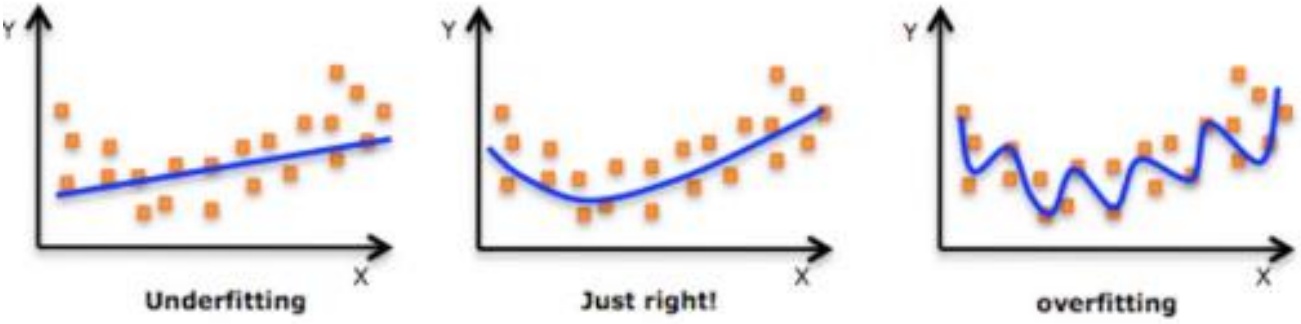
Student Name: Mahmud Omer ID: 19660

Week1_hw2

Q8 ==> The process of Machine Learning and using Overfitting to evaluate Linear Regression Model and Non-linear Regression

8. 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

ANS:



As can be seen from the picture above, models with higher degree has more overfitting issue as its trying to connect to any single data set. This can take more computational time, memory and add complexity to the algorithm.

- 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 $\hat{y}$ 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".
  - Optional: You may want to implement the following 3 programs:
    - Program 1: To implement Linear Regression Model 1
      Note:
      - This program is to use Real Data Set 1 to determine a1 and b1 based on Model 1.
      - The program can be used to fill part of the blank spaces in the 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

ANS:

Linear model Formula:

Regression equation($\hat{y}$) $=a1 + b1 * x$
Intercept(a) $= (\Sigma Y - b(\Sigma X)) / N$
Slope(b) $= (N\Sigma XY - (\Sigma X)*(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

Non-linear model Formula:

Regression Equation(y) $= a + bx^2$

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

Phase 1: Training Phase

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 $\underline{X} * Y$, $\underline{X}^2$

  See the below table.

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

  Step 3:

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

$\Sigma\underline{X}$	ΣY	$\Sigma\underline{X}Y$	$\Sigma\underline{X}^2$
31.8	32.5	120.8	121.34

- Step 4:

  Substitute in the above slope formula given.

  $\text{Slope(b)} = (N\Sigma\underline{X}Y - (\Sigma\underline{X})*(\Sigma Y)) / (N\Sigma\underline{X}^2 - (\Sigma\underline{X})^2)$
  $= ((10) *(120.8)-(31.8)*(32.5))/((10)*(121.34)-(31.8)^2)$
  $= 0.8632$

- Step 5:

  Now, again substitute in the above intercept formula given

  $\text{Intercept(a)} = (\Sigma Y - b(\Sigma\underline{X})) / N$
  $= (32.5- 0.8632(31.8))/10$
  $= 0.5051$

- Step 6:

  Then substitute these values in regression equation formula.

  $\text{Regression Equation(y)} = \underline{a} + \underline{b}x^2$

$= 0.5051 + 0.8632x^2$

Non-Linear Regression Model 2:

Step1 ~ 2:

N = 10
Regression Equation(y) = $a + bx^2$

Slope(b) = $(N\Sigma PY - (\Sigma P)(\Sigma Y)) / (N\Sigma P^2 - (\Sigma P)^2)$

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

Where P = X * X

X	Y	X*Y	X*X = P	P*P	PY
1	1.8	1.8	1	1	1.8
2	2.4	4.8	4	16	9.6
3.3	2.3	7.59	10.89	118.5921	25.047
4.3	3.8	16.34	18.49	341.8801	70.262
5.3	5.3	28.09	28.09	789.0481	148.877
1.4	1.5	2.1	1.96	3.8416	2.94
2.5	2.2	5.5	6.25	39.0625	13.75
2.8	3.8	10.64	7.84	61.4656	29.792
4.1	4	16.4	16.81	282.5761	67.24
5.1	5.4	27.54	26.01	676.5201	140.454

Step 3:

Find ΣX, ΣY, ΣXY, ΣX^2, ΣP, ΣPY, ΣP^2

ΣX	ΣY	ΣXY	ΣX^2	ΣP	ΣPY	ΣP^2
31.8	32.5	120.8	121.34	121.34	509.762	2329.986

- Step 4:

  Substitute in the above slope formula given.

  Slope(b) = $(N\Sigma PY - (\Sigma P) *(\Sigma Y)) / (N\Sigma P^2 - (\Sigma P)^2)$
  $= ((10) *(509.762) -(121.34)*(32.5))/((10)*(2329.986)-(121.34)^2)$
  $= 0.13456$

- Step 5:

  Now, again substitute in the above intercept formula given.

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

= (32.5- 0.13456(121.34))/10
= 1.6172197

- Step 6:

Then substitute these values in regression equation formula

Regression Equation(y) = a + bx^2
= 0.13456+ 1.6172197*x^2

Training Set Result:

x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2
1	**1.8**	1.368272655	1.751782112
2	**2.4**	2.231450336	2.155469346
3.3	**2.3**	3.353581322	3.08260436
4.3	**3.8**	4.216759003	4.105278687
5.3	**5.3**	5.079936684	5.397077836
1.4	**1.5**	1.713543728	1.880962027
2.5	**2.2**	2.663039177	2.458234771
2.8	**3.8**	2.921992481	2.672189005
4.1	**4**	4.044123467	3.879213836
5.1	**5.4**	4.907301148	5.11718802

Phase 2: Validation Phase

x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2
1.5	1.7	1.799861496	1.919985126
2.9	2.7	3.008310249	2.74888958
3.7	2.5	3.698852394	3.459379112
4.7	2.8	4.562030075	4.589703368
5.1	5.5	4.907301148	5.11718802
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X

Choosing the best model based on the root mean square error (MSE) method:

Training Set MSE:

Training MSE for Model 1 = [Σ (ŷ1 -y)^2]/N = 2.822549/10 = 0.2822549

Training MSE for Model 2 = [Σ (ŷ2 -y) ^2]/N = 2.356/10 = 0.2356

Validation set MSE:

Validation Set MSE Model 1 = [Σ (ŷ1 -y)^2]/N = 4.998317/5 = 0.999663

Validation Set MSE Model 2 = [Σ (ŷ2 -y)^2]/N = 4.320775084/5 = 0.864

Then the best model is chosen based on the formula

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

- Compare Model 1 and Model 2
  - Mode1

    0.999663 / 0.2822549 = 3.54

  - Model 2

    0.864/ 0.2356 = 3.66

- Conclusion
  - Model 1 is a better model as it has lower training set and validation MSE ratio.

Test Phase:

Model 1 equation along with its a1 and b1 values were used based on the result obtained in the validation phase

X		ŷ=a1 + b1 * x
1.4		1.713543728
2.5		**2.663039177**
3.6		**3.612534626**
4.5		**4.389394539**
5.4		**5.166254452**