Second Group Assignment

This homework assignment addresses the training of Hidden Markov models for the purpose of unsupervised anomaly detection using the electricity consumption dataset analyzed in the previous group assignment as train data. For household electricity consumption data, we can generally assume that normal instances are far more frequent than anomalies; that is, except for some inevitable noise, the datasets considered are largely comprised of normal data instances. Thus, the resulting models learned during training should be robust to relatively few anomalies, if any.

Please complete the assignment as described below and submit an electronic copy of your solution through the course page by November 3, 2018.

For the dataset assigned to your group, complete the following tasks using the R language and environment for statistical computing and graphics:

1. For Sunday mornings only, for the time window of 8:00 AM to 11:00 AM, choose three HMMs that represent normal electricity consumption behaviour. Based on their number of states, probabilistic parameter values and feature combinations different models will naturally perform the task differently. The goal is finding number of states and feature(s) that provide the best fit of the model to the data, avoiding both overfitting and underfitting. Include the obtained BIC value in your solution.

2. Please repeat the process described above for Sunday nights, for the time window of 9:00 PM to 12:00 AM. The same optimization criteria apply.

3. After extracting Sunday's time windows, you would have 52 Sunday mornings and 52 Sunday nights. In this part you should analyze these 52 weeks in order to find the differences in the general pattern of mornings and nights as well as seasonal patterns.

    A. For the same feature as you trained your HMM (in Question 1 and 2) calculate the average value of each data point through the Sunday morning of all weeks and create a new time series (i.e., the average value of that feature at 8:00 AM for all 52 weeks and then the average of 8:01 AM and then 8:02 AM…). Create the same time series for Sunday nights as well. Plot these two time series and explain your interpretation.

    B. Try to find some seasonality in the Sunday mornings. Meaning you compute different statistical measurements (e.g., min, max, average, standard deviation,…) for each week, for each month, and also for each season. Show the differences for each month and for each season. Repeat this for Sunday nights as well.