Project Report 1

Learning to Rank using Linear Regression

CS 574 – Fall 2016
Raj Jaysukh Patel:

(UBIT Name: rajjaysu UBIT : 50208278)

1. 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.

2. 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. Synthetic data set is the data set given which is generated using some mathematical formula.

$$y = f(\mathbf{x}) + \varepsilon$$

Here f(x) is a function unknown to us and ε is noise.

3. Hyper Parameter Tuning

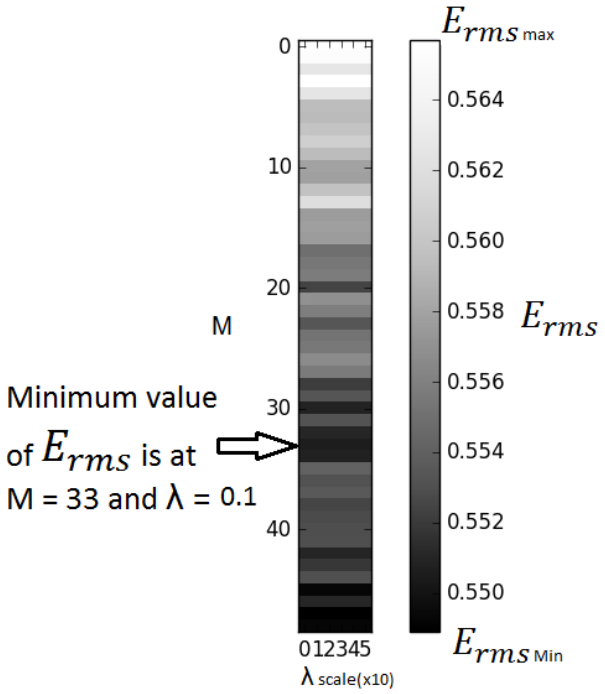
- ### Adjusting M and λ :

We can observe that as M increases, the polynomial will increasingly tune to noise and hence, Large value of M does not always give better function.

We have use λ for regularizing the polynomial or say for controlling the overfitting. That why we are validating values of λ and M based on validation set which is independent from Training set and then the function is checked on test set.

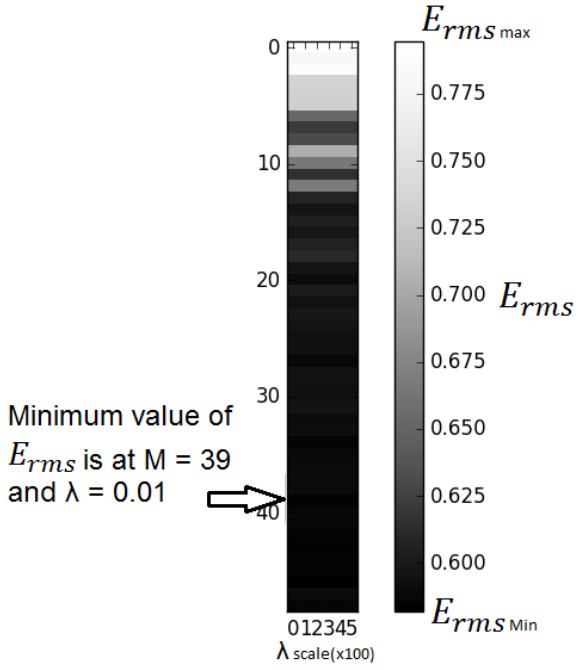
➢ For LeToR Data Set

I checked E_{rms} 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 E_{rms} at M = 33 and λ = 0.1. The graph for different values of E_{rms} for different values of M and λ is as follow.



➢ For Synthetic Data Set

I checked E_{rms} 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 E_{rms} at M = 39 and λ = 0.01. The graph for different values of E_{rms} for different values of M and λ is as follow.



- ## Calculating μ_j and Σ_j:

For calculating μ_j , I picked up M random points (M random rows from training set) and considered them as centroids. And for calculating Σ_j, 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.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_D^2 \end{pmatrix}$$

Where, $\sigma_i^2 = \frac{1}{10}\mathrm{var}_i\,(\mathbf{x})$

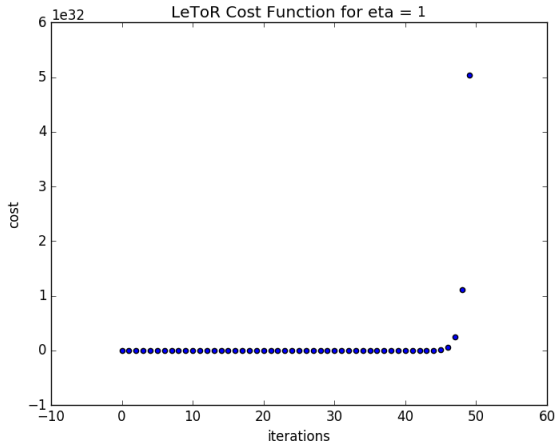
To calculate a D x D x M Sigma matrix we are replicating D x D plane M times.

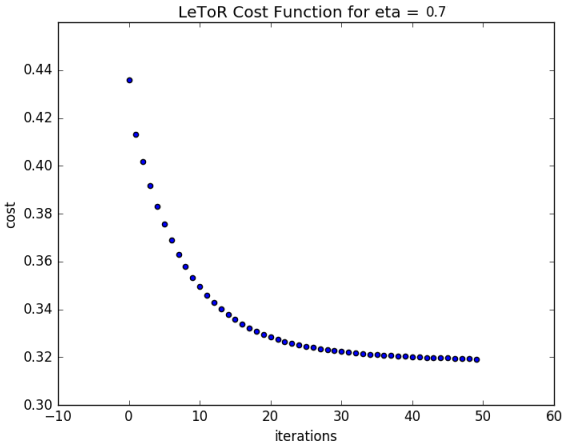
- ## Deciding Learning Rate, $\eta^{(\tau)}$:

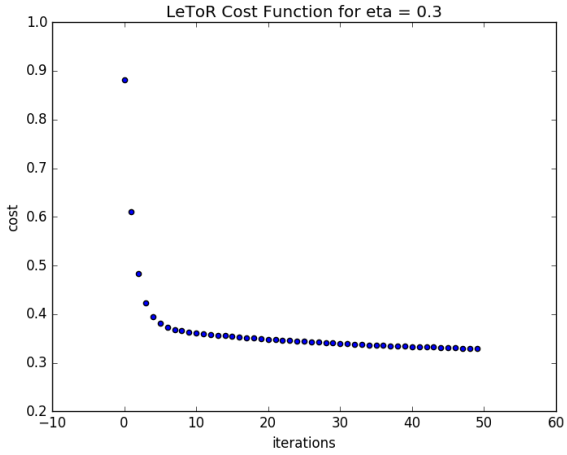
Learning rate decides, how fast the error function is reduced towards zero. At each step the weight vector is moved in the direction of the greatest rate of decrease of the error function.

➢ For LeToR Data Set

For deciding $\eta^{(\tau)}$, I started with 0.01 and checked the graph of cost(error) function against iterations by increasing $\eta^{(\tau)}$ in multiple of 3. For smaller values, of $\eta^{(\tau)}$ the cost(error) function decreases too slowly, so it will take longer time to reach the minimum value. And for higher values it cost(error) function jumps too long that it misses minimum values. Figures below shows that the ideal value of $\eta^{(\tau)}$ is 0.7. As for $\eta^{(\tau)}$ = 1, It is visible that it misses the minimum value. And for $\eta^{(\tau)}$ = 0.3 , it takes more time to reach minimum value.



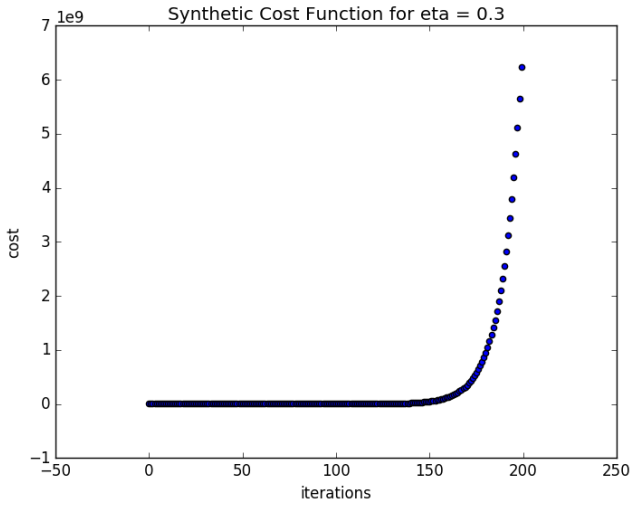


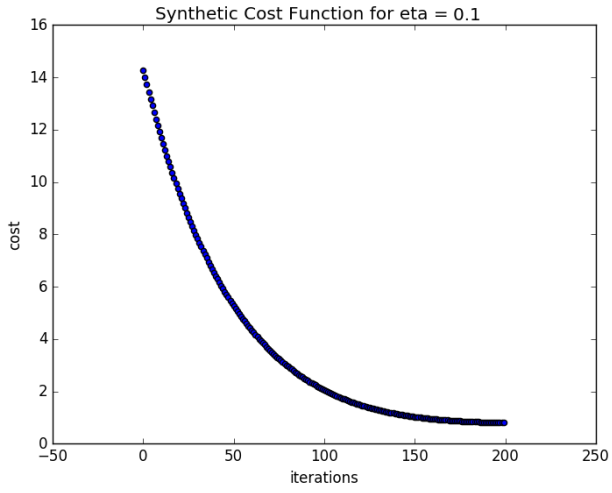


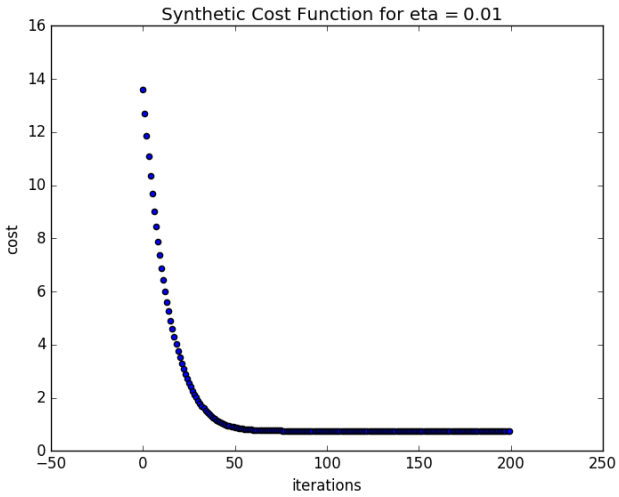
➢ For Synthetic Data Set

For deciding $\eta^{(\tau)}$, I started with 0.001 and checked the graph of cost(error) function against iterations by increasing $\eta^{(\tau)}$ in multiple of 3. For smaller values, of $\eta^{(\tau)}$ the cost(error)

function decreases too slowly, so it will take longer time to reach the minimum value. And for higher values it cost(error) function jumps too long that it misses minimum values. Figures below shows that the ideal value of $\eta^{(\tau)}$ is 0.1. As for $\eta^{(\tau)}$ = 0.3, It is visible that it misses the minimum value. And for $\eta^{(\tau)}$ = 0.01, it takes more time to reach minimum value.







4. Results and Evaluation

- ## Calculations

Basis function is a matrix of dimension NxM where N is number of training data samples. It is calculated as-

$$\mathbf{\Phi} = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}$$

Where,

$$\phi_j(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right)$$

- ## Closed Form Solution

Closed form solution is calculated using basis function, Sigma, lambda and M.

$$\mathbf{w}^* = (\lambda\mathbf{I} + \mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{t}$$

For each solution, we find Root Mean Square error on training, validation and test set as-

$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^*)/N_{\mathrm{V}}}$$

The training data is used to train the w(basis function). This trained basis function is validated for different values of λ and M by validation set. Then the result obtained is used to predict values test data set and compared with original values.

For LeToR Data Set

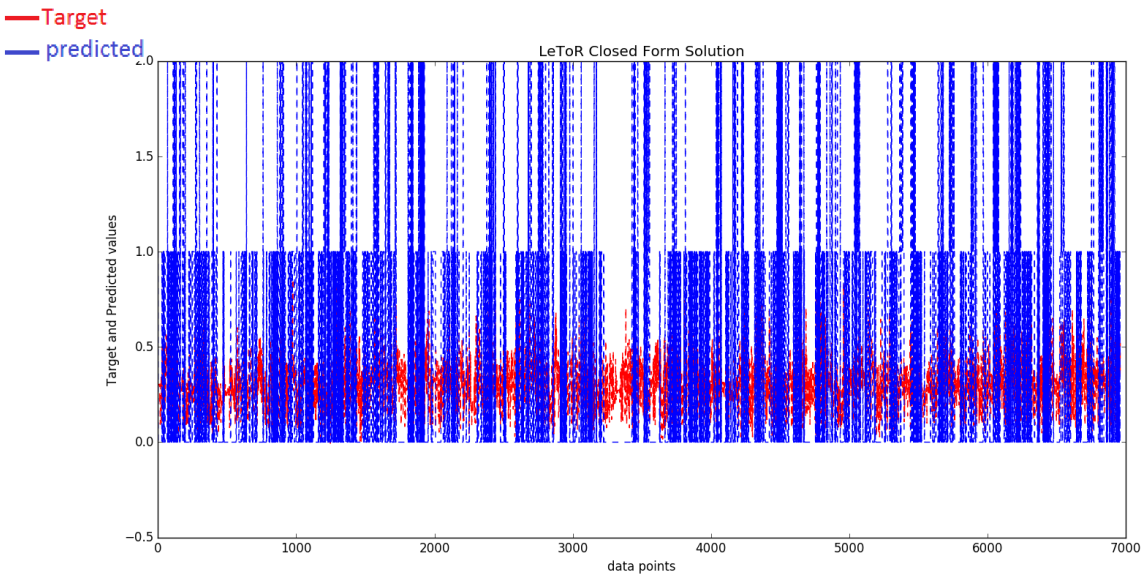
The E_{rms} values for different datasets are as follows:

Training Data Set E_{rms} = 0.5556

Validation Data Set E_{rms} = 0.5586

Test Data Set $E_{rms} = 0.5534$

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



For Synthetic Data Set

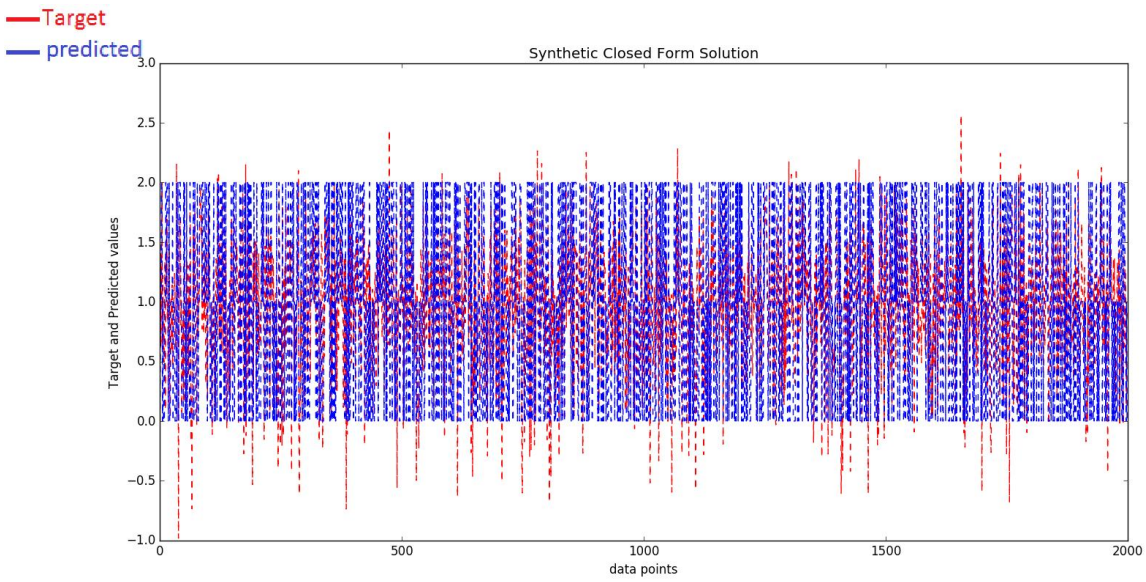
The E_{rms} values for different datasets are as follows:

Training Data Set E_{rms} = 0.5790

Validation Data Set E_{rms} = 0.5989

Test Data Set $E_{rms} = 0.6004$

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



- ## Stochastic Gradient Descent

Stochastic gradient descent works on single data sample at a time unlike Batch gradient descent. We start with an initial random value of solution. We modify this solution for each data sample.

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \Delta\mathbf{w}^{(\tau)}$$

Where,

$$\Delta\mathbf{w}^{(\tau)} = -\eta^{(\tau)}\nabla E$$

Where,

$$\nabla E_D = -(t_n - \mathbf{w}^{(\tau)\top}\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n)$$

$$\nabla E_W = \mathbf{w}^{(\tau)}$$

This process generates next solution which is used as previous solution for the next step. The process can be repeated E number of times where E is greater than N. That means we can iterate over the same data set multiple times to get accurate solution.

- ## For LeToR Data Set

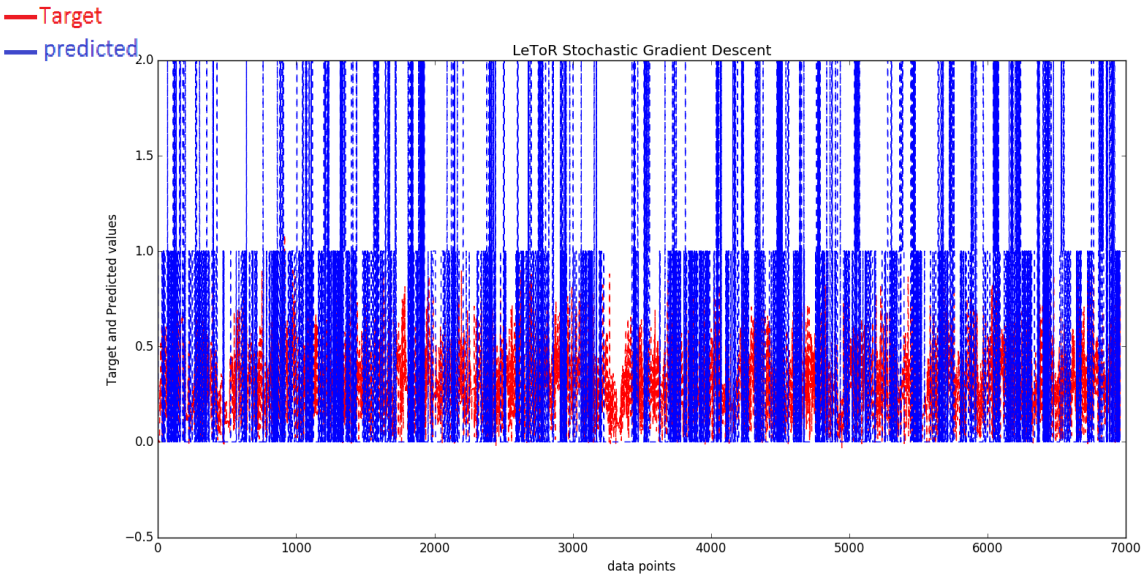
The E_{rms} values for different datasets are as follows:

Training Data Set E_{rms} = 0.5899

Validation Data Set E_{rms} = 0.5927

Test Set $E_{rms} = 0.5831$

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



➢ For Synthetic Data Set

The E_{rms} values for different datasets are as follows:

Training Data Set E_{rms} = 0.6376

Validation Data Set E_{rms} = 0.6507

Test Data Set $E_{rms} = 0.6412$

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