


Model Evaluation Using Overfitting

Farnaz Golnam MCS 19576

Introduction

Regression models are used to model a relationship between the dependant and independent variables. When data shows a **curvy trend** this relationship is **non-linear** otherwise the relationship is **linear**.

One important parameter to choose a model is overfitting. **Overfit regression models** correspond to training data too closely and therefore fail to generalize on test data.



Step1: organizing training and test data set

Training data

50% of the collected data

x	y
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
5.1	5.4

Validation data

25% of the collected data

x	y
1.5	1.7
2.9	2.7
3.7	2.5
4.7	2.8
5.1	5.5
X	X
X	X
X	X
X	X
X	X

Test data

25% of the collected data

x
1.4
2.5
3.6
4.5
5.4
X
X
X
X
X

Step2: Finding the linear equations using the training data

Linear Regression Equation(y) = $a + bx$

$$\text{Slope}(b) = (N\sum XY - (\sum X)(\sum Y)) / (N\sum X^2 - (\sum X)^2)$$

$$\text{Intercept}(a) = (\sum Y - b(\sum X)) / N$$

$$\sum XY = [(1*1.8)+(2*2.4)+(3.3*2.3)+(4.3*3.8)+(5.3*5.3)+(1.4*1.5)+(2.5*2.2)+(2.8*3.8)+(4.1*4)+(5.1*5.4)]=120.8$$

$$\sum X = [1+2+3.3+4.3+5.3+1.4+2.5+2.8+4.1+5.1] = 31.8$$

$$(\sum X)^2 = 1011.24$$

$$\sum X^2 = 121.34$$

$$\sum Y = [1.8+2.4+2.3+3.8+5.3+1.5+2.2+3.8+4+5.4] = 32.5$$

$$N = 10$$

$$b = (1208 - 1033.5) / (1213.4 - 1011.24) = 0.863$$

$$a = (32.5 - 0.863 * 31.8) / 10 = 2.74$$

x	y
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
5.1	5.4

Linear Regression Equation: $y = 2.74 + 0.83 x$

Step3: Finding the linear equations using the training data

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

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

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

$$\text{Where } \underline{P} = X * X$$

$$\Sigma PY = \Sigma (X^2 Y) = 509.762 \quad \Sigma P = 121.34 \quad (\Sigma P)^2 = (121.34)^2 = 14723.39 \quad \Sigma P^2 = 2329.986$$

$$\Sigma Y = [1.8+2.4+2.3+3.8+5.3+1.5+2.2+3.8+4+5.4] = 32.5 \quad N = 10$$

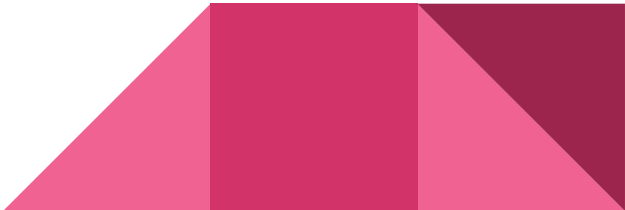
$$b = (5097.62 - 3943.55) / (23299.86 - 14723.39) = 1154.07 / 8576.47 = 0.13$$

$$a = (32.5 - 0.13 * 121.34) / 10 = 1.57$$

Non-Linear Regression Equation: $y = 1.57 + 0.13 x^2$



Training Data			
X	Y	2.74 + 0.83 X (Model1)	1.57 + 0.13 X ² (Model2)
1	1.8	3.57	1.7
2	2.4	4.4	2.09
3.3	2.3	5.47	2.98
4.3	3.8	6.3	3.97
5.3	3.3	7.13	5.22
1.4	1.5	3.9	1.82
2.5	2.2	4.81	2.38
2.8	3.8	5.06	2.58
4.1	4	6.14	3.75
5.1	5.4	6.97	4.95



Validation Data

X	Y	$2.74 + 0.83 X$	$1.57 + 0.13 X^2$
1.5	1.7	3.98	1.86
2.9	2.7	5.14	2.66
3.7	2.5	5.81	3.34
4.7	2.8	6.64	4.44
5.1	5.5	6.97	4.95

Test Data (unknownY, should be predicted)

X	$2.74 + 0.83 X$	$1.57 + 0.13 X^2$
1.4	3.9	1.82
2.5	4.81	2.38
3.6	5.72	3.25
4.5	6.47	4.2
5.4	7.22	5.36

Step4: Calculating MSE for Overfitting

The Mean Squared Error(MSE) is a measure of how close a fitted line is to data point, the smaller the MSE the closer the fit is to the data

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2.$$

In this step the MSE will be calculated for :

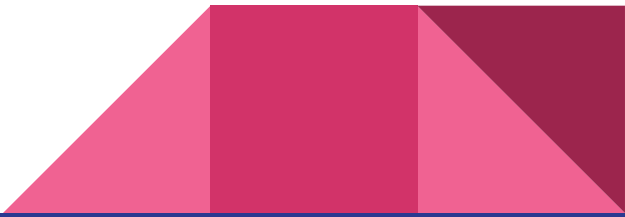
- Training data model 1
- Training data model 2
- Validation data model 1
- Validation data model 2



Training Data

X	Y	$2.74 + 0.83 X$ (Model1)	$(\hat{Y}_i - Y_i)^2$	MSE(Model1)
1	1.8	3.57	3.13	5.93
2	2.4	4.4	4	
3.3	2.3	5.47	10.04	
4.3	3.8	6.3	6.25	
5.3	3.3	7.13	14.66	
1.4	1.5	3.9	5.76	
2.5	2.2	4.81	6.81	
2.8	3.8	5.06	1.58	
4.1	4	6.14	4.57	
5.1	5.4	6.97	2.46	

Training Data				
X	Y	$1.57 + 0.13 X^2$ (Model2)	$(\hat{Y}_i - Y_i)^2$	MSE(Model2)
1	1.8	1.7	0.01	0.61
2	2.4	2.09	0.09	
3.3	2.3	2.98	0.46	
4.3	3.8	3.97	0.02	
5.3	3.3	5.22	3.68	
1.4	1.5	1.82	0.1	
2.5	2.2	2.38	0.03	
2.8	3.8	2.58	1.48	
4.1	4	3.75	0.06	
5.1	5.4	4.95	0.2	



Validation Data

X	Y	$2.74 + 0.83 X$ (model1)	$(\hat{Y}_i - Y_i)^2$	MSE(Model1)
1.5	1.7	3.98	5.19	3.9
2.9	2.7	5.14	5.95	
3.7	2.5	5.81	10.95	
4.7	2.8	6.64	14.74	
5.1	5.5	6.97	2.16	

Validation Data				
X	Y	$1.57 + 0.13 X^2$ (Model2)	$(\hat{Y}_i - Y_i)^2$.	MSE(Model2)
1.5	1.7	1.86	0.02	0.37
2.9	2.7	2.66	0.0	
3.7	2.5	3.34	0.7	
4.7	2.8	4.44	2.68	
5.1	5.5	4.95	0.3	

Step5: Comparing Models

If the **accuracy** over the **training data set increases**, but the **accuracy** over the **validation data set** stays the **same** or **decreases**, then the model has **overfitting** and **training** needs to be **stopped**. For our 2 models, the accuracy increases (MSE decreases) from training to validation data sets, so models do not have overfitting, therefore to choose the best model:

Best model = $\text{MIN}(\text{max}(\text{Training_Set_MSE}, \text{Validation_Set_MSE}) / \text{min}(\text{Training_Set_MSE}, \text{Validation_Set_MSE}))$

	Training_Set_MSE	Validation_Set_MSE	Max/Min
Model 1	5.93	3.9	$5.93 / 3.9 = 1.52$
Model 2	0.61	0.37	$0.61 / 0.37 = 1.64$

So Model 1 is the better model.