University at Buffalo
The State University of New York

INTRODUCTION TO MACHINE LEARNING
CSE 574

FALL-2016

PROJECT-2 REPORT

SUBMITTED BY:
ARUN CHANDRA PENDYALA(50207136)

LEARNING TO RANK USING LINEAR REGRESSION

1. Introduction

In this project, we train a linear regression model on the given datasets – Microsoft LeToR 4.0 dataset and synthetic dataset by using closed-form solution and Stochastic gradient descent(SGD). The model parameter w is trained on the training set by tuning various sets of hyper-parameters on training set and validation set such that the solution is optimized for both the cases. The model's performance is evaluated on the testing set and the performance of the model determines the generalization power of it.

2. Closed form solution:

Closed form solution involves the use of Gaussian radial basis functions to form the design matrix which is given by:

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

where μ_j is the centre of the basis function and Σ_j decides the spread of the basis function and the design matrix is given by:

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

The weight vector is learnt from the training samples and is given by

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

where λ is called regularization parameter which is introduced to the error function to avoid overfitting. The error function, which measures the misfit between the linear regression function for a given weight vector and training set data points , is given by:

$$E(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \quad , \quad E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top\boldsymbol{\phi}(\mathbf{x}_n)\}^2 \quad , \quad E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

For good model performance, the training and validation error should be low. This is done by tuning the hyper-parameter set on the training dataset and the final derived model is evaluated on the testing set.

3. Stochastic gradient descent:

This is another method to find weight vector for the linear regression function which takes a random initial weight vector and it is updated as follows for each new datasubset:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \Delta\mathbf{w}^{(\tau)} \quad , \quad \Delta\mathbf{w}^{(\tau)} = -\eta^{(\tau)}\nabla E \quad ,$$

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

$$\nabla E = \nabla E_D + \lambda \nabla E_W \qquad \nabla E_W = \mathbf{w}^{(\tau)}$$

,

The learning rate η should be chosen carefully such that it does not lead to divergence which is observed for large values of learning rate. By choosing a small learning rate, global minimum of the error rate can be attained effectively.

4. Closed form solution (CFS) on the LeToR dataset:

Microsoft LeToR 4.0 dataset:

The dataset consists of 46 input features and 1 output feature. The input values are real valued vectors(features from query-document pair) and the output or target values are relevance labels.

4.1. Methodology:

The input data is parsed from the querylevelnorm.txt and it is stored in an array and the relevance labels vector is also parsed from same file. The data is partitioned into three sets and the training set is used for computing the weight vectors and design matrix of size NxM which form the basis of linear regression function. The selection of hyperparameters is discussed below and the closed form solution is derived by using the various equations mentioned in (2).

4.2. Data Partition:

The data was partitioned into three sets – training(80 %) , validation(10 %) and testing set(10%) such that they do not overlap. So the LeToR input data set (query level norm set) which is of the size 69623 x 46 is partitioned as follows:

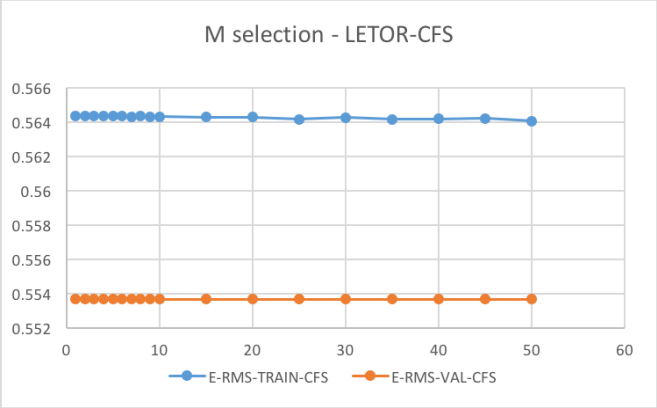
i) Training set – 55699 x 46 matrix
ii) Validation set – 6962 x 46 matrix
iii) Testing set – 6962 x 46 matrix

The target value set is of the size 69623 x 1 matrix which takes the values – 0 , 1 or 2 – the relevance labels

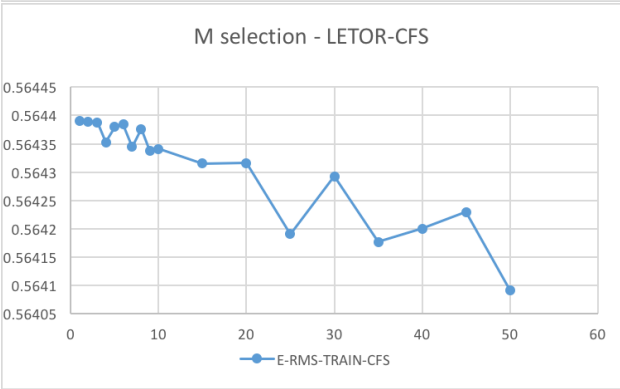
4.3. Hyper parameter selection:

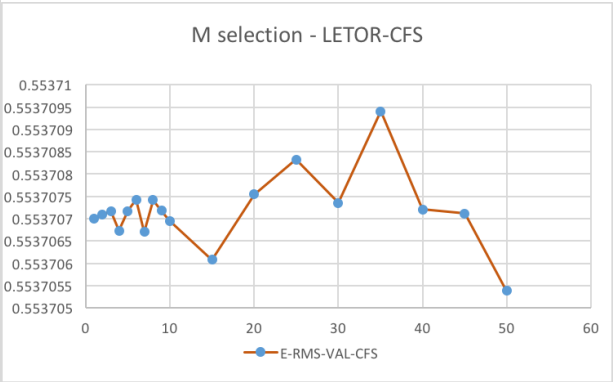
4.3.1. Selection of M and λ values:

M value decides the number of basis functions and λ value is used to avoid overfitting.
The following plots may help in deciding the values of these parameters:



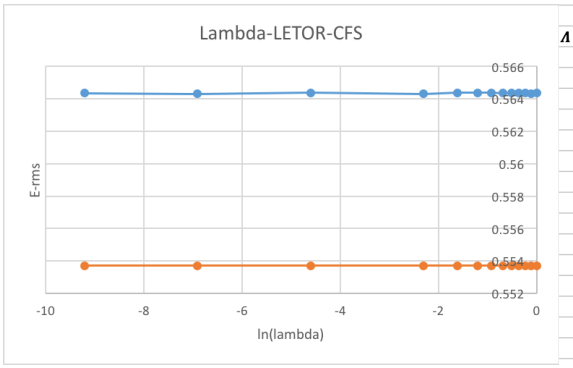
	lmda = 0.3		LETOR DATA	lmda = 0.3	
				eta = 0.0001	
M	E-RMS-TRAIN	E-RMS-VAL		E-RMS-TR-SGD	E-RMS-VAL-SGD
1	0.564390176	0.553707011		0.568082016	0.55624816
2	0.564389018	0.553707091		0.56808233	0.556248163
3	0.564387861	0.553707172		0.568082662	0.556248181
4	0.564352815	0.553706728		0.568080663	0.556248243
5	0.564379404	0.553707171		0.568082762	0.556248504
6	0.564384203	0.553707428		0.568084442	0.556248886
7	0.564344633	0.553706708		0.568080428	0.556248258
8	0.564375905	0.553707415		0.568083416	0.556248247
9	0.564337881	0.553707178		0.568082351	0.556248582
10	0.564341161	0.55370695		0.568081781	0.556248618
15	0.564314982	0.553706082		0.56808088	0.556249249
20	0.564315724	0.553707543		0.568084973	0.556250189
25	0.564190585	0.553708326		0.568087095	0.556250374
30	0.564292989	0.553707355		0.568083402	0.556249507
35	0.564177228	0.553709406		0.568090679	0.556250444
40	0.564200498	0.553707206		0.568083029	0.556252027
45	0.564229556	0.553707116		0.568082642	0.556251399
50	0.564092043	0.553705389		0.568075639	0.556251011



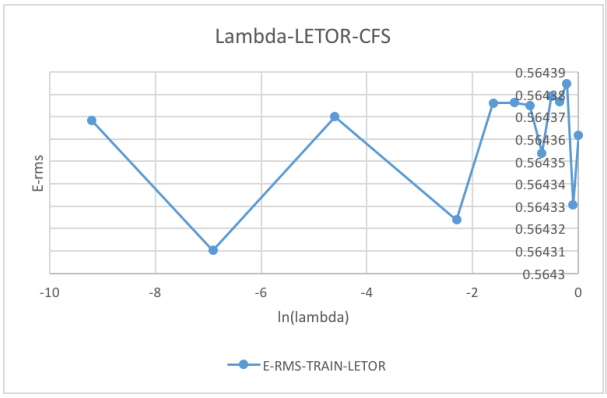


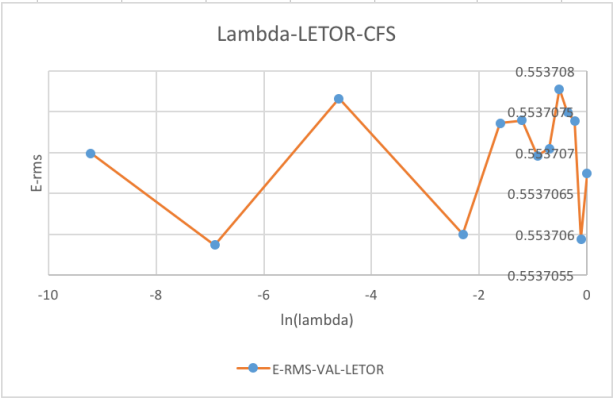
At the value of M= 7 , the training set and validation set errors are reasonably low and the difference between these two errors are also observed to be relatively low. At higher values of M such as 50 , the training and validation set errors may decrease but the time to compute the weights and deriving the regression function and model is becoming higher.

In order to choose the λ value, the plot between RMS value of error and ln(λ) is graphed. The results were tabulated and plotted in MS Excel. It is as below:



				LETOR DATA		
					M=7 , ETA=0.0001	
λ	ln(λ)	E-RMS-TRAIN	E-RMS-VAL		E-RMS-TR-SGD	E-RMS-VAL-SGD
0.0001	-9.2103404	0.56436846	0.55370699		0.564424316	0.55385257
0.001	-6.9077553	0.56431019	0.55370587		0.564403915	0.55384532
0.01	-4.6051702	0.56436996	0.55370766		0.564411585	0.55377249
0.1	-2.3025851	0.56432385	0.553706		0.564795309	0.55371961
0.2	-1.6094379	0.56437613	0.55370736		0.566176448	0.55467462
0.3	-1.2039728	0.56437621	0.55370739		0.568083677	0.55624868
0.4	-0.9162907	0.56437513	0.55370696		0.570267027	0.55815945
0.5	-0.6931472	0.56435376	0.55370705		0.572570854	0.56023529
0.6	-0.5108256	0.56437921	0.55370778		0.574898689	0.56237084
0.7	-0.3566749	0.56437688	0.55370749		0.577192327	0.56450186
0.8	-0.2231436	0.56438481	0.55370738		0.579419766	0.56658934
0.9	-0.1053605	0.5643307	0.55370594		0.581562648	0.56861124
1	0	0.56436163	0.55370674		0.583612415	0.57055488





From the above graphs, it is clear that for ln(λ) = -6.907755279 or λ = 0.01 , both the training and validation errors are relatively low. But it can be observed that this parameter does not have a significant effect on the model performance as it mainly influences at the point of overfitting as it discourages high values of weights in order to avoid overfitting problem.

4.3.2. Selection of μ and Σ:

The data is partitioned into 3 sets namely- training, validation and testing sets. The centres for radial Gaussian functions is found out by taking M data points. This is done by using numpy.random.randint() function which produces random values which are taken as indices of the input data matrix. By doing so , the entries in μ matrix are randomly picked data points.

Σ matrix is constrained to be a diagonal matrix as shown below and the entries in it, i.e , σ_i^2 values are chosen as (1/10) of $var_i(x)$ or ith dimension variance of the training data. I also chose σ_i^2 values as (1/20) , (1/5) and (1/1000) of $var_i(x)$ but it doesn't have significant effect on RMS values of error for both training and validation cases. In order to avoid singularity of matrix, the zeros in the Σ matrix have been taken as small values(i.e, 0.00001). If this matrix is singular, the basis function cannot be deduced.

5. Stochastic gradient descent (SGD) on the LeToR dataset:

5.1. Methodology:

In this, a random initial weight vector is assumed and it is updated iteratively by taking datapoints one at a time by using the equations discussed in (3). The step-size of each update is decided by learning rate or η. The value of η has been chosen as 0.001 to ensure convergence and to find the global minimum of error rate. I chose a fixed learning rate in this project.

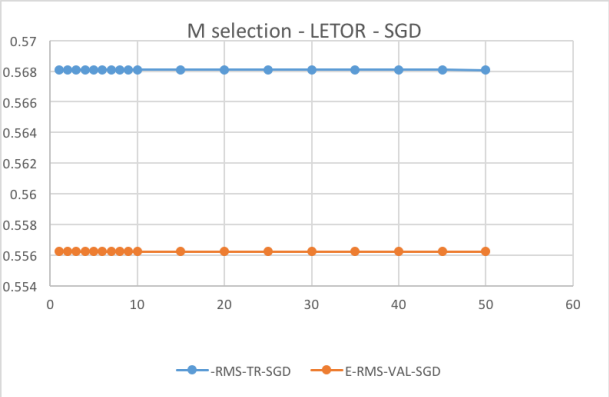
5.2. Data partition:

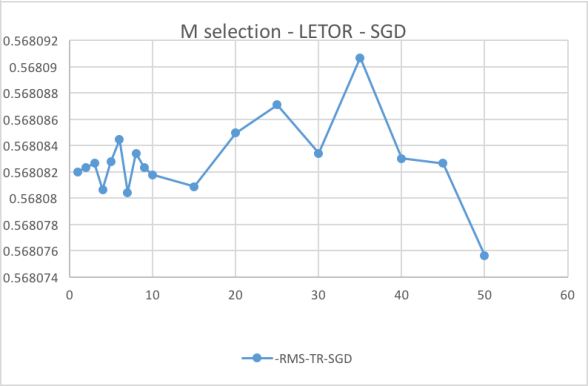
The data partition into three sets of training, validation and testing datasets is carried out exactly as discussed in 4.2.

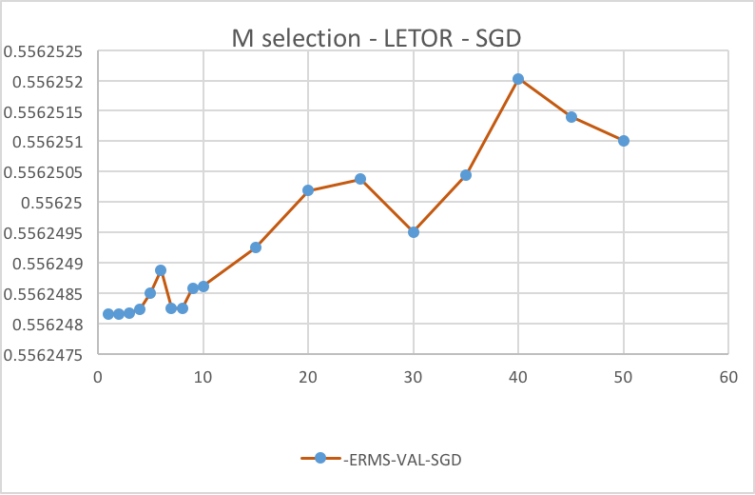
5.3. Hyper parameter Selection:

The hyper-parameters, which are used in the construction of the design matrix, are chosen as follows:

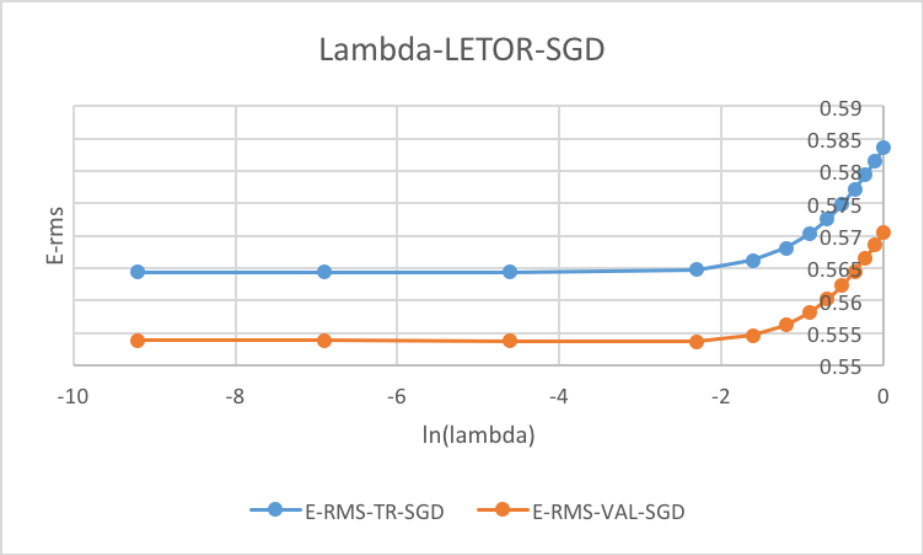
5.3.1. Selection of M and λ values:







At the value of M= 7, the training set and validation set errors are reasonably low and the difference between these two errors are also observed to be relatively low.



As discussed in 4.1. , the graph between E_{rms} and ln(λ) is plotted.
From the above graphs, it is clear that for ln(λ) = -6.907755279 or λ = 0.01 , both the training and validation errors are relatively low. But it can be observed that this parameter does not have a significant effect on the model performance.

5.3.2. Selection of μ and Σ:

These values are selected in the same manner as discussed in section 4.3.2.

5.3.3 Selection of Learning rate or η:

The learning rate is chosen by observing the plot between η and E_{rms} and it can be clearly seen that

6. Closed form solution (CFS) on the Synthetic dataset:

Synthetic dataset:

The dataset consists of 10 input features and 1 output feature. It has a total of 20,000 datapoints. The methodology adopted to deduce the models for this synthetic dataset is similar to the discussion in (4) and (5).

6.1. Data partition:

The data was partitioned into three sets – training(80 %) , validation(10 %) and testing set(10%) such that they do not overlap. So the synthetic input data set which is of the size 20000 x 10 is partitioned as follows:

i) Training set – 16000 x 10 matrix
ii) Validation set – 2000 x 10 matrix
iii) Testing set – 2000 x 10 matrix

The target value set is of the size 20000 x 1 matrix which takes the values – 0 , 1 or 2 .

6.2. Hyper parameter selection:

4.3.1. Selection of M and λ values:

M value decides the number of basis functions and λ value is used to avoid overfitting.
The following plots may help in deciding the values of these parameters:

RESULTS:

#######LETOR DATA#######

weights(cfs)
[0.30756594 -0.15378297 0.34621703 -0.15378297 -0.15378297 -0.15378297
-0.15378297]
RMS error-training error
0.564383654273
RMS error-validation error
0.55370708133
RMS error-testing error
0.624665049548
weights(sgd)

[0.15901166 0.00075134 0.00076269 0.00076178 0.00076161 0.0007611
  0.00076079]
RMS error - training error (SGD)
0.583612473114
RMS error-validation error-SGD
0.672934744069
RMS error-testing error(SGD)
0.655736782236
Hyperparameters
M
7
lambda
1
eta
0.0001
#######SYNTHETIC DATA#######
weights(cfs)
[0.96473685 -0.96377308 1.03422892 1.03422892 1.03422892 0.03522792
  0.03522792]
RMS error- training error
0.790099543448
RMS error-validation error
0.786029442572
RMS error-testing error
0.785350284425
weights(sgd)
[0.81132168 0.19963858 0.19978547 0.19979149 0.19983525 0.19970782
  0.19969409]
RMS error - training error (SGD)
0.805004087031
RMS error-testing error-SGD
0.800244978324
RMS error-testing error-SGD
0.797594015967
Hyperparameters for synthetic data
M
7
lambda
0.001
eta
0.0001