Regression Project 2 Report

Shivam Pandit

Problem Description

Given a dataset file named GPAData.txt that contains three tab-separated real values that represent minutes studying/week, ounces of beer/week, semester grade point average and the first line in the file is an Integer value indicating how many lines of data are in the file. We have to create a polynomial regression solution (y = w0 + w1x1 + w2x2 + w3x1x2 + w4x12 + w5x22) and program should prompt the user to enter values for minutes spent studying per week and ounces of beer consumed per week, and then predicts their semester GPA.

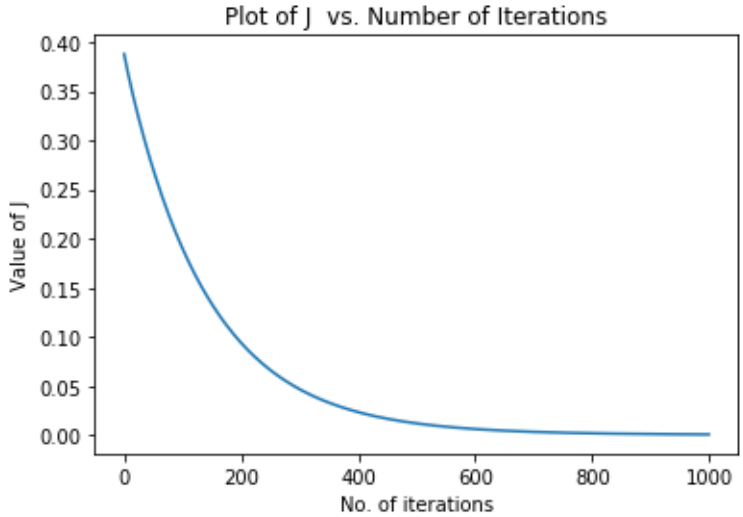
Initial Values for weights, alpha and J

Initially, I took random values for weights using random.seed(7) which gave weights as [0.08, 0.78, 0.44, 0.72, 0.98, 0.54] and I chose alpha as 0.1. Initial cost(J) came out to be 0.392

Final values for alpha, weights, number of iterations and final value of J on training set

After splitting dataset into 70% training and 30% test we had 210 examples in training set and 90 examples in test set. After training model with different values of alpha, I chose alpha as 0.01, number of iterations as 1000 as that gave me reasonable plots. My final values for weights after 1000 iterations were [-0.59, 0.10, -0.23, 0.15, 0.71, 0.01]. Final value of cost(J) on training data set came as 0.0011

Plot of *J* (vertical axis) vs. number of iterations (horizontal axis)



Feature Scaling

Since the features were not normally distributed, so I used standardization to scale down features using formula shown below. For all the fields in the dataset, mean was subtracted from them and then divided by standard deviation of the dataset. Standardizing the features so that they are centered around 0 with a standard deviation of 1 is important so that smaller features don't get neglected.

$$z = \frac{x - \mu}{\sigma}$$

Value of J in Test Dataset

Test Dataset had 90 examples or rows along with 90 true labels. Value of J on test set came out to be 0.00092 using final weights.