CS550 HOMEWORK 2

USING OVERFITTING TO EVALUATE LINEAR REGRESSION AND NON-LINEAR REGRESSION MODELS

Real data set 1 50% of the collected data		Model 1: Linear Regression	Model 2: Non-linear Regression
x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2
1	1.8		
2	2.4		
3.3	2.3		
4.3	3.8		
5.3	5.3		
1.4	1.5		
2.5	2.2		
2.8	3.8		
4.1	4.0		

Calculating the slope and intercept(a,b) to complete the table

For Linear regression:

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

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

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

Finding the values **ΣX, ΣY, ΣXY, ΣX^2, N**

N = 10(total number of values)
ΣX =31.8
ΣY = 32.5
ΣXY = 120.8
ΣX^2 = 121.34

Applying the formula.
Slope(b1) = (NΣXY - (ΣX)(ΣY)) / (NΣX^2 - $(\Sigma X)^2$) = 10 *(120.8) - (31.8)*(32.5)/10 *(121.34) - (31.8)^2
= 0.86

Intercept(a) = (ΣY - b(ΣX)) / N = (32.5) - 0.86(31.8)/10 = 0.51

For Nonlinear regression:

Slope(b) = (NΣPY - (ΣP)(ΣY)) / (NΣP^2 - $(\Sigma P)^2$)

Intercept(a) = (ΣY - b(ΣP)) / N

P = X * X

Finding the values

X	Y	P = X*X	PY	P*P
1	1.8	1	1.8	1
2	2.4	4	9.6	16
3.3	2.3	10.89	25.04	118.59
4.3	3.8	18.49	70.26	341.88
5.3	5.3	28.09	148.87	789.04
1.4	1.5	1.96	2.94	3.84
2.5	2.2	6.25	13.75	39.06
2.8	3.8	7.84	29.79	61.46
4.1	4.0	16.81	67.24	282.57
5.1	5.4	26.01	140.45	676.52

Finding the sum of the values:

N = 10(total number of values)
ΣP =121.34
ΣY = 32.5
ΣPY = 509.76
ΣP^2 = 2329.98

Applying the formula.
Slope(b2) = (NΣPY - (ΣP)(ΣY)) / (NΣP^2 - (ΣP)^2) = 10 *(509.76) - (121.34)*(32.5)/10 *(2329.98) - (121.94)^2
= 0.13

Intercept(a2) = (ΣY - b(ΣP1)) / N = (32.5) - 0.13(121.94)/10 = 1.66

The next step is substituting the values in the Regression equation formula for the training phase.

Regression equation is y = a + bx

Training Set Result

x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2
1	1.8	1.37	1.79
2	2.4	2.23	2.18
3.3	2.3	3.34	3.07
4.3	3.8	4.20	4.06
5.3	5.3	5.06	5.31
1.4	1.5	1.71	1.91
2.5	2.2	2.66	2.47
2.8	3.8	2.94	2.67
4.1	4.0	4.03	3.79
5.1	5.4	4.89	5.04

Validation Set Phase

x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$
1.5	1.7	1.80	1.95
2.9	2.7	3.0	2.75
3.7	2.5	3.69	3.43
4.7	2.8	4.55	4.53
5.1	5.5	4.89	5.04
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	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 or ŷ=a2 + b2 * x^2
1	1.8	1.37	1.79	1.5	1.7	1.80	1.95	1.4	
2	2.4	2.23	2.18	2.9	2.7	3.0	2.75	2.5	
3.3	2.3	3.34	3.07	3.7	2.5	3.69	3.43	3.6	
4.3	3.8	4.20	4.06	4.7	2.8	4.55	4.53	4.5	
5.3	5.3	5.06	5.31	5.1	5.5	4.89	5.04	5.4	
1.4	1.5	1.71	1.91	X	X	X	X	X	X
2.5	2.2	2.66	2.47	X	X	X	X	X	X
2.8	3.8	2.94	2.67	X	X	X	X	X	X
4.1	4.0	4.03	3.79	X	X	X	X	X	X
5.1	5.4	4.89	5.04	X	X	X	X	X	X

Choosing the best model based on the mean square error method.
Training set

Model 1
MSE =((1.37-1.80)^2 + (2.23 - 2.4)^2 + (3.34 - 2.3)^ 2 + (4.20 - 3.8)^2 + (5.06 - 5.30)^2) + (1.71 - 1.5)^2 + (2.66 - 2.2)^2 + (2.94 - 3.8)^2 + (4.03 - 4.00)^2 + (4.89 - 5.4)^2)/10
=(0.18 + 0.02 + 1.08 + 0.16 + 0.05 + 0.04 + 0.21 + 0.73 + 0.0009 + 0.26)/10
= 0.27

Model 2
MSE = ((1.79- 1.8)^2 + (2.18- 2.4)^2 + (3.07 - 2.3)^2 + (4.06 - 3.8)^2 + (5.31-5.3)^2) + (1.91 -1.5)^2 + (2.47 - 2.2)^2 + (2.67 - 3.8)^2 + (3.79 - 4.0)^2 + (5.04 - 5.4)^2)/10
= (0.0001 + 0.04 + 0.59 + 0.06 + 0.0001 + 0.16 + 0.07 +1.27 + 0.04 + 0.12)/10
= 0.23

Validation set

Model 1
MSE = ((1.7- 1.80)^2 + (2.7 - 3.0)^2 + (2.50 - 3.69)^2 + (2.8 - 4.55)^2 + (5.5 - 4.89)^2)/5
= (0.01 + 0.09 + 1.41 + 3.06 + 0.37)/5
= 0.98

Model 2
MSE = ((1.7 - 1.95)^2 + (2.7 - 2.75)^2 + (2.50 - 3.43)^2 + (2.8 - 4.53)^2 + (5.5 - 5.04)^2
= (0.0625 + 0.0025 + 0.86 + 2.99 + 0.21)/5
= 0.86

Comparing model 1 and model 2

Model 1 = 0.98/0.27
= 3.62

Model 2 = 0.86/0.23
= 3.73

In conclusion Model 1 is better it has a lower training set

Test Phase

x	**$\hat{y}=a1 + b1 * x$**
1.4	**1.71**
2.5	**2.66**
3.6	**3.60**
4.5	**4.38**
5.4	**5.15**