COURSE NAME: ARTIFICIAL INTELLIGENCE
COURSE CODE: CIS 412

SYED TANGIM PASHA

LECTURER,

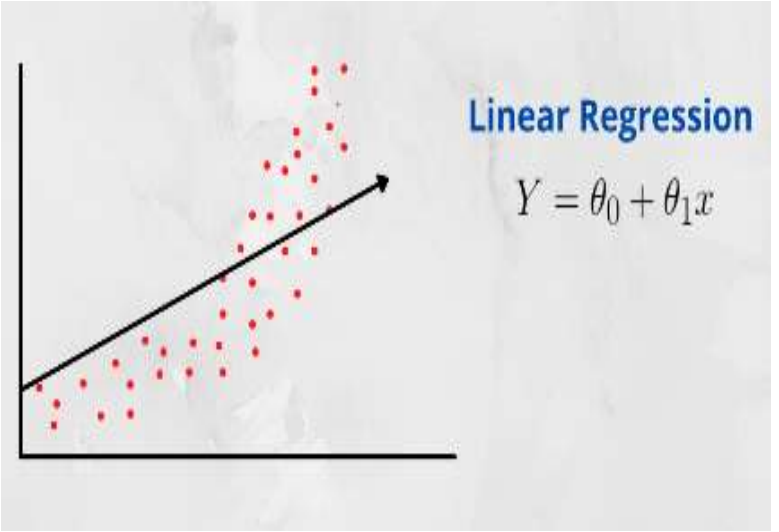
DEPARTMENT OF COMPUTING AND INFORMATION SYSTEM (CIS)

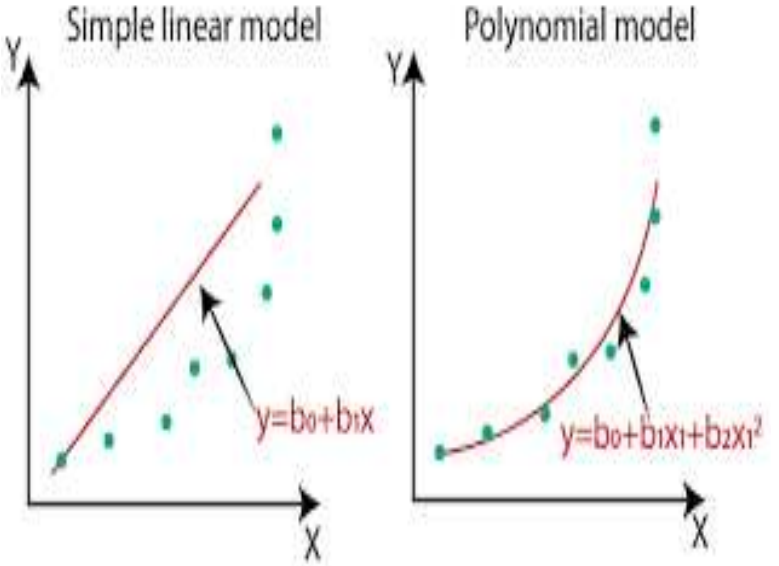
DAFFODIL INTERNATIONAL UNIVERSITY (DIU)

DHAKA,BANGLADESH

POLYNOMIAL REGRESSION

- **Polynomial Regression:** In Linear Regression, algorithm only works when the relationship between the data is linear but suppose if we have non-linear data then Linear Regression will not capable to draw a best-fit line and it fails in such conditions.
- Polynomial Regression is a regression algorithm that models the relationship between a dependent variable (Y) and independent variable (X) as nth degree polynomial.
- $y = a_0 + a_1x_1 + a_2x_1^2 + \ldots + a_nx_1^n$





POLYNOMIAL REGRESSION

Simple Linear Regression

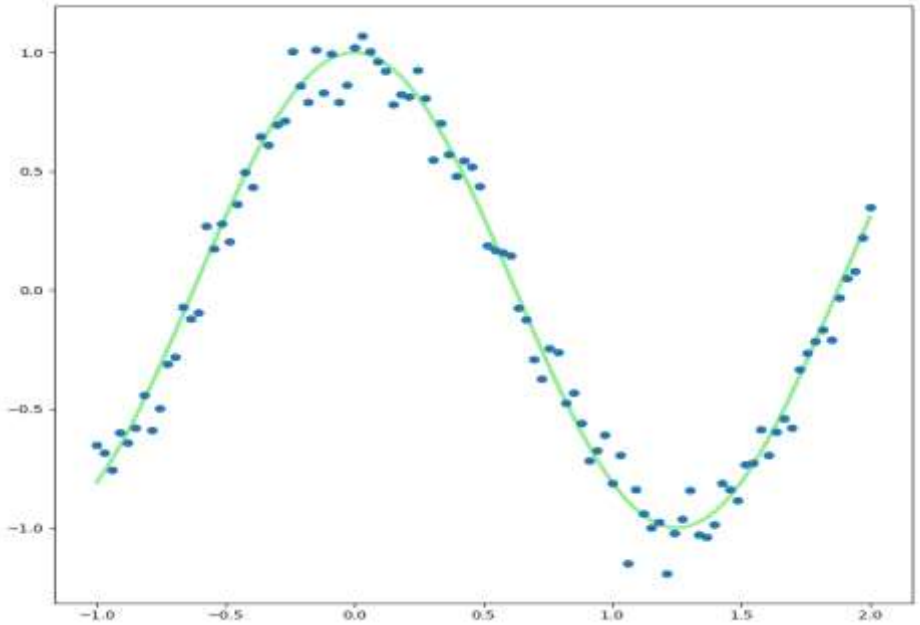
$$y = b_0 + b_1 x_1$$

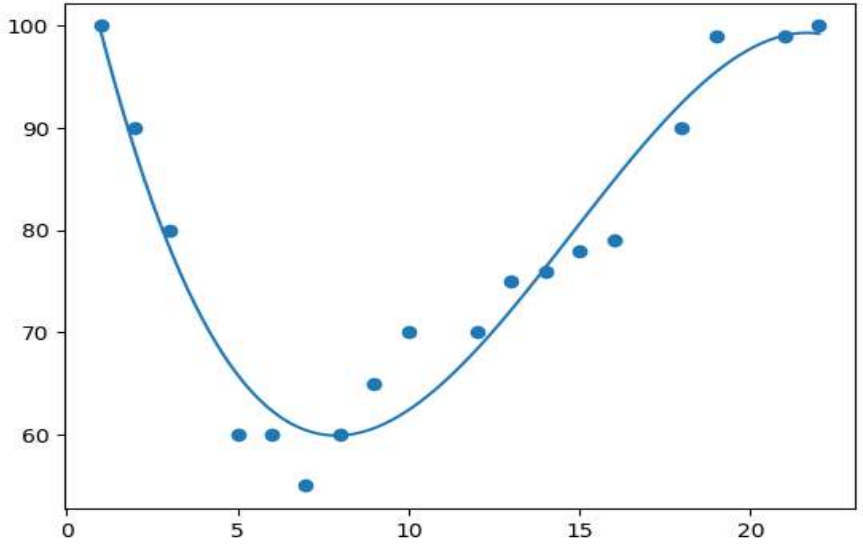
Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

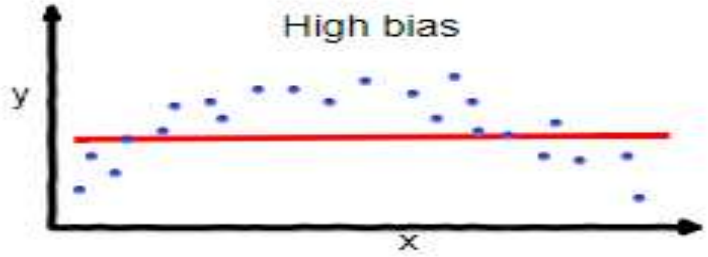




POLYNOMIAL REGRESSION CURVES

UNDERFITTING

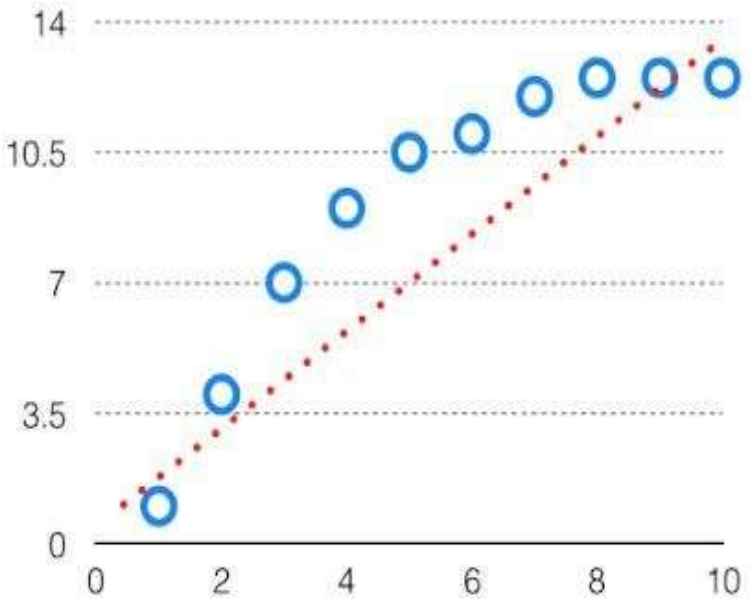
- **Underfitting:** Underfitting is a scenario in Data Science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data.
- Underfitting=high bias, low variance



underfitting

BIAS

- **Bias:** The bias is known as the difference between the prediction of the values by the ML model and the correct value. Bias gives a large error in training as well as testing data.
- By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set. Such fitting is known as underfitting of data. This happens when the hypothesis is too simple or linear in nature.
- Bias=error of the training data/Training error will be high

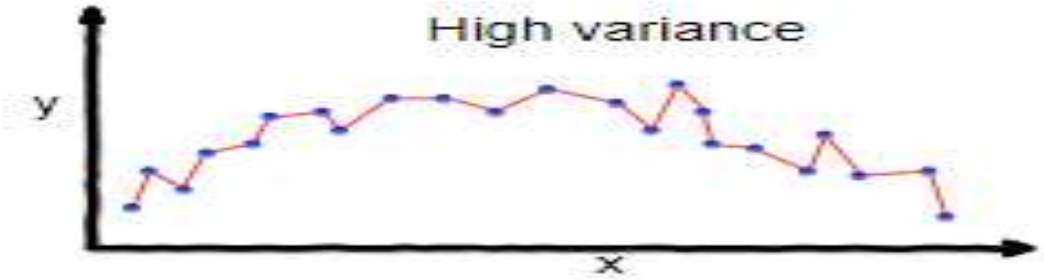


Reducing Underfitting

- Increase model complexity
- Increase the number of features
- Removing noise from the data
- Increase the number of epochs or increase the duration of training to get better results.

OVERFITTING

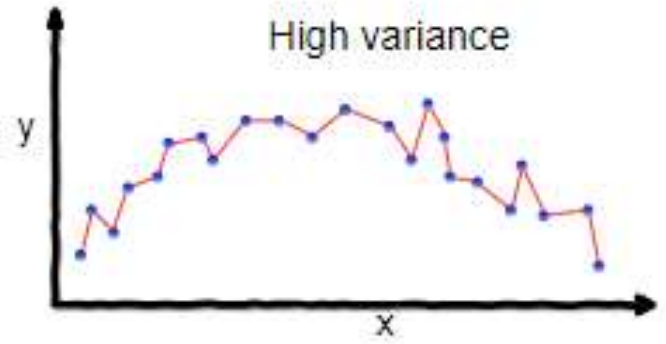
- **Overfitting:** if our model does much better on the training set than on the test set, then we're likely overfitting. For example, it would be a big red flat if our model has 99% accuracy on the training set but 55% accuracy on the test set.
- Overfitting=low bias, high variance



overfitting

VARIANCE

- **Variance:** The variability of model prediction for a given data point which tells us spread of our data is called the variance of the model. When a model is high on variance, it is then said to as overfitting of data. Overfitting is fitting the training set accurately via complex curve and high order hypothesis, but not the good solution for unseen data because high error on unseen data.
- Variance=error of the test data/test error high



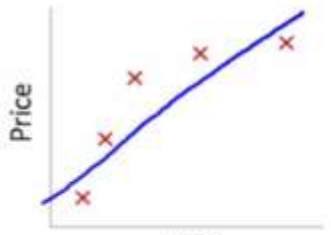
overfitting

Reducing Overfitting

- Increase training data.
- Reduce model complexity
- Regularization technique should apply

OVERFITTING FOR DEGREE OF POLYNOMIAL

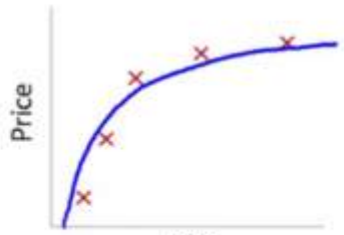
Bias/variance



$\theta_0 + \theta_1 x$

High bias
(underfit)

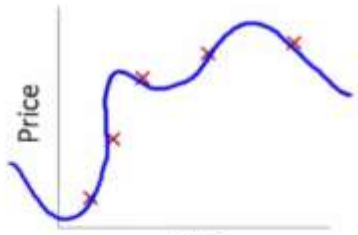
$d = 1$



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

$d = 2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

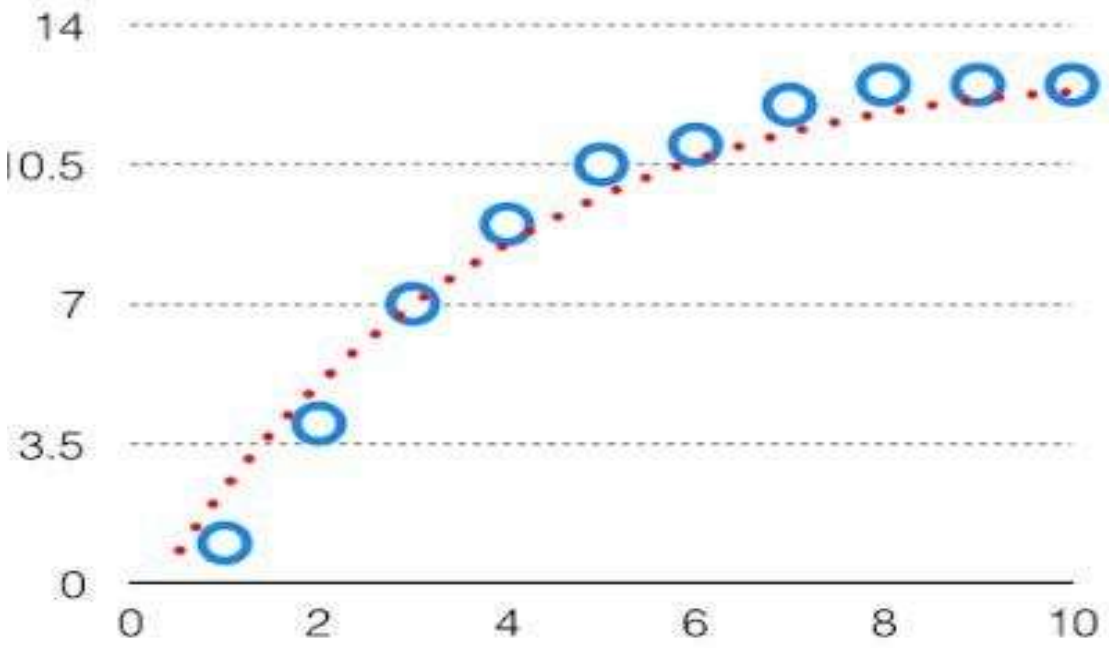
High variance
(overfit)

$d = 4$

Andrew Ng

BIAS-VARIANCE TRADE OFF

- **Bias-Variance Tradeoff:** if the algorithm is too simple(hypothesis linear) then it may be on high bias and low variance condition which is called underfitting. If the algorithm fit too complex(hypothesis with high degree) then it may be on high variance and low bias.
- So, there is a bias-variance tradeoff between bias and variance.
- So our target will be, make a model which is not more complex and not so less complex at the same time.



CROSS VALIDATION (CV)

- **Cross Validation (CV):** Cross-Validation is a technique used to assess how well our Machine Learning models perform on unseen data.
- Suppose you build a machine learning model to solve a problem, and you have trained the model on a given dataset. When you check the accuracy of the model on the training data, it is close to 95%. But that does not mean that this is the best model because of high accuracy!
- Because your model is trained on the given data, it knows the data well and has generalized very well over the given data. If you expose the model to completely new unseen data, it might not predict with the same accuracy and it might fail to generalize over the new data. This problem is called **overfitting**.

- **Cross-Validation** is a resampling technique with the fundamental idea of splitting the dataset into 2parts-training data and test data. Train data is used to train the model and the unseen test data is used for prediction. If the model performs well over the test data and gives good accuracy, it means the model hasn't over fitted the training data and can be used for prediction.

CROSS VALIDATION (CV)

- **K-Fold Cross-Validation:** In this resampling technique, the whole data is divided into k sets of almost equal sizes. The first set is selected as the test set and the model is trained on the remaining (k-1) sets. The test error rate is then calculated after fitting the model to the test data.

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$

CROSS VALIDATION (CV)

D1	D2	D3	D4	D5
20	20	20	20	20

Iteration (1 to K)	Training Set	Test Set	Performance Score
1	D2, D3, D4, D5	D1	S1
2	D1, D3, D4, D5	D2	S2
3	D1, D2, D4, D5	D3	S3
4	D1, D2, D3, D5	D4	S4
5	D1, D2, D3, D4	D5	S5

এখন তাহলে ফাইনাল মডেল ইভ্যালুয়েশন স্কোর হচ্ছে = $\frac{1}{k}\sum_{i=1}^{k} S_i$

অর্থাৎ, আমাদের এই উদাহরণের ক্ষেত্রে = $\frac{S1+S2+S3+S4+S5}{5}$