Use Overfitting To Evaluate Different Models

Submitted by Manish Bafna
Student Id: 19655
Instructor: Dr. Henry Chang
GIT: MachineLearning-Overfitting

Table of Content

Introduction

What Is Machine Learning?

- Machine Learning uses Data Mining techniques and other learning algorithms to build models of what is happening behind some data so that it can predict future outcomes.

<u>Types of Machine Learning Systems</u>

- There are so many different types of Machine Learning systems that it is useful to classify them in broad categories based on:
    - Whether or not they are trained with human supervision
        - supervised
        - unsupervised
        - semisupervised
        - Reinforcement Learning

Introduction

- Components of Machine Learning
  - Data
  - Machine Learning Algorithms (e.g., Best Fit, Deep Learning)
  - Model
  - Prediction
- Based on the accumulated data to generate a new model (i.e., the line) every day using Best Fit algorithm and to do better predicton.
- Data is like oil for a country
  - Countries having large population are easier to develope Machine Learning Industry.
  - Internet companies providing free services to collect data.

Design

Training Phase				Validation Phase				Test Phase	
Real Data Set 1 50% of the collcted data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collcted data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collcted data	The better model (Model 1 or Model 2) selected from the Validation Phase based on the analysis of overfitting will be used to calculate ŷ

- **After calculating a1, b1, a2, b2 in Training Phase, the values are not changed with the new Real Data Sets in Validation Phase and Test Phase.**
- **Only ŷ values are changed with the new Real Data Sets.**

x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$	x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$	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

Design

Suppose we collect a set of sample data and distribute the sample data by

Training phase: 50%

Validation phase: 25%

Test phase: 25%

After calculating a1, b1, a2, b2 in Training Phase, the values are not changed with the new Real Data Sets in Validation Phase and Test Phase.

Only $\hat{y}$ values are changed with the new Real Data Sets.

Design

Real Data Set 1 can be used to determine the formulas for Model 1: Linear Regression and Model 2: Non-Linear Regression. That is, to determine the values of a1, b1, a2, and b2 in the following formulas:

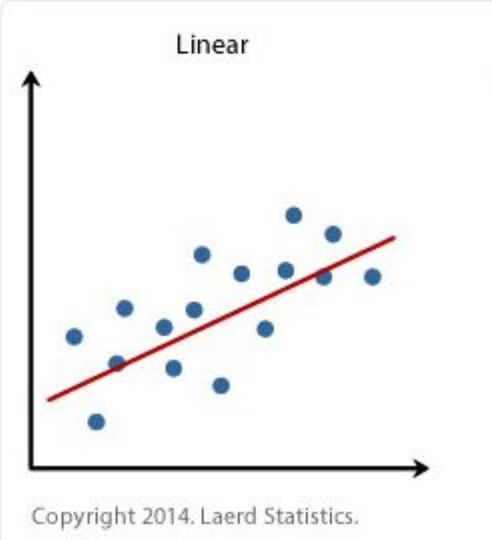
$\hat{y}=a1 + b1 * x$

$\hat{y}=a2 + b2 * x^2$

- After the formulas are determined, you can use the formulas to calculate the $\hat{y}$ values in the following phases:
  - Training Phase
  - Validation Phase
  - Test Phase
- Note: The values of "x" in "$\hat{y}=a1 + b1 * x$" and "$\hat{y}=a2 + b2 * x^2$" are the same as the "x" list on the "Real Data Set".

Implementation(Linear Regression)

1. We first calculate the slope and intercept of Linear regression for the data in Training set



Formula for Linear Regression(The Normal Equation)

Regression Equation(y) = a + bx

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

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

Where:

x and y are the variables.

b = The slope of the regression line

a = The intercept point of the regression line and the y axis.

N = Number of values or elements

Formula for Linear Regression(cont)

Where:

X = First Score

Y = Second Score

ΣXY = Sum of the product of first and Second Scores

ΣX = Sum of First Scores

ΣY = Sum of Second Scores

ΣX^2 = Sum of square First Scores

Implementation

Step 1:

Count the number of values. N=10

Step 2:

Find X * Y, X^2

Step 3:

Find ΣX, ΣY, ΣXY, ΣX^2.

Step 4:

Substitute in the above slope formula given.

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

Implementation

Step 5:

Now, again substitute in the above intercept formula given.

Intercept(a) = (ΣY - b(ΣX)) / N

Step 6:

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

Regression Equation(y) = a1 + b1x

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) = a1 + b1x

Implementation

	X Values	Y Values	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	16.4	16.81
	5.1	5.4	27.54	26.01
sum	**31.8**	**32.5**	**120.8**	**121.34**

Implementation

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

$=(10*(120.8)-(31.8)*(32.5))/((10)*(121.34)-(31.8*31.8))$

$=0.86$

Intercept(a1) = $(\Sigma Y - b(\Sigma X)) / N$

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

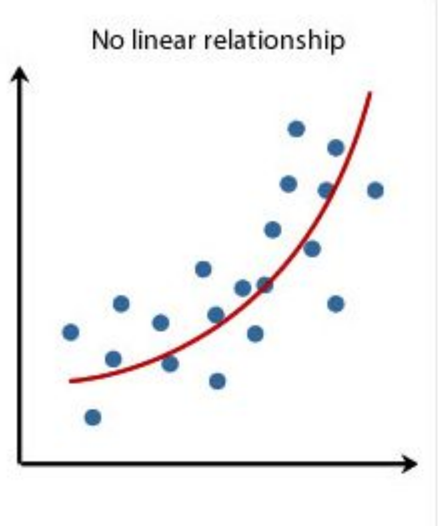
$=0.52$

Test(Linear Regression)

	Training Set			Validation Set			
	Real Data(Set 1)		Model 1(Linear Regression)	Real Data(Set 2)		Model 1(Linear Regression)	Real Data(Set 3)
x	x	y	ŷ=a1 + b1 * x	x	y	ŷ=a1 + b1 * x	x
1	1	1.8	1.38	1.5	1.7	1.81	1.4
2	2	2.4	2.24	2.9	2.7	3.014	2.5
3	3.3	2.3	3.358	3.7	2.5	3.702	3.6
4	4.3	3.8	4.218	4.7	2.8	4.562	4.5
5	5.3	5.3	5.078	5.1	5.5	4.906	5.4
6	1.4	1.5	1.724				
7	2.5	2.2	2.67				
8	2.8	3.8	2.928				
9	4.1	4	4.046				
10	5.1	5.4	4.906				

Implementation(Non-Linear Regression)

1. We first calculate the slope and intercept of Non-Linear regression for the data in Training set



Formula for Non-Linear Regression(The Normal Equation)

Regression Equation(y) = a + bx^2

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

Where:

x and y are the variables.

b = The slope of the regression line

a = The intercept point of the regression line and the y axis.

N = Number of values or elements

Formula for Linear Regression(cont)

Where:

X = First Score

Y = Second Score

ΣXY = Sum of the product of first and Second Scores

ΣX = Sum of First Scores

ΣY = Sum of Second Scores

ΣX^2 = Sum of square First Scores

Implementation

We can simply create X from X where X =X*X

Step 1:

Count the number of values. N=10

Step 2:

Find X * Y, X^2

Step 3:

Find ΣX, ΣY, ΣXY, ΣX^2.

Step 4:

Substitute in the above slope formula given.

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

Implementation

Step 5:

Now, again substitute in the above intercept formula given.

Intercept(a) = (ΣY - b(ΣX)) / N

Step 6:

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

Regression Equation(y) = a2 + b2x

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) = a2 + b2x

Implementation

	X	X Values	Y Values	X*Y	X*X
	1	1	1.8	1.8	1
	2	4	2.4	9.6	16
	3.3	10.89	2.3	25.047	118.5921
	4.3	18.49	3.8	70.262	341.8801
	5.3	28.09	5.3	148.877	789.0481
	1.4	1.96	1.5	2.94	3.8416
	2.5	6.25	2.2	13.75	39.0625
	2.8	7.84	3.8	29.792	61.4656
	4.1	16.81	4	67.24	282.5761
	5.1	26.01	5.4	140.454	676.5201
sum	31.8	121.34	32.5	509.762	2329.986

Implementation

Slope(b2) = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

=(10*(509.76)-(121.34)*(32.5))/((10) *(2330)-(121.34)*(121.34))

=0.13

Intercept(a2) = $(\Sigma Y - b(\Sigma X)) / N$

=(32.5-(0.13)*(121.34))/10

=1.67

Test(Non-Linear Regression)

	Training Set			Validation Set		
	Real Data(Set 1)		Model 2(Non Linear Regression)	Real Data(Set 2)		Model 2(Non Linear Regression)
x	x	y	ŷ=a2 + b2 * x^2	x	y	ŷ=a2 + b2 * x^2
1	1	1.8	1.8	1.5	1.7	1.9625
2	2	2.4	2.19	2.9	2.7	2.7633
3	3.3	2.3	3.0857	3.7	2.5	3.4497
4	4.3	3.8	4.0737	4.7	2.8	4.5417
5	5.3	5.3	5.3217	5.1	5.5	5.0513
6	1.4	1.5	1.9248			
7	2.5	2.2	2.4825			
8	2.8	3.8	2.6892			
9	4.1	4	3.8553			
10	5.1	5.4	5.0513			

Implementation

- The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points.
  - The smaller the MSE, the closer the fit is to the data.
- If $\hat{Y}$ is a vector of n predictions, and Y is the vector of the true values, then the (estimated) MSE of the predictor is:

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

Implementation

Calculate MSE:

Validation Set

Model 1(Linear Model)

MSE =[(1.7-1.8)^2+(2.7-3.0)^2+(2.5-3.7)^2+(2.8-4.6)^2+(5.5-4.9)^2)]

=5.01/5

=1.002

Similarly for

Model 2(Non-Linear Model)

MSE =10.87/5

=2.17

Test

The better model (<u>Model 1</u> or <u>Model 2</u>) selected from the Validation Phase based on the analysis of <u>overfitting</u> will be used to calculate ŷ

Model 1:

=1.0

Model 2

=2.1

Conclusion

The better model (Model 1 or Model 2) selected from the Validation Phase based on the analysis of overfitting will be used to calculate ŷ

The smaller the MSE, the closer the fit is to the data.

Since Model 1 MSE is smaller than Model 2 MSE

Hence Model 1 is better fit

Enhancement Ideas

Since Model 1 is better fit, we calculate the Regression Expression of Test Set data using Model 1 Regression expression i.e Linear Regression

	Training Set				**Validation Set**				**Test Set**	
	Real Data(Set 1)		**Model 1(Linear Regression)**	**Model 2(Non Linear Regression)**	**Real Data(Set 2)**		**Model 1(Linear Regression)**	**Model 2(Non Linear Regression)**	**Real Data(Set 3)**	**ŷ=a1 + b1 * x or ŷ=a2 + b2 * x^2**
x	x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	x	ŷ=a1 + b1 * x
1	1	1.8	1.38	1.8	1.5	1.7	1.81	1.9625	1.4	1.724
2	2	2.4	2.24	2.19	2.9	2.7	3.014	2.7633	2.5	2.67
3	3.3	2.3	3.358	3.0857	3.7	2.5	3.702	3.4497	3.6	3.616
4	4.3	3.8	4.218	4.0737	4.7	2.8	4.562	4.5417	4.5	4.39
5	5.3	5.3	5.078	5.3217	5.1	5.5	4.906	5.0513	5.4	5.164
6	1.4	1.5	1.724	1.9248						
7	2.5	2.2	2.67	2.4825						
8	2.8	3.8	2.928	2.6892						
9	4.1	4	4.046	3.8553						
10	5.1	5.4	4.906	5.0513						

References

https://hc.labnet.sfbu.edu/~henry/sfbu/course/data_science/algorithm/slide/overfit.html
https://hc.labnet.sfbu.edu/~henry/sfbu/course/data_science/algorithm/slide/linear_regression_example.html#lf
https://hc.labnet.sfbu.edu/~henry/sfbu/course/data_science/algorithm/slide/non_linear_regression_example.html#nl