CPSC 6820 Project 2 Report : Regression

Problem Description: Given a dataset with 300 records and 2 features, the job was to use meticulously gathered data to create a regression hypothesis function that will predict a student's grade point average based on how many minutes per week they study and how many ounces of beer they consume per week.

Data Description:

The goal was to create a polynomial regression solution (y = w0 + w1x1 + w2x2 + w3x1x2 + w4x12 + w5x22). The first line in the given file was an Integer value indicating how many lines of data are in the file. Each line after that contained three tab-separated real values that represented minutes studying/week, ounces of beer/week, and semester grade point average.
Data of 300 records was divided into a training set (70%) and test set (30%)

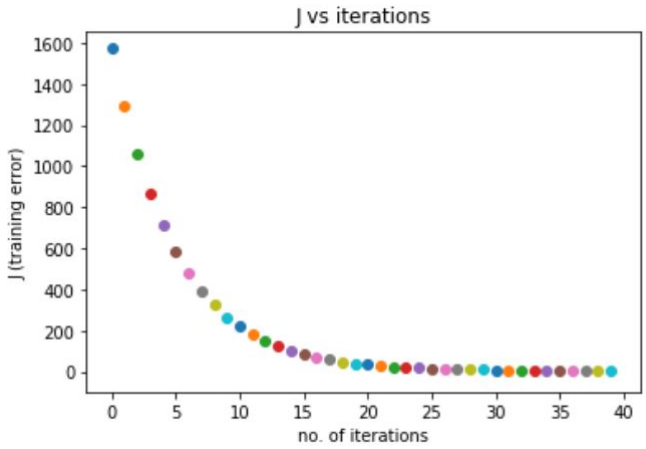
Initial Values:

Initial Values for weights, alpha were chosen as follows:
w0=5, w1=1, w2=1, w3=2, w4=3, w5=3, alpha=0.001
Value of J with these initial values: **1573.07**

Final Values:

After 40 iterations,when gradient descent started to converge, final values for weights, alpha were obtained as follows:
w0: 4.445, w1: -0.309, w2: -0.297, w3: -0.941, w4: 0.381, w5: 0.404, alpha=0.001
Final value of J on training set: **4.78**

Plot of J vs number of iterations for gradient descent:



After 40 iterations, the gradient descent converges and J values start to decrease very slowly, almost constant, hence 40 was chosen to get final values of weights and alpha.

Feature scaling:

Since the two features had a different range of values from GPA output range, I used standardization method of feature scaling and scaled all values to 0-4 range. (x1,x2 and y)
This way, the values were in the same range for training and higher value did not dominate.

Value of J on test set:

After using final weight values on the test set, the test error was observed : **0.90**