Assignment 2

Durbasmriti Saha
FUNDAMENTALS OF DATA SCIENCE AND MACHINE INTELLIGENCE

February 14, 2025

1) To prove the fact that the contribution of variance to error may be more than that of bias when number of datasets are less, we can take example of house size and its price.

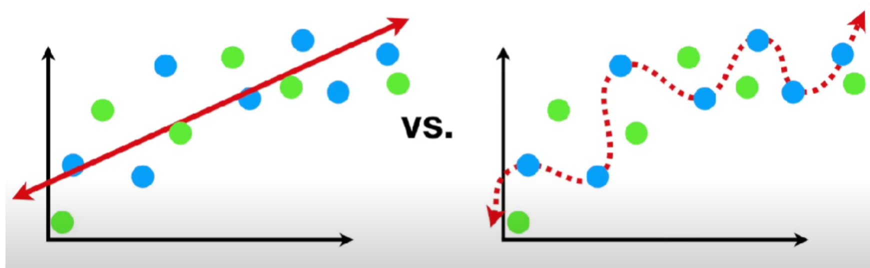

Figure 1: Comparison between simple model and complex model

Here blue dots are training data and green dots are testing data and we have only 13 data points.

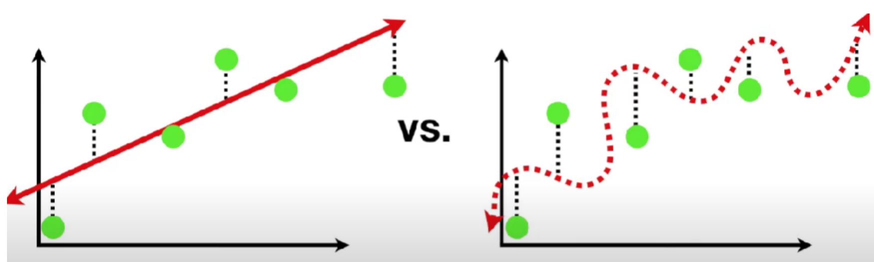

Figure 2: Comparison between simple model and complex model on testing data

The simple model has high bias but low variance. Whereas the complex model has low bias and high variance. The complex model is overfit in this dataset since it fits training datasets so well but not the testing data. Whereas, the simple model might not give great predictions but is consistently giving good predictions for any data points. Moreover, if there is/are outliers in training data, the complex model will fit well but for an ideal model, the outliers should have been discarded since they modifies the weights of model and give an overfit model. Hence, when a new data point will come, the simple model will predict better than the complex one. Hence the contribution of variance to error is more than that of bias in this example.

2) The given equation is of *regularized loss function*, which is used to control overfitting. The second term in the equation is **L2 regularization term**, and $\boldsymbol{\lambda}$ is the **regularization parameter**.
When the $\boldsymbol{\lambda}$ is increased :

- The regularization term becomes more significant.
- The model is penalized more heavily for having large weights, which forces the weights to shrink toward zero.
- The model becomes less flexible and simpler.
- Bias gets increased and the model may not capture the pattern perfectly.

When $\boldsymbol{\lambda}$ is decreased:

- The regularization term becomes less significant.
- The model is allowed to have larger weights, making it more flexible.
- The model becomes more complex and can better fit the training data.
- Bias is reduced and the model can capture the pattern well.

3) **a)** The simple linear regression model is y = c+mx, where c = intercept and m = slope of the line. Let $\bar{x}$ = mean of all x and $\bar{y}$ = mean of all y. Then

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{11211.00}{250} = 44.844$$

and

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{44520.80}{250} = 178.0832$$

and

$$m = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Putting all the values, we get

$$m = \frac{1996904.15 - 250 \cdot 44.844 \cdot 178.0832}{543503.00 - 250 \cdot (44.844)^2}$$

So, m = 0 and

$$c = \bar{y} - m\bar{x}$$

and putting all the values

$$c = 178.0832 - 0 \cdot 44.844 = 178.0832$$

So our final regression equation is

$$y = 178.0832 + 0 \cdot x$$

b) For x = 25: y = 178.0832. Hence the predicted weight is 178.0832 lbs.
c)

$$\text{Residual} = y_{\text{obs}} - \hat{y}$$

observed weight, $y_{\text{obs}} = 170$ lbs and predicted weight, $\hat{y} = 178.08$ lbs.
Residual = 170 - 178.08 = -8.08
d) Since the residual is negative (-8.08) means the observed weight is less than the predicted weight hence the prediction was an overestimate.

4) **a)** The simple linear regression model is y = c+mx, where c = intercept and m = slope of the line. Let $\bar{x}$ = mean of all x and $\bar{y}$ = mean of all y. Then

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{43}{14} = 3.0714$$

and

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{572}{14} = 40.8571$$

and

$$m = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Putting all the values, we get

$$m = \frac{1697.8 - 14 \cdot 3.07144 \cdot 40.8571}{157.42 - 14 \cdot (3.0714)^2}$$

So, m = -2.34 and

$$c = \bar{y} - m\bar{x}$$

and putting all the values

$$c = 40.8571 - (-2.34) \cdot 44.3.0714 = 48.0441$$

So our final regression equation is

$$y = 48.0441 - 2.34 \cdot x$$

$$variance = \sigma^2 = \frac{\text{SSE}}{n-2}$$

where SSE = sum of squared errors.

$$\text{SSE} = \sum_{i=1}^{n} y_i^2 - \beta_0 \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i y_i$$

putting all the values,

$$\text{SSE} = 23,530 - 48.0441 \cdot 572 - (-2.34) \cdot 1697.8 = 21.4$$

Hence,

$$variance = \sigma^2 = \frac{21.4}{14-2} = 1.7833$$

b) For x = 4.3 :

$$y = 48.0441 - 2.34 \cdot (4.3) = 37.9821$$

Hence the predicted permeability for x = 4.3 is 37.98.

c) Point estimation of mean permeability for x = 3.7 :

$$y = 48.0441 - 2.34 \cdot (3.7) = 39.3861$$

Hence, the point estimate of the mean permeability for x = 3.7 is 39.39.

d) The residual for x = 3.7 and y = 46.1

$$\text{Residual} = y_{\text{obs}} - \hat{y}$$

given, $y_{\text{obs}} = 46.1$ and $\hat{y} = 39.39$
So, Residual = 46.1 - 39.39 = 6.71

5) a) The given model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

The least squares normal equations are obtained by minimizing the sum of squared errors. Let's consider two predictors (x1 and x2), then the normal equations are

$$n\beta_0 + \beta_1 \sum x_{i1} + \beta_2 \sum x_{i2} = \sum y_i$$

$$\beta_0 \sum x_{i1} + \beta_1 \sum x_{i1}^2 + \beta_2 \sum x_{i1} x_{i2} = \sum x_{i1} y_i$$

$$\beta_0 \sum x_{i2} + \beta_1 \sum x_{i1} x_{i2} + \beta_2 \sum x_{i2}^2 = \sum x_{i2} y_i$$

Substituting the values,

$$10\beta_0 + 223\beta_1 + 553\beta_2 = 1916$$

$$10\beta_0 + 5200.9\beta_1 + 12352\beta_2 = 43550.8$$

$$553\beta_0 + 12352\beta_1 + 31729\beta_2 = 104736.8$$

b) The parameters are β_0, β_1 and β_2. The system of equations. we have :

$$10\beta_0 + 223\beta_1 + 553\beta_2 = 1916 - - - (1)$$

$$10\beta_0 + 5200.9\beta_1 + 12352\beta_2 = 43550.8 - - - (2)$$

$$553\beta_0 + 12352\beta_1 + 31729\beta_2 = 104736.8 - - - (3)$$

To solve this

$$\begin{bmatrix} 10 & 223 & 553 \\ 223 & 5200.9 & 12352 \\ 553 & 12352 & 31729 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1916 \\ 43550.8 \\ 104736.8 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 10 & 223 & 553 \\ 223 & 5200.9 & 12352 \\ 553 & 12352 & 31729 \end{bmatrix}^{-1} \begin{bmatrix} 1916 \\ 43550.8 \\ 104736.8 \end{bmatrix}$$

So we get

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 171.0555 \\ 3.7133 \\ -1.1259 \end{bmatrix}$$

c) Strength for x1 = 18 feet and x2 = 43% :

$$Y = 10.5 + 0.8x_1 + 1.2x_2$$

Substituting, Y = 10.5 + (0.8).(18) + (1.2).(43) = 76.5.
Hence, the predicted strength for x1 = 18 feet and x2 = 43% is 76.5.

6) The percent body fat = %BF.

$$\%BF = \beta_0 + \beta_1 \cdot \text{Height} + \beta_2 \cdot \text{Waist} + \epsilon$$

The regression coefficient = β and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 \end{bmatrix}^T = (X'X)^{-1}X'y$$

Substituting the given values:

$$\boldsymbol{\beta} = \begin{bmatrix} 2.9705 & -4.0042 \times 10^{-2} & -4.1679 \times 10^{-2} \\ -4.0042 \times 10^{-2} & 6.0774 \times 10^{-4} & -7.3875 \times 10^{-5} \\ -4.1679 \times 10^{-2} & -7.3875 \times 10^{-5} & 2.5766 \times 10^{-4} \end{bmatrix} \begin{bmatrix} 4757.9 \\ 334335.8 \\ 179706.7 \end{bmatrix}$$

To calculate β_0:

$$\beta_0 = (2.9705 \times 4757.9) + (-4.0042 \times 10^{-2} \times 334335.8) + (-4.1679 \times 10^{-2} \times 179706.7)$$

$$\beta_0 = 14133.5 - 13386.5 - 7488.6 = -6741.6$$

To calculate β_1 :

$$\beta_1 = \big(-4.0042 \times 10^{-2} \times 4757.9\big) + \big(6.0774 \times 10^{-4} \times 334335.8\big) + \big(-7.3875 \times 10^{-5} \times 179706.7\big)$$

$$\beta_1 = -190.5 + 203.2 - 13.3 = -0.6$$

To calculate β_2 :

$$\beta_2 = \big(-4.1679 \times 10^{-2} \times 4757.9\big) + \big(-7.3875 \times 10^{-5} \times 334335.8\big) + \big(2.5766 \times 10^{-4} \times 179706.7\big)$$

$$\beta_2 = -198.3 - 24.7 + 46.3 = -176.7$$

So the regression model is :

$$\%BF = -6741.6 - 0.6 \cdot \text{Height} + 1.767 \cdot \text{Waist} + \epsilon$$

7) Given quadratic model :

$$f(x_1, x_2) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2$$

The error function = E,

$$E = \sum_{i=1}^{N} \left[y_i - \left(w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i1} x_{i2} + w_4 x_{i1}^2 + w_5 x_{i2}^2 \right) \right]^2$$

Now, we have to minimize E, so we have to take partial deivatives of E with respect to each coefficient w_j, $\quad (j = 0, 1, 2, 3, 4, 5)$ and make them equal to zero.

$$\frac{\partial E}{\partial w_j} = 0 \quad \text{for } j = 0, 1, 2, 3, 4, 5.$$

The above equation can be written in matrix form as : $X'Xw = X'y$ where X is the matrix of size N x 6,

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} & x_{11}^2 & x_{12}^2 \\ 1 & x_{21} & x_{22} & x_{21}x_{22} & x_{21}^2 & x_{22}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} & x_{N1}x_{N2} & x_{N1}^2 & x_{N2}^2 \end{bmatrix}$$

and w is the vector of coefficients :

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{bmatrix}$$

and y is the vector of observed responses :

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

And the least squares estimates of the coefficients are given by :

$$w = (X^T X)^{-1} X^T y$$