12-752: Data-Driven Building Energy Management
Final Exam

December 15, 2017

1. Personal Information:

Name	
Andrew ID	

2. There should be a total of 4 pages in this exam (including this cover sheet).

3. You are free to use any books, handouts and computers. The only thing you are not allowed to do is to receive help from other human beings. Robots are OK.

4. All of your work will be done in the computer.

5. You are **required** to submit a Jupyter Notebook file that implements the steps you took to answer the questions. You will submit it by submitting to Canvas.

6. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting caught in the more difficult ones before you have answered the easier ones.

7. Good luck!

Question	Topic	Max. Score	Score
1	Concepts	40	
2	Anomaly Detection	60	

1 Concepts [40%]

Answer the following questions as best as you can, and include the answers in the Jupyter Notebook that you will submit with the exam. **Note:** If you answer "False" to any of the questions, please explain why this is the case.

1. [6%] The within cluster point scatter, as defined in the ESL book $W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d(x_i, x_{i'})$ has a $\frac{1}{2}$ in front so that the derivative of $W(C)$ cancels it out (i.e., since $d(x_i, x_{i'})$ is generally quadratic, its derivative will bring out a factor of 2 that can be cancelled out):

   (a) True

   (b) False

2. [6%] The k-Means algorithm finds the global optimum for the assignment of k cluster centroids to a set of data, as measured by a suitable loss function like the within cluster point scatter.

   (a) True

   (b) False

3. [6%] On your first day as a Data Scientist for the Department of Energy, you are asked to develop a data-driven model to categorize commercial buildings as being "Green" or "Wasteful" given data about their monthly power consumption for a year, as well as physical characteristics. Immediately, you decide to fit which of the following model classes? **Note:** Please explain your choice in detail.

   (a) Clustering

   (b) Linear Regression

   (c) Hidden Markov Models

4. [6%] Once all of its parameters are specified, a Hidden Markov Model can be used as a generative model and sequence of samples can be obtained from it. In other words, it is possible to sample a sequence of observations from an HMM.

   (a) True

   (b) False

5. [6%] How many parameters need to be specified for a Hidden Markov Model $\lambda = \{\pi, A, B\}$, which has 2 hidden states ($N = 2$), and each of the states can produce 10 distinct observation symbols ($M = 10$)? Note that we are asking how many values need to be specified to fill the matrices and vectors π, A and B.

   (a) 24

   (b) 20

   (c) 26

   (d) None of the above, my answer is:

2 Anomaly Detection [60%]

For this section, we will be re-using the `campusDemand.csv` file we used in Assignments #2 and #3, and focus on the main campus meter (i.e., "Electric kW Calculations - Main Campus kW").

We will be exploring ways to detect anomalous energy consumption patterns from the data using both linear regression and clustering.

2.1 Clustering: assumed partitioning [30%]

Let's begin by using clustering. Our intuition will be that there are typical patterns of weekly/daily consumption during the year and that if we know what these patterns are, then anomalies can be defined as days/weeks that deviate significantly from them.

Of course, these patterns will change throughout the year so we will do better if we find not just one prototypical pattern for the whole dataset, but a few of them corresponding to different times of the year. At first, we will assume that we know of a good partition (clustering) of the dataset, based on the fact that there should be a strong seasonal effect on the data (i.e., the energy demand is highly influenced by the weather).

Follow these steps:

1. Re-compute the daily load curves for the dataset (i.e., these are vectors of size 24×1 that contain average hourly consumption for a period of 24 hours).
2. In addition to this, generate weekly load curves (i.e., instead of a 24-hour vector, you'll have a 168-hour vector).
3. Using both daily load curves and weekly load curves, perform the following steps:
   (a) Separate the data into four seasons: winter, spring, summer and fall corresponding to the four calendar seasons.
   (b) For each season, find the centroid (i.e., imagine that you had partitioned the dataset into four clusters, one for each season, and you wanted to compute the cluster centroid). At first, do this using the Eucledian distance.
   (c) Within each season, compute the (Euclidean) distance between each load curve and the centroid for that season.
   (d) Again for each season, find the load curve with the highest distance from the centroid. We will call this an anomaly.
   (e) Finally, repeat this set of steps using a different distance measure (you pick).

If you followed the steps above correctly, you should end up with 4 anomalous days (one for each season) for each of the distance measures (Eucledian and your own). You also will have found four anomalous weeks (one for each season) for each of the distance measures. What do these anomalous days and weeks tell you? Are they meaningful (i.e., does commencement week, or Spring break show up?)?

2.2 Clustering: k-Means partitioning [20%]

Now let's try to do something similar but instead of assuming calendar seasons are the best partitioning function, let's let k-Means do the partitioning for us. Use the centroids you calculated in the previous part and use them as the initial guess for running k-Means (i.e., when you call `sklearn.cluster.kmeans` pass it an ndarray containing the centroids (one row per centroid) using the `init=centroids`, where `centroids` is your ndarray). Once k-Mean finishes, you will have a new partitioning of the dataset. Use it and re-do all of the steps of the previous section under the new partitioning. **Note:** you may want to provide special treatment to the samples at the end of the time series given that they have significantly less demand than the rest of the data. You can either modify them, eliminate them, or carefully consider their influence in your results.

After re-running this with k-Means partitioning, how did the results change? Can you think of a better way to implement this process?

2.3 Linear Regression [10%]

For this part, we will not ask that you implement anything but rather only provide a sketch of how you would go about solving the problem using regression. In particular, please explain how you would use a piece-wise linear model like the one we used in Assignment #3 to find anomalous consumption patterns throughout the dataset. **Hints:** think about using the residuals of your regression model, consider the training/testing separation, and include possible seasonal effects.

For all of the exam, please write down all of the reasoning and thought processes that you used to arrive at your conclusions.

Good luck!