Justin Pham

SID: 862114467

Email: [jpham079@ucr.edu](mailto:jpham079@ucr.edu)

Dec 9, 2021

Project 2 Machine Learning with Dr. Eamonn Keogh

My code is heavily based on the MatLab code provided by Dr. Keogh. All code is original except for functions from following Python libraries:

* **csv**, **os**, **pandas** and **numpy** were used to parse the provided text file
* **sys** and **math** were used to perform cross validation and calculate Euclidean distance respectively
* **time** was used to record how long the program took to run

**Premise**

Dr. Keogh assigned each student a small data set and a large data set. Our task is to use Forward Selection and Backward Elimination to find the most relevant features in the data sets.

**Results**

I was assigned Ver\_2\_CS170\_Fall\_2021\_Small\_data\_\_91.txt.

In Figure 1 below, we see the results of running Forward Selection on the Small data. Based on these results, the features [6,2] would be the most accurate at 97.4 % accuracy. I believe it may be over-fit. Therefore, I believe features [6,2,5] to be the best features, with accuracy of 95.8 %. It is only 1.6% less accurate, but has more features. The point is that it maintains basically the same accuracy while using more features, so there is less chance if it being an over-fit.

In Figure 2 below we see the results of running Backward Elimination on the Small data.

I was assigned Ver\_2\_CS170\_Fall\_2021\_LARGE\_data\_\_91.txt.