**CSE 574 Machine Learning**

**Project 1 – Linear Regression**

**(Learning to rank – LETOR Problem)**

Harshith Kumar Ramadev

UB # 50134007

[harshith@buffalo.edu](mailto:harshith@buffalo.edu)

**Overview:**

The agenda of this project is to train a model through linear regression in the context of a web search engine. The search engine would be exposed to a variety of Query-URL pairs as input. Consequently, each of these input pairs must generate unique search results to the end user. This method of generating appropriate search results by training the model is achieved through ranking the features in the feature space. The features constitute a characteristic of data like colour, shape, textual relevance and orientation of the desired search result. In learning to rank problem (LETOR), we rank each of these features as per the relevance to the query-URL pair and teach the model to generate appropriate search results.

The given data set comprises of the relevancy of web pages for a set of query. The relevancy values range from 0 to 2, where 2 denotes the highest relevancy and 0 denotes the least. The data contains 69623 query document pairs, each having 40 non zero features. Initially the given data is parsed to extract only the relevancy labels and the 40 feature values. The data is then split into three portions- Training data set (first 80%), the Validation data set (next 10%) and the Testing data set (remaining 10%).

We need to obtain a linear model that fits the given data and the model used in this project is- Gaussian Linear Basis Function Model. To get the best fit we need to find out the correct value for model complexity ‘M’ such that it does not leads to over fitting (the purpose of the Validation phase) and in order to handle it we apply regularization parameter lambda. The best fit is obtained by minimizing the error function that calculates the misfit between the predicted value and the target value. The metric used to calculate the error is Root Mean Square Error, which also corresponds to the precision of the linear regression model, i.e., lower the ERMS value better the fit.

The three main phases are as follows:

1. Training

The model selected was a Gaussian Linear regression model and it was trained on the training data set (first 80% of the input data). The Gaussian model has two parameters- mean and standard deviation and primarily the mean (m) was modified for different model complexities (M). The result of training is the weight vector (w), which constitute the weights of basis function. The Basis function, in this case Gaussian is the best approach to handle the over fitting in higher degree polynomials.

1. Validation

Here we validate the weight vector ‘w’ that was calculated in the previous step. The validation is one by finding the Root Square Error ERMS. This ERMS is calculated for different model complexities ‘M’, i.e., M is varied from 2 to 9. But to handle over fitting we need to need to also perform Regularization and it is done by looping the regularization parameter ‘lambda’ over 8 different values from 1 to 8. In order to get the best fit, we need to find an ordered pair of M and lambda which gives the least ERMS value. Henceforth, we finalize the values for Lambda and M which corresponds to the lease error rate. The ERMS value is computed to be 0.5574 for the model complexity M=9 and Lambda =8

1. Testing

In the test phase we check if the calculated weight vector ‘w’, the model complexity ‘M’ and the regularization parameter ‘lambda’ is providing the best fit for the test data. The corresponding test error is calculated, and is observed to be ERMS = 0.641 for model complexity M=8 and lambda=4 for the Linear Regression Model with a Gaussian basis function.

We have approached the problem with two methods. They are:

1. Maximum Likelihood Closed form solution
2. Stochastic Gradient solution:

In Maximum likelihood solution, we give a probabilistic approach to solve a given problem of regression using the target values of input data. Here, we use the Basis functions such as Gaussian distribution to determine the best values of M – model complexity, Regularization parameter – Lambda and Design matrix consisting of weights. All these calculations are taken place in training and validation phases. Hence, we plot a normal distribution of the target values over the input data which results in a predictive solution to the regression problem.

Post which, we plot the actual graph of featured target values over the input data. The difference between these graphs will result in Root mean square error of the problem.

In Stochastic gradient solution, we determine the deterministic solution to the regression problem. Here, we start off with a combination of weight vector, Model complexity and Lambda and then fine tune the model as per the Root mean square error differences so as to make the ERMS as small as possible.

The graph of Closed form solution in MLE approach is as shown below:

