Electricity Demand Prediction in California

Bill Carson
Stanford University
wcarson1@stanford.edu

Bettina Chen
Stanford University
bettinac@stanford.edu

Eric Glover
Stanford University
esglover@stanford.edu

Abstract

Electricity demand must be met at all times and there are economic penalties to the utility for incorrectly predicting the demand. Using a training set comprised of a year of data, we developed a learning algorithm to predict the electricity demand for the next day. The learning algorithm used input features of hourly temperature data and day of week to determine a model. It was found that a linear regression model of a first order polynomial gave the lowest error on the test set. The effectiveness of our model was then compared to the day-ahead market (DAM) prediction made by California Independent System Operator (CAISO). While the developed model had a higher test error than the CAISO prediction, the error is comparable. The difference can likely be attributed to the size of the training data set as well as to the overall complexity of the model. However, our process shows that a simple model using only hourly temperature data and day of week as features can effectively predict electricity demand.

1. Introduction

Electricity is the ultimate perishable resource. There is very little capacity for storage in the current grid. Supply and demand must be equal at all times. Any additional power produced is essentially wasted, and mismatches create undesirable frequency variations in the output supply. Therefore, it is important that there is an extremely accurate model for predicting demand of electricity. This demand depends on numerous factors including time of day, temperature, season, and day of week.

Demand forecasts are used by the DAM to schedule generation capacity. Certain types of generation, such as nuclear and coal-fired power plants, can produce energy very cheaply but can take a long time to ramp production up or down. Because of their slow response rate, these plants are usually operating at full capacity and are referred to as base load production. Other types of generation, such as natural gas plants, can ramp up or down very quickly, but these types also tend to be the most expensive methods of generating power. A more accurate model can allow for greater use of the cheapest types of generation and save significant costs in the production of electricity.

1.1. Data Sources

We used archived electricity demand data from Open Access Same-time Information System (OASIS), the database from the California Independent System Operator (CAISO) [1]. The control area of CAISO includes the majority of California outside of the Sacramento Municipal Utility District and the Los Angeles Department of Water and Power. We defined our project scope as the San Diego Gas & Electric (SDG&E) service territory since it covers a relatively small area of 4,100 square miles, but services 3.5 million people [2]. It is ideal to select a location with a large population in order to provide an adequate representation of electricity demand. Additionally, the benefit of a small geographic location, as opposed to all of CAISO's service territory, is that the weather data will be more uniform in a small control region than across the entire state. For our model, we gathered temperature data from Camp Elliot, California, a military base in the SDG&E territory [3].

2. Method

2.1. Data Preprocessing

Given the large data set of electricity demand from all the service providers in OASIS, we isolated the SDG&E electricity production. To balance out regional electricity differences, a service territory may need to buy or sell a portion of its electricity production to neighboring territories. However, we simplified our model by assuming that electricity production is equal to electricity demand for the SDG&E region. Using the temperature data repository, we filtered the data to select out both the average daily and hourly temperature.

2.2. Least Squares

We included five features in our first model: the maximum, minimum, and average daily temperature, hour of day, and day of week[1]. Day of week will distinguish between weekends and weekdays, since electricity demand on any weekday will follow a similar pattern, which may be different than the demand on a weekend. Using the hourly temperature data, we constructed our model as: $y = Ax$ where $y \varepsilon R^{(n*24)}$, $\mathrm{A}\varepsilon \mathrm{R}^{(\mathrm{n*24})\mathrm{x5}}$, $x \varepsilon R^{5}$.

We defined y as the actual hourly electricity demand, A as the matrix containing the input features for each hour, n as the number of days, and x as the parameter to optimize. We determined x using the least squares solution: $x = (A^T A)^{-1} A^T y$. Additionally, we tried a variant of this model by including higher order terms, i.e. average temperature squared. This allowed us to determine if a nonlinear model can better predict electricity demand.

[1] We defined day of week as follows: day of week = $\left\{ \begin{matrix} 1 \text{ if day=M-F} \\ 0 \text{ if day=Sa,Su} \end{matrix} \right\}$

2.3. Parameter Separation Least Squares

One potential flaw in our first attempt is the definition of x, our learned parameter. It gives the relative weighting between the different attributes, such as hour of day or average daily temperature. However, this introduces bias into the model since it assumes a direct mapping from the hour of the day to the energy demand based on some high order polynomial relationship. One way to address this was by redefining our initial problem into 24 separate optimization problems: one for each hour of the day. In other words, given hour j ={1,2,...,24}, we selected out the demand history and temperature features for hour j from the electricity and weather data. This required changing the temperature information from daily to hourly temperature statistics. After redefining the problem, the x values that are obtained as the least squares solution have a different meaning. Now, the x values represent the weighting during hour j among the hourly temperature statistics and the day of week. This allowed us to determine the weighting among the different features for each unique hour and eliminated the bias from the earlier attempt.

3. Results

3.1. Least Squares

Using least squares, we developed a linear model that had fair performance. We realized that the typical electricity load curve follows a cyclical pattern, with a valley in the early morning and peak in the late afternoon. Therefore, while a nonlinear model of the features improved performance compared to the linear combination, it was still not satisfactory. We determined that the cause of the unnecessary bias in the model was the arbitrary labeling of the hours of the day.

3.2. Parameter Separation Least Squares

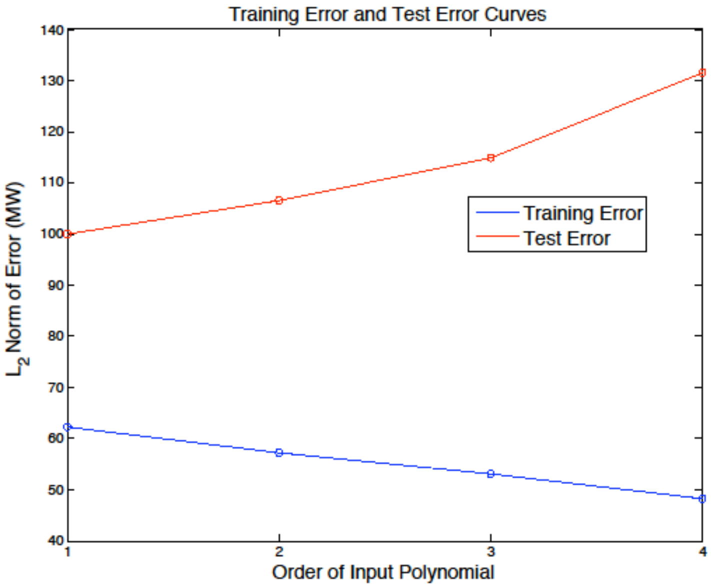

Figure 1: Test and Training Error vs Order of Input Polynomial

The second attempt used a more complicated model, which improved performance. We gathered a dataset of one full year of electricity demand and temperature data. A training set was generated by randomly selecting 2/3 of this data while the rest was reserved for a test set. This was used to determine the optimal order of the input features. Figure 1 shows that a first order model gives the lowest test error for this set of features as this optimizes the tradeoff between bias and variance.

3.3. Comparison to Utility Prediction

To quantify the effectiveness of our model, we compared the test error of our model's prediction with the DAM prediction made by CAISO for the day-ahead market. The test error was computed by taking the L2-norm of the difference between the actual and predicted energy demands and dividing by the total number of days in the test set to define the average error per day. By averaging over numerous iterations, we found the test error for our model to converge to 98.6 MW whereas the utility prediction had a test error of 42.6 MW. Figure 2 compares the predicted electricity demand for one day using our model and CAISO's to the actual demand.

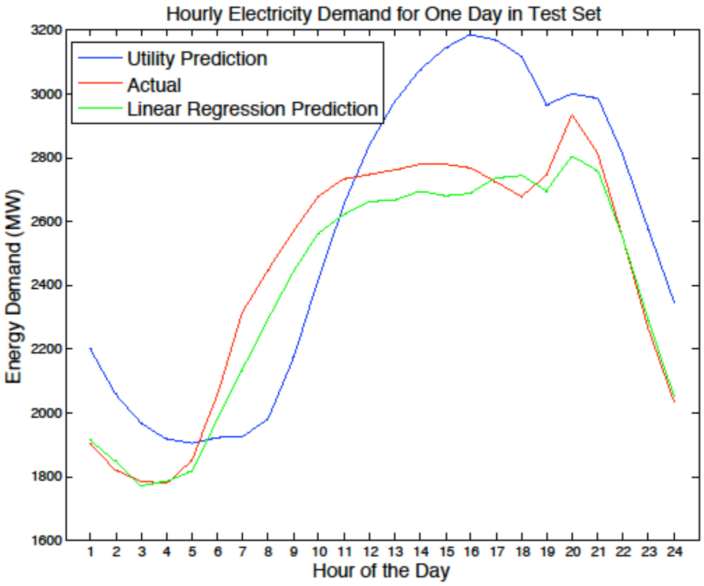

Figure 2: Comparison of Energy Predictions for Random Day In Test Set

4. Sources of Error/Areas for Improvement

We came up with three possible sources of error in our method: energy data, temperature data, and energy balance. Firstly, one disparity in test error is most likely attributable to the amount of training data each model uses. Our model is only based off of one year of data, while CAISO has collected many years of data. The additional data should allow the regression model to train its parameters more accurately. An additional source of error is the temperature measurements. We used the temperatures measurements from Camp Elliot as our sole source of temperature information and made the assumption that the temperature across the entire SDG&E service territory was uniform. Instead, we could supplement the Camp Elliot temperature data with hourly temperature measurements from each zip code within the SDG&E service territory [4]. Lastly, the final major source of error we recognized was the energy balance. One of our initial assumptions was that the energy demand in a given territory was equal to the energy produced in the same territory. In reality, CAISO would have information on which service territories are typically net exporters of electricity and which service territories are net importers of electricity. This would allow our model to add a feature to scale up or down its prediction of electricity demand based on the hourly import/export characteristics of the SDG&E service territory.

5. Conclusion

Using linear regression to determine a fixed weighting among all features for the entire day was not a successful algorithm. Instead, using linear regression for each hour of the day to determine feature weighting provided a more accurate prediction by eliminating the bias caused by the hourly labeling. When determining model size, using the first order of the features provided the best model, which we judged using cross validation and selecting the lowest test error. We found the predictions from our model to be comparable to the DAM predictions made by CAISO since the average test error for our model was on the same order of magnitude as that of the CAISO prediction.

6. References

[1] *California ISO OASIS*. [Online]. Available: http://oasis.caiso.com/

[2] *San Diego Gas & Electric Company Facts*. [Online]. Available: http://sdge.com/our-company/about-us/company-facts

[3] *Camp Elliot California*. [Online]. Available: http://www.wrcc.dri.edu/cgi-bin/rawMAIN.pl?caCCAE

[4] Weather Underground. [Online]. Available: http://www.wunderground.com/history/