Short-Term Load Forecasting Under Dynamic Pricing

Yu Xian Lim, Jonah Tang, De Wei Koh

1 Abstract

Short-term load forecasting of electrical load demand has become essential for power planning and operation, and is a well-researched area [1]. In this study, we propose a locally-weighted linear regression model for 24-hour-ahead short-term load forecasting of the electrical load for individual customers, taking into account dynamic pricing effects. As shown by the results, the model presented can work effectively.

2 Introduction

2.1 Data Set

We obtained the electrical load data from 101 residential customers (25 serviced by the Pacific Gas and Electric Company, and 76 by Southern California Edison) in the state of California under experimental dynamic pricing rates in the period dating April 2003 to September 2004. Each customer's data included the electrical load usage at 15-minute intervals (for a total of 96) per day spanning varying periods of 96 to 526 days. Of the 101 customers, 31 were subjected to variable time period critical peak pricing (CPP-V) rates and 23 were subjected to fixed time period critical peak pricing (CPP-F) rates. The remaining customers were on constant-tier rate schemes.

In the CPP-V rate scheme, ISOs provide consumers with 1-day notice of a Critical Peak Period (CPP) Event that may occur at any time for a variable amount of time, during which customer electrical rates are increased to a pre-determined (typically much higher) level. In the CPP-F scheme, customers have access to the list of upcoming CPP Events at the beginning of the month; these events have a fixed time window of 2 – 7pm. However, the actual price data was not made available, and an indicator variable was used to mark the presence of a dynamic pricing event occurring during a given time.

We also obtained the hourly temperature and humidity data collected by weather stations in each customer's residential area over the corresponding time period. Hence, the data set comprises a total of three quantitative variables (load, temperature, humidity) and one qualitative variable (dynamic pricing indicator, henceforth termed 'CPP indicator variable').

1.2 Data pre-processing

Simple initial diagnostic box plots of the data suggested large idiosyncratic load usage pattern differences across customers and for different times of the day. Therefore, model selection and fitting as described in the following section was applied independently to each individual customer. We considered the data for each customer as a twenty-four hourly series (over about 400 days).

To deal with missing points, we performed data imputation on each customer's data by interpolating missing values using the mean value of adjacent values for all 3 types of quantitative data. For temperature and humidity, the mean of values before and after missing values was taken; for electrical load, the mean was taken across load values from the same time of day from the days before and after. If the missing data was unable to be interpolated in this way (such as when the points were at the very start or end of the customer's data), the entire day was removed from consideration. We also removed 12 customers completely for having a significant proportion (>20%) of missing temperature, humidity or load values.

3 Methodology

3.1 Moving window linear regression

Given a specified feature set, we applied the following simple linear regression model:

$$y^{(h)} = \beta_0^{(h)} + \sum_{i=1}^{|\mathcal{F}|} \beta_i^{(h)} x_i^{(h)} \qquad (1)$$

where $y^{(h)}$ is the electrical load at hour h, $x_i^{(h)}$ is the i-th predictor variable, $\hat{\beta}_i^{(h)}$ the parameters, and $|\mathcal{F}|$ is the size of the feature set. Training sets were taken as moving 90-day windows, indexed by d, such that window d contains the training set $\{\mathbf{x}^{(j,h)}, y^{(j,h)}\}$ for days j = d–90 , … , d–1. This training set is used to compute the least squares estimate $\hat{\beta}_i^{(d,h)}$. (Note that the training set size was 90, but 100 days of data were needed, because the predictors may include autoregressive components up to lag 10.) From the least squares fit, the mean of the squared residuals over the training set, $\frac{1}{90}\sum_{j=d-90}^{j=d-1}\left(y^{(j,h)} - \widehat{\boldsymbol{\beta}}^{(d,h)T}\mathbf{x}^{(j,h)}\right)^2$, was taken as one observation $\varepsilon_{tr}^{(d,h)}$ of training error. The trained parameters $\hat{\beta}_i^{(d,h)}$ were then used to make a single prediction of the electrical load on day d at hour h, and the squared-difference between the true load and the predicted load was taken as one observation $\varepsilon_{gen}^{(d,h)}$ of generalization error. Another set of parameters $\hat{\beta}_i^{(d,h)}$ were then trained for the next value of h, and the process repeated for all h=1,2,…24. The entire window is then moved forward by one day to train the parameters $\hat{\beta}_i^{(d+1,h)}$, predictions made on day d+1, and so on until the last day for which data is available. The training error and estimated generalization error for this feature set $\mathcal{F}$, by hour, are taken as

$$\varepsilon_{tr}^{(h)}(\mathcal{F}) = \underset{d}{\text{mean}}\left\{\varepsilon_{tr}^{(d,h)}(\mathcal{F})\right\} \quad , \qquad \hat{\varepsilon}_{gen}^{(h)}(\mathcal{F}) = \underset{d}{\text{mean}}\left\{\hat{\varepsilon}_{gen}^{(d,h)}(\mathcal{F})\right\} \qquad (2)$$

and likewise we can define the overall training error and estimated generalization error as

$$\varepsilon_{tr}(\mathcal{F}) = \underset{h}{\text{mean}}\left\{\varepsilon_{tr}^{(h)}(\mathcal{F})\right\} \quad , \qquad \hat{\varepsilon}_{gen}(\mathcal{F}) = \underset{h}{\text{mean}}\left\{\hat{\varepsilon}_{gen}^{(h)}(\mathcal{F})\right\} \qquad (3)$$

3.2 Feature Set Selection

The features that go into our regression model are selected from: the variables CPP indicator, temperature, humidity on the current day and on previous days (up to n = 10 days); and also the load on previous days (autoregressive components), all corresponding to hour h. Therefore, there are about $4n$ possible features, or 2^{4n} possible feature sets. To pick an appropriate feature set, we use the forward search algorithm, evaluating feature sets $\mathcal{F}$ by their estimated generalization error $\hat{\varepsilon}_{gen}(\mathcal{F})$ as defined above. Our forward search algorithm terminates if the addition of the next feature fails to decrease the estimated generalization error.

3.3 Time-Binning Selection

For each d, The model equation (1) fits $24(1+|\mathcal{F}|)$ parameters to predict 24 load values. To guard against over-fitting to the fine 1-hour time binnings, we explored models with variable, coarser time binnings (i.e. varying combinations of bins). This can be formalized as

$$y^{(h)} = \beta_0^{(h)} + \sum_{i=1}^{|\mathcal{F}|} \beta_i^{(h)} \tilde{x}_i^{(h)} \qquad (4)$$

where we have defined

$$\tilde{x}_i^{(h)} = \begin{cases} \underset{h \in \text{bin}(h)}{\text{mode}} \left\{x_i^{(h)}\right\} & \text{if } x_i \text{ is an indicator} \\ \underset{h \in \text{bin}(h)}{\text{mean}} \left\{x_i^{(h)}\right\} & \text{otherwise.} \end{cases}$$

Roughly speaking, we have introduced features from adjacent h, but under the imposed constraint that that all $\hat{\beta}_i^{(d,h)}$ for the same bin are equal. For selecting an appropriate time-binning out of the possible 2^{24} binning combinations, a backward search algorithm was employed. The algorithm initialized with the finest binning, and at each step of the backward search picked the best pair of adjacent bins to coalesce. Here 'best' was measured as having the lowest estimated generalization error. (For a given binning $\mathcal{B}$, all the linear regression, moving window evaluation, and feature selection was performed to yield the estimated generalization error for binning $\mathcal{B}$). Our backward search terminates when generalization error failed to improve for several consecutive fusions. Overall the entire model selection procedure (feature selection and time-binning selection) can be viewed as a heuristic attempt to find optimal feature set $\mathcal{F}^*$ and optimal binning $\mathcal{B}^*$ to give the minimal generalization error $\hat{\varepsilon}_{gen}^*$:

$$(\mathcal{B}^*, \mathcal{F}^*) = \underset{(\mathcal{B}, \mathcal{F})}{\text{argmin}} \left\{\hat{\varepsilon}_{gen}(\mathcal{B}, \mathcal{F})\right\} \quad , \quad \hat{\varepsilon}_{gen}^* = \hat{\varepsilon}_{gen}(\mathcal{B}^*, \mathcal{F}^*) \qquad (5)$$

3.4 Locally weighted linear regression

Locally weighted linear regression (LWR) was also explored as a modification to the simple linear regression training scheme, as follows. The training examples in each moving window were weighted according to their distance from the test example (which recall is the load on the day immediately after the training window). Specifically, the weighting function for window d was chosen as

$$w^{(j,h)} = \frac{1}{1 + [\mathbf{x}^{(j,h)} - \mathbf{x}^{(d,h)}]^T \, \hat{\Sigma}^{-1} \, [\mathbf{x}^{(j,h)} - \mathbf{x}^{(d+1,h)}]} \qquad (6)$$

where $\mathbf{x}^{(j,h)}$ is the vector of predictor variables (features) for the j-th training data point of the 90-day training window d (i.e. j = d–90 , … , d–1) , and $\hat{\Sigma}$ is an estimate of the covariance of the predictor variables from the points in this training window. Just for this weight calculation, the constant term and indicator variable have been excluded from the feature vector. This form of weighting function has heavier tails than a Gaussian form, so heavier weights can be assigned to points at a greater distance; this helps in cases when the test example is an outlier (with no or very few training examples close to it in distance).

3.5 Comparing errors

Some customers have very high electricity usage while others have very low electricity usage. Since in general their loads would have different variances, and our training and estimated generalization errors are mean-squared errors, it is difficult to compare them directly between customers. We thus normalize the root-mean-square errors (RMSE) by dividing them by the mean load to obtain a form of noise-to-signal ratio for each customer (and implicitly we have assumed that the variances are roughly proportional to the square of the means). For purposes of comparing across customers, we define the hourly quantities $\gamma_{tr}^{(h)}$, $\gamma_{gen}^{(h)}$

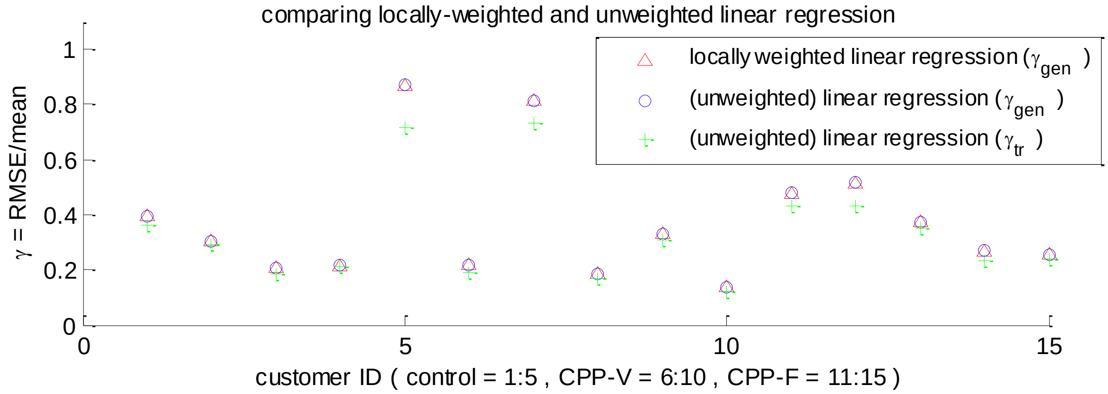
$$\gamma_{tr}^{(h)} = \frac{\sqrt{\hat{\varepsilon}_{tr}^{*\,(h)}}}{\underset{d}{\operatorname{mean}}\{y^{(d,h)}\}} \quad , \qquad \gamma_{gen}^{(h)} = \frac{\sqrt{\hat{\varepsilon}_{gen}^{*\,(h)}}}{\underset{d}{\operatorname{mean}}\{y^{(d,h)}\}} \tag{7}$$

and overall quantities γ_{tr} , γ_{gen}

$$\gamma_{tr} = \frac{\sqrt{\hat{\varepsilon}_{tr}^{*}}}{\underset{d,h}{\operatorname{mean}}\{y^{(d,h)}\}} \quad , \qquad \gamma_{gen} = \frac{\sqrt{\hat{\varepsilon}_{gen}^{*}}}{\underset{d,h}{\operatorname{mean}}\{y^{(d,h)}\}} \tag{8}$$

4 Results

Typical features selected during the forward search were the load 1 day ago, the load 7 days ago, and the current day's temperature.



5 Discussion & Future Directions

We have obtained some encouraging preliminary results in the application of machine learning to load forecasting under dynamic pricing. In future work, the following could be taken into consideration.

Firstly, we assumed a linear relationship between the variables. If the true relationship is nonlinear, appropriate transformations (e.g. taking log) might yield an improved fit.

The choice of training window width (90 days) was arbitrary; it could be included as a model parameter over which to minimize error.

We could also expand the set of features under consideration by higher order terms, as well as interaction terms between variables, in particular interaction terms between the CPP indicator variable and the other feature variables. Also, we have currently allowed for inclusion of features with other values of h in only a very rigid way (binning). This could be relaxed by, for instance, directly including features from other h into the linear regression model.

6 Acknowledgements

We would like to thank Dr Amit Narayan and Rajeev Singh for working with us on this project. We would also like to thank Professor Andrew Ng and our TAs for their guidance.

7 References

[1] A. Munoz, E. F. Sanchez-Ubeda, A. Cruz, J. Marin. Short-term Forecasting In Power Systems: A Guided Tour, Springer-Verlag Berlin Heidelberg 2010.