Project Report

Instructor: Sargur N Srihari

Arun Sharma

UB Person: 50206920

Overview

This project is to implement regression on web search ranking data set LETOR 4.0 as well as synthetic dataset. We are given a dataset of vectors and target values. There are 46 features for each sample in the dataset. We need to learn the weight of each feature which can give us the best result given a new sample. We use the technique of regression to learn these weights by using the samples given. We use the Gaussian Model of regression and Stocastic Gradient Descent and vary the degree or complexity of the model until it gives us the closest prediction we can get on a previously unseen keyword.

Dataset:

LeToR Dataset:

Features 46

Synthetic Dataset:

Features 10

Data Partition:

Data partition in both the sets are 80 percent training, 10 percent validation and 10 percent testing. This type for partition gives multiple samples to train out model.

Hyper-parameter Tuning:

Synthetic Dataset

Lamda for Training Set:

Lamda	Training	Validation	Testing
0	**0.78370123**	**0.80672099**	**0.78995315**
0.001	0.78370290	0.80672362	0.78996052
0.01	0.7837179	0.80674703	0.79002649
0.4	0.78430517	0.8074528	0.79238781

For the increase in the value of Lamda, i.e. the regularizer tends to underfit the model and avoid overfitting. The best value in lamda is 0 for synthetic

Value of M:

M	Training	Validation	Testing
8	0.7845099	0.80531023	0.80992794
9	0.7678919	0.87606776	0.85540088
10	**0.78370123**	**0.80672099**	**0.78995315**
11	0.76795661	0.82259386	0.83163069
12	0.7659175	0.83237046	0.82067336
13	0.78354906	0.91178741	0.88837565
14	0.77362844	0.84082739	0.84143535
15	0.76628688	0.83606991	0.8316191
16	0.77169746	0.82478792	0.83436477

M = 10 gives the most optimal value,

M is chosen based on it's the training and validation value. We check the global minimum i.e. whether the value is decreasing and then suddenly increasing in the next value of M.

LetoR Dataset:

Lamda:

Lamda	Training	Validation	Testing
0	0.56409924	0.55374635	0.63994679
0.001	0.56409928	0.55374650	0.63994651
0.01	0.56409964	0.55374778	0.63994406
0.4	0.56411339	0.55379605	0.63986867

Lamda gives the lowest error rate in Training and Testing, hence we choose Lamda as 0 since it gives most optimal value.

Value of M:

M	Training	Validation	Testing
10	0.56409924	0.55374635	0.63994679
11	0.56424387	0.55453878	0.63915331
12	0.56403575	0.5561158	0.63993193
13	0.56383188	0.55444959	0.64216796
14	0.56401028	0.55224101	0.64018118
15	0.5639949	0.55571141	0.64029152
16	**0.56304330**	**0.55527466**	**0.64030006**
17	0.56363322	0.55488564	0.64086109
18	0.56380966	0.5567085	0.64191984

M = 10 gives the most optimal value,

M is chosen based on it's the training and validation value. We check the global minimum i.e. whether the value is decreasing and then suddenly increasing in the next value of M.

Mean:

In this report, mean is calculated by partitioning of dataset into small subsets. However, kmeans was applied but partitioning method gives better results in validation and testing set. We divided the dataset into M partitions and calculated M means and stored into NxM matrix.

Covariance Matrix:

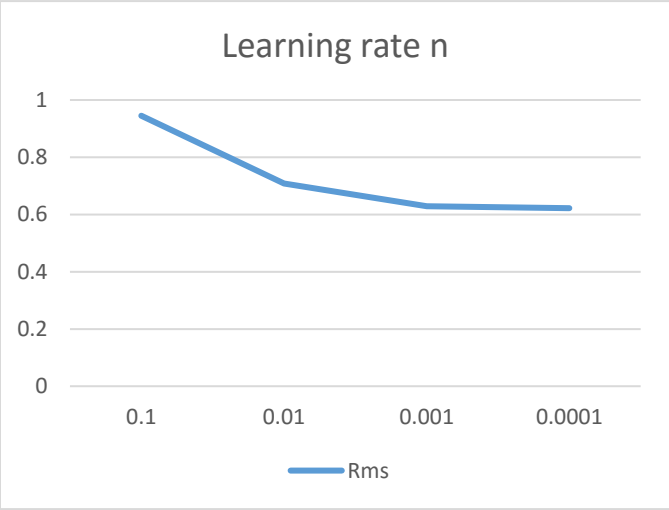
LeToR Dataset:

[0.055068636914840048, 0.065164223770869723, 0.11678013479320339, 0.095597517133717919, 0.055031751951812216, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.076950984554048391, 0.064519464015930492, 0.11764071441479139, 0.08646748548081612, 0.077349972961916605, 0.055864214542558743, 0.05783083656025851, 0.074847959472686063, 0.065229021076474325, 0.055832602598019872, 0.073176353604472527, 0.074814136917517884, 0.083094824920513358, 0.079023304094258895, 0.098502594747719593, 0.1010564851008962, 0.089277190978382742, 0.09850704385739785, 0.11173079479259054, 0.11540185933586918, 0.12829324483488044, 0.11391627148705047, 0.086659016765283134, 0.056233437555229505, 0.062422726579754118, 0.055872265660075179, 0.072232882661776146, 0.073954545250310325, 0.081073921770226653, 0.079054118205832538, 0.039866348694882744, 0.041997269359339055, 0.061995516787576858, 0.070663020407067983, 0.061907243640642193, 1.7953643686200027e-05]

Synthetic Dataset:
[0.0834848730096798, 0.084316895707118755, 0.082853583543096188, 0.084379245384526055, 0.084384214117963027, 0.083363123247479728, 0.08331644761094617, 0.083559162975482201, 0.082415155200886783, 0.083114192920330274]

Covariance matrix was calculated of each column of the dataset and stored in the identity matrix. No additional noise has been added to the matrix since it's pure form gives the best results. Covariance matrix tells us about the spread of the function. To calculate phi, for the synthetic dataset, we used identity matrix since the dataset was small hence there should be a low spread. Identity matrix gives the best result in case of Stocastic Gradient Descent in the synthetic Dataset. (Diagonal matrix is not printed due to limited space)

Learning Rate:



Learning rate value is on the x axis which give value of N. When the value converges and the change in the RMS is not much there then it will eventually stop. The optimal value of learning rate n is 0.0001

Evaluation and Results

Synthetic Dataset

RMS Values

Training	Validation	Testing
0.7845099	0.80531023	0.80992794
0.7678919	0.87606776	0.85540088
0.78370123	**0.80672099**	**0.78995315**
0.76795661	0.82259386	0.83163069
0.7659175	0.83237046	0.82067336
0.78354906	0.91178741	0.88837565
0.77362844	0.84082739	0.84143535
0.76628688	0.83606991	0.8316191
0.77169746	0.82478792	0.83436477

Best RMS Value

M	**Lamda**	**Training Set**	**Validaiton Set**	**Testing Set**
10	**0**	**0.78370123**	**0.80672099**	**0.78995315**

LeToR Dataset:

Training	Validation	Testing
0.56409924	0.55374635	0.63994679
0.56424387	0.55453878	0.63915331
0.56403575	0.5561158	0.63993193
0.56383188	0.55444959	0.64216796
0.56401028	0.55224101	0.64018118
0.5639949	0.55571141	0.64029152
0.56304330	**0.55527466**	**0.64030006**
0.56363322	0.55488564	0.64086109
0.56380966	0.5567085	0.64191984

Best RMS Value

M	**lamda**	**Training Set**	**Validation Set**	**Testing Set**
16	**0**	**0.56304330**	**0.55527466**	**0.64030006**