Building Energy Efficiency and Building Shapes

Oscar Wang
5562 Hobart St
Pittsburgh PA 15213
chingiw@andrew.cmu.edu

Chengcheng Mao
4716 Ellsworth Ave
Pittsburgh PA 15213
chengchm@andrew.cmu.edu

ABSTRACT
In this paper, we aim at find the relationships between energy consumption and the building shapes. We implement two machine learning methods, linear regression model and the Support Vector Regression (SVR) model to analyze the energy efficiency dataset from UCI Machine Learning Repository[1], which contains two responses as heating loads and cooling loads, and eight features regarding to building shapes of 768 different buildings simulated in AUTUDESK Ecotect. We trained the classifier using 70% the dataset, and then test the classifier using the rest 30% of the dataset. The test results for the two models are both quite good.

Keywords
Linear Regression; Support Vector Regression; Building Energy Consumption; Building Shape; Machine Learning

1. INTRODUCTION
Heating load and cooling load are good indicators for building energy efficiency. We are interested in finding relationships between building energy efficiency and building shapes. Building shapes are defined by many parameters, and is complicated to identify. We are also curious to see whether building shapes have different impact on cooling loads and heating loads. We implement two different methods, linear regression and SVR to analyze the pattern among building energy efficiency and building shapes, trained and tested our models to improve our model.

2. Proposed Approach
We propose to use two different models to analyze the dataset. One is the linear regression model, the other one is the SVR model. Linear regression method is one of the most widely used statistic analysis method. It is used when we have one scalar response, and one or more independent, explanatory inputs. If we use Y to denote the output, and X_1, X_2, …, X_i to denote the inputs, the value of the predicted Y is:

$$Y = b_0+b_1X_1+B_2X_2+ \dots + b_kX_k$$ [2]

SVR is a machine learning model which can perform both linear and non linear regression analysis. In this paper we use linear SVR model. A linear Support Vector Machine (SVM) uses hinge loss and linear hypothesis to optimize machine learning models. The rule is [3] :

$$\text{minimize}_\theta \quad \sum_{i=1}^{m} \max\{1 - y^{(i)} \cdot \theta^T x^{(i)}, 0\} + \frac{\lambda}{2}\|\theta\|_2^2$$

Unlike least squares, we solve these optimization problems by gradient descent to update the function loss.

3. Dataset
We use the Energy Efficiency Data Set generated by Angeliki Xifara from the UCI Machine Learning Repository. The dataset contains 768 instances of building performance data simulated in Ecotect. Each instance has 8 attributes regarding to building shapes and 2 responses (Heating Load and Cooling Load) indicating the building energy efficiency. The whole dataset is simulated based on 12 different shapes of buildings, each category has the same parameters on Relative Compactness, Surface Area, Wall Area, Roof Area and Overall Height. The author then assign different different orientation, glazing area, and glazing area distribution to each of the different shapes. The author use value 2, 3, 4, 5 to indicate four types of orientation. The four ratios of glazing area is represented by 0, 0.1, 0.25, and 0.4. There are 6 types of glazing area distribution which is indicated by integer range from 0 to 5. The attributes of shapes of real buildings would be much more complicated, but here since the data is based on the simulation in Ecotect, it is simplified, and less noisy. We think that the dataset contains good features that may affect the building energy efficiency, and it is very clean, so we perform analysis on this dataset.

4. Results
After cleaning the data frame, we first plotted the two outputs to see their ranges and patterns.

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
0	0.98	514.5	294.0	110.25	7.0	2.0	0.0	0.0	15.55	21.33
1	0.98	514.5	294.0	110.25	7.0	3.0	0.0	0.0	15.55	21.33
2	0.98	514.5	294.0	110.25	7.0	4.0	0.0	0.0	15.55	21.33
3	0.98	514.5	294.0	110.25	7.0	5.0	0.0	0.0	15.55	21.33
4	0.90	563.5	318.5	122.50	7.0	2.0	0.0	0.0	20.84	28.28

Figure 1: Data Frame Head

[1] Energy Efficiency Data Set, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Energy+efficiency]

[2] Introduction to linear regression analysis [http://people.duke.edu/~rnau/regintro.htm]

[3] Practical Data Science: Liner classification [http://www.datasciencecourse.org/linear_classification.pdf]

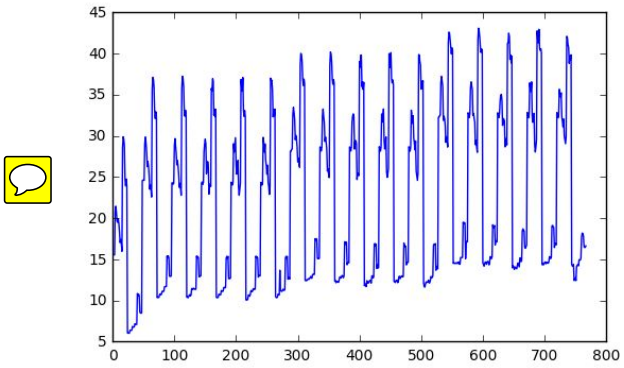

Figure 2: Pattern of Heating Loads

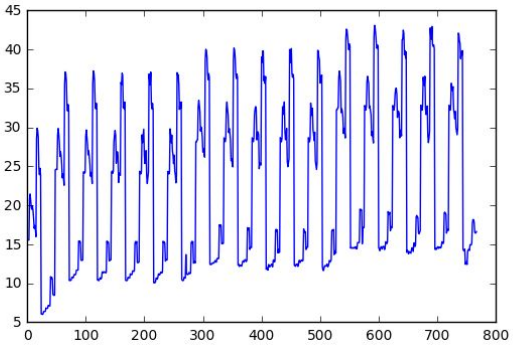

Figure 3: Pattern of Cooling Loads

After seeing the plots of heating loads and cooling loads, we think there is a good chance that the two outputs are in a linear relationship. We than scatter plot the two outputs.

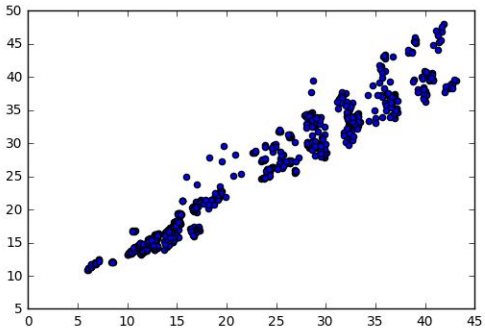

Figure 4: Scatter Plot of Heating Loads and Cooling Loads

Now we know that the two outputs are in a linear relationship. We use linear regression model and the SVR model to analyze the data. We use 70% of the dataset to train the models (train set), and test the model using the rest 30% of the dataset (validation set). We drop the input label, and create the label vector. In order to compare the predicted result with the actual output, we sort the validation set by the label value, so that we can plot the predicted results and the real outputs.

We first train and test the model for output Y1, heating loads. The R^2 for the Linear Regression model is 0.9166, and the score for the SVR model is 0.8964. The scores for the two models are similar, but the SVR model yield a better result because of the lower R^2 rate. In order to see the difference of the two models more clearly, we plotted the predicted results by the two models and the real output.

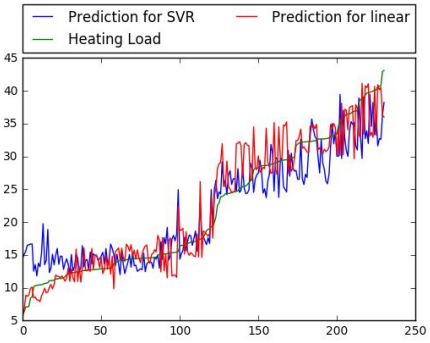

Figure 5: Predicted Results and Real Outputs for Heating Loads

According to the figure on the left, the real output is more stable than the predicted results. But the majority of the predicted results fall in certain error range of the real output. The overall pattern of the predicted results match with the pattern of the real heating method.

We then do the same analysis on output Y2, cooling loads. The score of the Linear Regression model for the cooling load model is 0.8855, and the score of the SVR model is 0.8843. The score of the models are almost the same. We also plot the two predicted outputs and the real cooling loads. The overall pattern of the predicted results match with the real result. But as the real cooling loads goes higher, the predicted results trend to get apart from the real outputs.

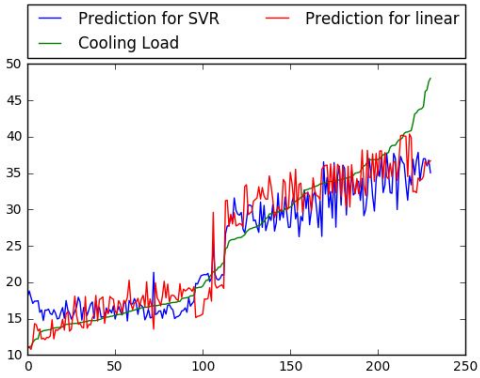

Figure 6: Predicted Results and Real Outputs for Cooling Loads

Since the two models are all linear, the most important features should have largest coefficient. In order to see which feature has the most power of impacting the outputs, we print out the coefficients of the attributes.

```
# coefficients of linear model
print (reg_lin.coef_)
```

```
[ -6.98002310e+01  -6.81438293e-02   2.60555604e-02  -4.70996949e-02
   3.98884423e+00   4.94366677e-02   1.50578572e+01  -2.25925048e-02]
```

Figure 7: Coefficients of Features

According to our model, feature X1 Relative Compactness and X7 Glazing Area are the most important features to the outputs.

5. Validation

Overall the test results for our models are quite good, except for the extreme conditions. Also in general, the test results of the SVR model is better than the results of the Linear Regression model. We also find out that among those eight features, relative compactness and glazing area are the two most important features that can impact the heating and cooling loads.

6. Discussion

One limitation of our dataset is that there are only 768 instances in that dataset. Among the 768 instances, samples are not distributed enough, for example, we only have two different overall heights, 3.5m and 7m. We trained the model using 70% of them, and tested using 30% of them. The models are not accurate enough since we don't have enough data. Another limitation of our dataset is that the data is collected from the simulation on Ecotect, we are not sure how accurate Ecotect is in simulating building performances based on the inputs. We want to collect data from real buildings, but it's very hard to get accurate representation of shapes of real buildings. So on one hand, the simulated dataset help us to quantify several important factors regarding building shapes. On the other hand, since the outputs are simulated results from Ecotect, they might not be accurate enough themselves. These reasons may contribute to the errors of our model.

Another limitation of our analysis is that we assume linear regression relationships for both models. The relationship of the building energy efficiency and the vector of building shapes is not linear. We use linear method to simplify the problem. This may be the reason of the error on extreme inputs.

7. Future Work

In the future, we plan to do the following tasks to enhance our model:

1. Use Ecotect to generate more data, add more distributed samples to the dataset, add more data of different features.
2. Implement non-linear models to analyze the dataset.

8. References

[1] Energy Efficiency Data Set, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Energy+efficiency]

[2] Introduction to linear regression analysis [http://people.duke.edu/~rnau/regintro.htm]

[3] Practical Data Science: Liner classification [http://www.datasciencecourse.org/linear_classification.pdf]