# Over-Fitting for Machine Learning Problem

Maryam Zubair

# Problem:

Given a problem with two models i.e., Linear and nonlinear for which we had to choose the best model based on how accurate it will make predictions.

In order to find predicted values, I have calculated parameters (Slope and Intercept) for each model and put these values in the relevant equation to see which line will best fit to the dataset. Then predicted which model is better based on the Mean Squared Error calculated for each model.

# Formulas:

## Linear Model:

Regression equation: $y = a + bx$

Slope: $b = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X2 - (\Sigma X)2)$

Intercept: $a = (\Sigma Y - b(\Sigma X)) / N$

# Formulas:

## Non- Linear Model:

Regression equation: $y = a + b(x)^2$

Slope: $b = (N\Sigma PY - (\Sigma P)(\Sigma Y)) / (N\Sigma P2 - (\Sigma P)2)$

Intercept: $a = (\Sigma Y - b(\Sigma P)) / N$

# Training Data

| Training Data | | Linear Model (M1) | Non-Linear Model (M2) |
|---|---|---|---|
| x | y | ŷ = (b1*X) + a1 | ŷ = (b2*x2) + a2 |
| 1 | 1.8 | 1.368272655 | 1.751782112 |
| 2 | 2.4 | 2.231450336 | 2.155469346 |
| 3.3 | 2.3 | 3.353581322 | 3.08260436 |
| 4.3 | 3.8 | 4.216759003 | 4.105278687 |
| 5.3 | 5.3 | 5.079936684 | 5.397077836 |
| 1.4 | 1.5 | 1.713543728 | 1.880962027 |
| 2.5 | 2.2 | 2.663039177 | 2.458234771 |
| 2.8 | 3.8 | 2.921992481 | 2.672189005 |
| 4.1 | 4 | 4.044123467 | 3.879213836 |
| 5.1 | 5.4 | 4.907301148 | 5.11718802 |

# Validation data

| Validation Data | | Linear Model (M1) | Non-Linear Model (M2) |
|---|---|---|---|
| x | y | ŷ = (b1*X) + a1 | ŷ = (b2*x2) + a2 |
| 1.5 | 1.7 | 1.799861496 | 1.919985126 |
| 2.9 | 2.7 | 3.008310249 | 2.74888958 |
| 3.7 | 2.5 | 3.698852394 | 3.459379112 |
| 4.7 | 2.8 | 4.562030075 | 4.589703368 |
| 5.1 | 5.5 | 4.907301148 | 5.11718802 |

# Mean Squared Error:

Mean squared error (MSE) is a commonly used metric to evaluate the performance of a predictive model.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

# Cost Function/ Mean Square Error for Training Data:

| MSE Linear (M1) | MSE Non-Linear(M2) |
|---|---|
| 0.1863885 | 0.002324965 |
| 0.028408989 | 0.059795241 |
| 1.110033601 | 0.612469585 |
| 0.173688066 | 0.093195077 |
| 0.048427863 | 0.009424106 |
| 0.045600924 | 0.145132066 |
| 0.214405279 | 0.066685197 |
| 0.770897203 | 1.271957639 |
| 0.00194688 | 0.014589297 |
| 0.242752159 | 0.079982616 |

MSE for Linear Model = 2.822/10 = 0.28
MSE for Non-Linear Model = 2.355/10 = 0.235

# • Comparison ( Linear Model Vs  Non-Linear Model)

**Model 1:**

Max (Training Set, Validation Set) / Min (Training Set, Validation Set) = 0.499/ 0.28 =1.77

**Model 2:**

Max (Training Set, Validation Set) / Min (Training Set, Validation Set) = 0.432/ 0.235 =1.83

# Result

“ As Mean Squared Error tells how close the fitted line to the data point is so smaller the MSE the better the Model. In this given question **Linear Regression is a better Model than Non- Linear Model.** “

A lower MSE indicates that the model is making predictions that are closer to the true values, and therefore the model is considered to be better.