The process of Machine Learning and using Overfitting to evaluate Linear Regression Model and Non-linear Regression .

- Please compare the following two Regression Models to see which one has more serious overfitting issue.
  - Linear Regression Model 1
  - Non-Linear Regression Model 2
- Suppose we collect a set of sample data and distribute the sample data by
-
- Training phase: 50%
- Validation phase: 25%
- Test phase: 25%

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

Note:

- Real Data Set 1 can be used to determine the formulas for Model 1: Linear Regression and Model 1: Linear Regression. That is, to determine the valuese 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
- Optional: You may want to implement the following 3 programs:
  - Program 1: To implement Linear Regression Model 1

    Note:
    - This program is to use RealData Set 1 to determine a1 and b1 based on Model 1.
    - The program can be used to fill part of the blank spaces in above table.
  - Program 2: Non-Linear Regression Model 2

    Note:
    - This program is to use RealData Set 1 to determine a2 and b2 based on Model 2.
    - The program can be used to fill part of the blank spaces in above table.
  - Program 3: Calculate MSE
- Adding the project to your portofolio
  . Please use Google Slides to document the project
  a. Please link your presentation on GitHub using this structure
  b.
  c. Machine Learning
  d. - Model Selection
  e. + Use Overfitting To Evaluate Different Models

- Submit
  . The URLs of the Google Slides and GitHub web pages related to this project.
  a. A PDF file of your Google Slides

Answer:

Training phase:

Linear regression:

N =10, we have 10 data

x	y	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
ΣX =	31.8		
ΣY =	32.5		
ΣXY =	120.8		
ΣX*X =	121.34		

We have formula Slope(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX2 - (ΣX)2)

So b1 = (10 * 120.8 – 31.8 * 32.5) / (10 * 121.34 – 31.8^2) = 0.86

We have formula Intercept(a) = (ΣY - b(ΣX)) / N

So a1 = (32.5 – 0.86 * 31.8) / 10 = 0.52

Thus the equation is y = 0.52 + 0.86x

Non-Linear regression:

N =10, we have 10 data

x	x	y	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	118.59
4.3	18.49	3.8	70.262	341.88
5.3	28.09	5.3	148.88	789.05
1.4	1.96	1.5	2.94	3.8416
2.5	6.25	2.2	13.75	39.063
2.8	7.84	3.8	29.792	61.466
4.1	16.81	4	67.24	282.58
5.1	26.01	5.4	140.45	676.52
	ΣX =	121.34		
	ΣY =	32.5		
	ΣXY =	509.76		
	ΣX*X =	2330		

We have formula Slope(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX*X - (ΣX)2)

So b2 = (10 * 509.76 – 121.34 * 32.5) / (10 * 2330 – 121.34^2) = 0.13

We have formula Intercept(a) = (ΣY - b(ΣX)) / N

So a2 = (32.5 – 0.13*121.34) / 10 = 1.67

Thus the equation is y = 1.67 + 0.13x^2

After we figure out those two equations, and we know the x value is stationary, then use the equations that I got (linear: y = 0.52 + 0.86x & nonlinear: y = 1.67 + 0.13x^2) to calculate the y value and fill the blank part of Training phase

Training Phase			
Real Data Set 1 50% of the collcted data		**Model 1: Linear Regression**	**Model 2: Non-Linear Regression**
x	**y**	**$\hat{y}$=a1 + b1 * x** y = 0.52 + 0.86x	**$\hat{y}$=a2 + b2 * x^2** y = 1.67 + 0.13x^2
1	**1.8**	**1.4**	**1.8**
2	**2.4**	**2.2**	**2.2**
3.3	**2.3**	**3.4**	**3.1**
4.3	**3.8**	**4.2**	**4.1**
5.3	**5.3**	**5.1**	**5.3**
1.4	**1.5**	**1.7**	**1.9**
2.5	**2.2**	**2.7**	**2.5**
2.8	**3.8**	**2.9**	**2.7**
4.1	**4.0**	**4.0**	**3.9**
5.1	**5.4**	**4.9**	**5.1**

Validation phase:

After we figure out those two equations, and we know the x value is stationary, then use the equations that I got (linear: y = 0.52 + 0.86x & nonlinear: y = 1.67 + 0.13x^2) to calculate the y value and fill the blank part of Validation phase

Validation Phase			
Real Data Set 2 25% of the collcted data		**Model 1: Linear Regression**	**Model 2: Non-Linear Regression**
x	**y**	**$\hat{y}$=a1 + b1 * x** y = 0.52 + 0.86x	**$\hat{y}$=a2 + b2 * x^2** y = 1.67 + 0.13x^2
1.5	**1.7**	**1.8**	**2.0**
2.9	**2.7**	**3.0**	**2.7**
3.7	**2.5**	**3.7**	**3.4**
4.7	**2.8**	**4.6**	**4.5**
5.1	**5.5**	**4.9**	**5.1**

We need calculate the MSE,

Training:

Model1:

((1.4 – 1.8)^2 + (2.2 – 2.4)^2 + (3.4 – 2.3)^2 + (4.2 – 3.8)^2 + (5.1 – 5.3)^2 + (1.7 – 1.5)^2 + (2.7 – 2.2)^2 + (2.9 – 3.8)^2 + (4.0 – 4.0)^2 + (4.9 – 5.4)^2))/ 10 = 0.296

Model2:

((1.8– 1.8)^2 + (2.2 – 2.4)^2 + (3.1 – 2.3)^2 + (4.1 – 3.8)^2 + (5.3 – 5.3)^2 + (1.9 – 1.5)^2 + (2.5 – 2.2)^2 + (2.7 – 3.8)^2 + (3.9 – 4.0)^2 + (5.1 – 5.4)^2))/ 10 = 0.233

Validation:

Model1:

((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 = 1.028

Model2:

((1.7 – 2.0)^2 + (2.7 – 2.8)^2 + (2.5 – 3.4)^2 + (2.8 – 4.5)^2 + (5.5 – 5.1)^2))/5 = 0.792

Use the formula :

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

Model1: 1.028 / 0.296 = 3.47297297297

Model2: 0.792/0.233 = 3.3991416309

Thus, Model2 is smaller, which is better

Test phase:

We use Model2 to calculate the y-values (it depending on the analysis of overfitting) to fill the table

Test Phase	
Real Data Set 3 25% of the collcted data	**The better model selected from Model 1 and Model 2 depending on the analysis of overfitting**
x	**$\hat{y}=a1 + b1 * x$ or $\hat{y}=a2 + b2 * x^2$** y = 1.67 + 0.13x^2
1.4	**1.9**
2.5	**2.5**
3.6	**3.4**
4.5	**4.3**
5.4	**5.5**

Finally, the whole table will be look like this:

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 selected from Model 1 and Model 2 depending on the analysis of overfitting
x	y	$\hat{y}=a1 + b1 * x$ $y = 0.52 + 0.86x$	$\hat{y}=a2 + b2 * x^2$ $y = 1.67 + 0.13x^2$	x	y	$\hat{y}=a1 + b1 * x$ $y = 0.52 + 0.86x$	$\hat{y}=a2 + b2 * x^2$ $y = 1.67 + 0.13x^2$	x	$\hat{y}=a1 + b1 * x$ or $\hat{y}=a2 + b2 * x^2$ $y = 1.67 + 0.13x^2$
1	1.8	1.4	1.8	1.5	1.7	1.8	2.0	1.4	1.9
2	2.4	2.2	2.2	2.9	2.7	3.0	2.7	2.5	2.5
3.3	2.3	3.4	3.1	3.7	2.5	3.7	3.4	3.6	3.4
4.3	3.8	4.2	4.1	4.7	2.8	4.6	4.5	4.5	4.3
5.3	5.3	5.1	5.3	5.1	5.5	4.9	5.1	5.4	5.5
1.4	1.5	1.7	1.9	X	X	X	X	X	X
2.5	2.2	2.7	2.5	X	X	X	X	X	X
2.8	3.8	2.9	2.7	X	X	X	X	X	X
4.1	4.0	4.0	3.9	X	X	X	X	X	X
5.1	5.4	4.9	5.1	X	X	X	X	X	X