12-752: Data-Driven Building Energy Management
Assignment#2

Due: Sunday Nov. 19th by 12:00pm on Canvas

November 8, 2017

Since we have established that you are comfortable programming in Python and completing tasks without much support, this assignment will not request for specific implementation methods to achieve the goals of each task. For instance, if the task asks for you to load a file, you are free to choose a specific approach (e.g., `read_csv` from Pandas, numpy's `loadtxt`).

You should implement your work in a Jupyter Notebook and it should contain proper documentation and be suitable to serve as a professional report (e.g., please be diligent in creating professional looking graphs including axis labels and units, as well as writing in a clear and technical language).

You will submit a `.ipynb` file through Canvas containing all of your solutions and comments to the tasks below.

1 Data Cleaning

We will begin our journey by re-importing the original 2014 electrical meter dataset for the CMU Pittsburgh Campus (`campusDemand.csv`).

Task #1 [5%]: *Load the campusDemand.csv, keeping only the columns of "Point name", "Time" and "Value".*

Instead of using the whole campus demand (as we did in the Lecture) we will focus on two other meters in the dataset for the remainder of this assignment.

Task #2 [5%]: *Remove all meters except for Porter Hall and Baker Hall from the loaded dataset.*

This time-series data has gaps and may even contain data points that are obviously wrong (e.g., negative power values, values that are obviously too high, etc.).

Task #3 [5%]: *Identify and remedy at least two examples of these problems in the dataset.*

We should now be ready to create the **load curves** that will serve as the input to most of the analysis that follows.

Task #4 [10%]: *Following the same normalization approach documented in Kwac et al. (i.e., Equation 1 from paper #1), create daily load curves for meters Porter Hall and Baker Hall (you can store these however you like, but you should be able to recover what meter each load curve came from).*

2 Exploratory Data Analysis

Before we dive into exploring the load curves themselves, let's first recreate some of the analysis that the paper presents regarding daily consumption. We only have two buildings, and they are on the same climate zone, so the analysis will reveal very different insights (it may not even be useful), but it is worth trying anyway.

Task #5 [10%]: *Using data from Porter Hall and Baker Hall, recreate Figure 2 from Kwac et al. (Zone 3 and Zone 13 from the paper would correspond to the two meters).*

Sometimes the PDF is not as useful as the CDF for revealing properties of the distribution.

Task #6 [10%]: *Plot the empirical Cumulative Distribution Function (CDF) of the densities you showed in the previous task*

Now that this is done, can you comment on what you can infer from these plots?

Task #7 [5%]: *Comment on what you have learned from the CDF and PDF plots that were generated for Porter Hall and Baker Hall.*

Task #8 [5%]: *If you were asked to suggest a probability distribution for this daily consumption data, which one would you suggest? Please provide* **some** *evidence for your claim.*

There are a few more things before we move on to load curves. For instance, let's investigate the difference in daily consupmtion betwen different days of the week for each of the meters.

Task #9 [5%]: *Create two figures (one for each meter), containing seven box plots corresponding to the daily consumption for the seven days of the week.*

Box plots tell a good part of the story, but sometimes the data distribution is more complex than what you can characterize with these robust statistics.

Task #10 [5%]: *Repeat the same process but now with Violin plots (see e.g.,*`seaborn.violinplot()`*)*

3 Clustering

To start, let's just plot the load curves directly so that we can get a better sense of the dataset. Again, since we are not going to be doing the analysis on a diverse sample of buildings across different climate zones, our clustering results and process will be different. Thus, is worth spending some time understanding the data first.

Task #11 [5%]: *Create one single figure showing line plots for all of the load curves for both meters. This plot should have, on the horizontal axis, the hours of the day (i.e., twenty four tick marks) and, on the vertical axis, the average power consumption in kW. By a line plot, we mean that each load curve will show interpolated values between the hours of the day.*

Let's improve this figure by making some changes that would allow us to better understand the differences between the buildings.

Task #12 [5%]: *Re-generate the figure by having load curves from Porter Hall show as light green, and curves from Baker Hall show as light red (think Christmas). Finish it by including in it an average load curve for each of the meters (i.e., the average hourly demand for all of the days recorded by the meter). Plot this average load curve in dark green (for Porter Hall) and dark red (Baker Hall).*

There is a lot of information on this figure so let's have you digest it.

Task #13 [5%]: *What did you learn from this? Are there some clear clusters of 24-dimensional vectors? Is this a useful way of thinking about clsuters of the data? Did any particular load curve stand out to you? If so, did you check what day it came from and what the causes for it may be?*

Now we are going to go into clustering. Let's take steps similar to what we did in class when reviewing the Clustering Analysis sub-chapter from the Elements of Statistical Learning.

Task #14 [5%]: *Compute the total scatter for the entire load curve dataset (i.e., both meters). For this task, and the following two, use the Euclidean distance as the dissimilarity measure.*

As you now know, this total scatter can be decomposed in to a "within" cluster scatter (i.e., $W(C)$) and a betwee-cluster scatter ($B(C)$) for any cluster assignment C. Since we know the data came from two separate buildings, it would make sense to ask what is the $W(C)$ and $B(C)$ for this dataset too, assuming that there are two clusters and the assignment is done by just assigning each load curve to the meter from which it came from.

Task #15 [5%]: *Calculate $W(C)$ and $B(C)$ as stated above.*

You also know that the total scatter T does not depend on the cluster assignment C. But you should make sure you trust this, so let's test it out.

Task #16 [5%]: *Create three random cluster assignment C_i with $i \in 3, 5$ and 10 clusters. Recompute T for each one of these cluster assignments, by computing $T_i = W(C_i) + B(C_i)$. Are they all the same?*

Now let's play with the k-Means clustering algorithm.

Task #17 [10%]: *Using the `sci-kit learn` k-Means implementation, perform clustering on the load curve dataset (combining data from Porter Hall and Baker Hall). Specifically, see if you can find cluster assignments that separate the load curves from both meters. Then see if you can apply a transformation to the load curve dataset to allow for clusters that are not based on the meters but rather on temporal patterns (i.e., different times of the year, different days of the week, etc).*