Activity Project 1 Report

DATA MINING

CSE 572: Fall 2019

Submitted to:

Professor Ayan Banerjee
Ira A. Fulton School of Engineering
Arizona State University

Submitted by:

V R S S Suryavamsi Tenneti (vtenneti@asu.edu)
Santhosh Kumar Bijinemula (sbijinem@asu.edu)
Sai Uttej Thunuguntla (sthunugu@asu.edu)
Aryan Prasad (aprasa29@asu.edu)

October 8, 2019

1. Introduction

The Project is a part of the course requirement for Data Mining (CSE 572) for the session of Fall 2019 at Arizona State University. The goal of the project is attempting to develop a system that can determine when a person has taken a meal. The data are collected from continuous glucose monitor (CMG).

2. Team Members

V R S S Suryavamsi Tenneti (vtenneti@asu.edu)
Santhosh Kumar Bijinemula (sbijinem@asu.edu)
Sai Uttej Thunuguntla (sthunugu@asu.edu)
Aryan Prasad (aprasa29@asu.edu)

3. Project Task1: Feature Extraction

In this phase, we have selected and implemented four feature extraction methods. The cgm data that we have is for every 5 min in an interval of 2.5 hours during a lunch meal for 33 days. So each cgm file has 33 rows and 31columns.

For feature extraction step, we have taken 16 features for each day using the techniques below, which would tell us about the most significant variations and patterns in the cgm levels during the lunch meal of 2.5 hours, which would help us in determining when the person has taken a meal.

The four feature extraction techniques that we have used are:

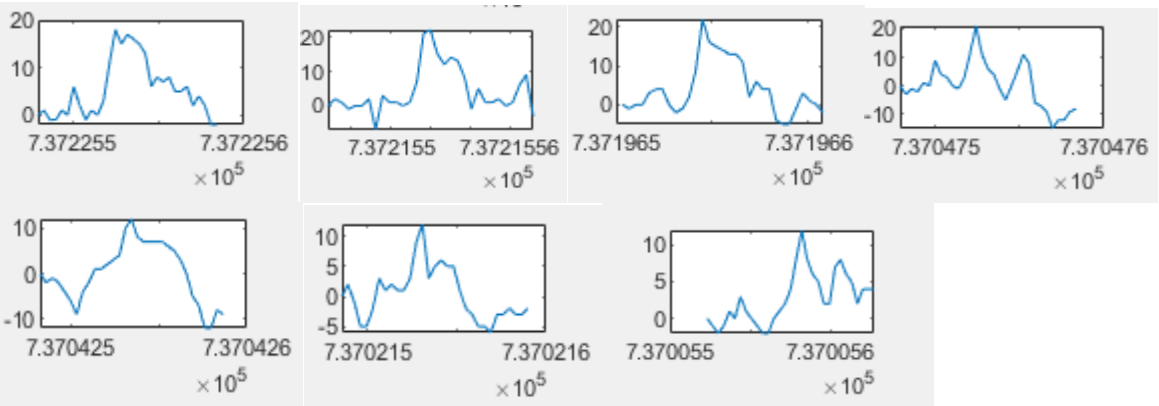
1. CGM Velocity
2. Moving RMS.
3. Discrete Wavelet Transform (DWT).
4. Power Spectral Density.

Intuition behind feature extraction

We will initially have a feature matrix with 33(rows) *31 columns (or features) in it. We have many features(values) which doesn't help us in detecting a meal like when there is not much a variation in the cgm values. So, we extract only those features which will help using in determining when the person has a taken a meal i.e. features having higher discrimination power. The contribution made by each feature extraction method and how this features are extracted is explained below.

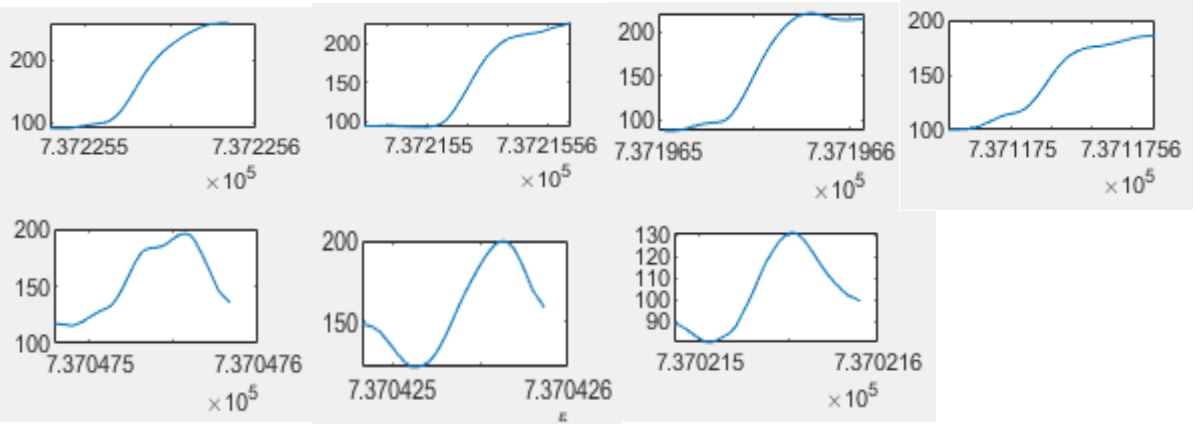
4.1 CGM Velocity

CGM Velocity measures the rate of change of glucose level for unit time. It measures the central tendency of the data sample. The difference between two consecutive glucose values are calculated and a graph is plotted with time values on X axis and glucose values on Y axis. Following are the results obtained:-



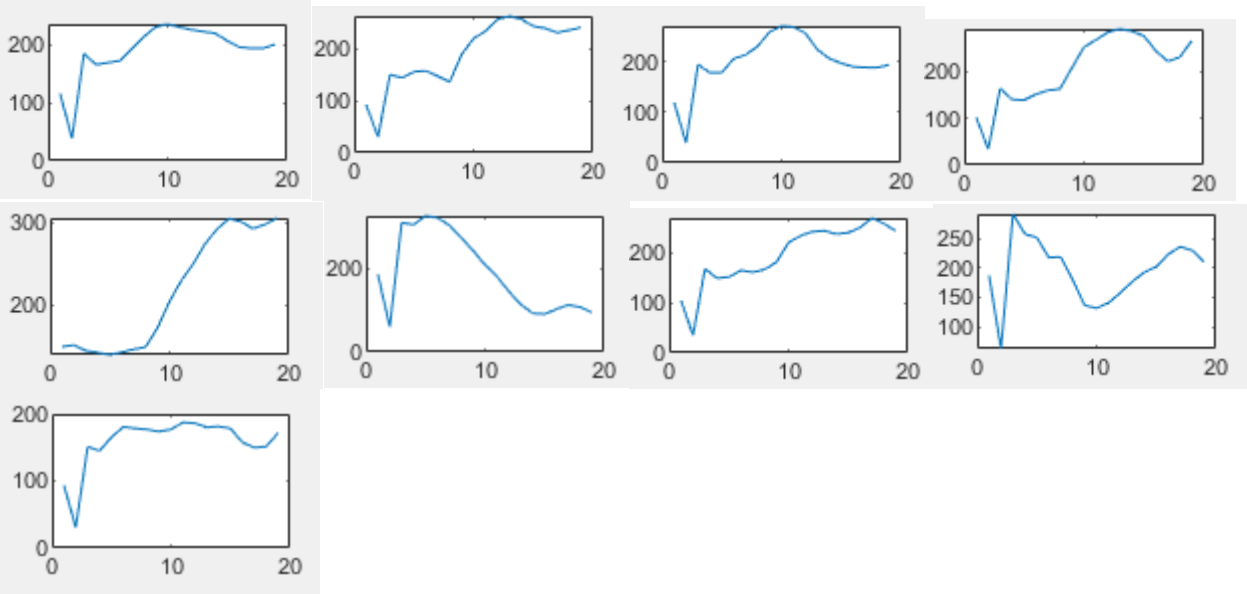
4.2 Moving RMS

a)Root Mean Squares is used to measure the magnitude of a data set. The RMS of a set of values is the square root of the arithmetic mean of the squares of the values. Moving RMS calculates RMS over a sliding window of length 5 across each neighboring element from each row. For the first value RMS of elements 1-5 is calculated, RMS of elements 2-6 is calculated for second value and so on. MATLAB RMS function with movemean() is used to calculate the moving RMS for all the 31 values of each patient. The obtained RMS values for each row of the patient along a time-series data are stored in an array. Then we have plotted a graph for times vs moving RMS values.



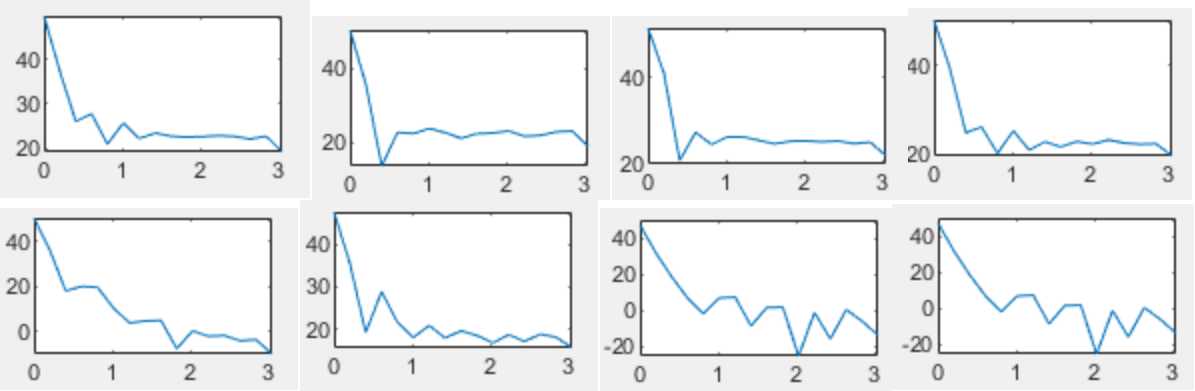
4.3 Discrete Wavelet Transform

The discrete wavelet transform (DWT) gives us an intuition of both frequency and time information of our data. It would give us an idea of most frequent occurrences along with info of when that event occurred. Here again, we used the in built Matlab dwt function on the data.



4.4 Power Spectral Density

Power spectral density function (PSD) shows the strength of the variations(energy) as a function of frequency. In other words, it shows at which frequencies variations are strong and at which frequencies variations are weak. So, we first applied fast fourier transform on the 31 data values for each row. Now, we followed the mathematical steps to compute the power spectral density of the output that we got from the fourier transform. So, this gives us where the variations of cgm values are strong.



4.6 Intuition about each feature and their outcomes:

We performed these four different feature extraction techniques namely: 1) CGM Velocity, 2) Moving Root Mean Square , 3)Discrete Wavelet Transform 4) Power Spectral Density. For CGM Velocity, the peak in graphs corresponds to maximum change in the glucose level for a given time interval and is an indicator for the meal consumed by the person. For moving RMS and DWT, as the glucose levels increases, the moving RMS and DWT values will also increase gradually and therefore the maxima in the graph indicates that person has taken the meal. To derive most from this features top 4 values for moving RMS and DWT are considered and feeded as input to feature vector for each row of the patient data. For Pwsd the time at which the variations are weak indicates the meal consumption by the person.

5. Feature Selection

5.1. Subtask 1: Arranging the feature matrix

Principal Component Analysis (PCA) takes input a feature matrix, in this case which is of dimensions 33 x 16. We take the best latent semantics which have the highest discrimination power, even among the ones selected during feature extraction process. We have performed PCA on the feature matrix that we created using the feature extraction process.

featureMatrix

33x16 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	2	2	2	1	160.0333	158.0633	156.4839	152.6860	231.9581	225.9441	220.7593	215.7907	47.2130	35.6383	25.9589	25.9585
2	11	4	4	3	131.2143	131.0198	130.2390	129.8237	187.4556	186.2315	181.4382	180.7277	48.2764	29.5073	26.6871	21.5944
3	6	6	6	5	127.4606	126.3448	125.5476	121.8975	185.9591	178.9518	164.1091	163.3566	46.6354	31.4593	23.9031	16.2497
4	21	20	19	12	189.2839	186.0441	179.8522	174.5428	291.3638	257.5233	251.4107	235.8287	49.6826	34.5200	28.7671	21.0520
5	3	3	3	3	182.7567	180.1277	177.9579	173.4756	268.5548	257.0782	249.7256	243.0985	49.7456	35.4858	22.0272	26.6620
6	13	11	9	9	173.6404	166.7513	160.1493	158.4134	275.7957	230.5005	212.8048	207.0918	49.5675	25.7920	29.5401	26.1595
7	18	14	10	10	122.8137	115.4513	115.2678	114.3818	203.4403	162.9410	159.8875	157.2336	46.4804	20.7273	29.2062	24.5095
8	6	5	5	5	131.2204	130.4500	130.0631	127.7169	187.9486	186.6623	175.1206	173.9495	47.2022	31.8414	18.6361	6.9463

5.2. Subtask 2: Execution of PCA

We have normalized the feature vector using normalize() function with 1 norm in Matlab

PCA decomposes a correlation matrix into a matrix with Principal Components and the resulting matrix contains the Principal Components in decreasing order of their variance.

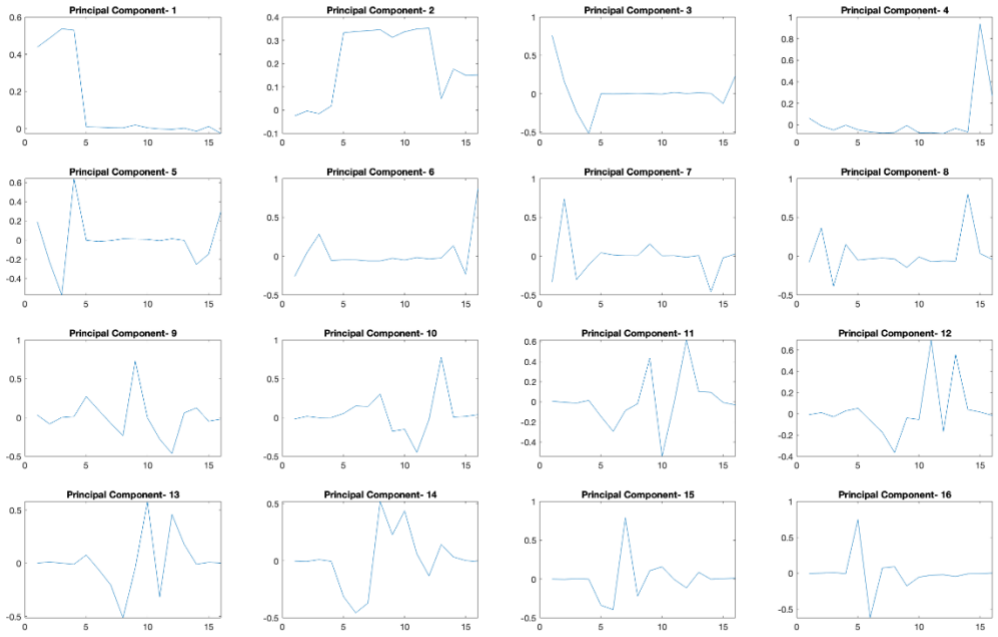
We pass the matrix obtained in Subtask 1 – 33 rows (31 rows corresponding to one lunch day meal) and 16 columns (selected features) - to PCA function of MATLAB.

PCA returns the following:

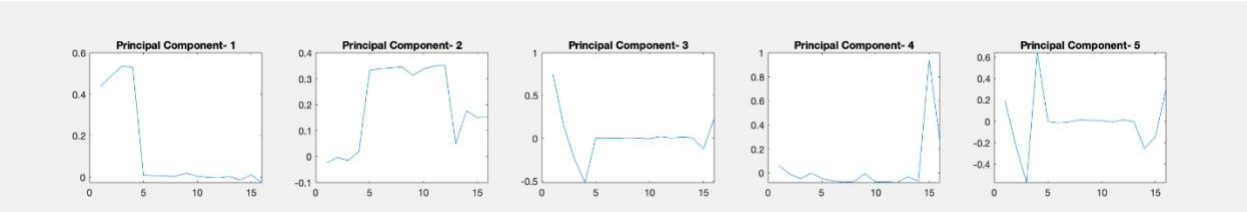
Coeff - a 33 x 16 matrix, representing the coefficients for Principal Components a.k.a Eigenvectors and the columns are in decreasing order of their variance (or eigenvalues).

Score - Principal Component scores are the transformed representation of the input matrix in the Principal Component Space.

The code for PCA has been included in Project1.m file. We have used PCA function of MATLAB to perform PCA.



5.4. Subtask 4: Results of PCA



These are the top 5 Principal components chosen as PCA returns the Eigen vectors in the decreasing order of the Eigen values of the components. In the following graphs, we can see that in each component there are certain features which have a greater importance.

For, example in the first plot, feature 3 and feature 4 which are the top 3rd and 4th maximum values of the CGM Velocity which kind of states that when the person took meal.

Also, in the second component, we have features 6 to 12 which are maximum values of RMS velocity, Discrete Wavelet Transform time series.

In the 4th principal component, feature 16 has a high value which also argues that the Power Spectral density is also considered the best feature.