Overfitting Problem Comparison Of Two Regression Models

Ademilton Marcelo da Cruz Nunes (19679)

Introduction

Machine learning supervised algorithms need a lot of labeled data to learn and make predictions. Supervised algorithms learn from given “right answers”. Regression is a type of supervised algorithm in which a model is trained to predict continuous values. The regression can predict things like stock prices, housing prices, and so on.

In order to make a model for predictions, regression algorithms take label data and split them into 3 groups:

- Training set: includes 50% of data
- Validation set: includes 25% of data
- Test set: includes 25% of data

Introduction

The regression model can trace a straight line among the given labeled data to learn and make predictions. This straight line is called **linear regression**. However, sometimes the labeled data can be distributed in a form that a straight line is unable to capture the data relationship, which is called **bias**. Therefore, there is a **non-linear regression**, which is when the relationship of data can be predicted in a non-straight line.

Even though non-linear regression can fit well the labeled data, in the other phase, some data can be far away from the model, causing an **overfitting model**.

An overfitting model is when a model performance on training data is good only in the training phase and causes imprecise prediction in phases like validation and testing. When a model overfits the training data, it is said it has **high variance**. The overfitting problem happens in the linear and non-linear regression models.

Introduction

The goal of this document is to apply a linear and non-linear regression model in a data set in two phases, training and validation phase. Moreover, compare which model is the best in relation to overfitting data using the Mean Squared Error (MSE). In addition, use the best model to predict values in the testing phase.

It is organized in:

- linear regression model determination using the training set,
- applying the linear and non-linear models in the training and validation sets,
- comparison of the best model by using MSE, and
- application of the best model in the test set.

Training set - Linear Model Determination

Training Phase	
X	**Y**
1.00	1.80
2.00	2.40
3.30	2.30
4.30	3.80
5.30	5.30
1.40	1.50
2.50	2.20
2.80	3.80
4.10	4.00
5.10	5.40

The math form for Linear Regression is:

$\hat{y} = a + bx$

Where:
a = the intercept point of the regression line and the y axis.
b = the slope point of the regression line and the y axis.

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

Intercept math formula is = $(\Sigma Y - b(\Sigma X)) / N$

Where:
N = Number of values or elements
b = slope formula
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

Training set - Linear Model - Slope

	Training Phase			
	X	**Y**	**X*Y**	**X*X**
	1.00	1.80	1.80	1.00
	2.00	2.40	4.80	4.00
	3.30	2.30	7.59	10.89
	4.30	3.80	16.34	18.49
	5.30	5.30	28.09	28.09
	1.40	1.50	2.10	1.96
	2.50	2.20	5.50	6.25
	2.80	3.80	10.64	7.84
	4.10	4.00	16.40	16.81
	5.10	5.40	27.54	26.01
Sum total	**31.80**	**32.50**	**120.80**	**121.34**

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

$N = 10$
$\Sigma XY = 120.80$
$\Sigma X = 31.80$
$\Sigma Y = 32.50$
$\Sigma X^2 = 121.34$

Slope= $(10*120.80 - (31.80)*(32.50)) / (10*121.34 - (31.80)^2)$
Slope = (1208 - 1033.5)/ (1213.4-1011,24)
Slope = 174.50 / 202,16
Slope = 0.863

Training set - Linear Model - Intercept

	Training Phase	
	X	**Y**
	1.00	1.80
	2.00	2.40
	3.30	2.30
	4.30	3.80
	5.30	5.30
	1.40	1.50
	2.50	2.20
	2.80	3.80
	4.10	4.00
	5.10	5.40
Sum total	**31.80**	**32.50**

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

N = 10
$\Sigma X = 31.80$
$\Sigma Y = 32.50$
b (slope) = 0.863

Intercept = $(32.50 - (0.863(31.80))) / 10$

Intercept = $(32.50 - (27.45)) / 10$

Intercept = $5.05/10$

Intercept = 0.50

Training set - Linear Model

Therefore, the linear model for the training set is

$$\hat{y} = a + bx$$

$$\mathbf{\hat{y} = 0.50 + (0.863 * x)}$$

Training set - Non-linear Model Determination

Training Phase	
X	**Y**
1.00	1.80
2.00	2.40
3.30	2.30
4.30	3.80
5.30	5.30
1.40	1.50
2.50	2.20
2.80	3.80
4.10	4.00
5.10	5.40

The math form for Linear Regression is:

$\hat{y} = a + bx^2$

Where:
a = the intercept point of the regression line and the y axis.
b = the slope point of the regression line and the y axis.

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

Intercept math formula is = $(\Sigma Y - b(\Sigma\underline{P})) / N$

Where:
N = Number of values or elements
b = slope formula
X = First Score
Y = Second Score
$\underline{P}$ = First Score squared (X*X)
ΣXY = Sum of the product of first and Second Scores
ΣX = Sum of First Scores
ΣY = Sum of Second Scores
$\Sigma\underline{P}$= Sum of First Scores Squared

Training set - Non-linear Model - Slope

	Training Phase				
	X	**Y**	**X*X**	**(X*X)*Y**	**(X*X)*(X*X)**
	1.00	1.80	1.00	1.80	1
	2.00	2.40	4.00	9.60	16
	3.30	2.30	10.89	25.05	118.5921
	4.30	3.80	18.49	70.26	341.8801
	5.30	5.30	28.09	148.88	789.0481
	1.40	1.50	1.96	2.94	3.8416
	2.50	2.20	6.25	13.75	39.0625
	2.80	3.80	7.84	29.79	61.4656
	4.10	4.00	16.81	67.24	282.5761
	5.10	5.40	26.01	140.45	676.5201
Sum total	**31.80**	**32.50**	**121.34**	**509.76**	**2329.9862**

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

$N = 10$
$\Sigma\underline{P} = 121.34$
$\Sigma\underline{P}Y = 509.76$
$\Sigma Y = 32.50$
$\Sigma\underline{P}^2 = 2329.9862$

$\text{Slope} = (10*509.76 - (121.34)*(32.50))/ (10*2329.9862 - (121.34)^2)$
$\text{Slope} = (5097.62 - 3943.55) / (23299.862 - 14723.3956)$
$\text{Slope} = 1154.07 / 8576.4664$
$\text{Slope} = 0.1346$

Training set - Non-linear Model - Intercept

	Training Phase		
	X	**Y**	**X*X**
	1.00	1.80	1.00
	2.00	2.40	4.00
	3.30	2.30	10.89
	4.30	3.80	18.49
	5.30	5.30	28.09
	1.40	1.50	1.96
	2.50	2.20	6.25
	2.80	3.80	7.84
	4.10	4.00	16.81
	5.10	5.40	26.01
Sum total	**31.80**	**32.50**	**121.34**

Intercept = (ΣY - b(Σ$\underline{P}$)) / N

N = 10
Σ$\underline{P}$= 121.34
ΣY = 32.50
B (slope) = 0.1346

Intercept = (32.50 - 0.1346 * 121.34) / 10

Intercept = (32.50 - 16.33) / 10

Intercept = 16.17/10

Intercept = 1.617

Training set - Non-linear Model

Therefore, the non-linear model for the training set is

$$\hat{y} = a + bx^2$$

$$\mathbf{\hat{y} = 1.617 + 0.1346 * x^2}$$

ŷ Determination - Training Set

Training Phase			
X	**Y**	**Linear**	**Non-Linear**
1.00	1.80	1.37	1.75
2.00	2.40	2.23	2.16
3.30	2.30	3.35	3.08
4.30	3.80	4.22	4.10
5.30	5.30	5.08	5.40
1.40	1.50	1.71	1.88
2.50	2.20	2.66	2.46
2.80	3.80	2.92	2.67
4.10	4.00	4.04	3.88
5.10	5.40	4.91	5.12

ŷ Determination - Validation Set

Validation			
x	**y**	**Linear**	**Non-Linear**
1.50	1.70	1.80	1.92
2.90	2.70	3.01	2.75
3.70	2.50	3.70	3.46
4.70	2.80	4.56	4.59
5.10	5.50	4.91	5.12

Mean Squared Error (MSE)

The Mean Squared Error has the formula:

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

Where:

$\hat{y}_i$= function prediction

y_i = target variables

n = Number of values or elements

Verifying MSE - Training Set

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

- **Linear**

MSE = $[(1.37-1.8)^2+(2.23-2.4)^2+(3.35-2.3)^2+(4.22-3.8)^2+(5.08-5.3)^2+(1.71-1.5)^2+(2.66-2.2)^2+(2.92-3.8)^2+(4.04-4)^2+(4.91-5.4)^2]/10$

MSE = [0.19+0.03+1.11+0.17+0.05+0.05+0.21+0.77+0+0.24]/10

MSE = 0.28

- **Non-Linear**

MSE = $[(1.75-1.8)^2+(2.16-2.4)^2+(3.08-2.3)^2+(4.1-3.8)^2+(5.4-5.3)^2+(1.88-1.5)^2+(2.46-2.2)^2+(2.67-3.8)^2+(3.88-4)^2+(5.12-5.4)^2]/10$

MSE = [0+0.06+0.61+0.09+0+0.14+0.07+1.27+0.01+0.08]/10

MSE = 0.24

Verifying MSE - Validation Set

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

- **Linear**

MSE = $[(1.8-1.70)^2+(3.01-2.7)^2+(3.7-2.5)^2+(4.56-2.8)^2+(4.91-5.5)^2]/5$

MSE = [0+0.1+1.43+3.10+0.35]/5

MSE = 0.99

- **Non-Linear**

MSE = $[(1.92-1.70)^2+(2.75-2.7)^2+(3.46-2.5)^2+(4.59-2.8)^2+(5.12-5.5)^2]/5$

MSE = [0.05+0+0.92+3.20+0.15]/5

MSE = 0.86

Comparison Of Two Regression Models

In order to compare MSE in the training set and validation set, we have to get the max value divided by the min value between them:

max(Training_Set_MSE, Validation_Set_MSE) /

min(Training_Set_MSE, Validation_Set_MSE)

Comparison Of Two Regression Models

- **Linear Model**

  MSE Training Set = 0.28

  MSE Validation Set = 0.99

  0.99 / 0.28 = 3.53

- **Non-linear Model**

  MSE Training Set = 0.24

  MSE Validation Set = 0.86

  0.86 / 0.24 = 3.59

Comparison Of Two Regression Models

The linear model won as it will provide a slightly better prediction compared to the non-linear model.

ŷ Determination - Test Set with the Linear Regression

Test	
x	**Linear**
1.40	1.71
2.50	2.66
3.60	3.61
4.50	4.39
5.40	5.17

References

Chang, H. (n.d.). *Use Overfitting To Evaluate Different Models*. labnet. Retrieved January 30, 2023, from https://hc.labnet.sfbu.edu/~henry/sfbu/course/data_science/algorithm/slide/overfit.html

Ng, A. (2022, 12 1). *#12 Machine Learning Specialization [Course 1, Week 1, Lesson 3]*. Youtube. Retrieved January 30, 2023, from https://www.youtube.com/watch?v=peNRqkfukYY&list=PLkDaE6sCZn6FNC6YRfRQc_FbeQrF8BwGI&index=13

Starmer, J. (2018, September 17). *Machine Learning Fundamentals: Bias and Variance*. YouTube. Retrieved January 30, 2023, from https://www.youtube.com/watch?v=EuBBz3bI-aA