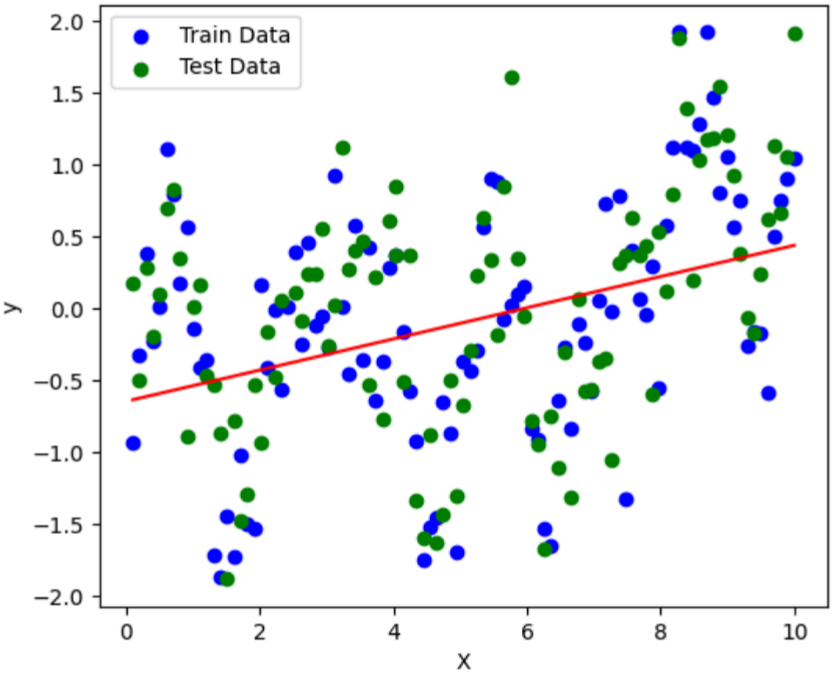
$$\theta^{(t+1)} = \theta^{(t)} - \mathbf{H}^{-1}\nabla L(\theta^{(t)})$$

Substituting the gradient and Hessian:

$$\theta^{(t+1)} = \theta^{(t)} - \left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(-\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\theta^{(t)})\right)$$

Experimental Studies

Here are the results of fitting the given data using Newton's method:



The analytical expression of the fitted line is:

$$y = 0.1083x - 0.6445$$

The mean squared error (MSE) of the traning data and the test data are:

$$MSE_{train} = 0.6195$$

$$MSE_{test} = 0.5990$$

Nonlinear Models

Based on the observation of the given data set, I found that the data has a certain periodicity, so I chose to use Fourier series for fitting.

Fourier Series

The Fourier series is a mathematical technique used to fit periodic data by representing it as a sum of sinusoidal functions (sines and cosines) with varying frequencies. Rooted in harmonic analysis, this method decomposes a periodic function into an infinite or finite series of basis functions.

Methodology

The goal is to approximate y using a Fourier series:

$$y = a_0 + \sum_{k=1}^{N} [a_k \cos(kx) + b_k \sin(kx)]$$

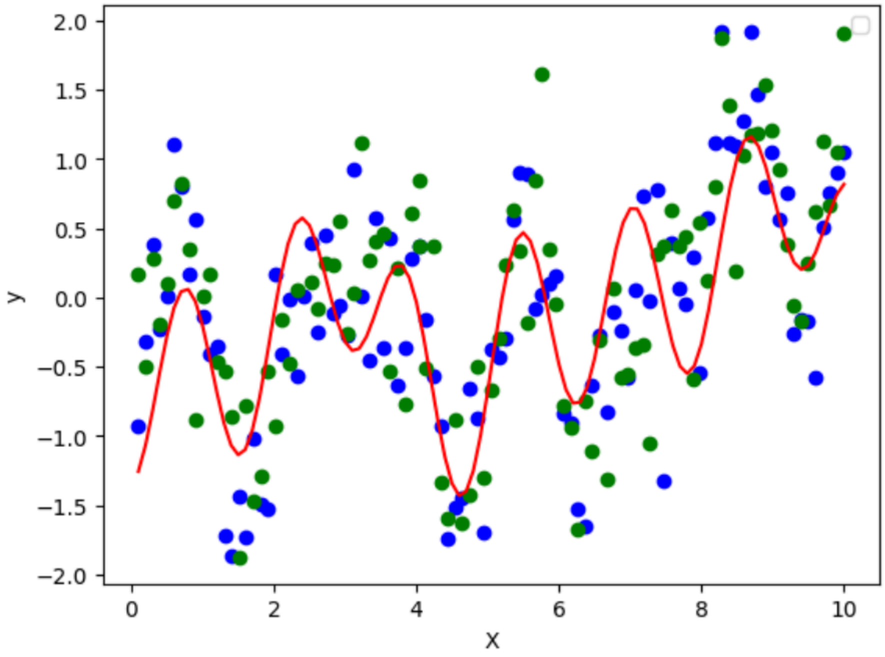
Where a_0 is the constant term, a_k and b_k are the coefficients for the cosine and sine terms at frequency k, k is a positive integer representing the harmonic number.

For the given dataset, I observed an overall upward trend in the data. As a result, I added a linear term to match this feature. The series becomes:

$$y = a_0 + \sum_{k=1}^{N} [a_k \cos(kx) + b_k \sin(kx)] + cx + d$$

Experimental Studies

I use the fourth-order Fourier series ($N = 4$) with a linear term for the fit. Here are the results of fitting the given data:



The parameters are as follows:

$$a_0 = 424.4184$$
$$a_1 = -0.1931$$
$$b_1 = 0.2539$$
$$a_2 = 0.2662$$
$$b_2 = -0.1432$$
$$a_3 = -0.1430$$
$$b_3 = -0.0264$$
$$a_4 = -0.6452$$
$$b_4 = 0.1121$$
$$c = 0.0936$$
$$d = -425.0566$$

The mean squared error (MSE) of the traning data and the test data are:

$$MSE_{train} = 0.3302$$

$$MSE_{test} = 0.4232$$

Conclusions

In summary, these three linear fitting methods can fit to similar results. By comparing their errors, we can find that their errors are similar. However, GD requires setting and adjusting parameters to converge to the optimal solution, while the training of the other two methods is more convenient.

	a	**b**	MSE_{train}	MSE_{test}
LSM	0.1083	-0.6445	0.6195	0.5990
GD	0.1057	-0.6271	0.6197	0.5982
Newton	0.1083	-0.6445	0.6195	0.5990
Fourier	-	-	0.3302	0.4232

For the given data, it can be seen that the nonlinear model has a better fitting effect. Therefore, for different data and purposes, different models should be used to achieve better fitting results.