**Inference from Electricity Consumption Data**

**Daniel Christopher Peric 25811460**

**Andisheh Partovi 22999825**

# Introduction

In this project, we have applied two statistical methods (a Maximum Likelihood Classification and a Hidden Markov Model (HMM)) to the power consumption data acquired from 9 houses to determine what fraction of the year each house is occupied and whether a house is a holiday house or not.

In this report, we first discuss some properties of the data and state some of our assumptions (The Data), then present the methods we used to perform the inference task (Methods) and the results each method yielded, in addition to describing the evaluation approaches we propose (Results and Analysis).

# The Data

The data comprises of power consumption values of 9 houses in Australia in 2013 every 30 minutes. The main inference task is to determine whether the residents are home or away the whole day. This raises the question of what is the definition of being away. A family may be away the whole day but come back at night and have dinner or be at home all day long. Because the other inference task is to find holiday houses, we decided to define being home as being home at least once during the course of the 24 hours. This way, we can clearly identify vacations as they will be consecutive ‘Away’ days.

We considered multiple options with regards to using all or parts of the observations. The naïve approach of using all the half-hourly data and seeing that at least one of the 48 inferred states is ‘Home’, is not feasible as first, its run time would have been very high and second, estimating transitional probabilities for the HMM would have been very hard.

Instead we considered the following two options: 1) taking the mean power consumption of the day and 2) choose a specific time of the day to be the representative of that day. In either of these methods, the number of observations are reduced from 17520 to 365.

For the first option, we only considered the observations from 5 a.m. - 12 p.m. (timesteps 10-48) so that the low sleeping-time consumption doesn’t affect the data statistics. The mean of the observations in this period was considered the observation value for that day.

For the second option, after investigating the data, we decide on the *dinner time* (timesteps 34-44 each day) as a good representative of the day. The other times of the day (breakfast, lunch or any other arbitrary time) yield a lot more false positives for “Away” as sometimes the power consumption is quite low. The mean of the observations in between 34 and 44 was considered the observation value for that day.

# Methods

Before applying our methods, we plotted the data for the houses to gain some insight into the data. Based on this observations, we came up with the assumption that none of the houses use solar panels because we don’t observe any negative power consumptions.

Simply plotting the means and standard deviations of the data and comparing the houses, reveals interesting information and we were able to identify the holiday house (house 3) at this stage before applying any complex models.

Figure 1. House 3 power distributions over the day in the whole year

Plotting the observations against the mean of the all observations in the year for that timestep, provides some indications regarding the state of the house and this is the intuition behind our first statistical method.

After this preliminary analysis of the data, we applied two method to determine the states of the houses: a simple Maximum Likelihood Classification and a Hidden Markov Model.

## Maximum Likelihood Classifier

The first method applied to the data is a Maximum Likelihood Classifier that has two class labels (‘Home’ and ‘Away’) and one continuous feature (‘Power Consumption’). The likelihood of each new observation belonging to the ‘Home’ or the ‘Away’ distributions are calculated and compared and the data is placed in the class with the maximum likelihood.

The intuition for this model, comes from observing the data that by simply comparing the observations to the mean, we can guess the state of the house.

### The Model

It is assumed that the data is generated by a single Gaussian distribution, therefore the parameters of the model are a mean and a standard deviation for each of the ‘Home’ and ‘Away’ distributions.

The mean and the standard deviation for both distributions were estimated in the data with the following procedure.

First, the empirical mean and standard deviation for each half-hour data was calculated over the course of the year.

Then the smallest values of and are found and used as the parameters of the ‘Away’ distribution. These represent the latent or background power usage and therefore we chose them to represent the ‘Away’ state[[1]](#footnote-1).

For the parameters of the ‘Home’ distribution, we removed the latent distribution from the data by removing the data that falls in the latent distribution and 3 times standard deviation away from the mean.

Then the empirical mean and standard deviation of this new data was calculated as before and set as . A sensitivity analysis over n was also conducted that will be discussed in the analysis section.

## Hidden Markov Model

The next utilized methods is a Hidden Markov Model with the following graphical representations. Two states were assumed for the model: ‘A’ for being Away and ‘H’ for being Home. The model can become more complicated by adding extra states such as one person being home, but we did not consider those in this version.

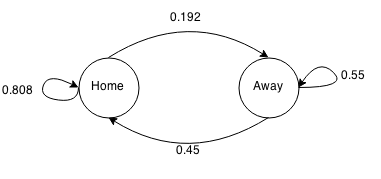


Figure 2. Transitional Model 2

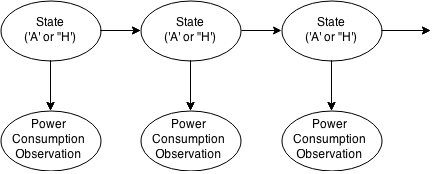


Figure 3. The dynamic Bayes Net representing the model

### The Model

Due to the time limitations of the project, no learning algorithm was implemented for this model and the parameters of the model were set for each house according to a series of assumptions and observations that will be explained shortly. If in later versions the Baum-Welch algorithm is implemented, the current model can serve as a good initial guess for the iterative leaning procedure. The values for the parameters of the model are presented in the Appendix I.

1. *The Initial Probabilities:* were considered uniform for all houses as there is no extra information to estimate a better model. Many Australians prefer to stay home on first of January and go away on vacation afterwards.
2. *The Transitional Probabilities:* were estimated by looking at the 2013 Australian calendar. We developed two sets of transitional probabilities, one assuming that families go on holiday only twice a year during the Christmas and Easter breaks and the rest are at home (*model 1*), and second, that they also go away on weekends(*model 2*). Based on this assumptions, sequences of 365 ‘A’s and ‘H’s were created and the transitional probabilities were calculated by simply counting and normalizing the four pairs (‘AA’, ‘AH’, ‘HA’, and ‘HH’). Comparing the results of the two models also provided a good sensitivity analysis on the effects of the transitional probabilities. A same procedure was employed for holiday houses but with the inverse of the previous sequence. These are generalized models that are not tailored for each individual house (other than distinguishing holiday and normal houses). Developing a more specialized model by learning the parameters from the data would have been a better approach.
3. The Emission Probabilities: were described with two single Gaussian distribution CDFs. The parameters of the distributions ( and ) were calculated using the data learnt in the course of the previous method.

The CDF can accurately describe the emission (observation) probabilities with its sigmoid shape. The larger the observation value is from the mean, the higher the likelihood of ‘Home’ state becomes.

# Results and Analysis

The classifier was ran for three different cases: 1) Dinner, where the dinner time is the representative of the whole day and we considered the mean of the observations in timesteps 34-44, 2) Whole day, where we considered the mean of the observations in timesteps 10-48. Table 1 shows the number of days classified as Home and Away with the two settings. The tables containing the class labels for every day is provided in separate documents titled “classifier\_allDay/dinner\_output\_summary”.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dinner Predictor | | |  | Day predictor | | |
| house | Days Away | Days Home |  | house | Days Away | Days Home |
| 1 | 178 | 187 |  | 1 | 167 | 198 |
| 2 | 170 | 195 |  | 2 | 182 | 183 |
| 3 | 270 | 95 |  | 3 | 252 | 113 |
| 4 | 51 | 314 |  | 4 | 44 | 321 |
| 5 | 54 | 311 |  | 5 | 30 | 335 |
| 6 | 16 | 349 |  | 6 | 13 | 352 |
| 7 | 90 | 275 |  | 7 | 67 | 298 |
| 8 | 12 | 353 |  | 8 | 42 | 323 |
| 9 | 0 | 365 |  | 9 | 0 | 365 |

Table 1. The number of days in each class

The HMM was ran with the two models that differ in transitional probabilities (see the Method Section). Table 2 shows the number of days with a specific state. The tables containing the states for every day is provided in separate documents titled “HMM\_output\_summary 1 and 2”.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| HMM Model 1 | |  |  | HMM Model 2 | |  |
| house | Days Away | Days Home |  | house | Days Away | Days Home |
| 1 | 52 | 313 |  | 1 | 105 | 260 |
| 2 | 119 | 246 |  | 2 | 137 | 228 |
| 3 | 170 | 195 |  | 3 | 272 | 93 |
| 4 | 33 | 332 |  | 4 | 38 | 327 |
| 5 | 8 | 357 |  | 5 | 27 | 338 |
| 6 | 8 | 357 |  | 6 | 10 | 355 |
| 7 | 0 | 365 |  | 7 | 29 | 336 |
| 8 | 0 | 365 |  | 8 | 4 | 361 |
| 9 | 0 | 365 |  | 9 | 0 | 365 |

Table 2. The number of days in each state

House 3 is a holiday house that is always occupied on the weekends and during the holidays. Houses 1, 2, and 7 seem to go out in the evenings a lot but seem to be main residences, the rest are all main residences. The people in house 2 seem to move out on day 284 and power is cut off on day 323 that would take 81 days from the away count and total.

The classifier is more sensitive than the HMM in the sense that HMM, because of its transitional probabilities, only picks ‘Away’s and ‘Home’s if they occur consecutively, whereas, the classifier assumes an independence between states and does not care about the preceding state.

## Evaluation

Without having labelled data, the evaluation of our systems is not straightforward. We devised two approaches for evaluation: 1) Using the 2013 calendar as the gold standard, and 2) Trying to predict the states for one of the months by looking at the rest of the year.

The calendar approach can only be employed in the classifier as the calendar data is used as part of the HMM model development. If we had used EM to learn model parameters, then this approach could have been valid for HMM as well. In this approach, we assume that the gold standard is being ‘Away’ on the Christmas and Easter breaks and also the weekends (based on the 2013 calendar), then compare all the results with this standard. The number of matched classes will give us the precision of the model. This evaluation method is far from perfect as there is no guarantee that in reality the residents always go out on weekends or breaks.

The second approach can be applied to both of our methods and it involves a high level prediction of the residents’ behaviour in a month. For instance, by looking at the results from January to November, we can make a guess about what will happen in December (without actually looking at December); if we guessed that the house is a holiday house, we predict the it will be occupied at least in the last ten days of December. This approach has its own imprecisions. Firstly, making a detailed prediction about each day is not accurate and secondly, if the results don’t support our prediction it does not necessarily indicate that the model was wrong.

## A Note on the Run Time

The following table summarizes the average and total running time for learning the parameters (shared between methods), running the classifier and the Viterbi algorithm, and the total for both (learning plus running). In both procedures, because we kept the number of observations to 365, the runs are very fast.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Learning parameters | Classifier Run | Total Classification (Learning + Running) | Viterbi Run | Total HMM  (Learning + Running) |
| Average | 0.044 | 0.570 | 0.614 | 3.560 | 3.604 |
| Total | 0.396 | 5.133 | 5.530 | 32.040 | 32.436 |

Table 3. Run times for different components

Comparing the two system on running time is not sensible as the learning procedure and the classifier were implemented with Python and were ran on a desktop and the HMM was implemented with C++ and ran on a laptop.

# Limitations

In reality, one of the main limitations inherent to this dataset is that we can never differentiate not being home and not consuming any power. We made the assumption that low power consumption during the day, means that the residents are away. Although the lowest power usage occurs at night, we assume that’s only because people are sleeping and most likely home. If it is true that a household uses only a little power at night, then we can assume that the power used when sleeping is similar to or the same as background power usage. This amount or lower indicates that the residents are not home.

# Appendix I – Parameters of the HMM

**1. Initial Probabilities for all Houses**

|  |  |
| --- | --- |
| A | H |
| 0.5 | 0.5 |

**2. Transitional Probabilities**

**Non-holiday Houses**

Model 1. Only away twice a year.

|  |  |  |
| --- | --- | --- |
| States | A | H |
| A | 0.867 | 0.133 |
| H | 0.006 | 0.994 |

Model 2. Away twice a year for a long period and also away on the weekends.

|  |  |  |
| --- | --- | --- |
| States | A | H |
| A | 0.550 | 0.450 |
| H | 0.192 | 0.808 |

**Holiday Houses**

Model 1. Only home twice a year.

|  |  |  |
| --- | --- | --- |
| States | A | H |
| A | 0.867 | 0.133 |
| H | 0.006 | 0.994 |

Model 2. Home twice a year for a long period and also on the weekends.

|  |  |  |
| --- | --- | --- |
| States | A | H |
| A | 0.550 | 0.450 |
| H | 0.192 | 0.808 |

**3. Sample emission hyper parameters**

House 1

|  |  |  |
| --- | --- | --- |
|  | Mean | Standard Deviation |
| Away | 103.066 | 46.729 |
| Home | 621.303 | 657.912 |

1. In reality, there is no way to distinguish between being away and having very low power consumption (e.g. when sleeping). See the Limitation Section for a discussion regarding this phenomenon. [↑](#footnote-ref-1)