Building energy systems: exploring relationship between power and climate data

In this project, you will explore relationship between power and climate data of the city of Pittsburgh, Pennsylvania. You will study the effects of climate factors on the power consumption.

Data

The data is named *power_data.csv*, and contains Pittsburgh data of power consumption and climate factors. The dataset consists of 35,061 data points. You have the data of Power, Temperature, Dewpoint, Humidity, Pressure, Windspeed, and weather condition (named as Condition).

Naive Bayes

In this part, you will build a naive Bayes classifier and do a 10-fold cross validation to predict the power consumption by looking at the observations of climate data.

Analysis 1. Divide the database into 10 datasets randomly (roughly same number of data for each dataset). Then, train your naive Bayes model on 9 of them and test on the remaining one. Do this iteratively, so each of the 10 folds will be a testing dataset once. Report the mean and standard deviation of both training accuracy and testing accuracy (you are doing this approach 10 times for each fold being a testing data and reporting the mean and standard deviation of the accuracies within those 10 experiments). At each time, you train the model based on 9 folds and your training accuracy would be testing your model on the 9 datasets you used for training the model and testing accuracy would be your accuracy on the 1 testing dataset.

What is your conclusion from this analysis?

Hidden Markov model

In this part, we are going to only use Temperature and Power consumption columns of the data. Imagine that your hidden states are Power consumption and your observations are the Temperature. So the goal is to predict the power given the observations of temperature.

Analysis 2. Use the first 70% of the data as training dataset to estimate the transition and emission probability matrices. To do so, loop over training dataset and count how many times each possible transition and emission happens and then normalize the matrices.

For example, transition matrix would be the dimension of $|S| \times |S|$, where $|S|$ is the number of states (i.e. power categories). Now the element $(1, 2)$ of the transition matrix, would be counting how many times, power at current step was in category 1, and power in the next step was in category 2. Same approach applies to emission probability matrix, which has the dimension of $|S| \times |Z|$, where $|Z|$ is the number of observations (i.e. temperature categories). Now the element $(1, 2)$ of the emission matrix, would be counting how many times, power at current step was in category 1, and temperature in current step was in category 2. **Note: After calculation and normalizations, the rows of your transition and emission matrices should sum to 1.**

Analysis 3. Once you built the transition and emission probability matrices, use the observations (i.e. temperature) of the remaining 30% of the data to do the following:

- Given the emission sequence of the testing dataset, return probabilities that emission is generated by hidden state, using Forward-Backward algorithm and visualize them. **Please note: You will**

also need the initial probability of states $p(S_0)$, fix this based on the final state that you observed in your training dataset.

- Given the emission sequence of the testing dataset, return the most likely sequence of hidden states, and its probability. The algorithm is called Viterbi algorithm.
- Now that you have the most likely sequence of states, and also the ground-truth power consumption (from the testing dataset), calculate the accuracy of the power consumption for the testing dataset. The accuracy would be the percentage of times that the estimated most probable hidden state in the testing data, is the same as the power category that has been recorded in the testing dataset.

What is your conclusion from this analysis?