Time Series Planning: Check-In

Overview: The purpose of this project is to explore the process of representation and classification of data. The data to be used for this project is a synthetic control data set from the University of California Irvine.

Objective Summary: Create two new representation data sets that reduce the number of samples needed to represent the data.

- Implement PAA on this data set (use Euclidian distance and Manhatten distance, compare)
- Implement SAX on this data set.

Engineering Design Process

Ask: The data is provided as a set of 60 columns of data (60-time points) with values between 0 and 100 and 6 different classes. It is grouped in order by its classification.

Research: I will need to research/ look back at previous lectures to understand how to implement PAA and SAX on a data set. Also, I must look at the data provided to determine its contents and how best to organize it.

Imagine: I will need to come up with a few ways to choose which data points need to be removed from the data set. One way could be to remove outliers, another can be to remove data that is obviously inconsistent.

Plan: Implement PAA, and SAX on the data set.

Create: Create functions the implement PAA and SAX, making sure that I can interchange between using Euclidian distance and Manhatten distance for PAA as desired. Also, add plenty of useful comments to help the reader understand the code.

Test: Run the PAA and SAX functions on the data set.

Improve: If the functions are leaving data that should be removed, or deleting important points, then I should run the debugger on my code to see where my code messes up.