Using Overfitting to evaluate different Models

CS550 Homework
Shoumya Singh
ID-19566

1. 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 (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	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 or ŷ=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 ŷ 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".
- 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
    - Please link your presentation on GitHub using this structure
        - Machine Learning
        - - Model Selection
        - + Use Overfitting to Evaluate Different Models

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

Training phase

Model 1: Linear Regression formulas:

```
Regression Equation(y) = a + bx
Slope(b) = (NΣXY - (ΣX) (ΣY)) / (NΣX² - (ΣX)²)
Intercept(a) = (ΣY - b(ΣX)) / 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
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
ΣX² = Sum of square First Scores
```

Training phase

- 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	4.1 * 4 = 16.4	4.1 * 4.1 = 16.81
5.1	5.4	5.1 * 5.4 = 27.54	5.1 * 5.1 = 26.01

- Step 3:

  Find $\Sigma X, \Sigma Y, \Sigma XY, \Sigma X^2$.

$$\Sigma X = 1+2+3.3+4.3+5.3+1.4+2.5+2.8+4.1+5.1$$
$$= 31.8$$
$$\Sigma Y = 1.8+2.4+2.3+3.8+5.3+1.5+2.2+3.8+4+5.4$$
$$= 32.5$$
$$\Sigma XY = 1.8+4.8+7.59+16.34+28.09+2.1+5.5+10.64+16.4+27.54$$
$$= 120.8$$
$$\Sigma X^2 = 1+4+10.89+18.49+28.09+1.96+6.25+7.84+16.81+26.01$$
$$= 121.34$$

- Step 4:

  Substitute in the above slope formula given.

$$\text{Slope}(b) = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$$
$$= ((10)*(120.8)-(31.8)*(32.5))/((10)*(121.34)-(31.8)^2)$$
$$= (1208 - 1033.5)/(1213.4 - 1011.24)$$
$$= 174.5/202.16$$
$$= 0.86$$

- Step 5:

  Now, again substitute in the above intercept formula given.

$$\text{Intercept}(a) = (\Sigma Y - b(\Sigma X)) / N$$
$$= (32.5 - 0.86(31.8))/10$$
$$= (32.5 - 27.348)/10$$
$$= 5.152/10$$
$$= 0.515$$

- Step 6:

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

$$\text{Regression Equation}(y) = a + bx$$
$$= 0.515 + 0.86x$$

- Step 7:

  Now we substitute the real data value of x in the above equation and get the values.

```
x = 1
Regression Equation(y) = a + bx
= 0.515 + 0.86(1)
= 0.515 + 0.86
= 1.375

x = 2
Regression Equation(y) = a + bx
= 0.515 + 0.86(2)
= 2.235

x = 3.3
Regression Equation(y) = a + bx
= 0.515 + 0.86(3.3)
= 3.353

x = 4.3
Regression Equation(y) = a + bx
= 0.515 + 0.86(4.3)
= 4.213

x = 5.3
Regression Equation(y) = a + bx
= 0.515 + 0.86(5.3)
= 5.073

x = 1.4
Regression Equation(y) = a + bx
= 0.515 + 0.86(1.4)
= 1.719

x = 2.5
Regression Equation(y) = a + bx
= 0.515 + 0.86(2.5)
= 2.665

x = 2.8
Regression Equation(y) = a + bx
= 0.515 + 0.86(2.8)
= 2.923

x = 4.1
Regression Equation(y) = a + bx
= 0.515 + 0.86(4.1)
= 4.041

x = 5.1
Regression Equation(y) = a + bx
= 0.515 + 0.86(5.1)
= 4.901
```

Model 2: Non - Linear Regression formulas:

- Linear Regression Formula:
  -
  - $\text{Regression Equation}(y) = a + bx$
  - $\text{Slope}(b) = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$
  - $\text{Intercept}(a) = (\Sigma Y - b(\Sigma X)) / N$

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
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

- **Non-linear Regression Formula:**
  - $\text{Regression Equation}(y) = a + bx^2$

We can still use Linear Regression formula

$\text{Slope}(b) = (N\Sigma PY - (\Sigma P)(\Sigma Y)) / (N\Sigma P^2 - (\Sigma P)^2)$
$\text{Intercept}(a) = (\Sigma Y - b(\Sigma P)) / N$

Where $P = X * X$

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

- Step 1:

  Count the number of values. N = 10

- Step 2:

  Find $\underline{P} * Y$, $\underline{P}^2$

  See the below table

P Value	Y Value	P*Y	P*P
1	1.8	1 * 1.8 = 1.8	1 * 1 = 1
4	2.4	4 * 2.4 = 9.6	4 * 4 = 16
10.89	2.3	10.89 * 2.3 = 25.04	10.89 * 10.89 = 118.59
18.49	3.8	18.49 * 3.8 = 70.26	18.49 * 18.49 = 341.88
28.09	5.3	28.09 * 5.3 = 148.87	28.09 * 28.09 = 789.04
1.96	1.5	1.96 * 1.5 = 2.94	1.96 * 1.96 = 3.84
6.25	2.2	6.25 * 2.2 = 13.75	6.25 * 6.25 = 39.06
7.84	3.8	7.84 * 3.8 = 29.79	7.84 * 7.84 = 61,46
16.81	4	16.81 * 4 = 67.24	16.81 * 16.81 = 282.57
26.01	5.4	26.01 * 5.4 = 140.45	26.01 * 26.01 = 676.52

- Step 3:

  Find $\Sigma\underline{X}$, ΣY, $\Sigma\underline{X}Y$, $\Sigma\underline{X}^2$.

$\Sigma\underline{P}$ = 1+4+10.89+18.49+28.09+1.96+6.25+7.84+16.81+26.01
= 121.34
ΣY = 1.8+2.4+2.3+3.8+5.3+1.5+2.2+3.8+4+5.4
= 32.5
$\Sigma\underline{P}Y$ = 1.8+9.6+25.04+70.26+148.87+2.94+13.75+29.79+67.24+140.45
= 509.74
$\Sigma\underline{P}^2$ = 1+16+118.59+341.88+789.04+3.84+39.06+61.46+282.57+676.52
= 2329.96

- Step 4:

  Substitute in the above slope formula given.

Slope(b) = $(N\Sigma\underline{P}Y - (\Sigma\underline{P})(\Sigma Y)) / (N\Sigma\underline{P}^2 - (\Sigma\underline{P})^2)$
= $((10) * (509.74) - (121.34) * (32.5)) / ((10) * (2329.96) - (121.34)^2)$
= (5097.4 - 3943.55) / (23299.6 - 14723.39)
= 1153.85/8576.21
= 0.134

- Step 5:

  Now, again substitute in the above intercept formula given.

```
Intercept(a) = (ΣY - b(ΣP̄)) / N
= (32.5 - 0.134(121.34))/10
= (32.5 - 16.25)/10
= 16.25/10
= 1.625
```

- Step 6:

  Then substitute these values in regression equation formula

```
Regression Equation(y) = a + bx²
= 1.625 + 0.134x²
```

- Step 7:

  Now we substitute the real data value of x in the above equation and get the values.

```
x = 1
Regression Equation(y) = a + bx²
= 1.625 + 0.134(1)
= 1.625 + 0.134
= 1.759

x = 4
Regression Equation(y) = a + bx²
= 1.625 + 0.134(4)
= 1.625 + 0.536
= 2.161

x = 3.3
Regression Equation(y) = a + bx²
= 1.625 + 0.134(10.89)
= 1.625 + 1.459
= 3.084

x = 4.3
Regression Equation(y) = a + bx²
= 1.625 + 0.134(18.49)
= 1.625 + 2.477
= 4.102

x = 5.3
Regression Equation(y) = a + bx²
= 1.625 + 0.134(28.09)
= 1.625 + 3.764
= 5.389
```

x = 1.4
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(1.96)
= 1.625 + 0.2626
= 1.887

x = 2.5
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(6.25)
= 1.625 + 0.8375
= 2.462

x = 2.8
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(7.84)
= 1.625 + 1.050
= 2.675

x = 4.1
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(16.81)
= 1.625 + 2.252
= 3.877

x = 5.1
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(26.81)
= 1.625 + 3.592
= 5.217

Validation Phase

Model 1: Linear Regression formulas:

- **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.**

a = 0.515 b = 0.86

- Step 1:

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

Regression Equation(y) = a + bx
= 0.515 + 0.86x

- Step 2:

  Now we substitute the real data value from validation phase of x in the above equation and get the values.

```
x = 1.5
Regression Equation(y) = a + bx
= 0.515 + 0.86(1.5)
= 0.515 + 1.29
= 1.805

x = 2.9
Regression Equation(y) = a + bx
= 0.515 + 0.86(2.9)
= 3.009

x = 3.7
Regression Equation(y) = a + bx
= 0.515 + 0.86(3.7)
= 3.697

x = 4.7
Regression Equation(y) = a + bx
= 0.515 + 0.86(4.7)
= 4.557

x = 5.1
Regression Equation(y) = a + bx
= 0.515 + 0.86(5.1)
= 4.901
```

Model 2: Non - Linear Regression formulas:

- **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.**

```
a = 1.625            b = 0.134
```

- Step 1:

  Then substitute these values in regression equation formula

$$\text{Regression Equation}(y) = a + bx^2$$
$$= 1.625 + 0.134x^2$$

- Step 2:

  Now we substitute the real data value of x from validation phase in the above equation and get the values.

```
x = 1.5
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(2.25)
= 1.625 + 0.301
= 1.926

x = 2.9
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(8.41)
= 1.625 + 1.126
= 2.751

x = 3.7
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(13.69)
= 1.625 + 1.834
= 3.459

x = 4.7
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(22.09)
= 1.625 + 2.960
= 4.585

x = 5.1
Regression Equation(y) = a + bx^2
= 1.625 + 0.134(26.01)
= 1.625 + 3.485
= 5.11
```

Test Phase

We need calculate the MSE, Training:

Model1:

((1.375 – 1.8)^2 + (2.235 – 2.4)^2 + (3.353 – 2.3)^2 + (4.213 – 3.8)^2 + (5.073 – 5.3)^2 + (1.719 – 1.5)^2 + (2.665 – 2.2)^2 + (2.923 – 3.8)^2 + (4.041 – 4.0)^2 + (4.901 – 5.4)^2))/ 10 = 0.2822

Model2:

((1.759– 1.8)^2 + (2.161 – 2.4)^2 + (3.084 – 2.3)^2 + (4.102 – 3.8)^2 + (5.389 – 5.3)^2 + (1.887 – 1.5)^2 + (2.462 – 2.2)^2 + (2.675 – 3.8)^2 + (3.877 – 4.0)^2 + (5.217 – 5.4)^2))/ 10 = 0.230

Validation:

Model1:

((1.7 – 1.805)^2 + (2.7 – 3.009)^2 + (2.5 – 3.697)^2 + (2.8 – 4.557)^2 + (5.5 – 4.901)^2)) /5 = 0.997

Model2:

((1.7 – 1.926)^2 + (2.7 – 2.751)^2 + (2.5 – 3.459)^2 + (2.8 – 4.585)^2 + (5.5 – 5.11)^2))/5 = 0.862

Use the formula : MSE = max(Training_Set_MSE, Validation_Set_MSE) / min(Training_Set_MSE, Validation_Set_MSE)

Model1: 0.997 / 0.2822 = 3.532

Model2: 0.862/0.230 = 3.747

Thus, Model 1 is smaller, which is better

So, we select the Model 1 equation $\hat{y}=a1 + b1 * x$

```
Regression Equation(y) = a + bx
= 0.515 + 0.86x

x = 1.4
Regression Equation(y) = a + bx
= 0.515 + 0.86(1.4)
= 0.515 + 1.204
= 1.719

x = 2.5
Regression Equation(y) = a + bx
= 0.515 + 0.86(2.5)
= 2.665

x = 3.6
Regression Equation(y) = a + bx
= 0.515 + 0.86(3.6)
= 3.611

x = 4.5
Regression Equation(y) = a + bx
= 0.515 + 0.86(4.5)
= 4.385

x = 5.4
Regression Equation(y) = a + bx
= 0.515 + 0.86(5.4)
= 5.159
```

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 (Model 1 or Model 2) selected from the Validation Phase based on the analysis of overfitting will be used to calculate ŷ
x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$	x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$		
		a = 0.515 b = 0.86	a = 1.625 b = 0.134			a = 0.515 b = 0.86	a = 1.625 b = 0.134	x	Model 1 is better $\hat{y}=a1 + b1 * x$
1	1.8	1.375	1.759	1.5	1.7	1.805	1.926	1.4	1.719
2	2.4	2.235	2.161	2.9	2.7	3.009	2.751	2.5	2.665
3.3	2.3	3.353	3.084	3.7	2.5	3.697	3.459	3.6	3.611
4.3	3.8	4.213	4.102	4.7	2.8	4.557	4.585	4.5	4.385
5.3	5.3	5.073	5.389	5.1	5.5	4.901	5.11	5.4	5.159
1.4	1.5	1.719	1.887	X	X	X	X	X	X
2.5	2.2	2.665	2.462	X	X	X	X	X	X
2.8	3.8	2.923	2.675	X	X	X	X	X	X
4.1	4	4.041	3.877	X	X	X	X	X	X
5.1	5.4	4.901	5.217	X	X	X	X	X	X