## 2.1 Task 1 : Linear Regression

Linear Regression works on the principle of ordinary least squares. It fits a linear model with coefficients $w = (w_1, w_2, \ldots, w_k)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

It is generally used as LinearRegression().fit(X, y) where X is an array of training data and y represents the array of target values.

It returns a fitted estimator or predictor object.

It calculates the optimal values for the model's coefficients using least squares technique.

This method is an essential step by which a model can learn from data and make predictions on new data.

## 2.2 Task 2 : Gradient Descent

Let's take the case where there is one independent variable 'x' and one dependent variable 'y'.

The linear regression takes slope-intercept form

$$y = w_1 x + b$$

We use the MSE (Mean Squared Error) as our cost function. If we minimize MSE, we can improve the accuracy of our model.

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (y_i - (w_1 x_i + b))^2$$

N is the total number of data points

We use gradient descent to minimize MSE.

We look at the error our current weight gives by computing the derivative of the cost function to find the gradient. The gradient is the slope of the cost function using our current weight.

We change our weight to move in the direction opposite of the gradient.

For example, if our gradient points up the slope, we know that if we move in the opposite direction it will decrease our error.

Let $f(w_1, b) = \frac{1}{N} \sum_{i=1}^{n} (y_i - (w_1 x_i + b))^2$ represent cost function.

Then, gradient is

$$f'(w_1, b) = \begin{bmatrix} \frac{df}{dw_1} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N} \sum -x_i \cdot 2(y_i - (mx_i + b)) \\ \frac{1}{N} \sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

Gradient descent updates the coefficients using the gradient to ensure the model is moving in the direction of steepest descent, or minimum error. Overall, it minimizes the difference between the predicted values and actual values in the training data.

## 2.3 Task 3: Calculating Bias and Variance

### 2.3.2 Task: How bias and variance change as you vary your function classes

Bias: It is typically observed that bias decreases from degree 1 to degree 4. It stays steady and low from degrees 4 to 10 and then rapidly increases from degrees 10 to 15.

Explanation: Because the test and training data is different, when the degree of the polynomial is low (1–4), the models are underfit. They are relatively simple and have high bias. However, as we go from degree 1 to 4, they are less sensitive, and have stable performance across different training and test data sets. Thus, bias decreases.

For degrees 5–10, the models fit the training data well.

As the degree increases beyond 10, the models become very complex and specific to the training data. They are very sensitive to unseen test data. This explains the rapid increase in bias.

Variance: It is typically observed that variance stays low and nearly constant till degree 12 and then rapidly increases.

Explanation: Typically, variance should increase with degree. But, if the size of the training data is not sufficient to represent model complexity, the variance may remain relatively constant. As the degree goes beyond 12, the models may start to capture random noise in training data leading to large ~~function~~ fluctuations. This leads to increased sensitivity and an increase in variance.

## 2.4 Task 4: Calculating Irreducible Error

Irreducible error is the error that cannot be reduced by creating good models. It is a measure of the amount of noise in the data. No matter how good we make our model, our data has a certain amount of noise or irreducible error that cannot be removed.

It is a fundamental limit on the accuracy that can be achieved.

The irreducible error once determined remains fixed and is independent of the model being used. It does not vary with the degree of the polynomial.

## 2.5 Task 5: Plotting Bias$^2$ - Variance graph

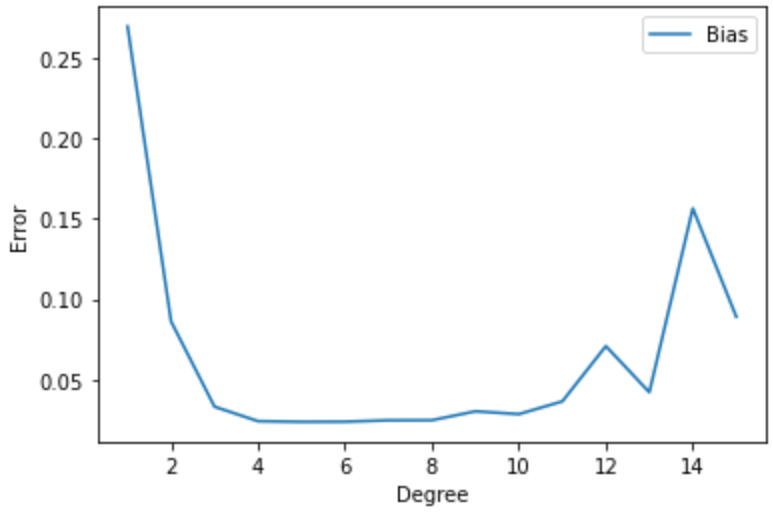
1. Underfitting: Typically characterized by high bias and low variance. In the plot, the underfit models (low complexity) have a high total error, additionally the bias$^2$ curve is high and the variance curve is low. The models are not able to accurately capture the patterns in the data and are too simple.

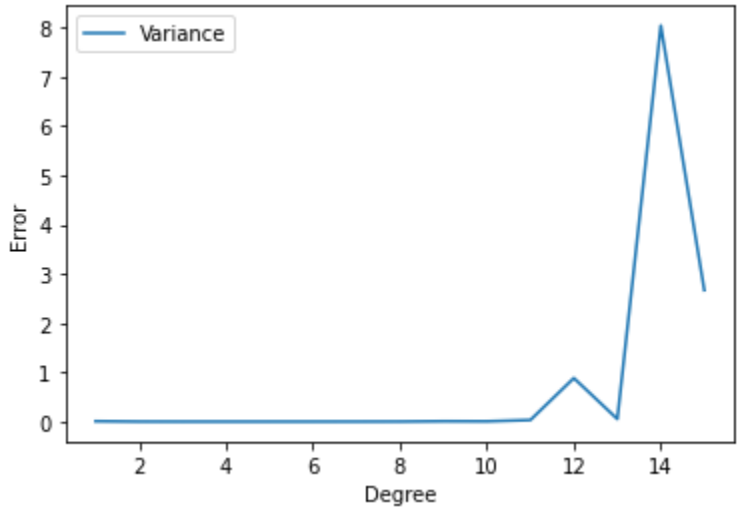
2. Overfitting: Typically characterized by low bias and high variance. In the plot, the overfit models (high complexity) have a high total error (for the last degrees), the bias$^2$ curve is low, and the variance curve is high. The model is too complex and is fitting to the noise in the data.

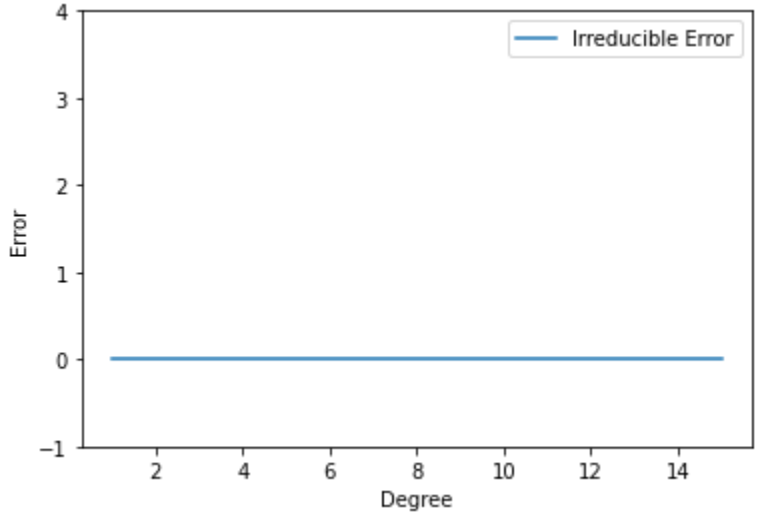
3. Type of data: The data seems to fit well with a model of degree 4-7 since it has both low bias and low variance.

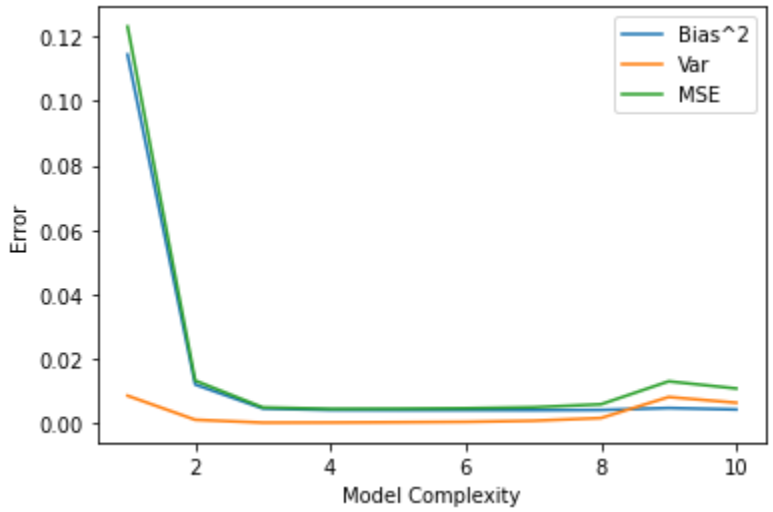
| Degree | Bias | Variance |
|---|---|---|
| 1 | 0.26939817862780674 | 0.00868095161648306 |
| 2 | 0.08625653626329853 | 0.0012243578446454934 |
| 3 | 0.03327182734458062 | 0.00033733942997528067 |
| 4 | 0.024282630552664566 | 0.0003669994779255661 |
| 5 | 0.023879318687144505 | 0.0004619378654759538 |
| 6 | 0.02395536905119038 | 0.0005815200868688666 |
| 7 | 0.024830004382677826 | 0.0009167961009386626 |
| 8 | 0.02488738938572835 | 0.001760919326943077 |
| 9 | 0.030418442333611097 | 0.00827682989084708 |
| 10 | 0.028663868273116302 | 0.006500850638034614 |
| 11 | 0.036590316991408585 | 0.032692754258296854 |
| 12 | 0.07091716133479918 | 0.8844942855270055 |
| 13 | 0.04224905208082726 | 0.05431696550337798 |
| 14 | 0.15642824853292725 | 8.043341910091595 |
| 15 | 0.08912936275561478 | 2.6695050294802156 |

| Degree | Irreducible Error |
|---|---|
| 1 | -2.949029909160572e-18 |
| 2 | -2.654126918244515e-18 |
| 3 | 7.135134497127105e-18 |
| 4 | 1.695692197767329e-18 |
| 5 | -3.527993869267831e-18 |
| 6 | 2.2594773274597912e-18 |
| 7 | -3.2786273695961652e-18 |
| 8 | 3.677613769070831e-18 |
| 9 | -4.163336342344337e-18 |
| 10 | -8.326672684688674e-19 |
| 11 | -5.551115123125783e-18 |
| 12 | -1.9539925233402755e-16 |
| 13 | -1.5543122344752193e-17 |
| 14 | -1.1368683772161603e-15 |
| 15 | 2.842170943040401e-16 |


Bias
0.25
0.20
0.15
0.10
0.05
Error
2
4
6
8
10
12
14
Degree


Variance
8
7
6
5
4
3
2
1
0
Error
2
4
6
8
10
12
14
Degree


Irreducible Error
4
3
2
1
0
−1
Error
2
4
6
8
10
12
14
Degree


Bias^2
Var
MSE
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Error
2
4
6
8
10
Model Complexity