Fitness Analysis of Electricity Consumption Regression Based on Weekends and Weekdays

Erjia Guan
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
eguan@andrew.cmu.edu

Yiming Chen
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
yimingc1@andrew.cmu.edu

ABSTRACT
The objective of this project is to evaluate the fitness of regression tree towards the database of DRED[1]. And, by comparing the cross-validation scores of different sub datasets refined by weekends or clusters or two features, the feature of weekend or weekday is not valid in this case for the regression analysis. And, of course, the cluster obviously demonstrates the best performance in regression.

Keywords
Regression Tree; Feature; Weekends; Cluster; Cross validation.

1. INTRODUCTION
Since lots of data analysis for electricity consumption have been done, in terms of the feature of weekend or weekday. We intend to use the DRED database to discuss whether it's an appropriate feature for electricity consumption to separate the whole dataset into several more uniform sub datasets, which means that we are able to make more precise regression analysis for electricity consumption.

If the result of our project demonstrates that weekend or weekday is a good feature for electricity consumption, the previous data analysis is much more reliable. However, if it shows opposite result, the previous work may have another more efficient feature to do the regression analysis.

2. DATASETS
The datasets which is called Dutch Residential Energy Dataset used in this project are provided by a group of Embedded Software in Delft University of Technology in Netherland. This dataset mainly focuses on providing details on the deployment of sensors that monitor energy consumption of a household in the Netherlands. The deployment consists of several sensors measuring electricity power, occupancy and environmental facts in a household. The data was collected in 1 Hz or 1/60 Hz over a period of 6 months from 5th July to 15th December 2015. This dataset includes:

- Electricity monitoring - aggregated energy consumption and appliance level energy consumption
- Environmental information - room-level indoor temperature, outdoor temperature, environmental parameters (wind speed, humidity and precipitation).
- Occupancy information - room-level location information of occupants, Wi-Fi and BT RSSI information for localization.
- Household information - house layout, number of appliance monitored, appliance - location mapping, etc.

3. DATA SELECTION
The mainly part of the data being used in this project is the outdoor temperature and the aggregated energy consumption. The temperature dataset is sampled in a 1-minute interval from July 20th, 2015 to December 15th, 2015. And the power consumption is sampled in the 1-second interval from July 5th, 2015 to December 5th, 2015.

At first, we noticed that some data was missing or invalid. For instance, there were some temperature data changing fiercely in 1 minute. We just use an average window of 1 hour to decrease the influence of these invalid data. So after resampling the dataset at the frequency of 1 hour, we acquire average outdoor temperature and average aggregated electricity consumption. In terms of the dataset shape, the Null data points in the dataset are not consecutive so that we just interpolated them with the same time interval.

After that, we join two datasets internally for temperature and power consumption. So our prepared data is from July 20th, 2015 20:00:00 to 29th November 2015 14:00:00. Then we add the day of the week, hour and date of each data point for further implementation. So the attribute of the data points is the outdoor temperature, aggregated power consumption, the day of week, hour and date.

4. DATASET PATTERN
Before implement the approaches for analysis the dataset, it's important for us to figure out the pattern of our dataset.

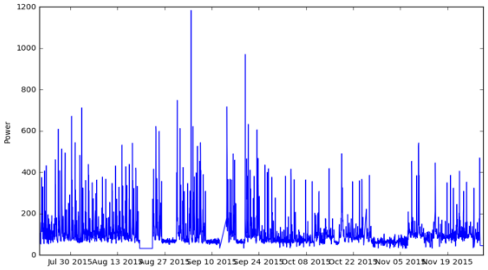

Figure 1. Dataset of Time Domain

The first thing we want to figure out is the distribution of the whole dataset as shown in Figure 1, Figure 2. In Figure 1, the

[1] http://www.st.ewi.tudelft.nl/~akshay/dred/

power consumption is demonstrated in the whole time domain of the dataset and implies that the higher power consumption mainly lies between August and September. Also, there is a period of notably low power consumption at the end of the October to about 5th November, which means several potential periods that the host of this apartment may leave.

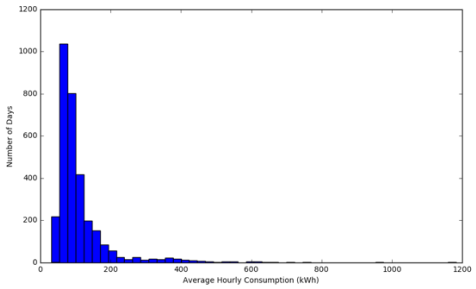

Figure 2. Distribution of Dataset

Also, in Figure 2, it exhibits that the distribution of the whole dataset. And the majority data points is in the area that the power consumption is from 0 to 200 W, because it is a small apartment and the aggregated data may not include all appliances.

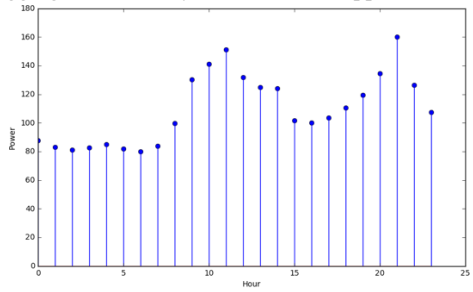

Figure 3. Daily Pattern (24 Hours Average Electricity Consumption)

In Figure 3, when we convert the data distribution into one day, it clearly displays the pattern of electricity consumption in 24 hours that there are two peaks of the power consumption at noon around 13:00 PM and around 21:00 PM, which are cooking time and the leisure time such as watching TV and playing computer games.

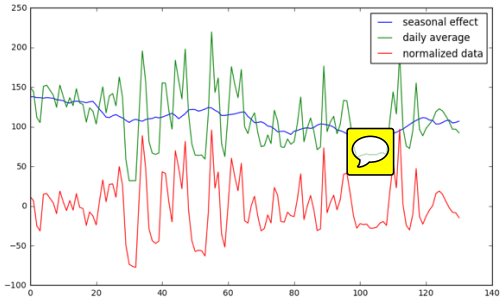

Figure 4. Seasonal Effect of Dataset and Normalized Data (131 Days)

Besides, since we are interested in the relation between the feature weekday and weekend, after extracting the seasonal influence on the entire dataset by the average window, the pattern of the dataset is much more apparently shown in Figure 4 and Figure 5. The peaks of daily electricity consumption happen at around 10:00 AM and 19:00 PM.

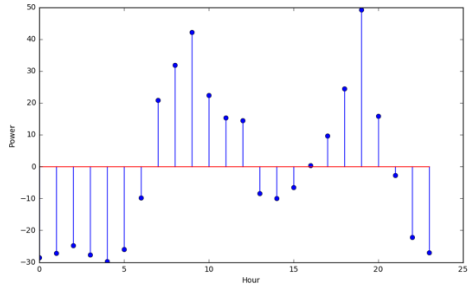

Figure 5. Daily Pattern of Normalized Data

5. THEOREM

The methodology being used in this project is k-means cluster and regression tree and cross validation.

5.1 K-means cluster

The algorithm aims at minimizing the objective function known as squared error function which is implemented in the Python library sklearn given by: [2]

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\| x_i - v_j \| \right)^2$$

where,

$\|x_i - v_j\|$ is the Euclidean distance between x_i and v_j.

c_i is the number of data points in ith cluster.

c is the number of cluster centers.

Algorithmic steps for k-means clustering:

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers. And in our project each x has the 24 data points of the day. So the following steps is

1) Randomly select c cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using: [4]

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, c_i represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

5.2 Regression tree model

Because we intend to obtain the multiple output, the model construction for our target follows the equation below which is adequately illustrated in [6]:

$$(x_k, t^k) = \arg\min \sum_{\{i | x_i \in R_L\}} \| y_i - \overline{y_L} \|_l^2 + \sum_{\{i | x_i \in R_R\}} \| y_i - \overline{y_R} \|_l^2$$

$x_i := [x_i^1, ..., x_i^n]^T \in R^n$,

Where $y_i := [y_i^1, ..., y_i^p]^T \in R^p$, $y_i \in R$ and where $y_i \in R$ and y_L $i \in \{1,2,...., N\}$.

and y_R are the mean outputs of all the data points in R_L and R_R, respectively.

The optimal split at each node of the regression tree is then determined by minimizing the sum of mean square error in both the branches, which is built in sklearn function.

5.3 Cross validation

The basic principle in k-fold cross-validation is that the original sample is randomly partitioned into *k* equal sized subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k − 1* subsamples are used as training data. [5]

6. PROPOSED APPROACH

Since we want to figure out whether it's a better way to fit the dataset in terms of weekends and weekdays or clustered by the inner program. We will discuss the reliability of regression by four different sets of this dataset.

- The entire dataset ①
- Two datasets separated by weekday and weekend ②
- Two datasets separated by clusters ③
- The whole dataset with two features(Temperature and Weekend) ④

First, we set the maximum depth of all regression trees as 5. And by comparing the cross-validation score of weekday regression tree with the score of the entire dataset, we are able to say whether the dataset only including weekday data has more uniform pattern or not.

Then by comparing the scores above with the score of a two-feature dataset, we can figure out whether it's a better way for regression to implement another feature.

Finally, we change the max depth of the second and the third datasets to 4, which means we manually build the first level of the entire dataset by weekdays or cluster. And the depth of these trees is similar as the first one. Therefore, we are able to compare these scores to figure out which is the best way for making regression towards the data of this apartment.

7. RESULTS

In Figure 6 below, the pattern of daily electricity consumptions in weekends (the second graph) and the pattern of daily electricity consumptions in the weekday (the third graph) are extracted from the entire dataset.

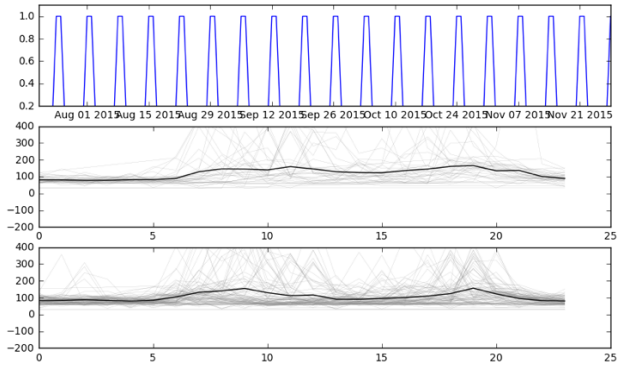

Figure 6. The Pattern of Separated Dataset in Weekends or Weekdays

Two clusters are extracted from the dataset by the K-Means cluster function above, which is the internal function embedded in the python sklearn module. And the patterns of daily electricity consumptions are shown in Figure 7 below.

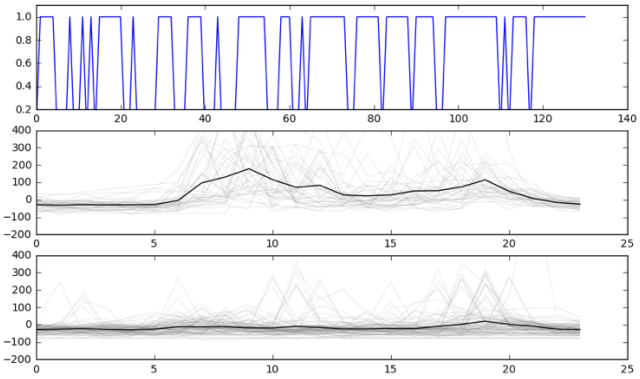

Figure 7. The Pattern of Separated Dataset of Clusters

It's relatively clearly that the difference between the patterns of two clusters is more obvious than the difference between the patterns of weekends and weekdays. So, hypothetically speaking, we think it may not be the best way to separate the entire dataset by weekends and weekdays.

In order to go further to theoretically inspect the results above, we build the regression trees for different datasets and set the maximum depth for all the trees as 5, the theory for which has been explained above. Then, we put our trees, features and daily patterns of electricity consumption into the cross-validation score, which are shown below. In terms of the limited number of data, we set the number of folds as 4 to calculate the cross-validation scores.

Table 1. Cross-validation Scores for All Datasets

	①	②		③		④
	The Entire Dataset	**Weekend Dataset**	**Weekday Dataset**	**Cluster 1**	**Cluster 2**	**Two Feature Dataset**
Number	131	37	94	40	91	131
Cross Validation Score	-1.016755108	-1.567172923	-1.368722625	-0.734156615	-0.297475512	-1.18781438

The potential reason for negative scores is that the entire dataset only has 131-pair data, which will lead to the regression tree over fitting our dataset.

However, the maximum score still corresponds the best fitting regression tree. It's obviously that the best regression tree is the third dataset derived from clusters. And this result supports the phenomenon that we find in Figure 6 and the Figure 7.

And it's worth noting that the cross-validation score for the entire dataset is higher than the second dataset and the last one, which are respectively consisted of weekends or weekdays and of two features (weekends and temperature). There are mainly two reasons for these results.

The first reason is that the criterion for dividing the dataset by weekends or weekdays is not a good method in this case. It will lead to more variance towards the entire dataset. The second reason is that the limited number of data can not precisely demonstrate the pattern of real electricity consumption.

8. VALIDATION AND DISCUSSION

At first, since the dataset is collected in Netherland, the temperature in this area is more moderate than the most of the area in the USA. So this result may only just correspond the area of North California, because the climate there is similar with Netherland. And as the data resource is an apartment, it's not valid to generalize global results.

However, this analysis method for estimating the validation of the feature corresponding to the multiple outputs can be applied to all cases. Therefore, to some degrees, the generalizability of this case is competitive.

9. FUTURE WORK

In the future, we intend to expand our dataset of samples to multiple households in order to acquire different patterns of daily consumption. Also, the choice of the additional feature needs to be deliberately considered in the future. Furthermore, since the appliance data was collected in the DRED dataset which is not addressed in our project, it is worth to research various power consumption patterns in different rooms of the household for future study, which will contribute the further research in power supply distribution of the residential building.

10. CONCLUSION

According to the results, weekends or weekdays is not a reliable feature to do the regression analysis, because the regression result of clusters is much more efficient in this case. However, it's not reasonable to generalize this conclusion into all datasets, because it's a specific case of an apartment in Netherland.

Furthermore, this process of checking the reliability of features in the dataset is worth for other datasets. Only with implement of reliable features, can we acquire the best fitting and efficient regression tree, which is important when the volumes of databases keep growing.

11. ACKNOWLEDGMENTS

Our thanks to Professor Mario Bergés and T.A. Henning Lange for instructing us in the Data-driven Building Energy Management.

Our thanks to DRED for allowing us to use their dataset they had collected.

12. REFERENCES

[1] Abreu, J. M., Pereira, F. C., & Ferrão, P. (2012). Using pattern recognition to identify habitual behavior in residential electricity consumption. Energy and Buildings, 49, 479-487. DOI=10.1016/j.enbuild.2012.02.044

[2] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., & Wu, A. Y. (2000). The analysis of a simple k-means clustering algorithm. Proceedings of the Sixteenth Annual Symposium on Computational Geometry - SCG '00. DOI=10.1145/336154.336189

[3] Ortega, J. P., Pazos, R. R., Hidalgo, M., Almanza, N., Díaz-Parra, O., Santaolaya, R., & Caballero, V. (2015). An improvement to the K-means algorithm oriented to big data. DOI=10.1063/1.4913021

[4] Matsukawa, I. (2016). Effects of In-home Displays on Residential Electricity Consumption. Consumer Energy Conservation Behavior After Fukushima SpringerBriefs in Economics, 45-79. DOI=10.1007/978-981-10-1097-2_4

[5] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[6] Achin Jain and Rahul Mangharam, Madhur Behl, Data Predictive Control for Peak Power Reduction, Buildsys 2016