Course: CS550 - Machine Learning and Business Intelligence
Faculty: Prof. Henry Chang

Machine Learning: Overfitting to evaluate Linear Regression Model and Non-linear Regression

Name: Haripriya A
ID: 19579

NORTHWESTERN POLYTECHNIC
UNIVERSITY

Table of Contents

Introduction

- Machine learning is a method of data analysis investigation that automates logical model structure. Machine learning algorithms are largely used to predict, classify, or cluster.
- Machine Learning

  * Supervised Learning

    - Regression : Linear Regression Algorithm
    - Classification : KNN Algorithm

  * Unsupervised Learning

    - Clustering : K-Mean Algorithm

About Regression

- A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables.
- We compare the following two Regression Models to see which one has more serious overfitting issue in the coming slides
  - Linear Regression Model 1
  - Non-Linear Regression Model 2

Input Data

- Training phase: 50%, Validation phase: 25%, Test phase: 25%

Training Phase				Validation Phase				Test Phase	
Real Data Set 1 50% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collected data	The better model selected from Model 1 and Model 2 depending on the analysis of overfitting
X	Y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$	X	Y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x2$	X	$\hat{y}=a1 + b1 * x$ or $\hat{y}=a2 + b2 * x^2$
1	1.8			1.5	1.7			1.4	
2	2.4			2.9	2.7			2.5	
3.3	2.3			3.7	2.5			3.6	
4.3	3.8			4.7	2.8			4.5	
5.3	5.3			5.1	5.5			5.4	
1.4	1.5			X	X	X	X	X	X
2.5	2.2			X	X	X	X	X	X
2.8	3.8			X	X	X	X	X	X
4.1	4.0			X	X	X	X	X	X
5.1	5.4			X	X	X	X	X	X

Model 1: Linear Regression

To find regression equation, we will first find slope, intercept and use it to form regression equation.

- Step 1:Count the number of values. N = 10
- Step 2:Find X * Y, X^2 See the below table

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

Linear Regression Model 1

Step 3:

Find ΣX, ΣY, ΣXY, ΣX^2.

ΣX = 31.8

ΣY = 32.5

ΣXY = 120.8

ΣX^2 = 121.34

Step 4:

Substitute in the above slope formula given.

Slope(b1) = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

= ((10*120.8)-(31.8*32.5) / ((10*121.34)-(31.8)$^{2)}$

= 0.86

Step 5:

Now, again substitute in the above intercept formula given.

Intercept(a1) = (ΣY - b(ΣX)) / N

= (32.5 - (0.86*31.8))/10

= (32.5-27.348)/10

= 0.51

Linear Regression Model 1

Step 6:

Then substitute Intercept(a) and Slope(b) in regression equation formula

Regression Equation(y) = a + bx

= 0.51+0.86x

Step 7:

Suppose if we want to know the approximate y value for the variable x = 64. Then we can substitute the value in the above equation.

Regression Equation(y) = a + bx

= 0.51+0.86x

= 0.51+0.86(1)

= 1.37

Linear Regression Model 1

- Training phase Model 1 ‘y’ Calculation in the table below

Training Phase				Validation Phase				Test Phase	
Real Data Set 1 50% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regressio n	Real Data Set 2 25% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collected data	The better model selected from Model 1 and Model 2 depending on the analysis of overfitting
X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	X	ŷ=a1 + b1 * x or ŷ=a2 + b2 * x^2
1	1.8	0.51+0.86(1)=1.37		1.5	1.7			1.4	
2	2.4	0.51+0.86(2)=2.74		2.9	2.7			2.5	
3.3	2.3	0.51+0.86(3.3)=3.348		3.7	2.5			3.6	
4.3	3.8	0.51+0.86(4.3)=4.208		4.7	2.8			4.5	
5.3	5.3	0.51+0.86(5.3)=5.068		5.1	5.5			5.4	
1.4	1.5	0.51+0.86(1.4)=1.714		X	X	X	X	X	X
2.5	2.2	0.51+0.86(2.5)=2.66		X	X	X	X	X	X
2.8	3.8	0.51+0.86(2.8)=2.918		X	X	X	X	X	X
4.1	4.0	0.51+0.86(4.1)=4.036		X	X	X	X	X	X
5.1	5.4	0.51+0.86(5.1)=4.896		X	X	X	X	X	X

Linear Regression Model 1

- Validation phase Model 1 ‘y’ Calculation in the table below

Training Phase				Validation Phase				Test Phase	
Real Data Set 1 50% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collected data	The better model selected from Model 1 and Model 2 depending on the analysis of overfitting
X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	X	ŷ=a1 + b1 * x or ŷ=a2 + b2 * x^2
1	1.8	0.51+0.86(1)=1.37		1.5	1.7	0.51+0.86(1.5)=1.8		1.4	
2	2.4	0.51+0.86(2)=2.74		2.9	2.7	0.51+0.86(2.9)=3.004		2.5	
3.3	2.3	0.51+0.86(3.3)=3.348		3.7	2.5	0.51+0.86(3.7)=3.692		3.6	
4.3	3.8	0.51+0.86(4.3)=4.208		4.7	2.8	0.51+0.86(4.7)=4.552		4.5	
5.3	5.3	0.51+0.86(5.3)=5.068		5.1	5.5	0.51+0.86(5.1)=4.896		5.4	
1.4	1.5	0.51+0.86(1.4)=1.714		X	X	X	X	X	X
2.5	2.2	0.51+0.86(2.5)=2.66		X	X	X	X	X	X
2.8	3.8	0.51+0.86(2.8)=2.918		X	X	X	X	X	X
4.1	4.0	0.51+0.86(4.1)=4.036		X	X	X	X	X	X
5.1	5.4	0.51+0.86(5.1)=4.896		X	X	X	X	X	X

Non-linear Regression Model 2

Non-linear Regression Formula:

Slope(b) = $(N\Sigma \underline{P}Y - (\Sigma \underline{P})(\Sigma Y)) / (N\Sigma \underline{P}^2 - (\Sigma \underline{P})^2)$

Intercept(a) = $(\Sigma Y - b(\Sigma \underline{P})) / N$ Where $\underline{P} = X * X$

Slope(b) = $(N\Sigma \underline{X}^2Y - (\Sigma \underline{X}^2)(\Sigma Y)) / (N\Sigma \underline{X}^4 - (\Sigma \underline{X}^2)^2)$

Intercept(a) = $(\Sigma Y - b(\Sigma \underline{X}^2)) / N$

- Step 1:Count the number of values. N = 10
- Step 2:Find X * Y, $X^{2,}$

  See the below table

Slope(b2) = $(N\Sigma \underline{X}^2Y - (\Sigma \underline{X}^2)(\Sigma Y)) / (N\Sigma \underline{X}^4 - (\Sigma \underline{X}^2)^2)$

=(10(509.762)-(121.34)(32.5))/(10(2329.9862)-(121.34*121.34))

=(5097.62-3943.55)/(23299.862-14723.3956)

=1154.07/8576.4664, 10211340

=0.13

Intercept(a2) = $(\Sigma Y - b(\Sigma \underline{X}^2)) / N$

=(32.5-0.13(121.34))/10

=1.67

Non-linear Regression Model 2

- Training phase, Validation Model 2 'y' Calculation in the table below

Training Phase				Validation Phase				Test Phase	
Real Data Set 1 50% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collected data	The better model selected from Model 1 and Model 2 depending on the analysis of overfitting
X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	X	ŷ=a1 + b1 * x or ŷ=a2 + b2 * x^2
2	2.4	0.51+0.86(2)=2.74	$1.67+0.13(2)^2=2.19$	2.9	2.7	0.51+0.86(2.9)=3.004	$1.67+0.13(2.9)^2=2.7633$	2.5	
3.3	2.3	0.51+0.86(3.3)=3.348	$1.67+0.13(3.3)^2=3.0857$	3.7	2.5	0.51+0.86(3.7)=3.692	$1.67+0.13(3.7)^2=3.4497$	3.6	
4.3	3.8	0.51+0.86(4.3)=4.208	$1.67+0.13(4.3)^2=4.0737$	4.7	2.8	0.51+0.86(4.7)=4.552	$1.67+0.13(4.7)^2=4.5417$	4.5	
5.3	5.3	0.51+0.86(5.3)=5.068	$1.67+0.13(5.3)^2=5.3217$	5.1	5.5	0.51+0.86(5.1)=4.896	$1.67+0.13(5.1)^2=5.0513$	5.4	
1.4	1.5	0.51+0.86(1.4)=1.714	$1.67+0.13(1.4)^2=1.9248$	X	X	X	X	X	X
2.5	2.2	0.51+0.86(2.5)=2.66	$1.67+0.13(2.5)^2=2.4825$	X	X	X	X	X	X
2.8	3.8	0.51+0.86(2.8)=2.918	$1.67+0.13(2.8)^2=2.6892$	X	X	X	X	X	X
4.1	4.0	0.51+0.86(4.1)=4.036	$1.67+0.13(4.1)^2=3.8553$	X	X	X	X	X	X
5.1	5.4	0.51+0.86(5.1)=4.896	$1.67+0.13(5.1)^2=5.0513$	X	X	X	X	X	X

Overfitting to compare Models

Training Set:

Model 1

$=(((1.8-1.37)^2+(2.4-2.23)^2+(2.3-3.348)^2+(3.8-4.208)^2+(5.3-5.068)^2+(1.5-1.714)^2+(2.2-2.66)^2+(3.8-2.918)^2+(4-4.036)^2+(5.4-4.896)^2)/10)$

$=2.823024/10$

$=0.2823024$

Model 2

$=(((1.8-1.8)^2+(2.4-2.19)^2+(2.3-3.0857)^2+(3.8-4.0737)^2+(5.3-5.3217)^2+(1.5-1.9248)^2+(2.2-2.4825)^2+(3.8-2.6892)^2+(4-3.8553)^2+(5.4-5.0513)^2)/10)$

$=2.37347478/10$

$=0.237347478$

Overfitting to compare Models

Validation Set:

Model 1

$(((1.7-1.8)^2+(2.7-3.004)^2+(2.5-3.692)^2+(2.8-4.552)^2+(5.5-4.896)^2)/5)$

$=4.9576/5$

$=0.99$

Model 2

$(((1.7-1.9625)^2+(2.7-2.7633)^2+(2.5-3.4497)^2+(2.8-4.5417)^2+(5.5-5.0513)^2)/5)$

$=4.20969381/5$

$=0.84193876$

Overfitting to compare Models

max(Training_Set_MSE, Validation_Set_MSE) / min(Training_Set_MSE, Validation_Set_MSE)

- Compare Model 1 and Model 2
- Model 1

  0.99 / 0.2823024 = 3.50687773

- Model 2

  0.84193876 / 0.237347478 = 3.547283

- Conclusion
  - Model 1 is a better model

Test Phase Calculation

Training Phase				Validation Phase				Test Phase	
Real Data Set 1 50% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collected data	The better model selected from Model 1 and Model 2 depending on the analysis of overfitting
X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	X	ŷ=a1 + b1 * x/ ŷ=a2 + b2 * x^2
1	1.8	0.51+0.86(1)=1.37	$1.67+0.13(1)^2=1.8$	1.5	1.7	0.51+0.86(1.5)=1.8	$1.67+0.13(1.5)^2=1.9625$	1.4	0.51+0.86(1.4)=1.714
2	2.4	0.51+0.86(2)=2.74	$1.67+0.13(2)^2=2.19$	2.9	2.7	0.51+0.86(2.9)=3.004	$1.67+0.13(2.9)^2=2.7633$	2.5	0.51+0.86(2.5)=2.66
3.3	2.3	0.51+0.86(3.3)=3.348	$1.67+0.13(3.3)^2=3.0857$	3.7	2.5	0.51+0.86(3.7)=3.692	$1.67+0.13(3.7)^2=3.4497$	3.6	0.51+0.86(3.6)=3.606
4.3	3.8	0.51+0.86(4.3)=4.208	$1.67+0.13(4.3)^2=4.0737$	4.7	2.8	0.51+0.86(4.7)=4.552	$1.67+0.13(4.7)^2=4.5417$	4.5	0.51+0.86(4.5)=4.38
5.3	5.3	0.51+0.86(5.3)=5.068	$1.67+0.13(5.3)^2=5.3217$	5.1	5.5	0.51+0.86(5.1)=4.896	$1.67+0.13(5.1)^2=5.0513$	5.4	0.51+0.86(5.4)=5.154
1.4	1.5	0.51+0.86(1.4)=1.714	$1.67+0.13(1.4)^2=1.9248$	X	X	X	X	X	X
2.5	2.2	0.51+0.86(2.5)=2.66	$1.67+0.13(2.5)^2=2.4825$	X	X	X	X	X	X
2.8	3.8	0.51+0.86(2.8)=2.918	$1.67+0.13(2.8)^2=2.6892$	X	X	X	X	X	X
4.1	4.0	0.51+0.86(4.1)=4.036	$1.67+0.13(4.1)^2=3.8553$	X	X	X	X	X	X
5.1	5.4	0.51+0.86(5.1)=4.896	$1.67+0.13(5.1)^2=5.0513$	X	X	X	X	X	X

Conclusion

- Conclusion
    - Model 1 is a better model and the Test values are calculated using Model 1 which is Linear regression.

References

- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/non_linear_regression_example.html
- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/non_linear_regression_example.html#nl
- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/overfit.html

Google slides URL :

https://docs.google.com/presentation/d/1D-MQA2TPTqSJ58bevw2EzOi-o6yeQWBp1TNz60yYz18/edit?usp=sharing