

INTRODUCTION TO MACHINE LEARNING

PROJECT-2 REPORT

Prof. Dr. Srihari

GROUP MEMBERS

NITHISH SHOKEEN (UB Person Number: 50247681)

MAHALAKSHMI PADMA SRI HARSHA MADDU (UB Person Number: 50246769)

CHARANYA SUDHARSANAN (UB person Number: 50245956)

Problem Statement: The learn to Rank data set has 69623 query-document pairs (rows) with 46 features for each row. Linear Regression is fit to the data set using closed-form solution and Stochastic Gradient Descent Solution. Minimum Difference of ERMS value between the training and the validation set is determined and the the test error is reported. The same is done for the Synthetic data set.

Step1: Data Partition:

The LeToR data set is imported using 'genfromtext' function from the library 'NUMPY'. The data is divided according to the following:

Training Data Set – 80%

Validation Data Set – 10%

Test Data Set – 10%

Step2: Closed form Solution

The following computations are done for the determination of the closed form solution.

- We find the clusters for the given data.
- We find the spreads for the given data.
- We compute the design matrix.
- Random values for hyper parameters are fixed and are updated after the hyper parameter tuning is done.
- Closed form solution is determined using the model and the hyper parameters.
- A function for finding the ERMS is defined.
- ERMS for the training, validation and test is determined.

Step3: Stochastic Gradient Solution

- Design matrix that has been computed earlier.
- A functions for determining the model parameters has been defined. This function uses early stopping algorithm.
- Using the model parameters, we predict the target values of the test data.
- We find ERMS using the predicted target values.

Step4: Hyper Parameter Tuning

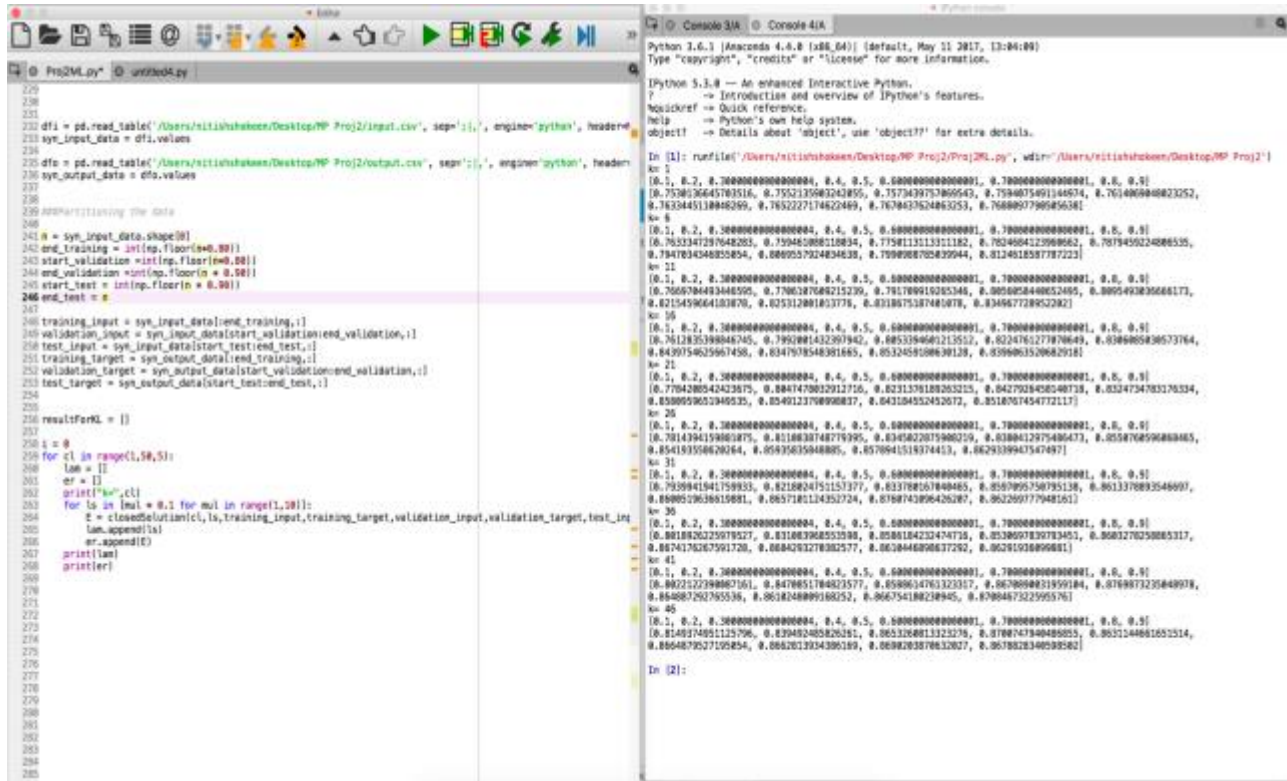
- Parameter Tuning for both Closed Form and SGS is done and the values are updated.
- Parameter Tuning is done by finding the min value of the difference between ERMS values of Training and Validation data sets. When the values for hyper parameters are obtained, ERMS for the test data is determined and is reported.

Tuning the Hyper Parameters:

In order to tune the hyper parameters, we have found out the difference between ERMS of Training and the Validation Set by using grid search on K and Lambda values. The following figure indicates the same.

The best combination of K and lambda have been determined.

The following figure uses the Synthetic Data.



```
Python 3.6.1 [Anaconda 4.4.0 (x86_64)] (default, May 11 2017, 13:04:06)
Type "copyright", "credits" or "license()" for more information.

Python 3.3.0 -- An enhanced Interactive Python.
?              -> Introduction and overview of Python's features.
help()         -> Quick reference.
help()         -> Python's own help system.
object?        -> Details about 'object', use 'object??' for extra details.

In [1]: runfile('/Users/nitishshukla/Desktop/MP Proj2/Proj2ML.py', wdir='/Users/nitishshukla/Desktop/MP Proj2')
x= 1
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.7538936445783516, 0.7552135861242855, 0.7573439757805643, 0.7594075491146674, 0.7614066848823252,
0.7633445138862865, 0.7652227174622469, 0.7670457624863353, 0.768899779654561)
x= 8
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.7653347297648283, 0.759462886118834, 0.7758113113311182, 0.7824604123968662, 0.787945022488535,
0.7947834348855854, 0.8009557924834639, 0.798988768339944, 0.812451837787223)
x= 11
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.766678443448595, 0.786387686215239, 0.781789916285346, 0.805685448852465, 0.809549383666173,
0.821549664183876, 0.825312891833778, 0.8318575387481878, 0.834667738952382)
x= 16
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.7612835398846745, 0.7992881432397942, 0.805339464213512, 0.824761277878649, 0.838885838573764,
0.8439754625967458, 0.847978548381665, 0.8532459108639128, 0.859863520682918)
x= 21
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.7704288542423875, 0.8047478032912716, 0.8231378189263215, 0.8427932458548713, 0.8524734783176334,
0.858959652946525, 0.8549123708986837, 0.843184552452672, 0.8518767454772117)
x= 26
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.7814394158851875, 0.8118838748774995, 0.8345822875980219, 0.8388412975486473, 0.8558768596888485,
0.854193558626264, 0.85435635848885, 0.8570841519374413, 0.8629339947547497)
x= 31
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.7939841941759933, 0.8218824751515777, 0.833788167848485, 0.8567895758795138, 0.8613378893546697,
0.868951963851981, 0.8657182114352724, 0.8788743896426287, 0.882269777948161)
x= 36
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.8018826225979527, 0.831887988553598, 0.8586184232474715, 0.8538697830783451, 0.8683276250885317,
0.867417627591728, 0.8684282278382577, 0.863846898637262, 0.8629193899681)
x= 41
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.8022127398827161, 0.8478851784823577, 0.8588634761323317, 0.8678809821959184, 0.8769873235849978,
0.884887292765536, 0.88182488899188252, 0.866754188238945, 0.8708467322595576)
x= 46
(0.1, 0.2, 0.30000000000000004, 0.4, 0.5, 0.6000000000000001, 0.7000000000000001, 0.8, 0.9)
(0.8149174951125796, 0.839482485826281, 0.8653268813323278, 0.8788747484848855, 0.883144861652514,
0.8864879527105854, 0.8862613934395189, 0.8888283878832827, 0.8878828348593582)
In [2]:
```

The same procedure has been followed for the LeToR Data and the hyper parameters have been determined.

The following hyper parameters are chosen after tuning with the validation set.

Linear Regression Model:

1.LeToR Data Set:

Hyper Parameters	Symbol	Value
Number of clusters	M/K	11
Regularization Term	λ	0.6
Learning Rate	$\eta^{(\tau)}$	0.05

2.Synthetic Data Set:

Hyper Parameters	Symbol	Value
Number of clusters	M/K	6
Regularization Term	λ	0.1
Learning Rate	$\eta^{(\tau)}$	0.05

ERMS Values:

1. LeToR Data Set

Data Set	Closed Form Solution	Stochastic Gradient Solution
Training	0.5634459614939566	0.5660121702119933
Validation	0.5525819199422247	0.5562475023625585
Test	0.6385530710060436	0.64142338488311

2. Synthetic Data Set

Data Set	Closed Form Solution	Stochastic Gradient Solution
Training	0.7391846044329995	0.7842058356432198
Validation	0.7504407958243514	0.7800265344713896
Test	0.7507374948090173	0.7813034199321781

Evaluation:

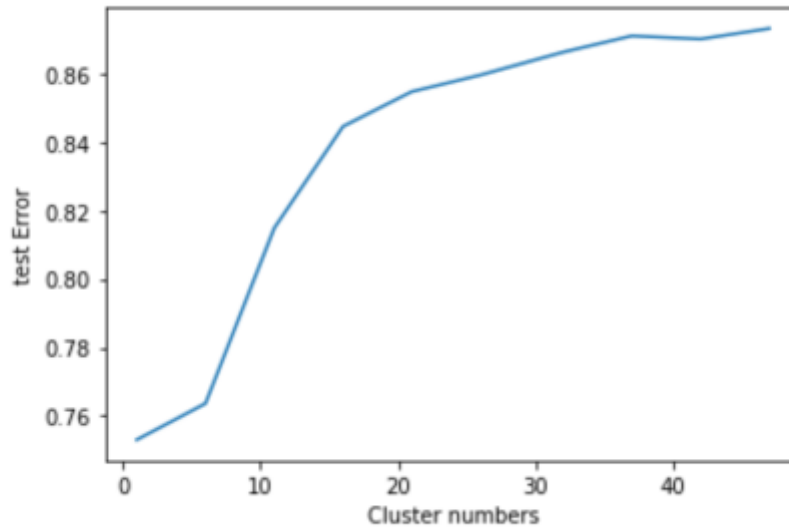
We performed linear regression and stochastic gradient descent for the given input data and basis functions.

By using these techniques, the target value can be predicted for new unseen input.

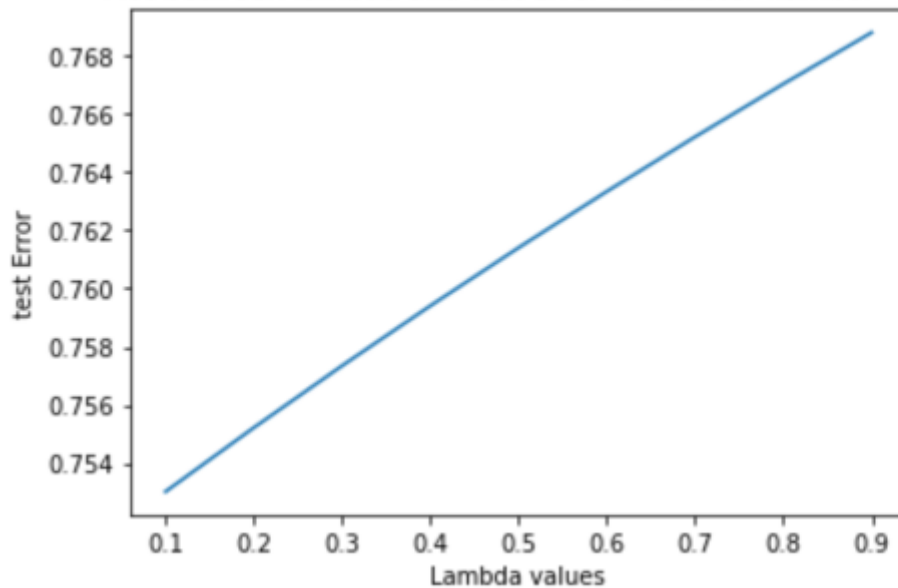
Also, performing the tuning of model parameters and hyper parameters makes the accuracy better.

Following are the screenshots of the plots obtained for the result:

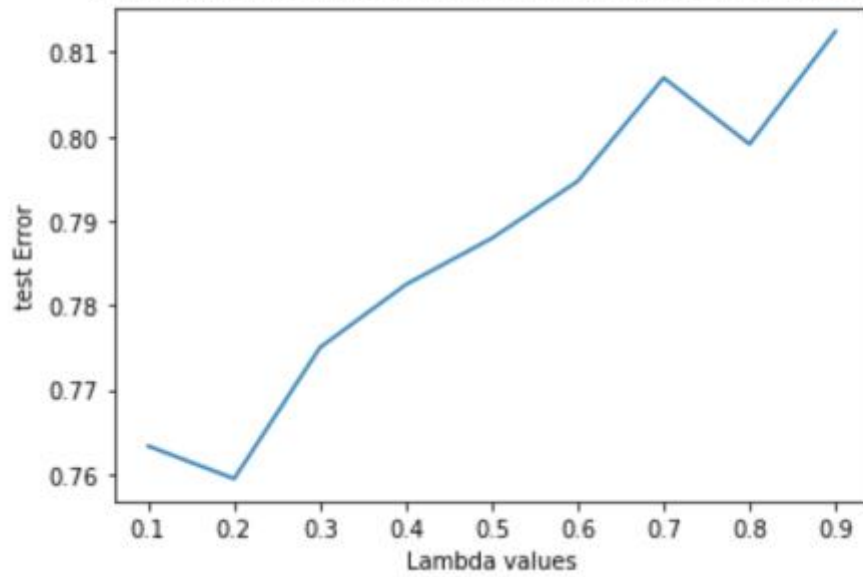
Plot between Clusternumbers and test error for Cluster number incrementing by 5



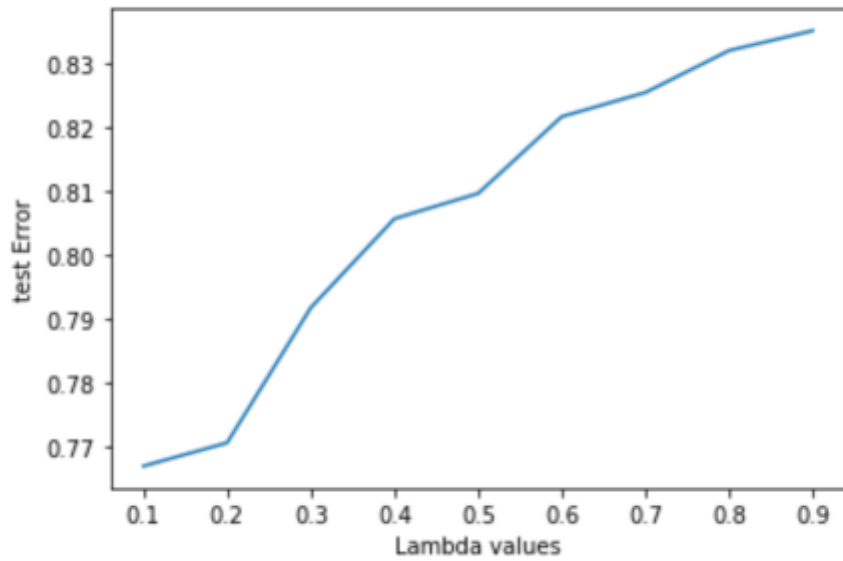
Plot between lambda and test error for Cluster number = 1



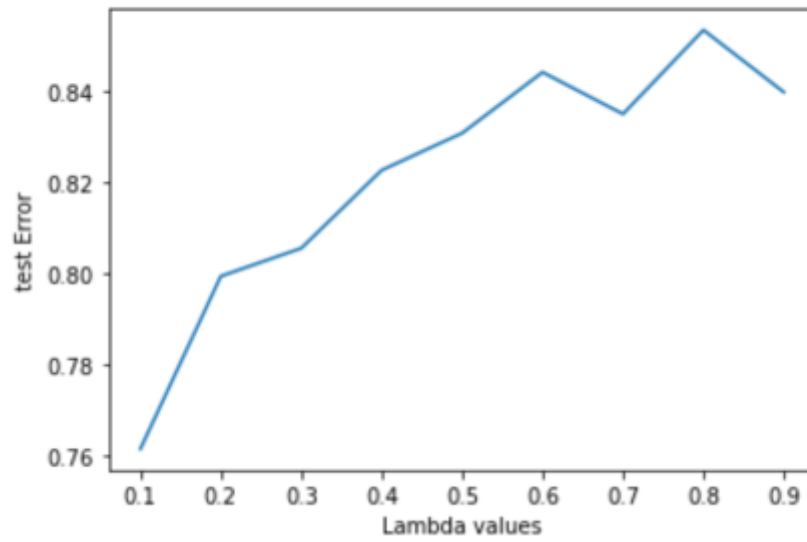
Plot between lambda and test error for Cluster number = 6



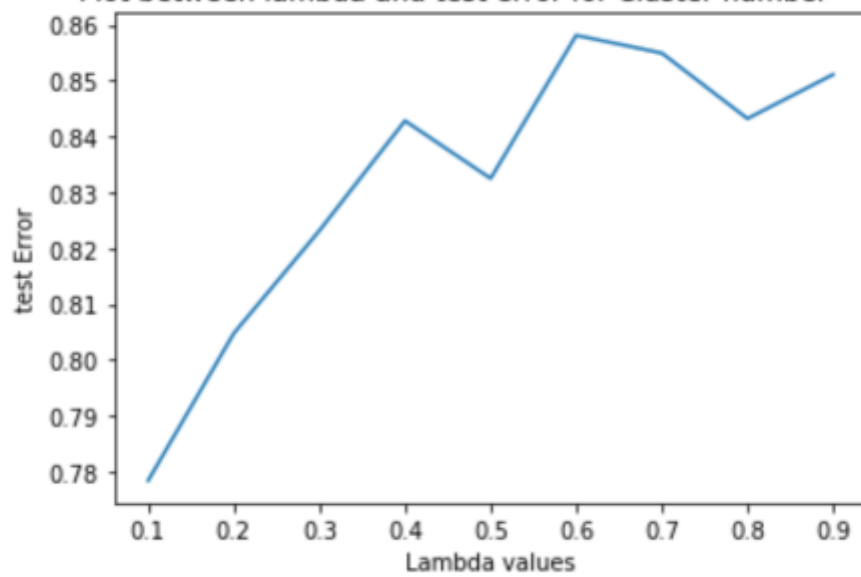
Plot between lambda and test error for Cluster number = 11



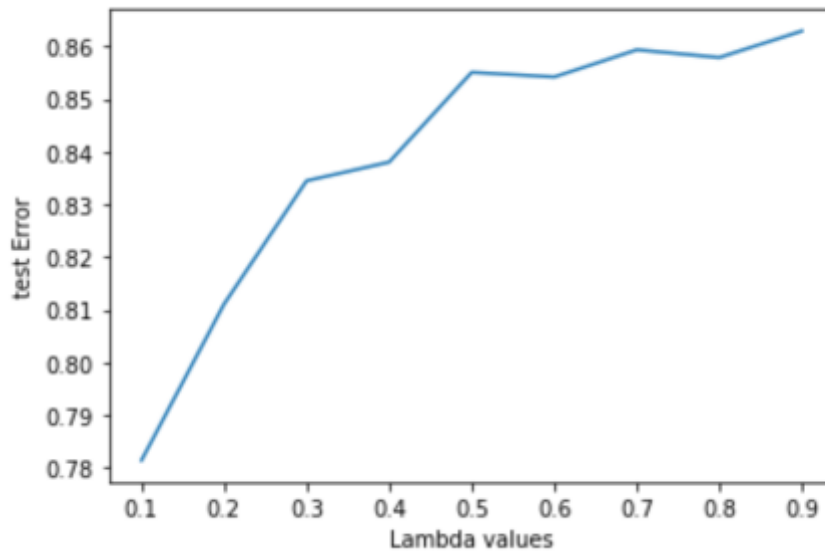
Plot between lambda and test error for Cluster number = 16



Plot between lambda and test error for Cluster number = 21



Plot between lambda and test error for Cluster number = 26



Plot between lambda and test error for Cluster number = 31

