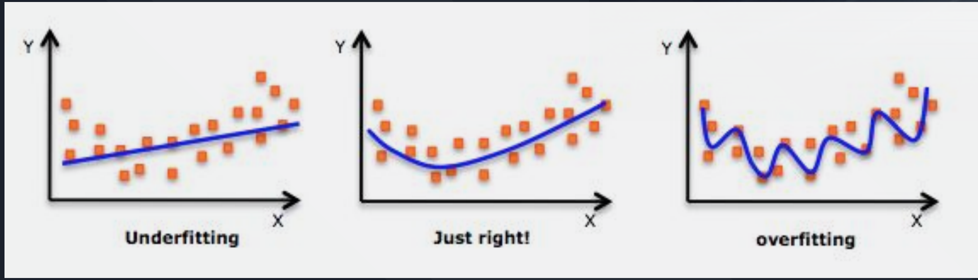
Machine Learning
Model Selection - Use Overfitting To Evaluate Different Models



Shoumya Singh

Table of Content

Introduction-Overfitting to evaluate Linear Regression and Non-Linear Regression Models

- A **regression** is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables.
- **Linear regression** is a **linear** approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).
- **Non-linear regression** is a form of **regression** analysis in which observational data are modeled by a function which is a **nonlinear** combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

Regression Equation(y) = a + bx
Slope(b) = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$
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

Table of Content

Project Description

We have collected a set of sample data and distribute the sample data by

- Training phase: 50%
- Validation phase: 25%
- Test phase: 25%

Evaluate which model is better model - Model 1 or Model 2.

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 $\hat{y}$
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 $\hat{y}=a2+b2*x2$
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			X	X	X	X	X	X
5.1	5.4			X	X	X	X	X	X

Find the values of a1,b1,a2,b2 and $\hat{y}$.

Project Description

1. Initially we will start with Training phase and calculate the value for a1 ,b1,a2 and b2 .
2. Then substitute these values in the linear regression and non-linear regression equations with real data values of x and calculate **ŷ.**

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

3. 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.
4. Only ŷ values are changed with the new Real Data Sets.
5. In the last test phase the better model is selected with MSE

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

6. The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points.

Table of Content

Implementation: Model 1 - Linear Regression

1. Count the number of values. $N = 10$
2. Find $X * Y$, $X * X$
3. Find ΣX, ΣY, ΣXY, ΣX^2
4. Slope(b) = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$
5. Intercept(a) = $(\Sigma Y - b(\Sigma X)) / N$
6. Regression Equation(y) = $a + bx$

Implementation: Model 2 - Non-Linear Regression

1. Count the number of values. N = 10
2. Find X * Y, X * X
3. P = X * X

   Find ΣP, ΣY, ΣPY, ΣP^2
4. Slope(b) = $(N\Sigma PY - (\Sigma P)(\Sigma Y)) / (N\Sigma P^2 - (\Sigma P)^2)$
5. Intercept(a) = $(\Sigma Y - b(\Sigma P)) / N$
6. Regression Equation(y) = a + bx

Table of Content

Training Phase

Model 1: Linear Regression	Model 2: Non - Linear Regression
$N = 10$ $\Sigma X = 1+2+3.3+4.3+5.3+1.4+2.5+2.8+4.1+5.1$ $= 31.8$ $\Sigma Y = 1.8+2.4+2.3+3.8+5.3+1.5+2.2+3.8+4+5.4$ $= 32.5$ $\Sigma XY = 1.8+4.8+7.59+16.34+28.09+2.1+5.5+10.64+16.4+27.54$ $= 120.8$ $\Sigma X^2 = 1+4+10.89+18.49+28.09+1.96+6.25+7.84+16.81+26.01$ $= 121.34$	$N=10$ $\Sigma \underline{P} = 1+4+10.89+18.49+28.09+1.96+6.25+7.84+16.81+26.01$ $= 121.34$ $\Sigma Y = 1.8+2.4+2.3+3.8+5.3+1.5+2.2+3.8+4+5.4$ $= 32.5$ $\Sigma \underline{P}Y = 1.8+9.6+25.04+70.26+148.87+2.94+13.75+29.79+67.24$ $+140.45 = 509.74$ $\Sigma \underline{P}^2 = 1+16+118.59+341.88+789.04+3.84+39.06+61.46+282.57+$ $676.52 = 2329.96$
Slope(b) $= (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$ $= ((10)*(120.8)-(31.8)*(32.5))/((10)*(121.34)-(31.8)^2)$ $= 0.86$ Intercept(a) $= (\Sigma Y - b(\Sigma X)) / N$ $= (32.5 - 0.86(31.8))/10$ $= 0.515$	Slope(b) $= (N\Sigma \underline{P}Y - (\Sigma \underline{P})(\Sigma Y)) / (N\Sigma \underline{P}^2 - (\Sigma \underline{P})^2)$ $= ((10) *(509.74) -(121.34) * (32.5))/ ((10) *(2329.96) - (121.34)^2)$ $= 0.134$ Intercept(a) $= (\Sigma Y - b(\Sigma \underline{P})) / N$ $= (32.5 - 0.134(121.34))/10$ $= 1.625$

Validation Phase

Training Phase				Validation 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
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
1	1.8	1.375	1.759	1.5	1.7	1.805	1.926
2	2.4	2.235	2.161	2.9	2.7	3.009	2.751
3.3	2.3	3.353	3.084	3.7	2.5	3.697	3.459
4.3	3.8	4.213	4.102	4.7	2.8	4.557	4.585
5.3	5.3	5.073	5.389	5.1	5.5	4.901	5.11
1.4	1.5	1.719	1.887	X	X	X	X
2.5	2.2	2.665	2.462	X	X	X	X
2.8	3.8	2.923	2.675	X	X	X	X
4.1	4	4.041	3.877	X	X	X	X
5.1	5.4	4.901	5.217	X	X	X	X

- 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.
- Substituting the values in the equations in both the models we get all the values for ŷ.

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

Test Phase

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

Model 1: 0.997 / 0.2822 = 3.532

Model 2: 0.862/0.230 = 3.747

Model 1 is smaller, which is better

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

Test Phase	
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	Model 1 is better ŷ=a1 + b1 * x
1.4	1.719
2.5	2.665
3.6	3.611
4.5	4.385
5.4	5.159
X	X
X	X
X	X
X	X
X	X

Table of Content

Conclusion

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 $\hat{y}$
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

Bibliography/References

- https://en.wikipedia.org/wiki/Nonlinear_regression
- https://en.wikipedia.org/wiki/linear_regression
- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/linear_regression_example.html
- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/overfit.html