BUAN 6341: Applied Machine Learning

Project 1 Report

Linear Regression on Bike Sharing dataset

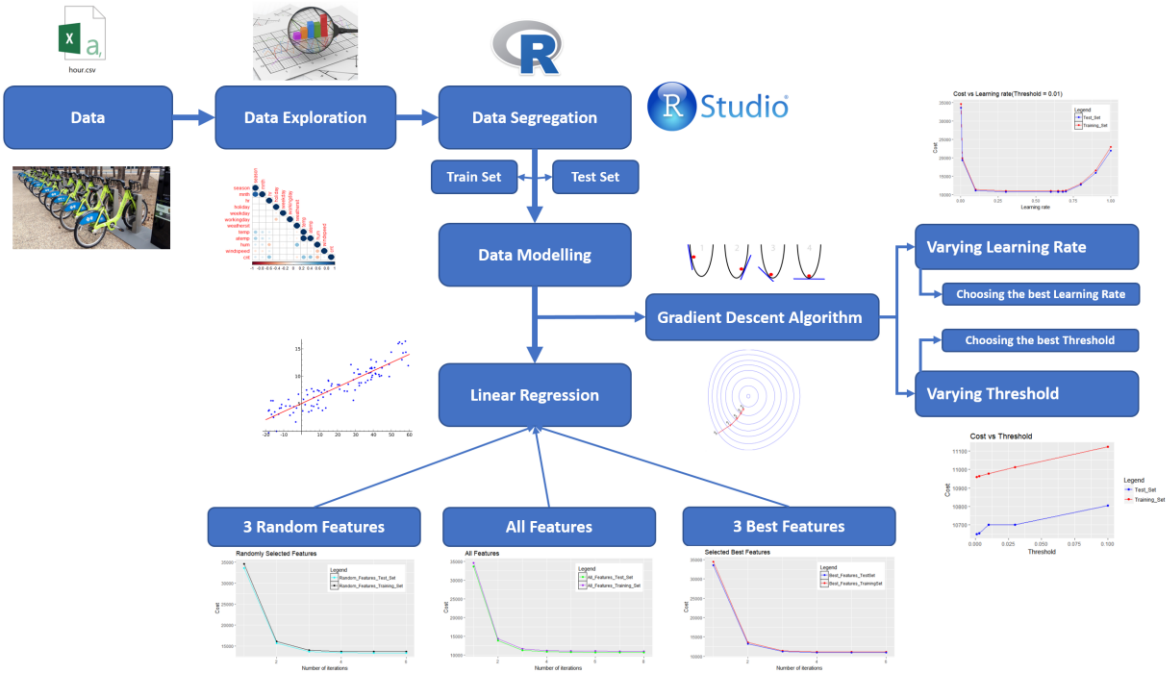
Using Gradient Descent Algorithm

Name: Aji Somaraj
NetID: axs161031

Introduction

The project aims to build a predictive linear regression model to predict the total number of bike rentals during a given hour. This project uses the gradient descent algorithm with batch update i.e. all training observation used at once. The dataset was loaded into R, basic data exploration was done to get the structure and summary of the dataset, feature scaling was done to standardize the range of independent variables. The dataset was divided into a Training set (70%) and a Test set (30%). Then the project experiments with various values of threshold and learning rate (alpha). Patterns are plotted on how the cost function changes for the test and train sets as the values of alpha and thresholds are changed. The project then experiments on selecting three random features from the data set and retraining the model. The results of using three random features and all features were compared. Finally, the project aims to pick the 3 best features and compares the result with the random and all features model.

The workflow of the project is shown below:

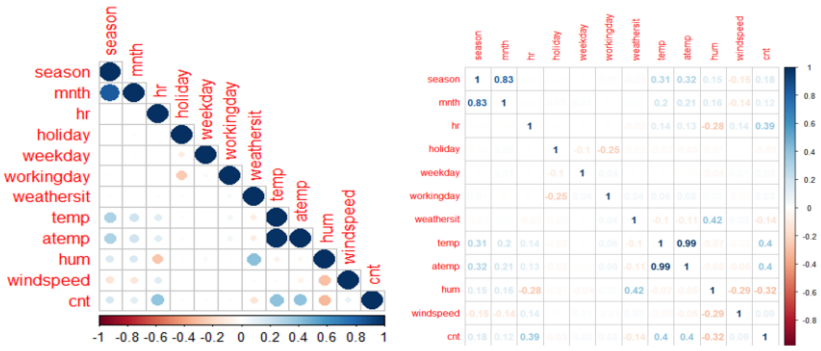


Data Exploration

The structure of the dataset is shown below:

```
> str(data)
'data.frame':   17379 obs. of  12 variables:
 $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ mnth      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ hr        : int  0 1 2 3 4 5 6 7 8 9 ...
 $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ weekday   : int  6 6 6 6 6 6 6 6 6 6 ...
 $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
 $ weathersit: int  1 1 1 1 1 2 1 1 1 1 ...
 $ temp      : num  0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...
 $ atemp     : num  0.288 0.273 0.273 0.288 0.288 ...
 $ hum       : num  0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...
 $ windspeed : num  0 0 0 0 0 0.0896 0 0 0 0 ...
 $ cnt       : int  16 40 32 13 1 1 2 3 8 14 ...
```

There are 11 features or independent variables and 1 dependent variable (cnt : Count of total rental bikes including both casual and registered) and there correlation is shown below:



The aim of the project is to predict the count of total rental bikes and we can see that features like temp, atemp, hr(hour), hum, season has relatively high correlation with the count(cnt), here **temp and atemp** are Normalized temperature in Celsius and Normalized feeling temperature in Celsius respectively since they both are same entities, the correlation between them is very high which gives us the intuition of using one of the temperature values. season and month also have high correlation, so we can include one among them. hum is the Normalized humidity, which has a negative correlation with cnt.

Data Modelling

Linear Regression is used to create the predictive model where we use the gradient descent algorithm with batch update rule. The linear Regression model considering **all features without** considering **correlations** are shown below:

$$cnt \approx \beta_0 X_0 + \beta_1 \times season + \beta_2 \times mnth + \beta_3 \times hr + \beta_4 \times holiday + \beta_5 \times weekday + \beta_6 \times workingday + \beta_7 \times weathersit + \beta_8 \times temp + \beta_9 \times atemp + \beta_{10} \times hum + \beta_{11} \times windspeed$$

At a threshold of $\mathbf{10^{-5}}$ and learning rate of 0.648, the linear regression model is:

$$cnt \approx 190.14 X_0 + 19.426 \times season + 0.69 \times mnth + 51.39 \times hr - 4.14 \times holiday + 3.46 \times weekday + 1.752 \times workingday + 0.214 \times weathersit + 27.79 \times temp + 30.98 \times atemp - 43.61 \times hum + 2.24 \times windspeed$$

Experimentation

1. **Experimenting with various values of learning rate(Threshold=0.01)**

A series of learning rate ranging from 0.001 to 1 was used in experimenting how the error varies with the test and train datasets. The error was maximum when alpha was 0.001 for both test and train data. Then the slope goes down as alpha is increased, reaching a minimum value of **10976.46 at alpha=0.648** for train set and 10672.22 at alpha=0.5 for test set. Gradually the error increases as the alpha is increased further. In sum, at a certain threshold, when the learning rate is low, the convergence is slow, and at certain points

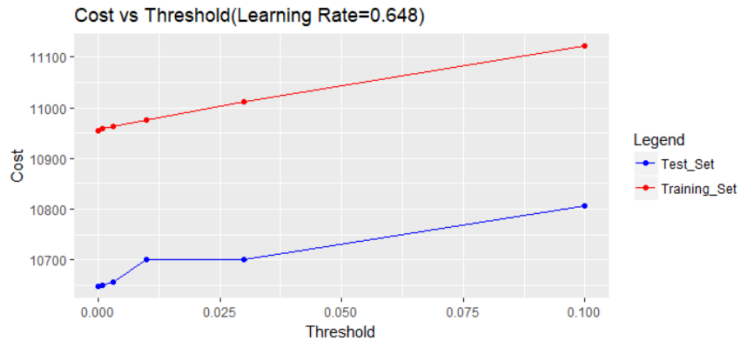
alpha gives minimum values of cost but as it is high the convergence won't occur as the large alpha will increase chances for non-convergence. The plot of cost or error versus the learning rate is shown below:

Alpha	Train Error	Test Error
0.001	34563.44	33600.89
0.01	19919.67	19368.35
0.1	11406.57	11078.14
0.3	11031.99	10716.92
0.4	11009.7	10695.43
0.5	10985.23	10672.22
0.6	10986.85	10674.24
0.648	10976.46	10700.47
0.65	10977.68	10703.08
0.68	11012.54	10705.81
0.7	11065.13	10760.83
0.8	12954.46	12629.24
0.9	16583.94	15939.37
1	22919.12	21896.18

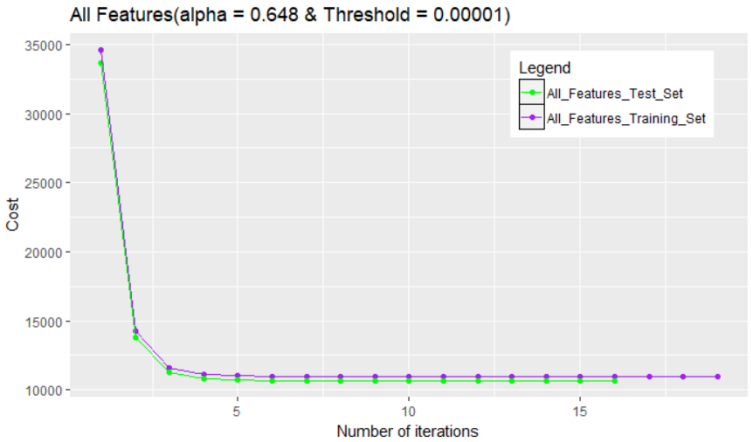
Since the minimum cost or error is observed when learning rate is **0.648** for the train set so we will choose this alpha as **the best alpha**. As values which are above and below 0.678 tends to increase the error or cost.

2. Experimenting with various Thresholds (Learning rate =0.648)

Thresholds varying from 10^{-6} to 0.1 were used in this study at a learning rate of 0.648. As threshold increases the error increases for both the test and train datasets. The minimum error observed and plot of error vs threshold is shown:



Threshold	Num_iters	Train Error	Num_iters	Test Error
0.000001	94	10950.9	27	10645.35
0.00001	19	10952.56	16	10645.72
0.0001	11	10954.28	10	10647.23
0.001	8	10958.52	8	10650.39
0.003	7	10963.72	7	10655.3
0.01	6	10976.46	5	10700.47
0.03	5	11011.15	5	10700.47
0.1	4	11122.52	4	10805.5

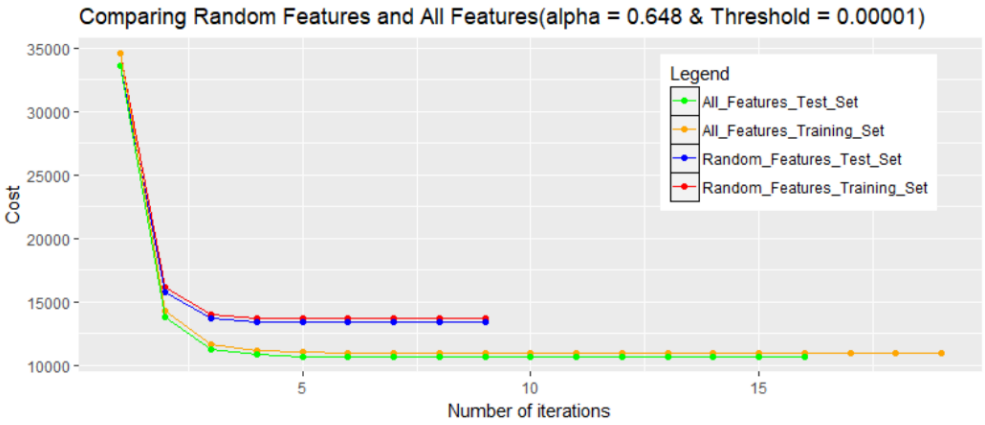


For a constant learning rate as the Threshold is decreased the error decreases, at the same time iterations increases. Here we consider $\mathbf{10^{-5}}$ as **best** even though the cost decreases further, as the change of error is very low.

Using threshold=$\mathbf{10^{-5}}$ and alpha= **0.648** we plot the cost function versus the number of iterations. We can see that train converges to minimum cost at the 19th iteration (**10952.56**) and test at the 16th iteration (**10645.72**)

Num_Iters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
All_Features_Train Error	34621.42	14257.02	11588.34	11122.52	11011.15	10976.46	10963.72	10958.52	10956.17	10954.97	10954.28	10953.83	10953.5	10953.25	10953.06	10952.9	10952.77	10952.66	10952.56
All_Features_Test Error	33657.11	13805.74	11245.47	10805.5	10700.47	10667.46	10655.3	10650.39	10648.26	10647.23	10646.68	10646.34	10646.11	10645.94	10645.82	10645.72			

Experiment 3: Choosing Three Features Randomly



Three features were selected randomly using the sample function they were weathersit, season, atemp. Model was retrained using these features with an alpha of **0.648** and threshold of **10^{-5}.** Since the features was selected randomly they show less correlation with the output variable cnt. Hence the random features model, both the train and test set converges at the 9th iteration with a **cost of 13663.7 and 13398.67** respectively

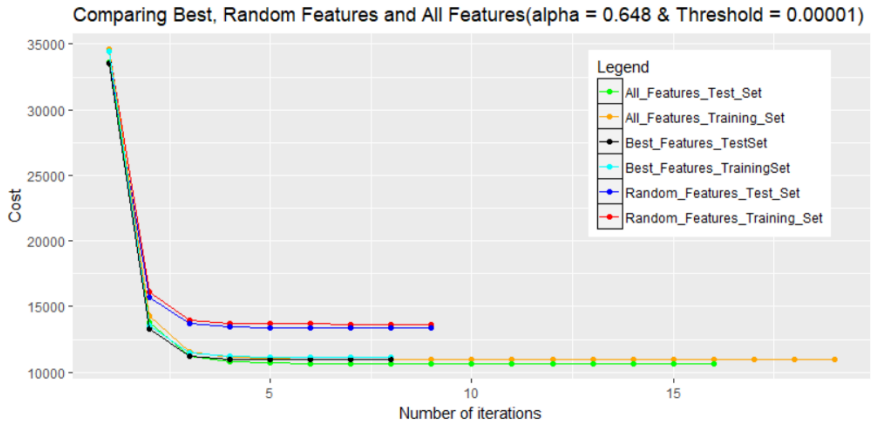
$$cnt \approx \beta_0 X_0 + \beta_1 \times weathersit + \beta_2 \times season + \beta_3 \times atemp$$

$$cnt \approx 190.09 X_0 - 17.65 \times weathersit + 9.57 \times season + 69.24 \times atemp$$

Num_Iters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
All_Features_Train Error	34621.42	14257.02	11588.34	11122.52	11011.15	10976.46	10963.72	10958.52	10956.17	10954.97	10954.28	10953.83	10953.5	10953.25	10953.06	10952.9	10952.77	10952.66	10952.56
All_Features_Test Error	33657.11	13805.74	11245.47	10805.5	10700.47	10667.46	10655.3	10650.39	10648.26	10647.23	10646.68	10646.34	10646.11	10645.94	10645.82	10645.72			
Random_Features_Train_Error	34591.37	16110.98	13996.13	13715.58	13673.55	13665.99	13664.29	13663.84	13663.7										
Random_Features_Test_Error	33628.36	15724.19	13703.23	13442.9	13406.24	13400.26	13399.06	13398.76	13398.67										

Experiment 4: Selecting three best features

3 features were selected based on the correlation plot. Features: **hr, atemp, hum** was selected since it was correlated the most with the dependent variable cnt and also, they are not correlated with each other The selected best model was compared with random features model and all features model taking an alpha of 0.648 and threshold of 0.00001, the results are shown below



$$cnt \approx \beta_0 X_0 + \beta_1 \times hr + \beta_2 \times atemp + \beta_3 \times hum$$

$$cnt \approx 190.013 X_0 + 51.417 \times hr + 65.08 \times atemp - 41.433 \times hum$$

Num_Iters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
All_Features_Train Error	34621.42	14257.02	11588.34	11122.52	11011.15	10976.46	10963.72	10958.52	10956.17	10954.97	10954.28	10953.83	10953.5	10953.25	10953.06	10952.9	10952.77	10952.66	10952.56
All_Features_Test Error	33657.11	13805.74	11245.47	10805.5	10700.47	10667.46	10655.3	10650.39	10648.26	10647.23	10646.68	10646.34	10646.11	10645.94	10645.82	10645.72			
Random_Features_Train_Error	34591.37	16110.98	13996.13	13715.58	13673.55	13665.99	13664.29	13663.84	13663.7										
Random_Features_Test_Error	33628.36	15724.19	13703.23	13442.9	13406.24	13400.26	13399.06	13398.76	13398.67										
Selected_Best_Features_Train_Error	34484.86	13555.85	11439.78	11185.96	11154.45	11150.5	11150	11149.94											
Selected_Best_Features_Test_Error	33519.97	13258.92	11218.89	10972.67	10942.12	10938.31	10937.83	10937.77											

In the selected best features model the cost converges at the **8th iteration** for both train and test sets at **11149.94** and 10937.77 respectively. The **selected best features model** was performing **better than the random features model** because the features selected was based on correlation with cnt hence those features could explain the model better than the randomly selected features. But the selected features model **was not able to explain the model better than the all features** as **more number of features could explain** the scenario of estimating the count of bikes **better than 3 features.**

Discussion

In experiment 1 we tested, how error changes at a fixed threshold when the alpha is changing. We understood that alpha should neither be too small nor too large, it should be at a point in between where cost is minimum.

In Experiment 2 we tested, how the error changes at a fixed alpha when the threshold is changing. As the threshold decreases the error decreases, but it was observed that at very low thresholds the number of iterations was found to be very high therefore at low thresholds the model takes more time for convergence. So, a suitable threshold should be selected based on alpha.

In Experiment 3 and 4 we retrained the model with 3 random features and 3 best selected features. The 3 best selected features model was better than the random features model but when compared to all features model the results was not that good. These observations remind us the importance of feature selection as the performance of the model depends on its features.

The important things that matter in predicting the number of bike rentals are (1) A proper learning rate should be selected (2) A threshold should be chosen in such a way that the cost or error should be minimized in less number of iterations. (3) The features should be selected based on the correlation (here (atemp, temp) and (season, month) had high correlation so anyone of these could have been used.) and using methods like backward elimination, intuition based on prior knowledge. (4) Feature Scaling: all independent variables are scaled similarly.

To improve results (1) On further study of the bike sharing scenario, we could have derived more new features like demographic details, financial variables like income, market etc., (2) We could have used algorithms like Principal Component Analysis(PCA) or Linear Discriminant Analysis (LDA) to reduce the dimensionality and get components that reflected most characteristics of the features, and use those components as independent variables for the model. (3) Feature Engineering, creating new variables in addition to the existing variables. These variables will improve the predictability of the model.

The importance of feature selection can be explained in this model by choosing 7 features **hr','atemp','hum','season','holiday','weekday',"windspeed"** rather than using 11 features. These features were selected based on the correlation plots and feature elimination methods (i.e. using all features initially and then eliminating them based on the reduced cost or error values).

$$cnt \approx \beta_0 X_0 + \beta_1 \times hr + \beta_2 \times atemp + \beta_3 \times hum + \beta_4 \times season + \beta_5 \times holiday + \beta_6 \times weekday + \beta_7 \times windspeed$$

$$cnt \approx 190.09 \times X_0 + 51.41 \times hr + 58.51 \times atemp - 43.664 \times hum + 20.36 \times season - 4.468 \times holiday + 3.66 \times weekday + 3.387 \times windspeed$$

Num_Iters	All_Features_Train Error	All_Features_Test Error	Final_Selected_Train_Error	Final_Selected_Test_Error
1	34621.42	33657.11	34528.74	33565.66
2	14257.02	13805.74	13298.15	12922.54
3	11588.34	11245.47	11242.7	10936.96
4	11122.52	10805.5	10991.08	10696.6
5	11011.15	10700.47	10958.71	10666.46
6	10976.46	10667.46	10954.32	10662.6
7	10963.72	10655.3	10953.67	10662.09
8	10958.52	10650.39	10953.56	10662.02
9	10956.17	10648.26	10953.53	
10	10954.97	10647.23		
11	10954.28	10646.68		
12	10953.83	10646.34		
13	10953.5	10646.11		
14	10953.25	10645.94		
15	10953.06	10645.82		
16	10952.9	10645.72		
17	10952.77			
18	10952.66			
19	10952.56			

We can see that using **7 features**, the cost converged to **10953.53** in the **9th iterations** but the all features model converged to **10952.56** at the **19th iterations** which gives us the impression that if the features are selected correctly with suitable alpha and threshold the model will perform better than using all features.