Week 1 HW 2:
Using Overfitting to evaluate Models

Khoi Duong
CS550
Prof. Chang
1/31/2023

Table of content

1. Explain and understand the problem

For this exercise, given a dataset, we have to compare the two regression models: Linear Regression and Nonlinear Regression to see which one has more serious overfitting issue. In order to do this, first, we have to distribute the sample data by 3 phases (Training Phase: 50%, Validation Phase: 25%, and Test Phase: 25%). Then, we have to determine the equation of the 2 Regression Models. Next, we fill different values of y into the data table. After filling up the table, we have to calculate the MSE and compare the overfitting issue.

2. Distribute sample data by 3 phases (Training, Validation, and Test phases)

Following to the given problem, the training, validation, and test phases will have the following values:

x	y
1	1.8
2	2.4
3.3	2.3
4.3	3.8
5.3	5.3
1.4	1.5
2.5	2.2
2.8	3.8
4.1	4.0
5.1	5.4

Training Phase

x	y
1.5	1.7
2.9	2.7
3.7	2.5
4.7	2.8
5.1	5.5

Validation Phase

x
1.4
2.5
3.6
4.5
5.4

Test Phase

3. Define the formulas for Linear Regression & Nonlinear Regression Models

Model 1: Linear Regression

Number of values: N = 10

We have the following table:

X value	Y value	X * Y	X * X
1	1.8	1.8	1
2	2.4	4.8	4
3.3	2.3	7.59	10.89
4.3	3.8	16.34	18.49
5.3	5.3	28.09	28.09
1.4	1.5	2.1	1.96
2.5	2.2	5.5	6.25
2.8	3.8	10.64	7.84
4.1	4.0	16.4	16.81
5.1	5.4	27.54	26.01

Calculating the total:

$$\sum X = 31.8,\ \sum Y = 32.5, \sum XY = 120.8,\ \sum X^2 = 121.34$$

We have the equation for linear regression: y = a + bx

$$\text{Slope (b)} = \frac{N * \sum XY - (\sum X)(\sum Y)}{N * \sum X^2 - (\sum X)^2} = \frac{10*120.8 - (31.8)(32.5)}{10*121.34 - (31.8)^2} = \frac{174.5}{202.16} \approx 0.863$$

$$\text{Intercept (a)} = \frac{\sum Y - b*\sum X}{N} = \frac{32.5 - 0.863*31.8}{10} = \frac{5.05}{10} = 0.505$$

Therefore, we have the regression formula: y = 0.505 + 0.863x

Model 2: Non-linear Regression

Number of values: N = 10

We have the table below:

X Values	$\underline{X}$ Values	Y Values
1	1	1.8
2	4	2.4
3.3	10.89	2.3
4.3	18.49	3.8
5.3	28.09	5.3
1.4	1.96	1.5
2.5	6.25	2.2
2.8	7.84	3.8
4.1	16.81	4
5.1	26.01	5.4

We have the following table:

$\underline{X}$ value	Y value	$\underline{X} * Y$	$\underline{X} * \underline{X}$
1	1.8	1.8	1
4	2.4	9.6	16
10.89	2.3	25.047	118.5921
18.49	3.8	70.262	341.8801
28.09	5.3	148.877	789.0481
1.96	1.5	2.94	3.8416
6.25	2.2	13.75	39.0625
7.84	3.8	29.792	61.4656
16.81	4.0	67.24	282.5761
26.01	5.4	140.454	676.5201

Calculating the total:

$$\sum \underline{X} = 121.34,\ \sum Y = 32.5, \sum \underline{XY} = 509.762,\ \sum \underline{X}^2 = 2329.9862$$

We have the equation for non-linear regression: $y = a + bx^2$

$$\text{Slope (b)} = \frac{N * \sum \underline{XY} - (\sum \underline{X})(\sum Y)}{N * \sum \underline{X}^2 - (\sum \underline{X})^2} = \frac{10*509.762 - (121.34)(32.5)}{10*2329.9862 - (121.34)^2} = \frac{1154.07}{8576.47} \approx 0.135$$

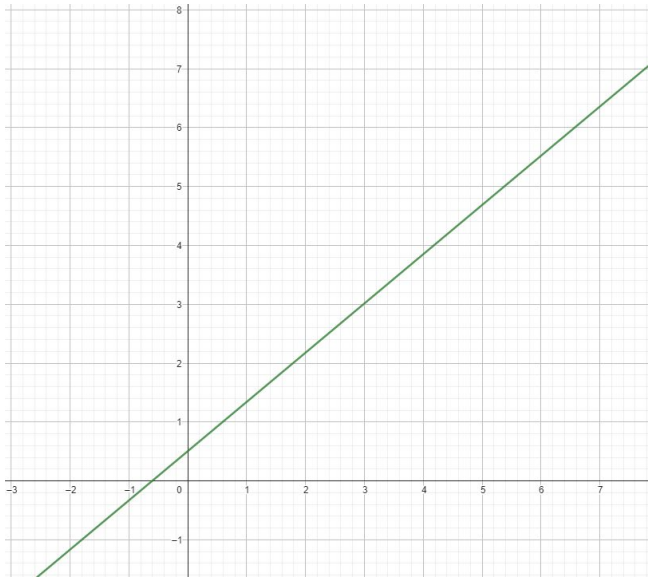
$$\text{Intercept (a)} = \frac{\sum Y - b*\sum \underline{X}}{N} = \frac{32.5 - 0.135*121.34}{10} = \frac{16.17}{10} = 1.617$$

Therefore, we have the regression formula: $y = 1.617 + 0.135x^2$

The equations for 2 Regression Models

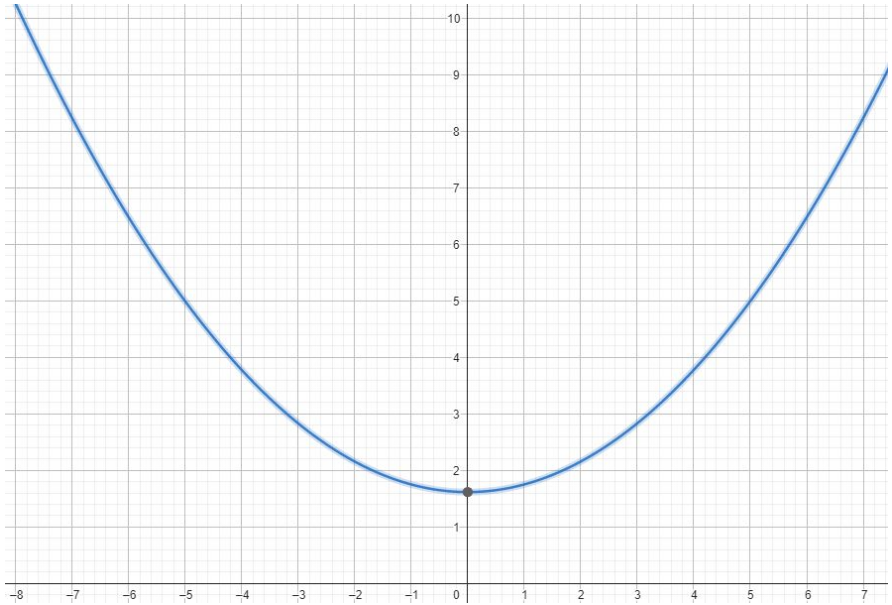
Model 1: Linear Regression

$$y = 0.505 + 0.836x$$



Model 2: Non-linear Regression

$$y = 1.617 + 0.135x^2$$



4. Fill out the table for training and validation phases

Real Data Set 1: Training Phase 50% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression
x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$
1	1.8	1.368	1.752
2	2.4	2.231	2.157
3.3	2.3	3.3529	3.08715
4.3	3.8	4.2159	4.11315
5.3	5.3	5.0789	5.40915
1.4	1.5	1.7132	1.8816
2.5	2.2	2.6625	2.46075
2.8	3.8	2.9214	2.6754
4.1	4	4.0433	3.88635
5.1	5.4	4.9063	5.12835

Real Data Set 2: Validation Phase 25% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression
x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$
1.5	1.8	1.7995	1.92075
2.9	2.4	3.0077	2.75235
3.7	2.3	3.6981	3.46515
4.7	3.8	4.5611	4.59915
5.1	5.3	4.9063	5.12835
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X

5. Calculate the MSE (Mean Squared Error) & Fill Test Phase

Real Data Set 1: Training Phase 50% of the collected data		Model 1: Linear Regression	Distance
x	y	ŷ=a1 + b1 * x	(ŷ-y)^2
1	1.8	1.368	0.186624
2	2.4	2.231	0.028561
3.3	2.3	3.3529	1.10859841
4.3	3.8	4.2159	0.17297281
5.3	5.3	5.0789	0.04888521
1.4	1.5	1.7132	0.04545424
2.5	2.2	2.6625	0.21390625
2.8	3.8	2.9214	0.77193796
4.1	4	4.0433	0.00187489
5.1	5.4	4.9063	0.24373969
		Total	**2.82255446**
		MSE	**0.282255446**

Training Phase

- Model 1: Linear Regression

MSE = 0.282

Real Data Set 2: Validation Phase 25% of the collected data		Model 1: Linear Regression	Distance
x	y	ŷ=a1 + b1 * x	(ŷ-y)^2
1.5	1.8	1.7995	0.00000025
2.9	2.4	3.0077	0.36929929
3.7	2.3	3.6981	1.95468361
4.7	3.8	4.5611	0.57927321
5.1	5.3	4.9063	0.15499969
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
		Total	**3.05825605**
		MSE	**0.305825605**

Validation Phase

- Model 1: Linear Regression

MSE = 0.306

Real Data Set 1: Training Phase 50% of the collected data		Model 2: Non-Linear Regression	Distance
x	y	ŷ=a2 + b2 * x^2	(ŷ-y)^2
1	1.8	1.752	0.002304
2	2.4	2.157	0.059049
3.3	2.3	3.08715	0.6196051225
4.3	3.8	4.11315	0.0980629225
5.3	5.3	5.40915	0.0119137225
1.4	1.5	1.8816	0.14561856
2.5	2.2	2.46075	0.0679905625
2.8	3.8	2.6754	1.26472516
4.1	4	3.88635	0.0129163225
5.1	5.4	5.12835	0.0737937225
		Total	**2.355979095**
		MSE	**0.2355979095**

Training Phase

- Model 2: Non-linear Regression

MSE = 0.236

Real Data Set 2: Validation Phase 25% of the collected data		Model 1: Linear Regression	Distance
x	y	ŷ=a1 + b1 * x	(ŷ-y)^2
1.5	1.8	1.92075	0.0145805625
2.9	2.4	2.75235	0.1241505225
3.7	2.3	3.46515	1.357574523
4.7	3.8	4.59915	0.6386407225
5.1	5.3	5.12835	0.0294637225
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
		Total	**2.164410053**
		MSE	**0.2164410053**

Validation Phase

- Model 2: Non-linear Regression

MSE = 0.216

6. Compare the 2 Models to see which one has more serious overfitting issue

- Model 1: 0.306/0.282 = 1.084
- Model 2: 0.236/0.216 = 1.089
- Compare Model 1 and Model 2: Model 1 is the better model since 1.084 < 1.089

7. Fill test phase data

Real Data Set 2: 25% of the collected data	The better model (Model 1) selected from Validation phase
x	$\hat{y}=a1 + b1 * x$
1.4	1.7132
2.5	2.6625
3.6	3.6118
4.5	4.3885
5.4	5.1652
X	X
X	X
X	X
X	X
X	X

8. Conclusion

Model 1 is the better model, this means that Model 2 has more serious overfitting issues since the quotient of the 2 MSE between validation & training data. Thus, for the test phase, we will use the Model 1 to calculate the value of y to get the best and closest estimated value.

9. Reference

Thangarani Prabhu - TA, Summer 2018

Use Overfitting to evaluate different Models

R and Linear/Non-Linear Regression - Lena Lee, Fall 2015