Machine, Data and Learning

Report

ASSIGNMENT - 1

Done by:

1) PUSHKAR TALWALKAR
2) BHAVESH SHUKLA

Algorithm Implemented

- Q1:
    1. Load the data from binary file given
    2. Split the data into training and testing data.
    3. Split the training data into 10 random and equal sized training sets
    4. For every polynomial degree(1 to 9):
       For every 1/10 training set:
        a) Convert the 1/10 training set into apt polynomial training set (reshaping the array).
        b) Fit a linear regression on the polynomial training set.
        c) Predict the values using the model just trained.
    5. Calculate the bias using the formula given.
    6. Calculate the variance using the formula given.

- Q2:
    1. Load the data from given binary file.
    2. Extract the training data from the data loaded.
    3. For each degree polynomial(1 to 9):
       For every training set (size 1/20 of all training samples)
        a) Convert the training set into apt polynomial training set
        b) Fit a linear regression on the polynomial training set.
        c) Predict values using the above trained model.
    4. Calculate the bias using the given formula
    5. Calculate the variance using the given formula
    6. Plot the graph of bias and variance against complexity of the model.

Results

For Question 1, we got the following results.

TABULATING THE VALUES OF BIAS^2, VARIANCE AND THEIR SUM:

Degree	*Bias^2*	*Variance*	*Sum = Bias^2 + Variance (3 decimal places,round)*
1	30.393825444417008	0.11781878190889924	30.511
2	6.326702434299205	0.038388254659950646	6.365
3	5.455872797039968	0.04135933973745604	5.497
4	3.3452673364819776	0.022783174560308362	3.368
5	3.1293300862005315	0.02667508665542843	3.156
6	2.634928336657051	0.02804777395901377	2.6629
7	2.533163882579966	0.037177013789658785	2.570
8	2.4758797631124936	0.044100762795187584	2.520
9	2.4513072740460897	0.04815798122721843	2.499

For Question 1, the X-axis is : x values

i.e. for all (a,b) in data, X axis plots all the a's.

, the Y-axis is : y values

I.e for all (a,b) in data, X axis plots all the b's.

Here,

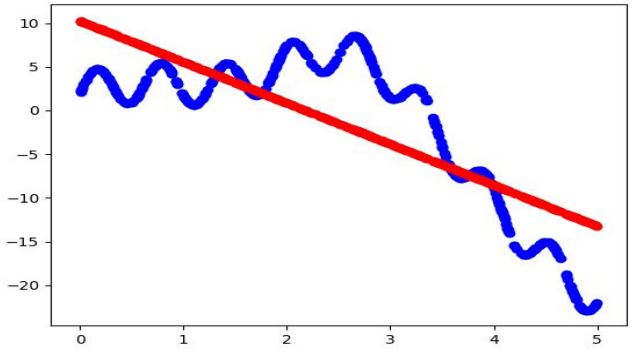
The blue line is the equation of all the points in the data set.

The red line is the polynomial equation plotted (linear, quadratic, cubic,...power of 9)

PLOTS OF EQUATIONS:

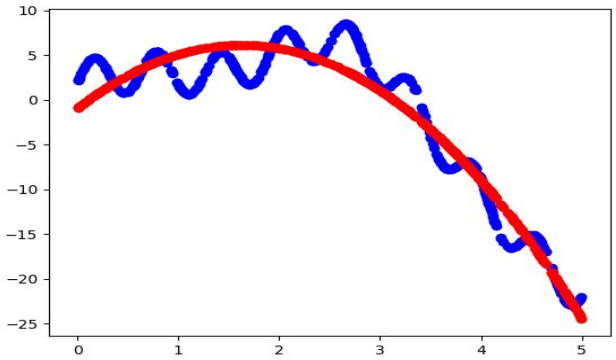
Equation : y=mx+c

Degree : 1



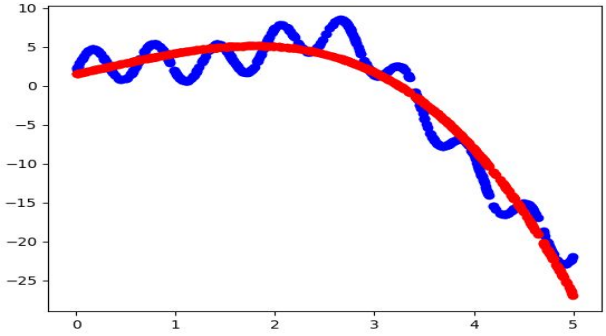
Equation : $y = ax^2 + bx + c$

Degree : 2

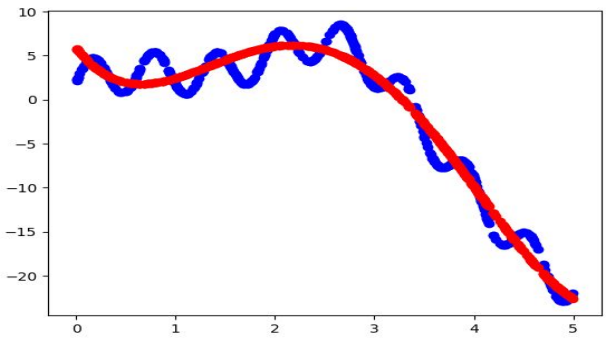


Equation : $y = ax^3 + bx^2 + cx + d$

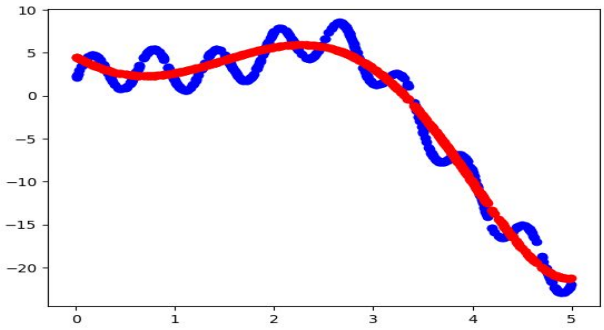
Degree : 3



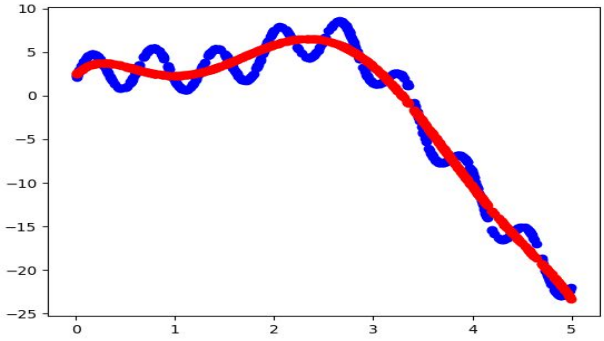
Equation : $y = ax^4 + bx^3 + cx^2 + dx + e$
Degree : 4



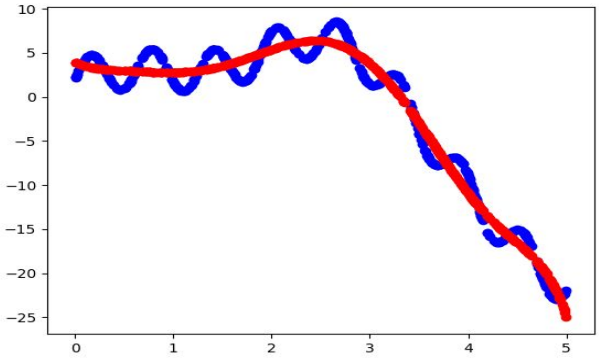
Equation : $y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$
Degree : 5



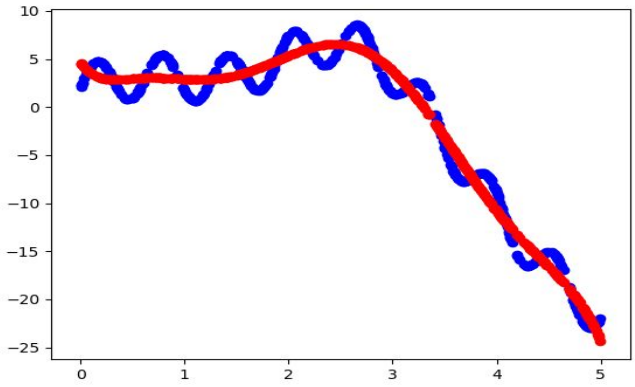
Equation : $y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$
Degree : 6



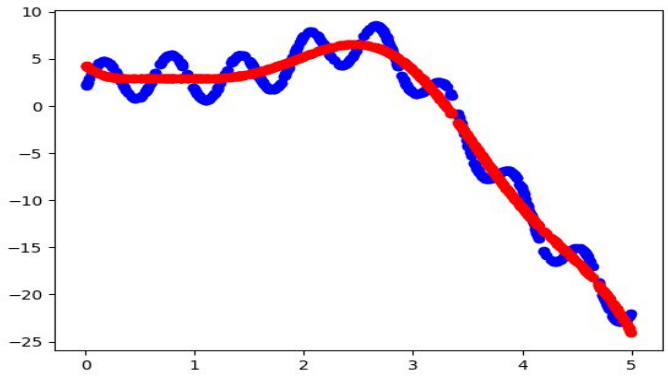
Equation : $y = ax^7 + bx^6 + cx^5 + dx^4 + ex^3 + fx^2 + gx + h$
Degree : 7



Equation : $y = ax^8 + bx^7 + cx^6 + dx^5 + ex^4 + fx^3 + gx^2 + hx + i$
Degree : 8



Equation : $y = ax^9 + bx^8 + cx^7 + dx^6 + ex^5 + fx^4 + gx^3 + hx^2 + ix + j$
Degree : 9



For Question 2, we got the following results.

For Question 2, the X-axis is : The polynomial degree

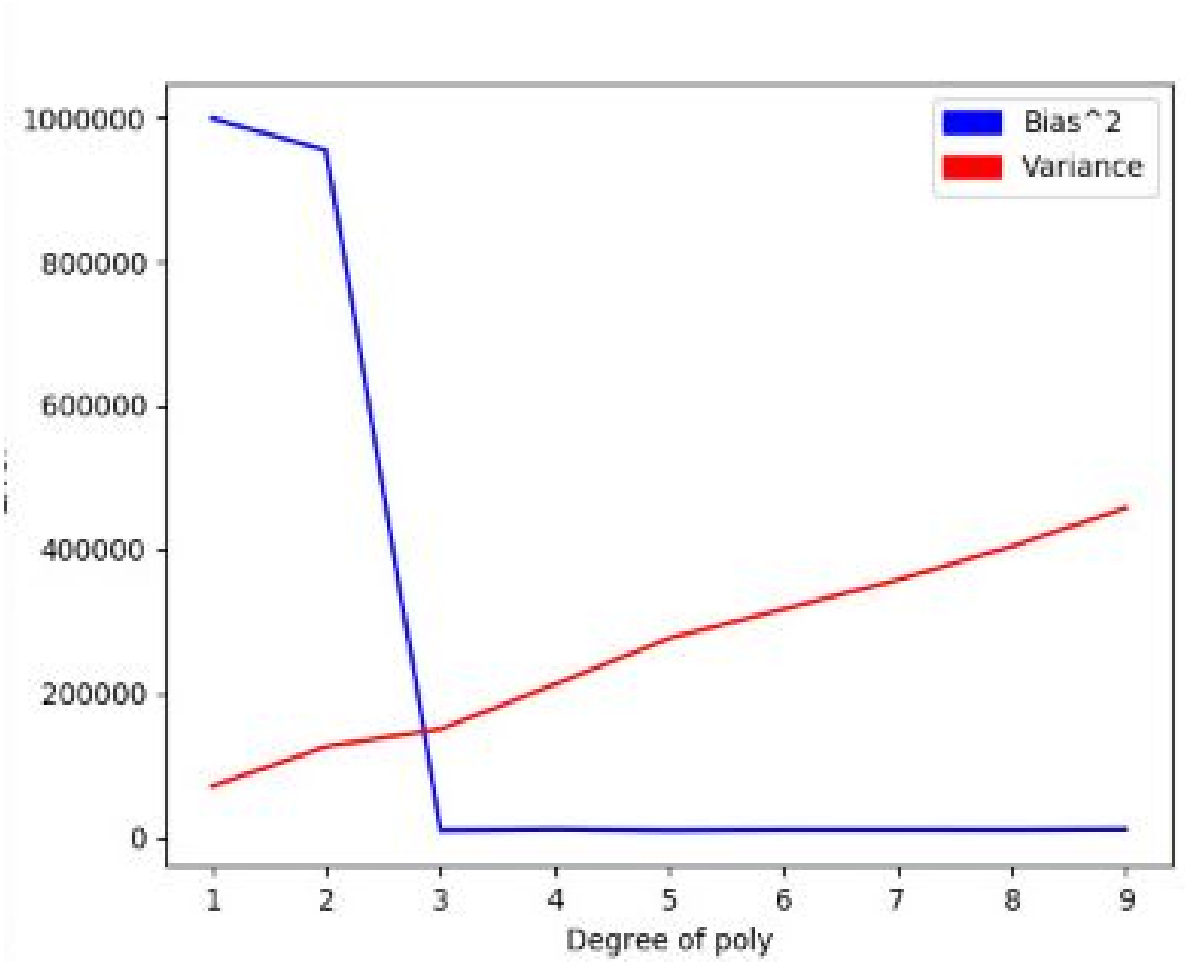
, the Y-axis is : Variance or the Bias^2 values

It depends on the color as shown in graph,

Red : Variance vs Degree of polynomial

Blue: Bias^2 vs Degree of polynomial

PLOT OF GRAPH:



TABULATING THE VALUES OF BIAS^2, VARIANCE AND THEIR SUM:

Degree	*Bias^2*	*Variance*	*Sum = Bias2 + Variance (3 decimal places,rounded)*
1	999228.3968719237	70545.48914575046	1069773.886
2	954619.273794425	125870.85554877335	1080490.129
3	9389.730116791214	150073.7395464768	159463.470
4	10907.34813407133	212235.70832526154	223143.056
5	9339.194291326017	276388.4802547406	285727.675
6	10248.585941147872	316863.49843748985	327112.084
7	10335.2758616491	357510.98475735466	367846.261
8	10149.419243937262	404286.670685786	414436.090
9	10815.487036574234	459132.37837248633	469947.865

We know that,

$$E[(y - f'(x))^2] = (Bias(f'(x)))^2 + Variance(f'(x)) + \sigma^2$$

where E[X] : Expectation of random variable X

y : The polynomial to be estimated. y= f(x) + ε

f'(x) : Polynomial which estimates y = f(x) + ε

σ^2 : The variance of the function y = f(x) + ε

The mean of y = f(x) + ε is 0.

$E[(y - f'(x))^2]$ is the mean squared error of f'(x) with respect to y.

In this model the σ^2 is fixed for the given data distribution.

Thus, the mean squared error of the model over degrees of polynomial varies only by the sum, $(\text{Bias}(f'(x)))^2 + \text{Variance}(f'(x))$

From the tabulated values it is clear that the optimum polynomial function is the cubic one, $\mathbf{y = ax^3 + bx^2 + cx + d}$, the coefficients are determined by the program.

The linear and quadratic functions are UNDERFITTING the data given.
The cubic function is BEST FIT for the data given.
The polynomial functions of degree 4, 5, 6, 7, 8, 9 are all OVERFITTING the data.

It is characterised perfectly by:
OVERFITTING : HIGH VARIANCE, LOW BIAS.
UNDERFITTING : LOW VARIANCE, HIGH BIAS

Observations on the bias and variance:

QUESTION 1:

- Bias^2 is steadily decreasing with an increase in the degree of the polynomial modelling the data.
- Variance is steadily increasing with an increase in the degree of the polynomial modelling the data. Except for linear to quadratic polynomial where it decreases. This is due to the randomness of the data.
- The model with the least sum of the two is the polynomial of degree 9, from all the polynomials to be considered. (1..9) However we cannot conclude that the best fit is the polynomial of degree 9 as the polynomials of higher degrees may have lower sums.

QUESTION 2:

- $Bias^2$ is steadily decreasing with an increase in the degree of the polynomial modelling the data until the cubic polynomial. From degree 4 and up, there are exceptions, but a trend is that the $Bias^2$ is steadily increasing.
- Variance is steadily increasing with an increase in the degree of the polynomial modelling the data. No exceptions.
- The model with the least sum of the two is the polynomial of degree 3, from all the polynomials to be considered. Models with lesser degree generate models which underfit the data. The models with greater degree generate models which overfit the data. Degree 3 is the best fit.