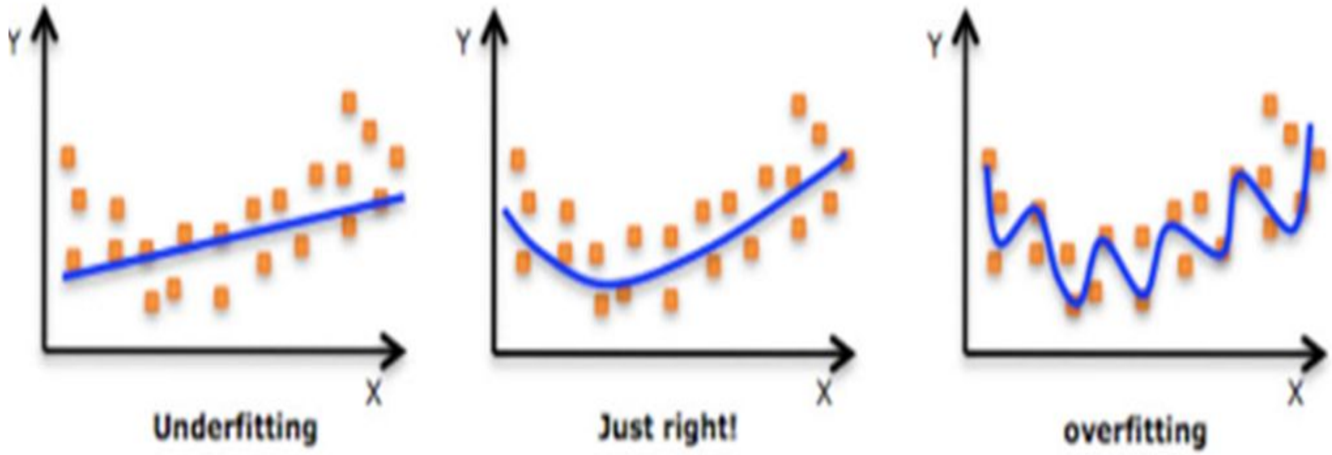
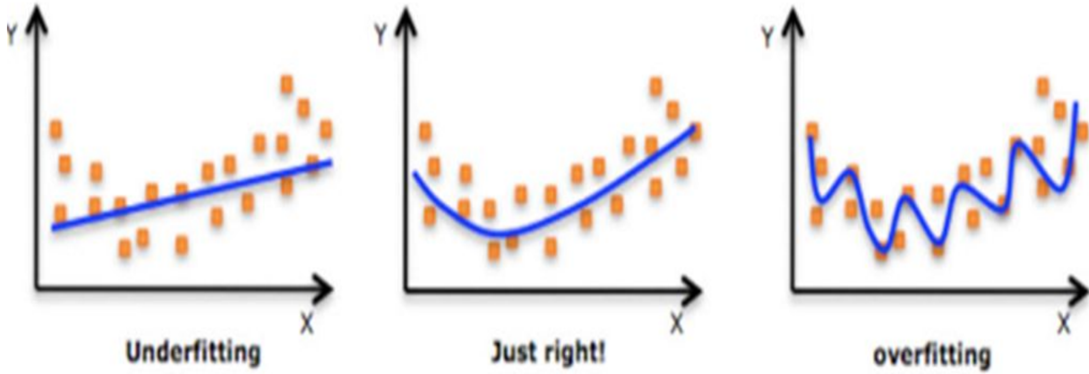
Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression



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.

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

Suppose we collect a set of sample data and <u>distribute</u> 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 $\hat{y}$
		• 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.							
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

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 values of a1, b1, a2, and b2 in the following formulas:

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

ANS:

Linear model Formula:

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

Non-linear model Formula:

Regression Equation(y) = a + bx^2

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

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 1: Training Phase : Linear 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 X * Y, 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

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 1: Training Phase

Step 3: Find ΣX, ΣY, ΣXY, ΣX^2 .

ΣX	ΣY	ΣXY	ΣX^2
31.8	32.5	120.8	121.34

Step 4: Substitute in the above slope formula given.

Slope(b) = (NΣXY - (ΣX)*(ΣY)) / (NΣX^2 - (ΣX)^2)
= ((10) *(120.8)-(31.8)*(32.5))/((10)*(121.34)-(31.8)^2)

= 0.8632

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 1: Training Phase

- Step 5:

  Now, again substitute in the above intercept formula given

  Intercept(a) = (ΣY - b(ΣX)) / N
  = (32.5- 0.8632(31.8))/10
  = 0.5051

- Step 6:

  Then substitute these values in regression equation formula.

  Regression Equation(y) = $a + bx^2$
  = $0.5051 + 0.8632x^2$

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 1: Training Phase : Non-Linear Regression Model 2:

Step 1: N = 10

Step 2: 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 $\underline{P} = X * X$

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 1: Training Phase : Non-Linear Regression Model 2:

Step 2: 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

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

Note: Real Data Set 1 can be used to determine the formulas for <u>Model 1: Linear Regression</u> and <u>Model 1: Linear Regression</u>. That is, to determine the values of a1, b1, a2, and b2 in the following formulas:

ŷ=a1 + b1 * x
ŷ=a2 + b2 * x^2

After the formulas are determined, you can use the formulas to calculate the ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. Test Phase

Note: The values of "x" in "ŷ=a1 + b1 * x" and "ŷ=a2 + b2 * x^2" are the same as the "x" list on the "<u>Real Data Set</u>".

<u>Phase 1: Training Phase : Non-Linear Regression Model 2:</u>

$\Sigma\underline{X}$	ΣY	$\Sigma\underline{X}Y$	$\Sigma\underline{X}^2$	ΣP	ΣPY	ΣP^2
31.8	32.5	120.8	121.34	121.34	509.762	2329.98 6

Step 3: Find $\Sigma\underline{X}$, ΣY, $\Sigma\underline{X}Y$, $\Sigma\underline{X}^2$, ΣP, ΣPY, ΣP^2

Step 4: Substitute in the above slope formula given.

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

= 0.13456

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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:

a. Training Phase
b. Validation Phase
c. 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".

Phase 1: Training Phase : Non-Linear Regression Model 2:

- Step 5:

  Now, again substitute in the above intercept formula given.

  Intercept(a) = (ΣY - b(ΣP)) / 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

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 1: Training Phase : Non-Linear Regression Model 2

Training Set Result:

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

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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:

a. Training Phase
b. Validation Phase
c. 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".

Phase 2: Validation Phase

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

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 2: Validation Phase

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

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 2: Validation Phase

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.***

Using Overfitting
To Evaluate Linear Regression Model and Non-linear Regression

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 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 ŷ values in the following phases:

a. Training Phase
b. Validation Phase
c. 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".

Phase 3: 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