POORNIMA BYREGOWDA
8/31/16

HOMEWORK1

1) Linear regression.

a. Model Y1 as a linear function of X1. Use linear regression to learn the model parameters.

$Y_1 = \theta_1 + \theta_2 * X$

b. Predict output for X1 = 4.10 and X1 = 6.5.

The theta values are

theta1(θ1) = 1.124344

theta2(θ2) = 1.978690

The prediction values y1 for x1 4.100000 is 9.236975

The prediction values y1 for x1 6.500000 is 13.985832

c. Repeat gradient descent learning for α = 0.01, 0.1, 1.0, and 100. Plot J for the learning duration.

- **For alpha value 0.01**

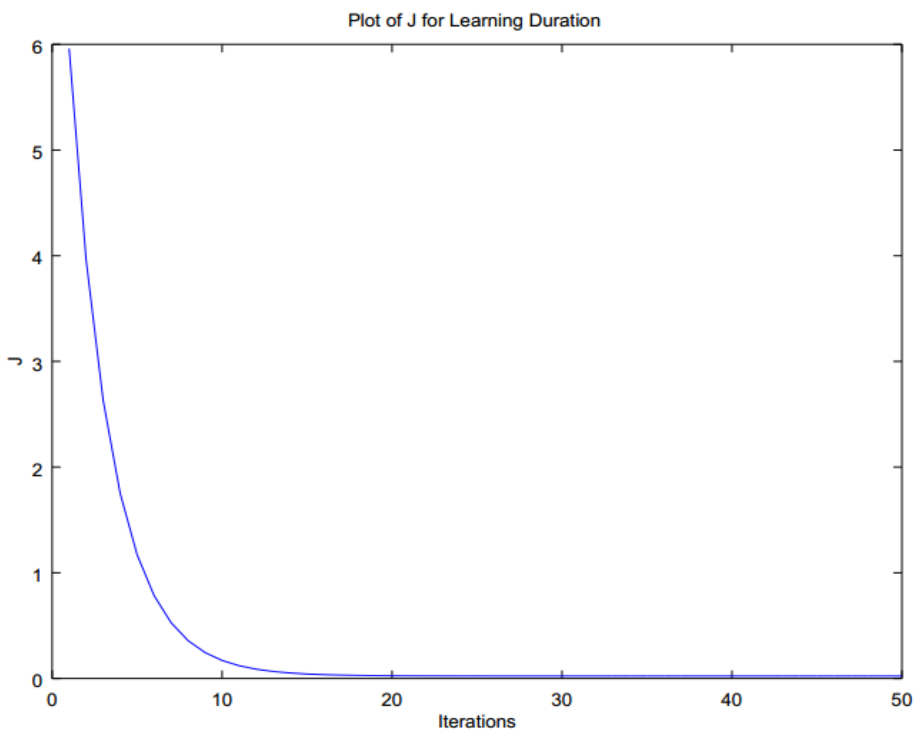

Fig 1: Plot of the “J” learning values for alpha 0.01

HOMEWORK1

- **For alpha value 0.1**

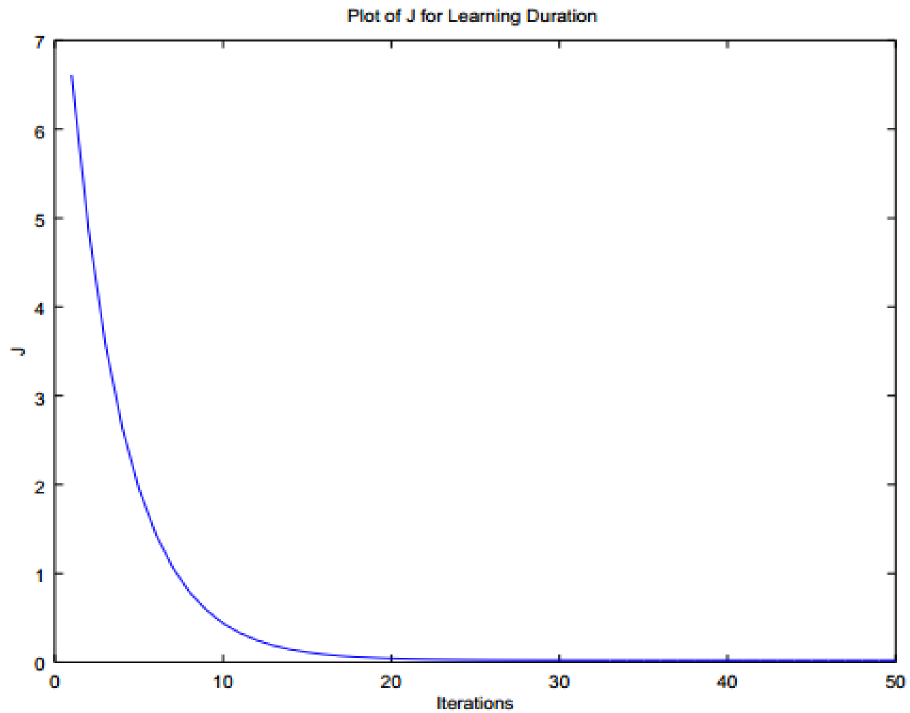

Fig 1: Plot of the “J” learning values for alpha 0.1

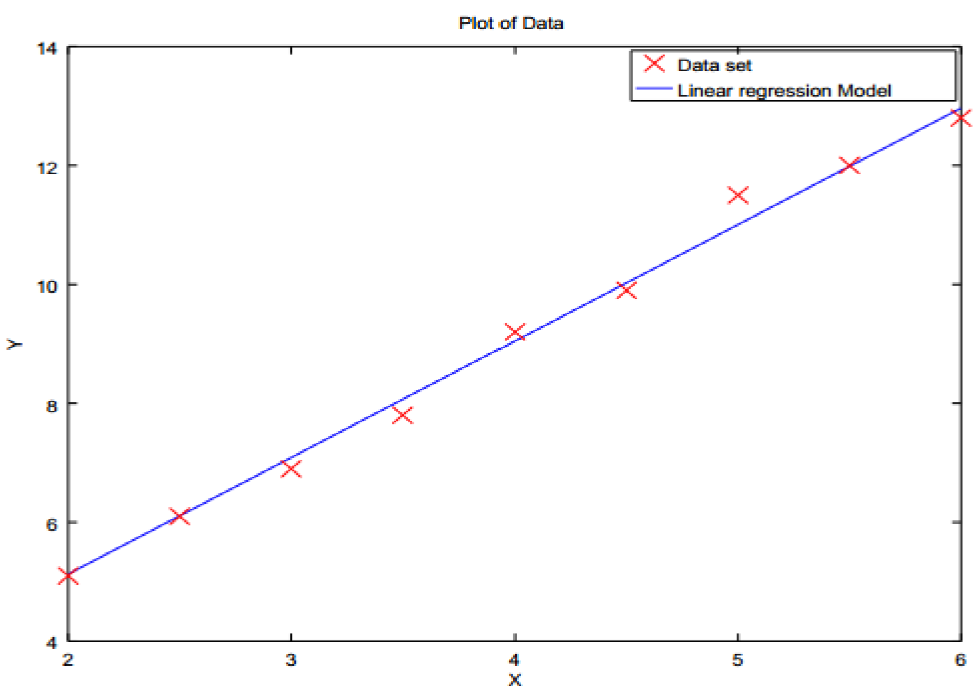

Fig 2: Plot of the Linear regression model vrs Actual Data set for alpha 0.01 and 0.1

HOMEWORK1

- **For alpha value 1**

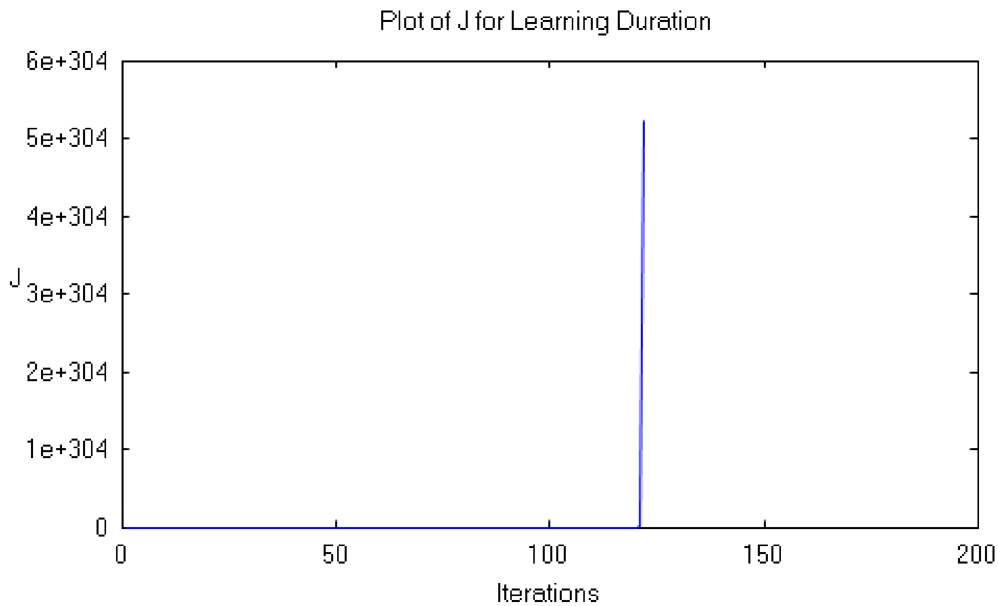

Fig 3: Plot of the "J" learning values for alpha 1

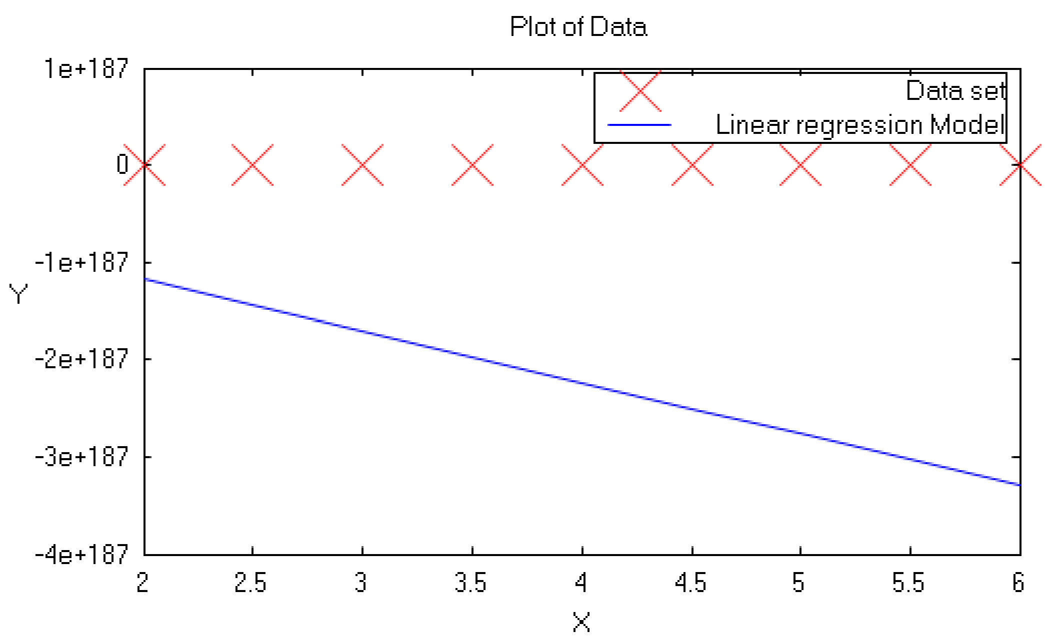

Fig 4: Plot of the Linear regression model vrs Actual Data set for alpha 1

HOMEWORK1

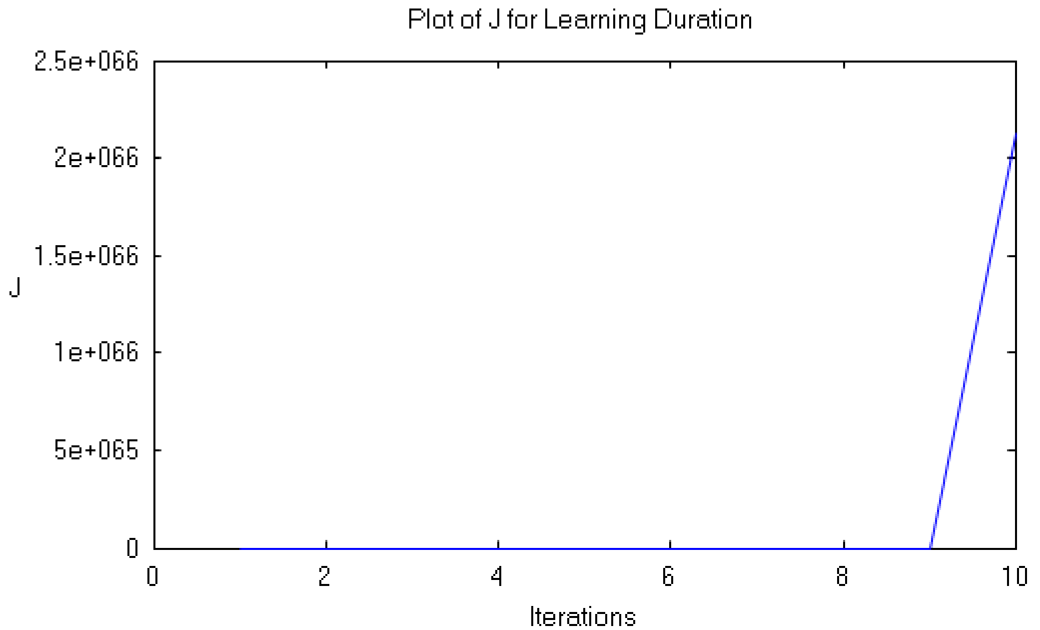

Fig 5: Plot of the “J” learning values for alpha 100

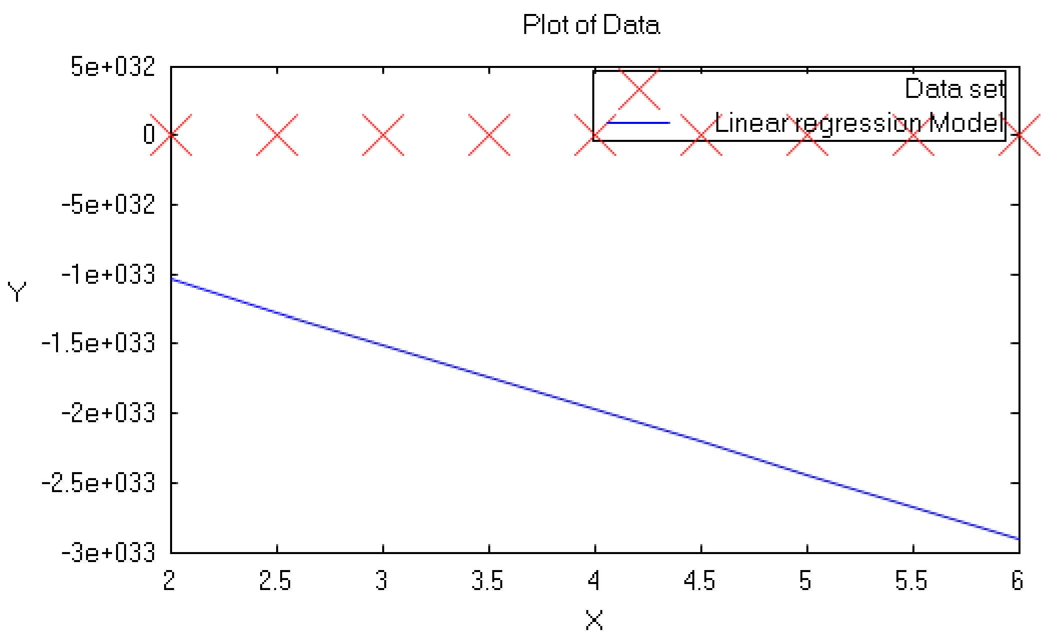

Fig 6: Plot of the Linear regression model vrs Actual Data set for alpha 100

HOMEWORK1

Observation:

For the alpha value of range 0.01 to 0.1. The J learning value in gradient descent drops to almost 0 in 20^{th} iteration only. That indicates θ1 and θ2 values are close to the convergence point.

For the alpha value greater (alpha >= 1). Gradient will overshoot the minimum J value. Hence it may fail to converge or even diverge. θ1 and θ2 values obtained are not correct. If we see the J plot for alpha 1 and 100. The value of J tends to reach very high value. Clearly indication the error is high.

Multiple Linear regression

1) Model Y2 as a linear function of X1 and X2. Use linear regression to learn the model parameters without scaling features.

$Y_2 = \theta_1 + \theta_2 * X_1 + \theta_3 * X_2$

a) Without Scaling

For the optimal alpha value **0.000001**, the initial value θ1, θ2, θ3 is considered to be 0.

The theta values θ1, θ2, θ3 are 0.001596 ,0.007428 and 1.006808 respectively.

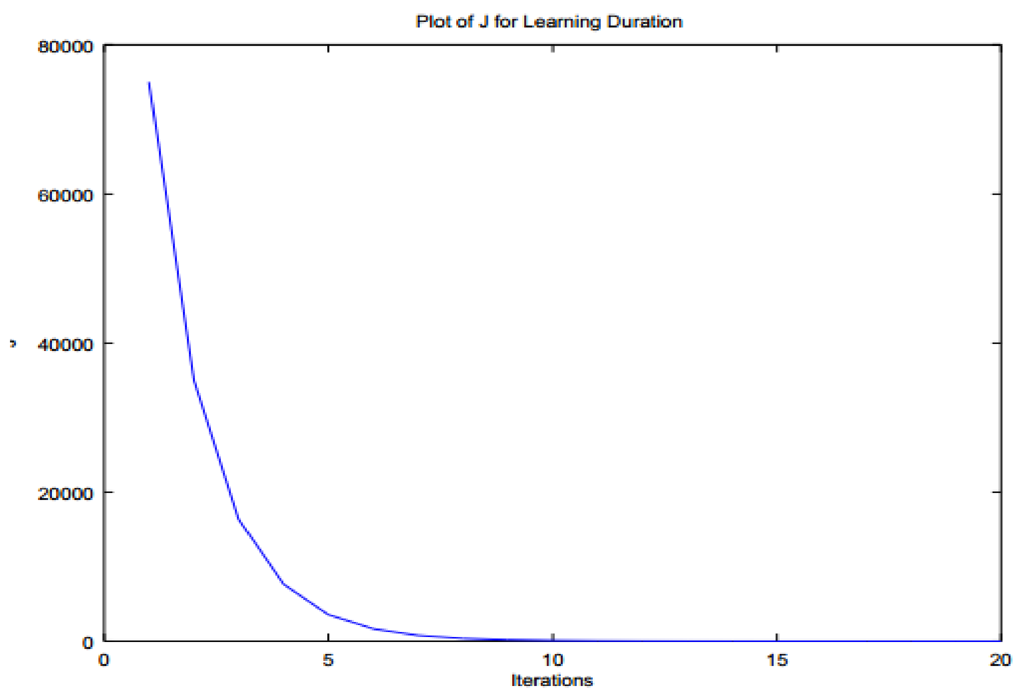

Fig 7: Plot of the “J” learning values for alpha 0.000001 without scaling

HOMEWORK1

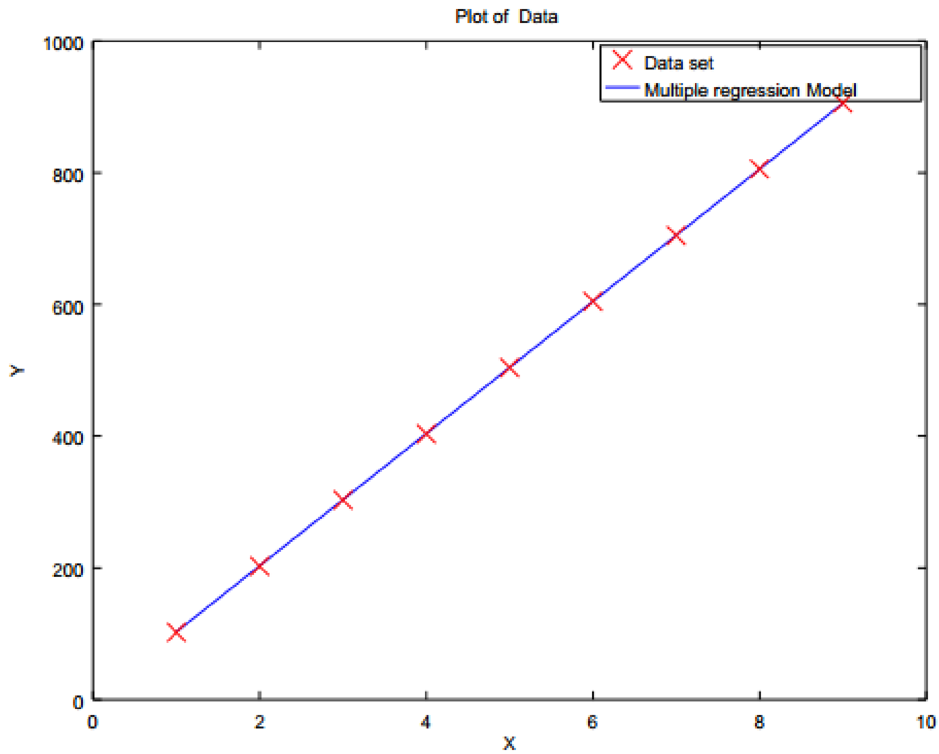

Fig 8: Plot of the Multiple regression model vrs Actual Data set for alpha 0.000001 without scaling.

Observation:

The initial value of θ and alpha chosen plays a key role. For the optimal alpha value of 0.000001 and initial value of θ to 0.

The obtained θ values are very closer to the global minima. Hence with number of less iteration, the J learning values tends to reach minimal value as seen in the Fig 7.

HOMEWORK1

b) With Scaling

Scaled X = (X – Mean) / Standard deviation.

$Y_2 = \theta_1 + \theta_2 * X_1(\text{scaled}) + \theta_3 * X_2(\text{scaled})$.

Note: Scaling is usually applied to the feature values of higher range. Here algorithm works correctly even if scaling in not applied to X_1 values.

For the optimal **alpha** value **0.01**. The theta values θ1, θ2, θ3 are 503.997750 ,137.599317 and 137.599317 respectively for scaled value of X_1 and X_2.

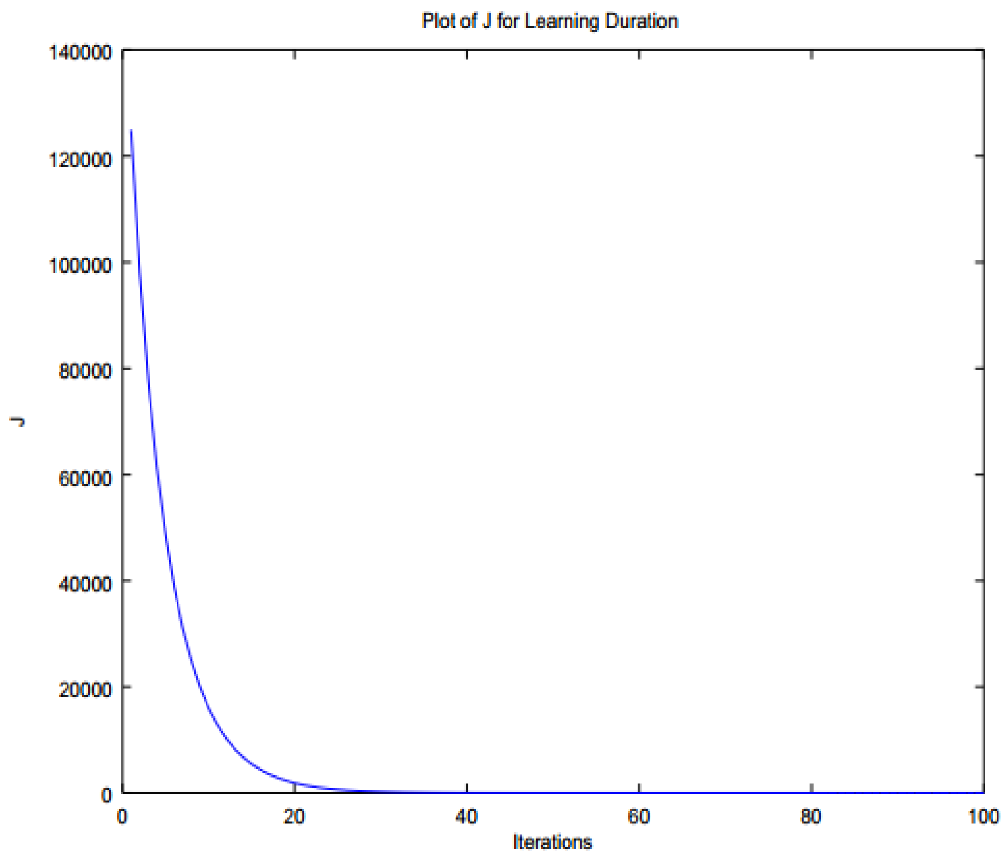

Fig 9: Plot of the "J" learning values for alpha 0.01 with scaling

HOMEWORK1

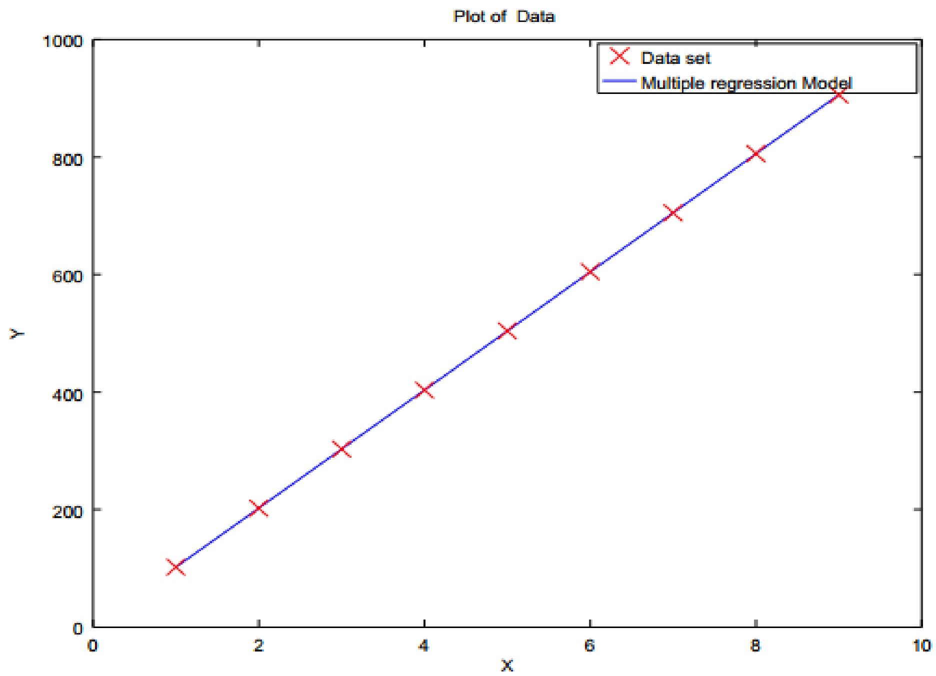

Fig 10: Plot of the Multiple regression model vrs Actual Data set for alpha 0. 01 with scaling.

Observation:

Gradient descent converges much faster with scaling. The contour graph is less skewed and look like circle instead of elliptical. The gradient descent finds the much more direct path to the Global minima in less number of iterations rather than taking convoluted path. The gradient decent will take around 30 iterations to reach the global minima for the alpha value 0.1.

c) Mathematical approach.

$\theta = (X^T X)^{-1} X^T Y$

The theta values are θ1, θ2, θ3 are 0.480595, 0.725899 and 1.001254.

Observation:

The mathematical approach finds the most optimal theta value. There is no requirement of alpha and feature scaling in this approach. However, when a large amount of data of different features are involved, the mathematical approach may require large computing time and becomes slower.

HOMEWORK1

2) Polynomial regression

a) Model Y3 as a quadratic function of X1. Use regression to learn the model parameters.
Using Multiple regression
$Y_3 = \theta_1 + \theta_2 X$(scaled) $+ \theta_3 X^2$ (scaled).
Alpha value = 0.1
The theta values are θ1, θ2, θ3 are 17.823955, 5.586512 and 5.586512 for scaling X values.

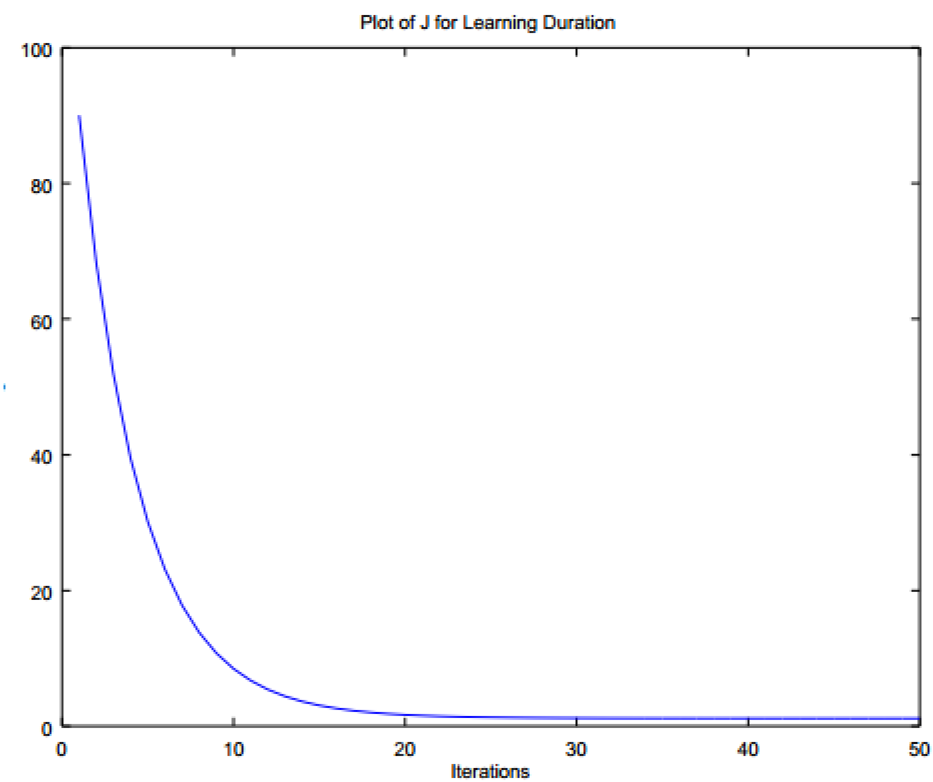

Fig 10: Plot of the "J" learning values for alpha 0.1 with scaling

HOMEWORK1

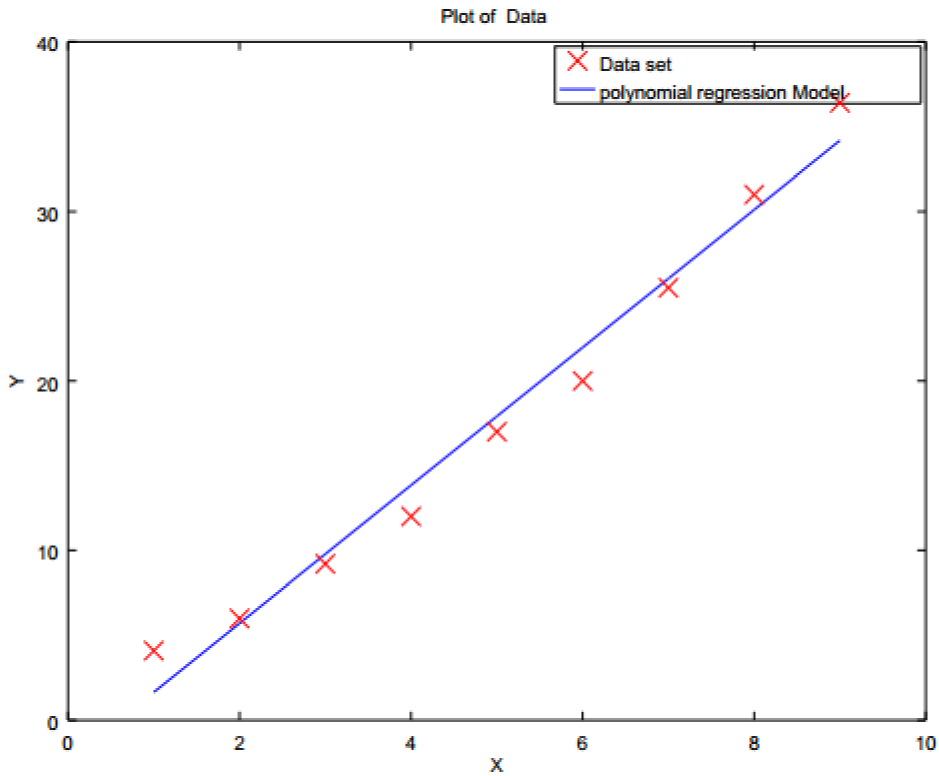

Fig 12: Plot of the Polynomial regression model vrs Actual Data set for alpha 0. 1 with scaling

Observation:

Polynomial regression can be reduced to linear regression in a high dimensional feature space.

From Fig 12. The obtained Model does not completely connect all the data set in the graph. Hence the obtained result is not completely optimal to all values of X. There might be large differences for particular value of X. For example, for x1= 2 the obtained y value is 1.6. but the actual y value is 4.1.

Conclusion:

The alpha value has to be chosen carefully. If alpha value is less it might take very large number of iterations to get the global minima of J value for the θ $values$. If alpha value is large, cost value diverges, the gradient descent will never reach converging point of J. Choosing the θ $values$ closer to local minima sometimes may result in mapping to the local minima J values. The θ value found in this case is not completely optimal, as there exists a global minima also. Hence it's better to try with different θ values.

HOMEWORK1