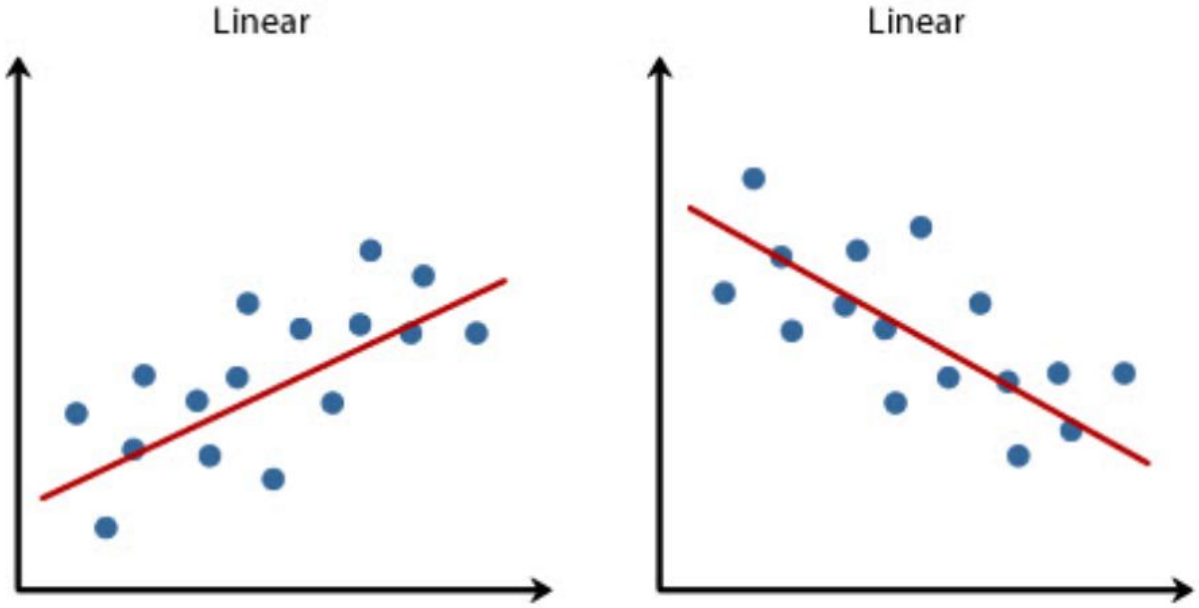
# The Process of Machine Learning

## Using Overfitting to evaluate Linear Regression and Non-Linear Regression Model

Ke (Kathy) She | CS550

# Linear Regression Model



# Linear Regression Model

## To calculate, use the formula:

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
$\Sigma XY$ = Sum of the product of first and Second Scores
$\Sigma X$ = Sum of First Scores
$\Sigma Y$ = Sum of Second Scores
$\Sigma X^2$ = Sum of square First Scores

# Non-Linear Regression Model

# Non-Linear Regression Model

To calculate, use the formula:

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:

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

# The Calculation

| | | | | | | | Linear | Non-Linear |
|---|---|---|---|---|---|---|---|---|
| **X** | **Y** | **XY** | **X^2 (P)** | **N** | **PY** | **P^2** | **y=a1+b1*X** | **y=a2+b2*X^2** |
| 1 | 1.8 | 1.8 | 1 | 10 | 1.80 | 1.00 | 1.37 | 1.75 |
| 2 | 2.4 | 4.8 | 4 | | 9.60 | 16.00 | 2.23 | 2.16 |
| 3.3 | 2.3 | 7.59 | 10.89 | | 25.05 | 118.59 | 3.35 | 3.08 |
| 4.3 | 3.8 | 16.34 | 18.49 | | 70.26 | 341.88 | 4.22 | 4.11 |
| 5.3 | 5.3 | 28.09 | 28.09 | | 148.88 | 789.05 | 5.08 | 5.40 |
| 1.4 | 1.5 | 2.1 | 1.96 | | 2.94 | 3.84 | 1.71 | 1.88 |
| 2.5 | 2.2 | 5.5 | 6.25 | | 13.75 | 39.06 | 2.66 | 2.46 |
| 2.8 | 3.8 | 10.64 | 7.84 | | 29.79 | 61.47 | 2.92 | 2.67 |
| 4.1 | 4 | 16.4 | 16.81 | | 67.24 | 282.58 | 4.04 | 3.88 |
| 5.1 | 5.4 | 27.54 | 26.01 | | 140.45 | 676.52 | 4.91 | 5.12 |
| | | | | | | | | |
| **Sigma X** | **Sigma Y** | **Sigma XY** | **Sigma X^2 (P)** | | **Sigma PY** | **Sigma P^2** | | |
| 31.8 | 32.5 | 120.8 | 121.34 | | 509.76 | 2329.99 | | |
| **(Sigma X)^2** | | **N Sigma XY** | **N Sigma X^2** | | **N Sigma PY** | **N Sigma P^2** | | |
| 1011.24 | | 1208 | 1213.4 | | 5097.62 | 23299.86 | | |
| | | | | | | | | |
| **b1** | **a1** | **b2** | **a2** | | | | | |
| 0.86 | 0.51 | 0.135 | 1.617 | | | | | |

# Validation Phase

| Validation Phase | | | |
|---|---|---|---|
| X | Y | Linear Delta Y | Non-Linear Delta Y |
| 1.5 | 1.7 | 1.80 | 1.92 |
| 2.9 | 2.7 | 3.01 | 2.75 |
| 3.7 | 2.5 | 3.70 | 3.46 |
| 4.7 | 2.8 | 4.56 | 4.59 |
| 5.1 | 5.5 | 4.91 | 5.12 |

# Use Overfitting (MSE) to evaluate different 2 models

| MSE | | | |
|---|---|---|---|
| Training-L | Training-NL | Validation-L | Validation-NL |
| -0.43 | -0.05 | 0.10 | 0.22 |
| -0.17 | -0.24 | 0.31 | 0.05 |
| 1.05 | 0.78 | 1.20 | 0.96 |
| 0.42 | 0.31 | 1.76 | 1.79 |
| -0.22 | 0.10 | -0.59 | -0.38 |
| 0.21 | 0.38 | | |
| 0.46 | 0.26 | | |
| -0.88 | -1.13 | | |
| 0.04 | -0.12 | | |
| -0.49 | -0.28 | | |
| MSE | | | |
| 0.28 | 0.24 | 1.00 | 0.86 |
| Linear | | | |
| **3.54** | **(Slightly better)** | | |
| Non-Linear | | | |
| 3.67 | | | |

# Training Phase

| Test Phase | |
|---|---|
| **X** | **Delta Y** |
| 1.4 | 1.71 |
| 2.5 | 2.66 |
| 3.6 | 3.61 |
| 4.5 | 4.39 |
| 5.4 | 5.17 |

# The Full Calculation

| Training Phase | | | | Validation Phase | | | | Test Phase | |
|---|---|---|---|---|---|---|---|---|---|
| Real Data Set 1 (50%) | | Linear Regression | Non-Linear Regression | Real Data Set 2 (25%) | | Linear Regression | Non-Linear Regression | Real Data Set 3 (25%) | The better Model |
| 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 | $\hat{y}$ |
| 1 | 1.8 | 1.37 | 1.75 | 1.5 | 1.7 | 1.80 | 1.92 | 1.4 | 1.71 |
| 2 | 2.4 | 2.23 | 2.16 | 2.9 | 2.7 | 3.01 | 2.75 | 2.5 | 2.66 |
| 3.3 | 2.3 | 3.35 | 3.08 | 3.7 | 2.5 | 3.70 | 3.46 | 3.6 | 3.61 |
| 4.3 | 3.8 | 4.22 | 4.11 | 4.7 | 2.8 | 4.56 | 4.59 | 4.5 | 4.39 |
| 5.3 | 5.3 | 5.08 | 5.40 | 5.1 | 5.5 | 4.91 | 5.12 | 5.4 | 5.17 |
| 1.4 | 1.5 | 1.71 | 1.88 | x | x | x | x | x | x |
| 2.5 | 2.2 | 2.66 | 2.46 | x | x | x | x | x | x |
| 2.8 | 3.8 | 2.92 | 2.67 | x | x | x | x | x | x |
| 4.1 | 4 | 4.04 | 3.88 | x | x | x | x | x | x |
| 5.1 | 5.4 | 4.91 | 5.12 | x | x | x | x | x | x |

# Conclusion

Linear Regression Model is a slightly better model for this example.