APPLIANCES ENERGY PREDICTION USING LINEAR AND LOGISTIC REGRESSION

SUDHARSANA RAJASEKARAN

APPLIED MACHINE LEARNING

Assignment 1

Table of Contents

Introduction:

The Appliances Energy prediction dataset was derived from the UCI Machine learning repository. The dataset consists of house temperature and humidity conditions that were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods, which then becomes each of the rows in the dataset. The aim of the project is predicting the energy usage of the appliances using Linear and Logistic Regression by hyperparameter tuning of parameters.

Exploratory Data Analysis:

Initial dataset analysis yielded the following results,

1. Number of Transactions: The dataset contains 19735 observation and 29 variables.
2. Missing Values: There are no missing values in the dataset.
3. Correlation Plot: The correlation plot tells us how far the dependent variables are correlated. So, can avoid multicollinearity, which greatly the standard errors of the coefficients.

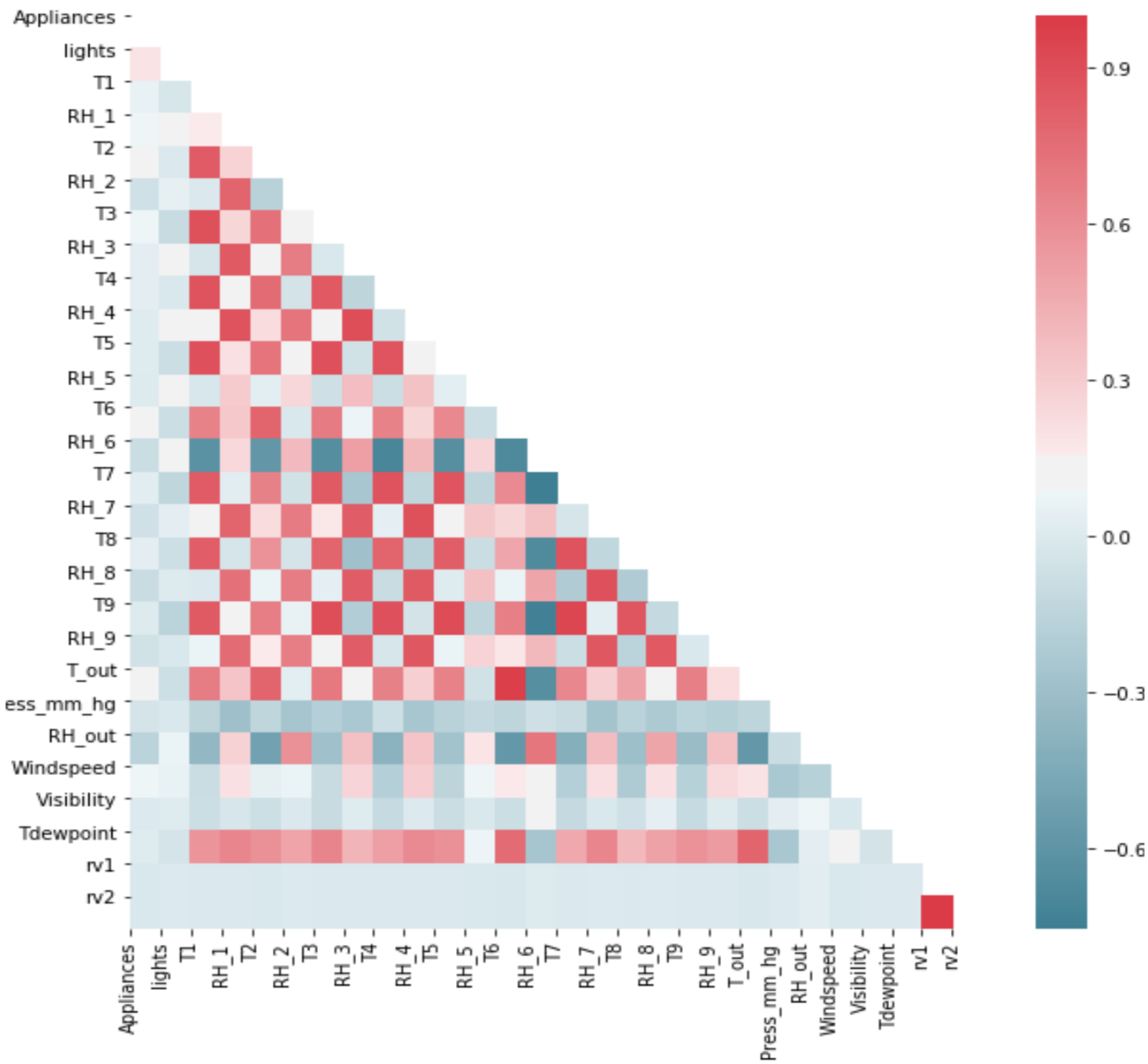

Figure 1. Correlation Plot

It can be observed from the correlation plot that Temperatures are highly positively correlated with one another. The humidities are negatively correlated with one another.

Experimentation:

1. Changing Learning Rates:

a. Effect of change of Learning Rate on Linear Regression:

Linear Regression Equation:

Appliances = $\beta_0 + \beta_1$ lights + β_2 T1 + β_3 RH_1 + β_4 T2 + β_5 RH_2 + β_6 T3 + β_7 RH_3 + β_8 T4 + β_9 RH_4 + β_{10} T5 + β_{11} RH_5 + β_{12} T6 + β_{13} RH_6 + β_{14} T7 + β_{15} RH_7

Appliances = (96.63) + (20.43) lights + (-2.53) T1 + (20.28) RH_1 + (6.31) T2 + (-10.80) RH_2 + (15.52) T3 + (8.40) RH_3 + (-10.45) T4 + (-3.21) RH_4 + (-8.1) T5 + (0.46) RH_5 + (8.21) T6 + (-5.72) RH_6 + (-3.08) T7 + (-19.65) RH_7

The effect of on Training/Test error due to different learning rates is studied below,

Learning Rate	Cost	Iterations	Train RMSE	Test RMSE
0.1000	4382.102350	1489	93.617331	99.326507
0.0500	4382.386870	2585	93.620370	99.334039
0.0100	4384.656597	8362	93.644611	99.376386
0.0050	4387.470375	12820	93.674654	99.418631
0.0010	4404.050942	27659	93.851488	99.612131
0.0005	4415.623708	38901	93.974717	99.722800

Figure 2. Effect of change of Learning rate on Error

It can be observed that, as the learning rate decreases,

- The minimum number of iterations before the algorithm convergence increases. This is because the step size decreases taken by the algorithm decreases taking a much longer time to converge.
- The cost function increases as the algorithm converge before it can reach the global minimum on account of smaller step size.
- As a result of increased cost, the difference between actual and the predicted increases resulting in increased training and test errors.

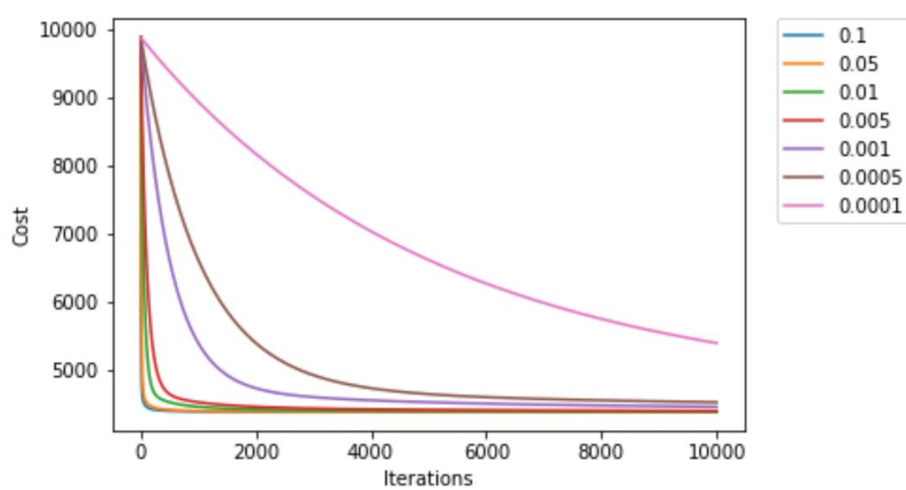

Figure 3. Effect of change of learning rate on cost

b. Effect of Change of Learning Rate on Logistic Regression:

Appliances = (-0.51) + (0.24) lights + (0.04) T1 + (0.11) RH_1 + (0.09) T2 + (-0.06) RH_2 + (0.03) T3 + (0.02) RH_3 + (-0.02) T4 + (-0.0008) RH_4 + (-0.02) T5 + (0.05) RH_5 + (0.09) T6 + (-0.08) RH_6 + (-0.02) T7 + (-0.12) RH_7

The effect of on training/test error due to different learning rates is studied below,

Learning Rate	Cost	Iterations	Train Accuracy	Test Accuracy
0.5000	0.467126	375	73.587535	78.008614
0.1000	0.471902	725	73.587535	77.324550
0.0500	0.474860	1033	73.587535	76.969851
0.0100	0.489273	2028	73.587535	75.728401
0.0050	0.499656	2571	73.587535	75.475044
0.0010	0.534134	5028	73.587535	74.638966
0.0005	0.562220	6117	73.587535	74.436281

Figure 4. Effect of change of Learning rate on Error (Logistic Regression)

It can be observed that, as the learning rate decreases,

- Like Linear Regression, the minimum number of iterations before the algorithm convergence increases.
- The cost function increases as the algorithm converges before it can reach the global minimum due to smaller step size.
- As a result of increased cost, the difference between actual and the predicted increases resulting in decreased test accuracy.

2. Changing Convergence Threshold

Convergence Threshold is the stopping criteria for us to understand has the algorithm has already taken necessary number steps for convergence.

The effect of on training/test error due to different convergence thresholds is studied below,

Learning Rate	Convergence Threshold	Cost	Iterations	Train RMSE	Test RMSE
0.10	0.10000	4403.964024	276	93.850562	99.611444
0.10	0.01000	4384.642982	836	93.644466	99.376171
0.10	0.00100	4382.102350	1489	93.617331	99.326507
0.10	0.00010	4381.847663	2143	93.614611	99.316292
0.10	0.00001	4381.822230	2796	93.614339	99.313631
0.05	0.10000	4415.476684	389	93.973152	99.721508
0.05	0.01000	4387.456207	1282	93.674502	99.418440
0.05	0.00100	4382.386870	2585	93.620370	99.334039
0.05	0.00010	4381.876144	3893	93.614915	99.318050
0.05	0.00001	4381.825071	5201	93.614369	99.314121
0.01	0.10000	4465.424336	843	94.503168	100.202814
0.01	0.01000	4404.046649	2765	93.851443	99.612105
0.01	0.00100	4384.656597	8362	93.644611	99.376386
0.01	0.00010	4382.103429	14900	93.617343	99.326539
0.01	0.00001	4381.847803	21442	93.614612	99.316302

Figure 5. Effect of Convergence Threshold on Error

It can be observed that, as the learning threshold decreases,

- The minimum number of iterations before the algorithm convergence increases. As the threshold decreases, the level of difference between two cost function decreases, increasing number of iterations.
- The cost function decreases as the algorithm takes large number of steps before it converges. This is due to increase number of iterations/steps taken with decrease in threshold.
- As a result of decreased cost, the difference between actual and the predicted decreased resulting in training/test errors getting decreased. Hence the low Training/Test error with decrease in convergence threshold.

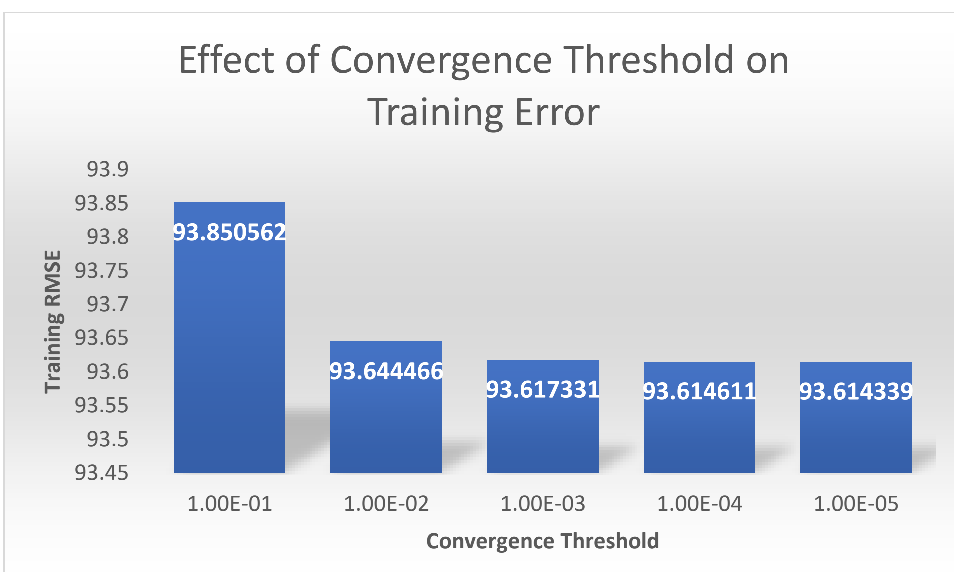

Figure 6. Effect of Convergence Threshold on Training Error

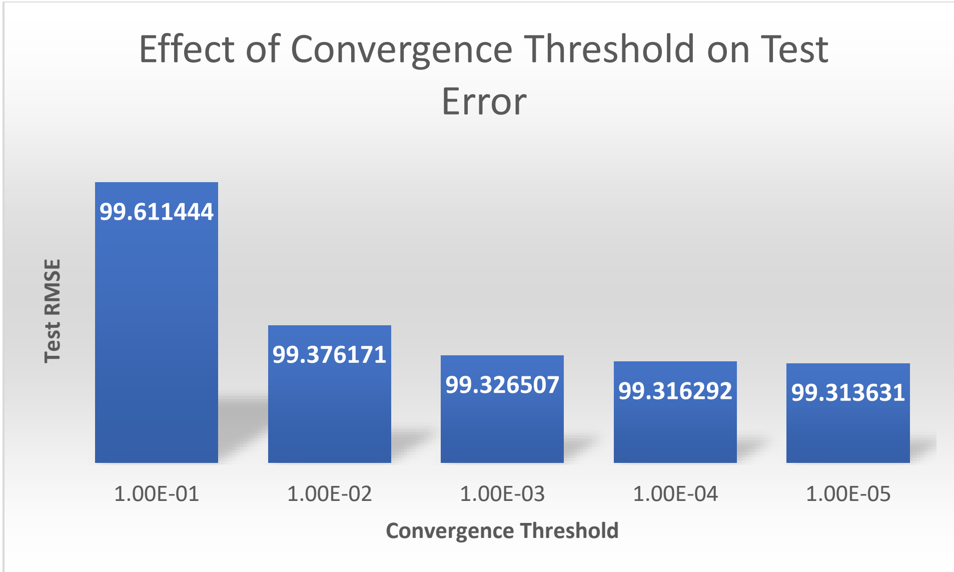

Figure 7. Effect of Convergence Threshold on Test Error

As the number of iterations increases, the training/test error increases this is because with each increase in the number of steps that leads to opposite to global minimum increases.

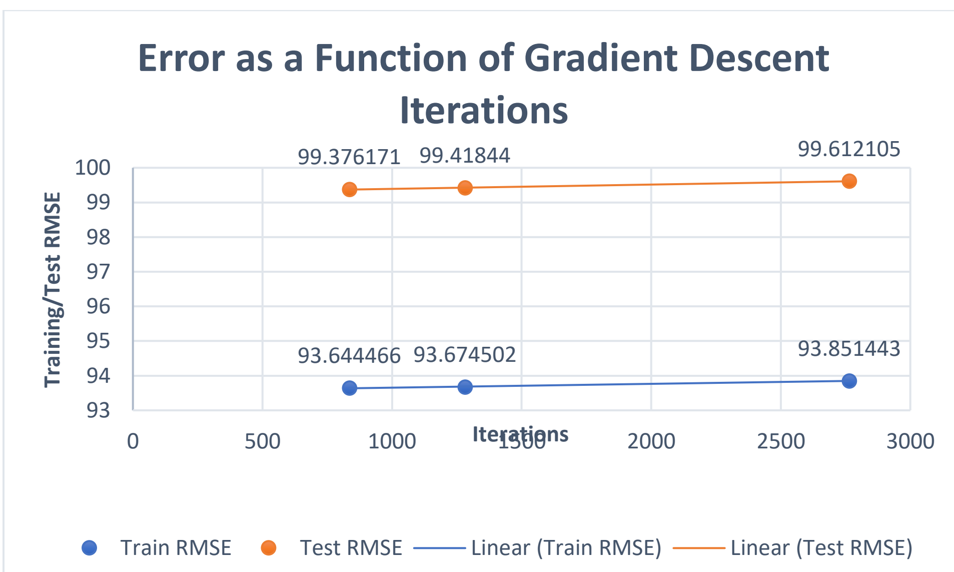

Figure 8. Error as a function of Number of Iterations

3. Changing Number of Variables:

Appliances = $\beta_0 + \beta_1$ T1+ β_2 RH_1 + β_3 T2 + β_4 RH_4 + β_5 T5 + β_6 T7 + β_7 Windspeed + β_8 Visibility+ β_9 Tdewpoint + β_{10} rv1

Appliances = (97.28) + (4.73) T1+ (37.63) RH_1 + (23.40) T2 + (-33.34) RH_4 + (-12.23) T5 + (-5.15) T7 + (8.70) Windspeed + (1.01) Visibility+ (-9.61) Tdewpoint + (-0.95) rv1

The effect of on training/test error due to different number of variables is studied below,

Learning Rate	Cost	Iterations	Train RMSE	Test RMSE
0.1000	4382.102350	1489	93.617331	99.326507
0.0500	4382.386870	2585	93.620370	99.334039
0.0100	4384.656597	8362	93.644611	99.376386
0.0050	4387.470375	12820	93.674654	99.418631
0.0010	4404.050942	27659	93.851488	99.612131
0.0005	4415.623708	38901	93.974717	99.722800

Figure 9. Effect on Error as function of Number of variables (15 Variables)

Learning Rate	Cost	Iterations	Train RMSE	Test RMSE
0.10	4834.904951	459	98.335192	103.329286
0.05	4834.971656	827	98.335870	103.330025
0.01	4835.490650	3091	98.341148	103.337312

Figure 10. Effect on Error as function of Number of variables (10 Variables)

It can be observed that, as the number of variables decreases,

- The minimum number of iterations before the algorithm convergence decreases. This is because the more variables the more it takes to reach the global minimum.
- The cost function decreases as the cumulative error from a smaller number of variables is less than its counterpart with a greater number of variables.
- Training/test errors increases, as because the more the variables the more the variation explained by explained by the variables hence the prediction is as close to actual with more variables.

4. Selection of Important Features Vs Randomly selected features

Appliances = β_0 + β_1 RH_8+ β_2 rv2+ β_3 rv1 + β_4 RH_5 + β_5 RH_9 + β_6 T6 + β_7 RH_out + β_8 Tdewpoint + β_9 Press_mm_hg + β_{10} RH_6

Appliances = (97.28) + (-21.09) RH_8+ (-0.34) rv2+ (-0.34) rv1 + (5.27) RH_5 + (3.09) RH_9 + (23.53)T6 + (-8.69) RH_out + (-2.82) Tdewpoint + (-3.10) Press_mm_hg + (20.35) RH_6

The effect of on training/test error due with inclusion of important variables is shown below,

Learning Rate	Cost	Iterations	Train RMSE	Test RMSE
0.10	4919.860754	687	99.195370	104.449000
0.05	4920.038002	1129	99.197157	104.453011
0.01	4921.362329	2921	99.210507	104.470245

Figure 11. Effect on Error as function of Important variables (10 Variables)

It can be observed that, as the inclusion of important variables,

- Training/test errors increases, as because the important variables are separated on different planes hence the increased test and training errors.
- The Error for important features are more than randomly selected features.