Project Report 1

Learning to Rank using Linear Regression

CS 574 – Fall 2016

Raj Jaysukh Patel:

(UBIT Name: rajjaysu UBIT : 50208278)

# Introduction

This project is about using Linear Regression for predicting target values for given input data. For this model is trained by using two techniques: Closed Form Solution and Stochastic Gradient Descent. The same procedure is repeated for two data sets: LeToR data set and Synthetic data set.

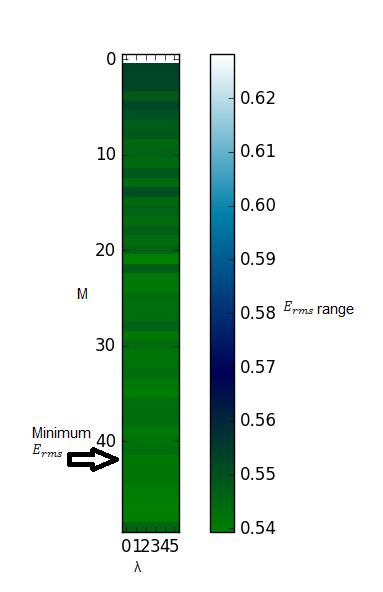
# Data Partition

The LeToR data set is read from QueryLevelNorm.txt. Synthetic data set is read from input.csv and output.csv files. First the input values and output values are stored in X and Y matrix respectively. The data is scrambled before dividing it into Training set (80%), validation set (10%) and test set (10%). I did this by randomizing integers from 0 to N-1, and using them as index for taking 80% rows in training set, then 10% in validation set and remaining 10% in test case.

# Hyper Parameter Tuning for LeToR Data

## Adjusting M and λ :

I checked value for validation set after training w by training data set for M ranging from 1 to 50 and λ ranging from 0 to 0.6 and found minimum at M = 35 and λ = 0.1. The graph for different values of for different values of M and λ is as follow.

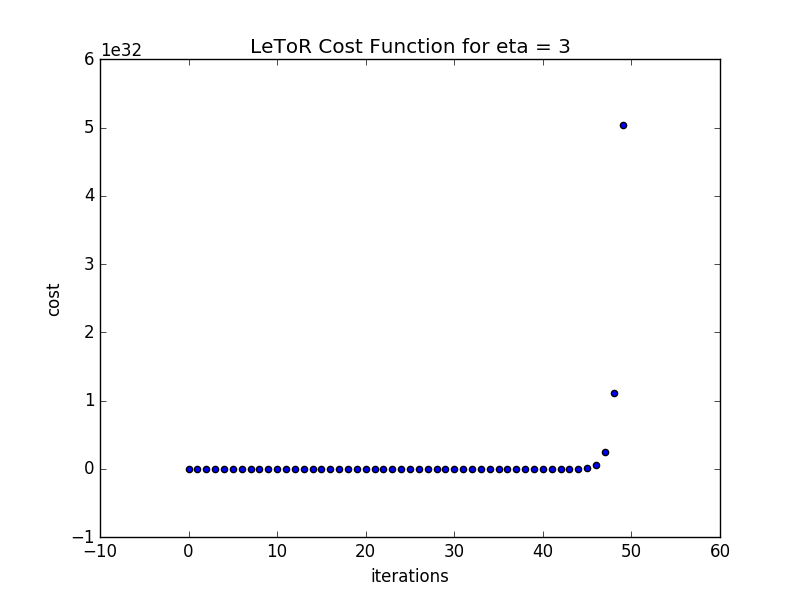


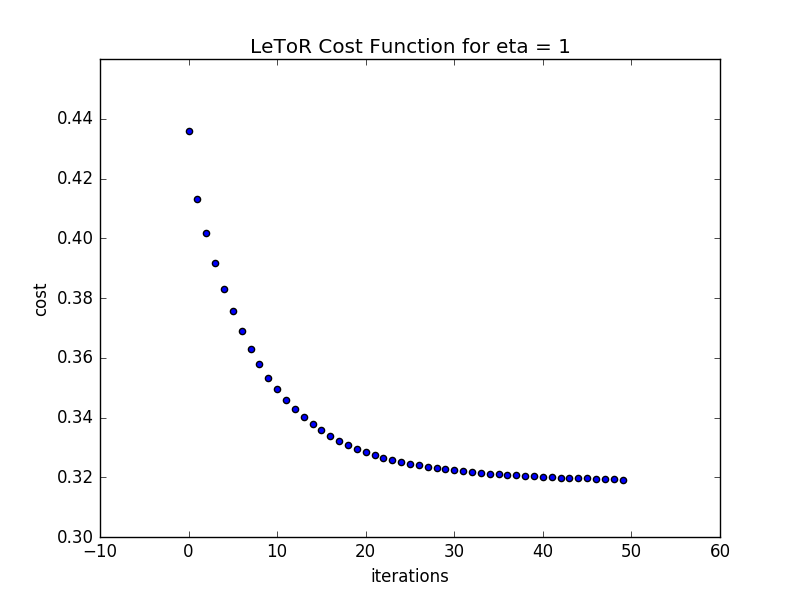
## Calculating   and :

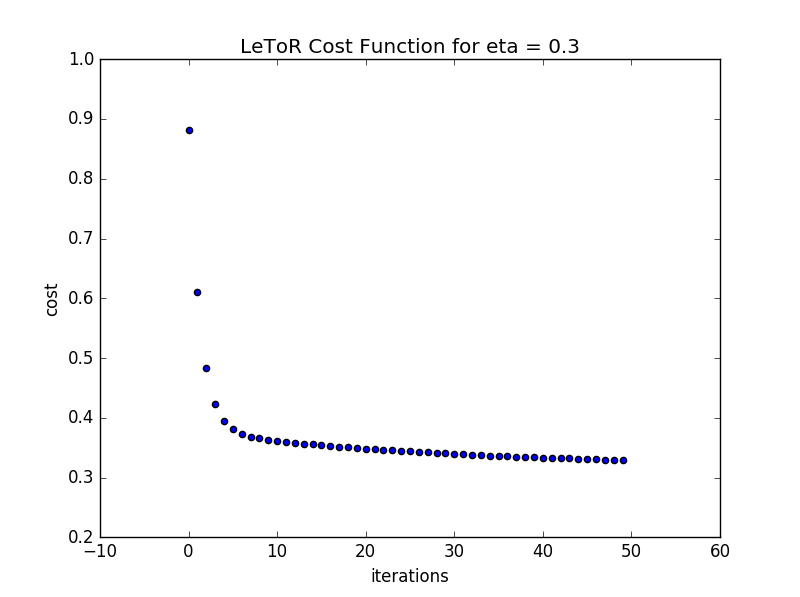
For calculating , I picked up M random points (M random rows from training set) and considered them as centroids. And for calculating , I calculated variance for each feature in training data set and created a d X d diagonal matrix where d is number of features, with variances as diagonal vector.

## Deciding :

For deciding , I started with 0.01 and checked the graph of cost function against iterations by increasing in multiple of 3. For smaller values, of the cost function decreases too slowly, so it will take longer time to reach the minimum value. And for higher values it cost function jumps too long that it misses minimum values. Figures below shows that the ideal value of is 1. As for = 3, It is visible that it misses the minimum value. And for = 0.3 , it takes more time to reach minimum value.







# Results and Evaluation

## Closed Form Solution

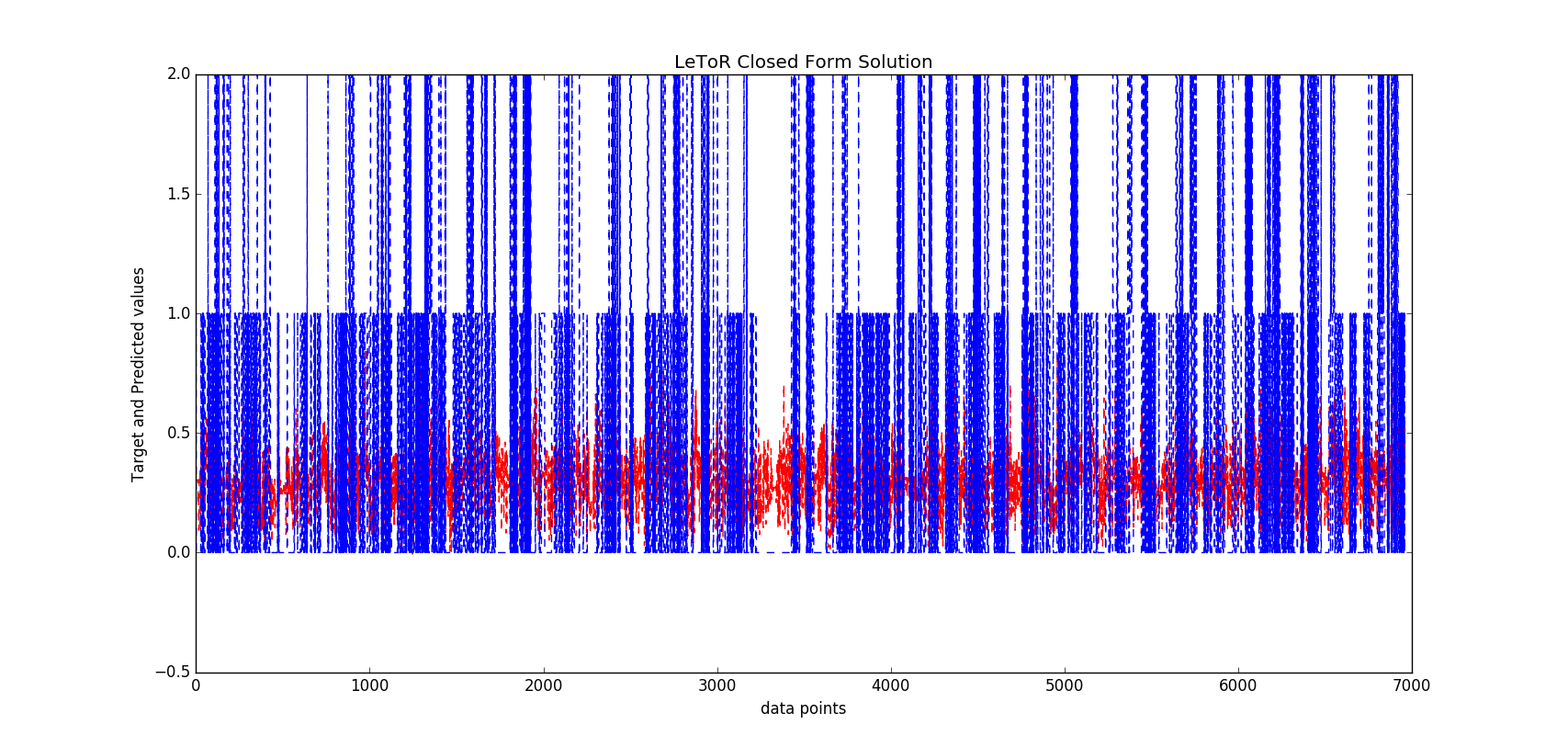
The values for different datasets are as follows:

Training Data Set = 0.5556

Validation Data Set = 0.5586

Test Data Set

The figure below shows the variations between original values and predicted values.



## Stochastic Gradient Descent

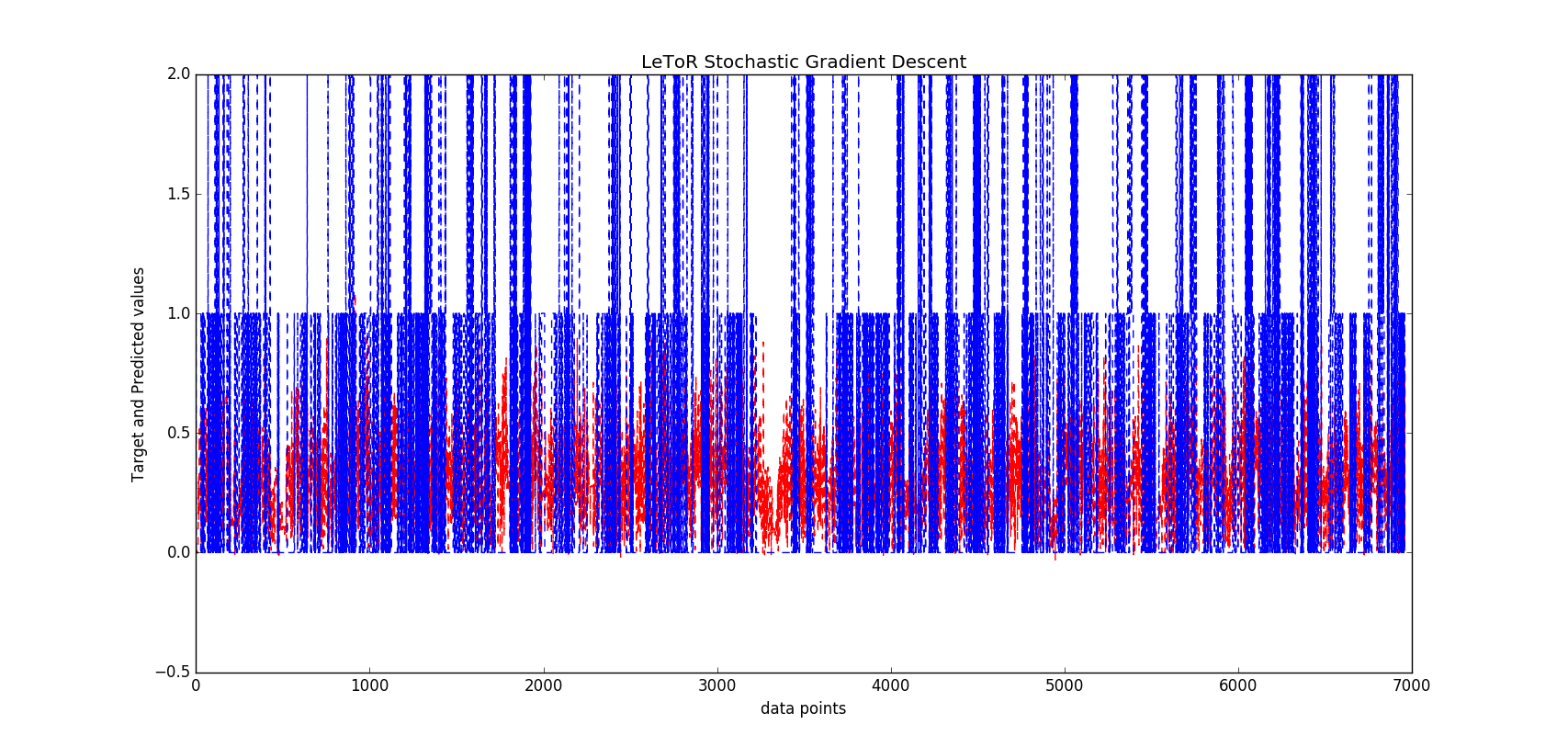
The values for different datasets are as follows:

Training Data Set = 0.5899

Validation Data Set = 0.5927

Test Set

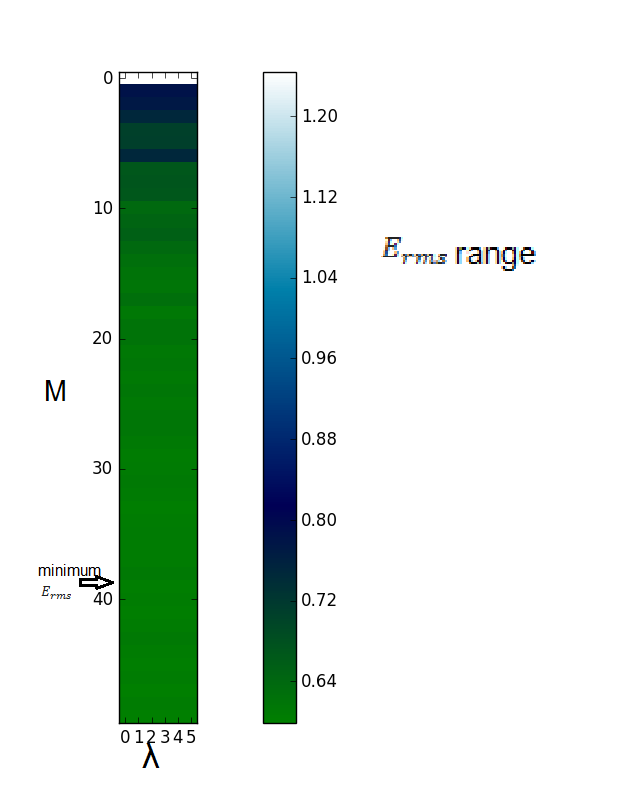
The figure below shows the variations between original values and predicted values.



# Hyper Parameter Tuning for Synthetic Data

## Adjusting M and λ :

I checked value for validation set after training w by training data set for M ranging from 0 to 50 and λ ranging from 0 to 0.06 and found minimum at M = 39 and λ = 0.01. The graph for different values of for different values of M and λ is as follow.

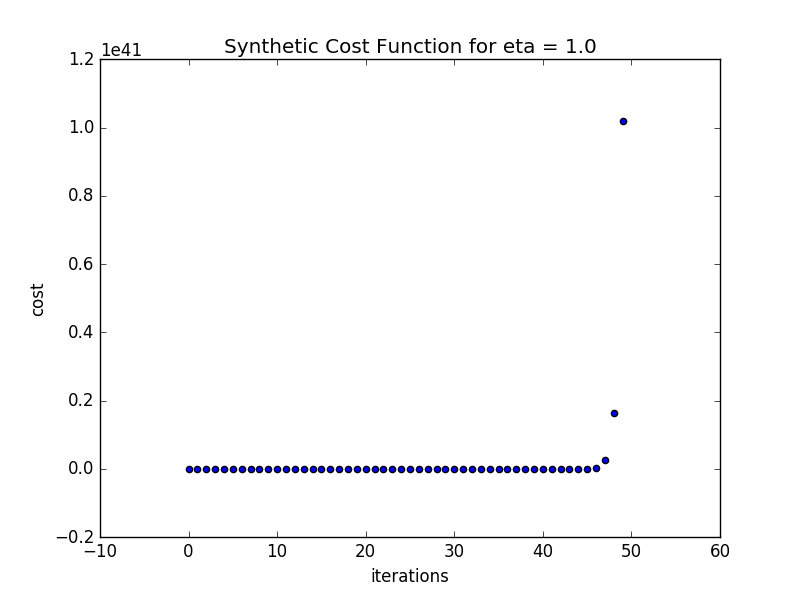


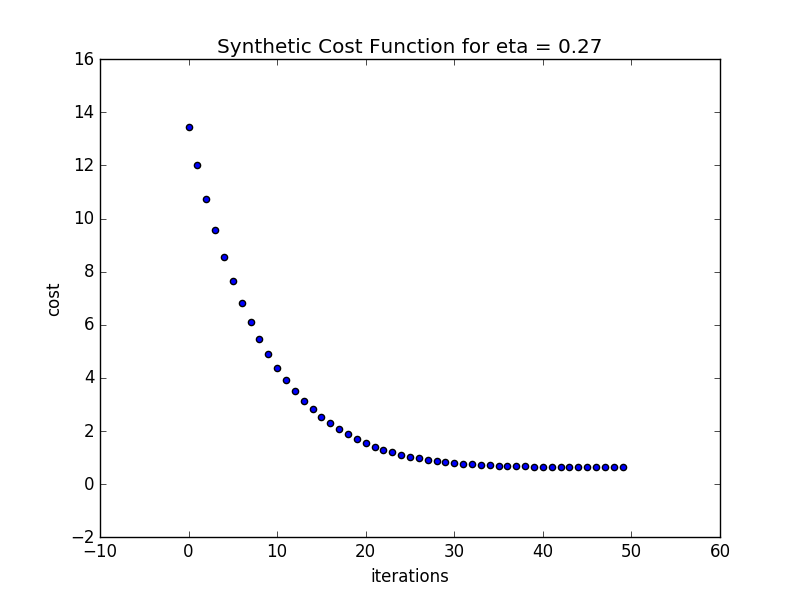
## Calculating   and :

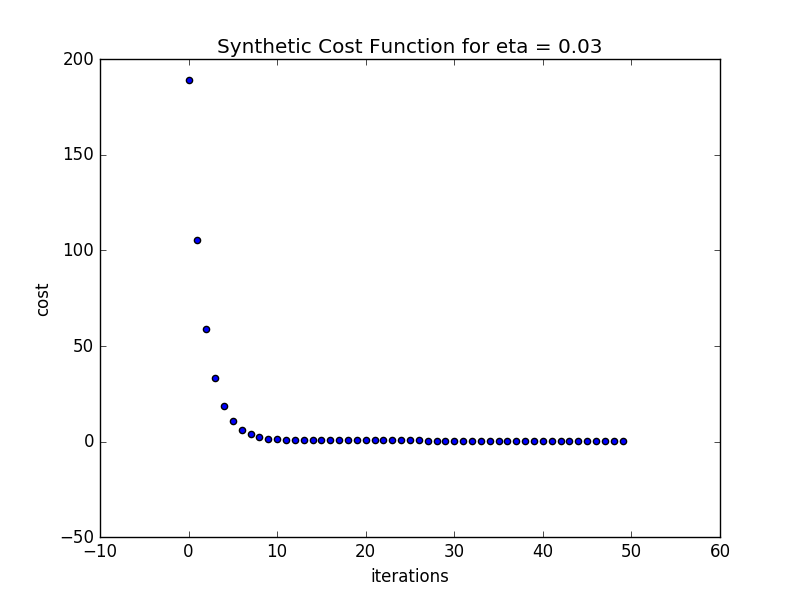
For calculating , I picked up M random points (M random rows from training set) and considered them as centroids. And for calculating , I calculated variance for each feature in training data set and created a d X d diagonal matrix where d is number of features, with variances as diagonal vector.

## Deciding :

For deciding , I started with 0.001 and checked the graph of cost function against iterations by increasing in multiple of 3. For smaller values, of the cost function decreases too slowly, so it will take longer time to reach the minimum value. And for higher values it cost function jumps too long that it misses minimum values. Figures below shows that the ideal value of is 0.27. As for = 1.0, It is visible that it misses the minimum value. And for = 0.03, it takes more time to reach minimum value.







# Results and Evaluation for Synthetic Data

## Closed Form Solution

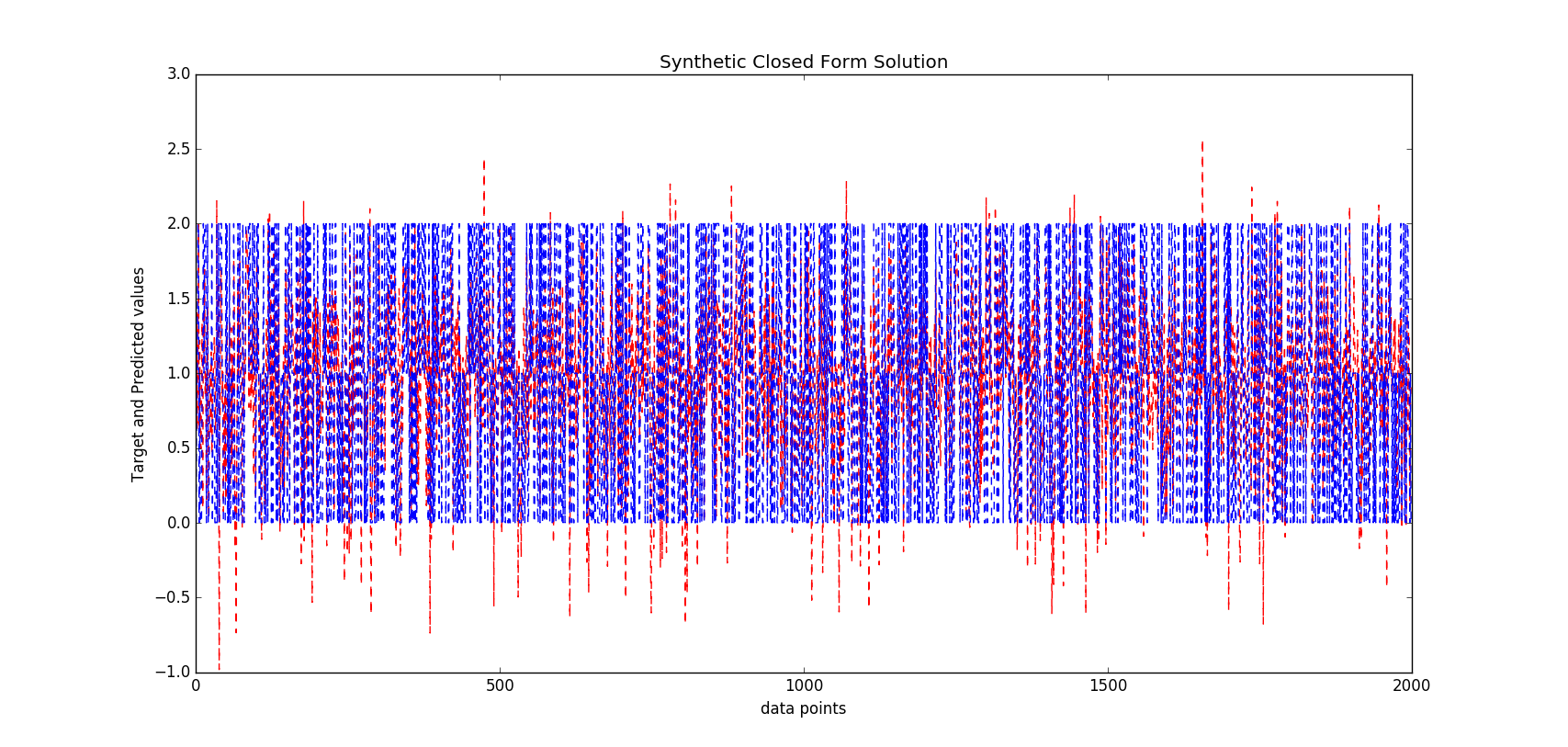
The values for different datasets are as follows:

Training Data Set = 0.5790

Validation Data Set = 0.5989

Test Set

The figure below shows the variations between original values and predicted values.



## Stochastic Gradient Descent

The values for different datasets are as follows:

Training Data Set = 0.6376

Validation Data Set = 0.6507

Test Set

The figure below shows the variations between original values and predicted values.

