CSE 572 Data Mining
Project 1

Tasks

Task a

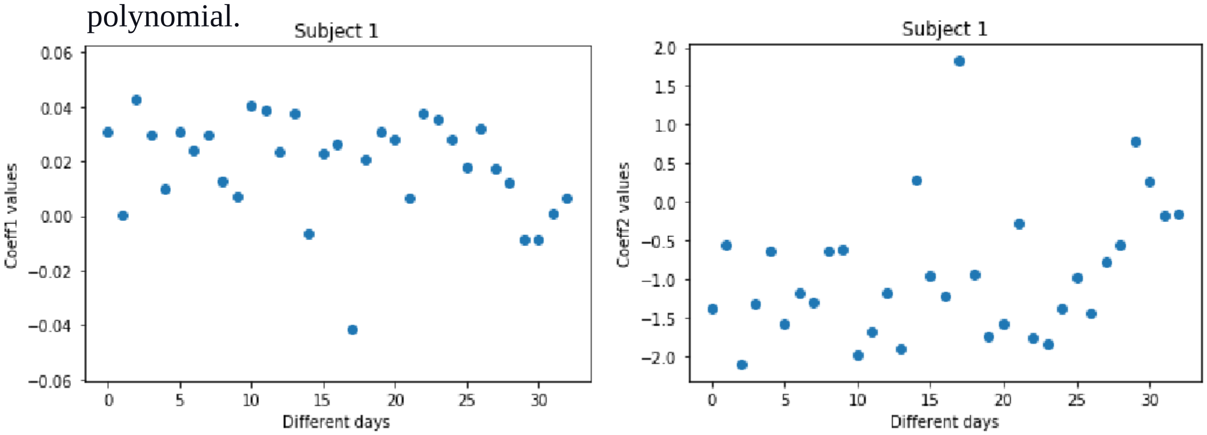
The four different time series features that I extracted using the CGM values and CGM timestamp data are:

1. **Curve Fitting**

   I used the polynomial curves of various orders on the CGM data. Below are the steps used to extract feature values:

   - The code is implemented in Python using *the* ***Polyfit*** function from the **Numpy** library.
   - **Polyfit** function takes a sequence of values as input and then outputs coefficients corresponding to the polynomial function that nearly suits those values.
   - I tried to plot and visualize polynomials of 1,2,3,4,5 degree for 2.5-hour duration time series data of 5 subjects.
   - According to the observations, the polynomial of degree 4 would better represent the data in most of the cases.

   Hence our feature values would be the five coefficients that result in the fourth-degree polynomial.



X^4 X^3

Figure1: Values of coefficient x^3 and x^4 for subject 1 across all days

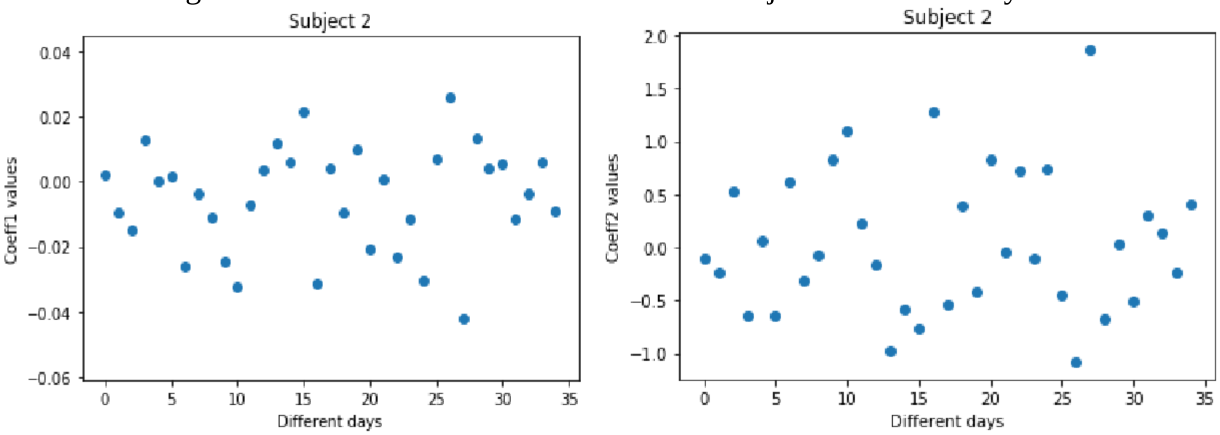


Figure2: Values of coefficient x^3 and x^4 for subject 2
across all days

2. Series Frequency

The next feature I used on the data is to perform Fast Fourier Transform (FFT). I used the Scipy.fftpack library for FFT function the python. The absolute values of the four FFT values from the top were used as features.

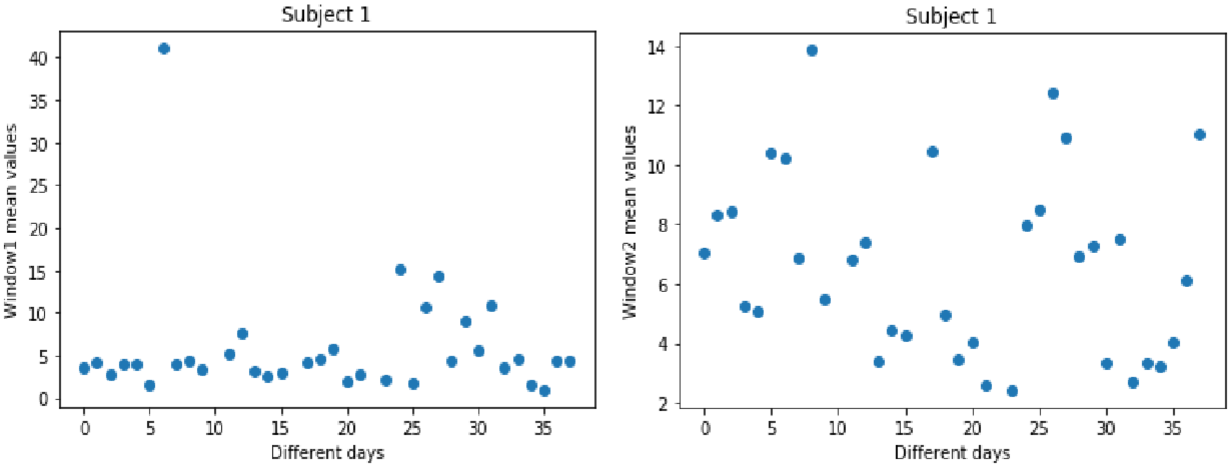


Figure3: First 2 FFT values for Subject 2
plot over all days

3. Maximum Excursion Window

After identifying the time window in which the glucose values begin at a low value and rises to a maximum value. Let's call this the maximum excursion window. To identify this window by first finding the max value in the given 2.5 hr. period and then finding the global smallest value before that maximum value and use this as the window. In this window of maximum excursion, the following are the extracted three different features from the window:

- The time gaps
- Mean of the values
- The standard deviation of the values

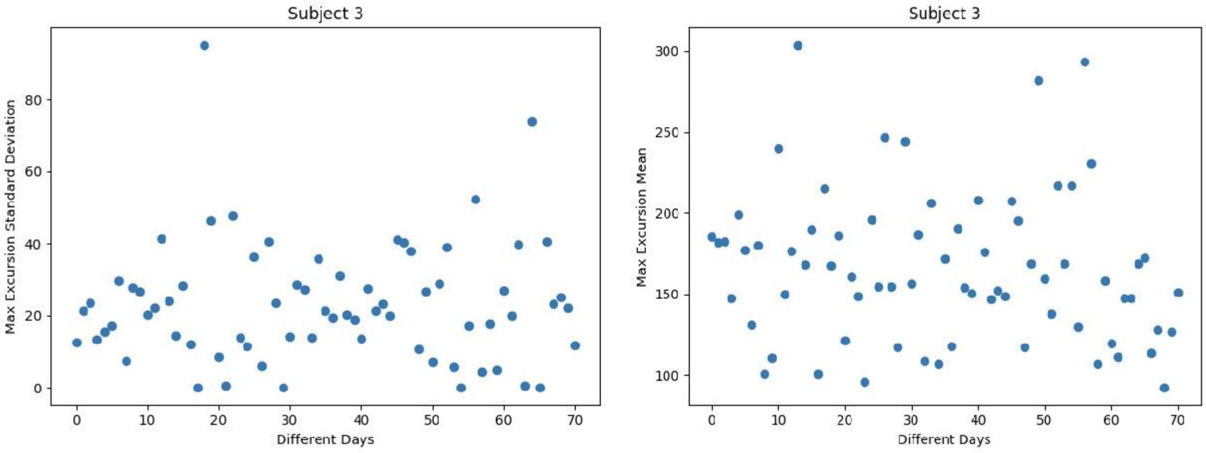


Figure4: Mean and Standard Deviation values for subject 3 in maximum excursion window plotted over various days

4. **Rate of change (CGM velocity)**

Calculated the CGM velocity by taking the difference between consecutive values. Then, by taking four overlapping windows and then calculate the mean and variance of CGM velocity. Thus, I got 2 features for each of the 4 windows making it 8.

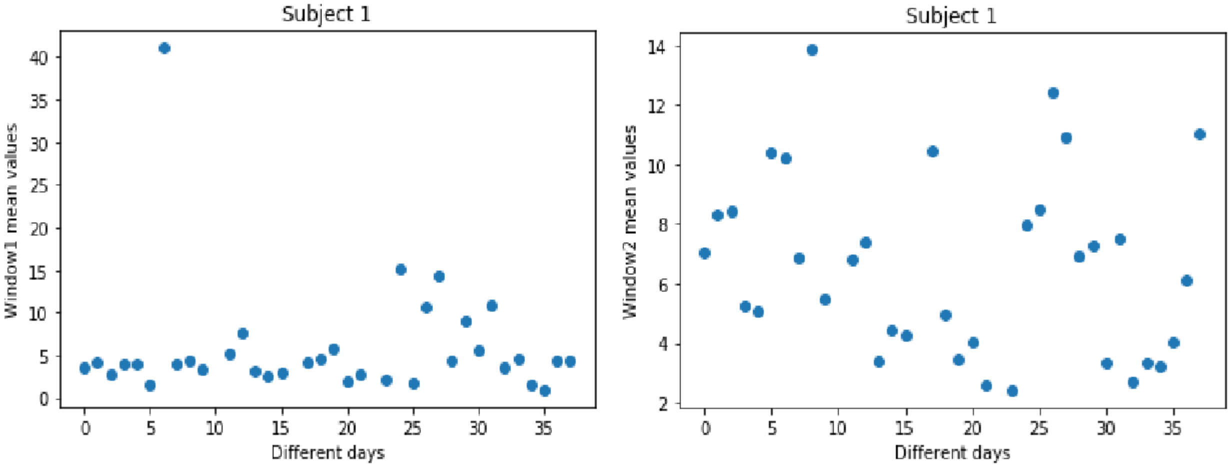


Figure5: Mean CGM velocity values for subject 1 for 2 different windows

Task b

Reason to choose the above features:

1. Curve Fitting

By outlining the CGM sensor and timestamp data, the CGM value plot changes its course (slope from negative to positive or from positive to negative) when a person has a meal or when insulin is introduced. Particularly, when a meal or any kind of glucose intake occurs, then the plot may have an increasing peak (glucose level rises and then declines). To obtain a characteristic of a meal, a polynomial curve best describes such turns in the CGM data. Hence, to fit polynomials of various degrees like 1, 2, 3, 4, 5 and discovered that polynomial of degree 4 fits the data better for different subjects. A n degree polynomial will be like $\mathbf{a_0x^n + a_1x^{n-1} + ... + a_n}$ and with n+1 coefficient. Hence, a polynomial degree of 4 would have 5 coefficients. These coefficients will be the feature values.

2. Series Frequency

Since, series data can be translated as a signal, treating the CGM time series data as the form of a signal. So, I observed the CGM signal in the frequency domain.

Frequency of a signal indicates the number of times a pattern has occurred in unit time. In this case the occurrence of meal makes the glucose value fires up and after a while when insulin begins working the glucose value begins declining and the structure of the graph looks like there is a pattern in terms of periodicity of the curve. Hence, it is important to recognize such patterns by extracting the frequency domain features. To accomplish this, a Fast Fourier Transform on the CGM sensor values was performed.

3. Maximum Excursion Window

Maximum excursion window attempts to catch the characteristics of the impact on the glucose level of a person by the meal. Also, this window expresses us the maximum glucose level is attained due to the meal and depends how much time it took to obtain that value. As this window communicates the influence of a meal on a person, lets consider using statistical measures like mean, variance for this window and use them as features.

4. Rate of change (CGM velocity)

The rate of change of values in the curve can be defined as CGM Velocity. To classify the characteristics of a meal it is important to find the point at which the CGM data start rising. At this point, preferably, we should see a change in the sign of CGM velocity data. Hence, by considering mean of a window data, we observe that the meal's mean value has a higher value than that of a meal without a window. Similarly, the value remains high until the curve starts dropping which indicates the starting of insulin effect. Thus, windowed variance and mean values of CGM velocity will be a good feature.

Task c

Values and intuition

1. Curve Fitting

   To validate the intuition behind whether the picked feature is useful or not, I plotted a scatter plot of all found polynomial coefficients across all days for all 5 subjects. The coefficient values for each subject fall within a certain range. For example:

   - From **Figure1**, for subject1 almost all the x^4 coefficients are within the range of -0.15 to 0.05 and the x^3 coefficients lie in the range of -2.0 to 0.0.
   - Similarly, from **Figure2**, for subject2 the x^4 coefficients lie in the range of -0.035 to 0.02 and the x^3 coefficients lie in the range of -1.0 to 1.5.

   As, the data resembles to a meal data and the variance of the features for each person is less considering these coefficients as good features for branding a meal.

2. Series Frequency

   From **Figure3**, there is only a unique outlier value and the other Fast Fourier Transform values of the subjects are packed tightly. To get better results, I eliminated such outliers. Otherwise, Fast Fourier Transform could be considered as a good feature to represent the CGM meal data.

3. Maximum Excursion Window

   From **Figure4**, by observing the values of standard deviation in the maximum excursion window lie within an interval of 0 - 50 along with few outliers that had standard deviation value higher than 60. Almost similar plots were developed for other subjects and the same pattern was observed. Therefore, the standard deviation of values in the maximum excursion window can be used as a preferable feature. Although, if we see that the values of mean are scattered in the maximum excursion window. The scattered values of the mean are in a large range of 0-300 which neglects mean to be a good feature value to characterize the CGM meal data.

4. Rate of change (CGM velocity)

   In **Figure5**, observing the mean values for window 1 are comparatively low and is in a specific area since the glucose values are constant in that window as there is no meal activity. However, the values in the next window are dispersed and higher which is also expected with the fact that our meal happens generally at 30 mins after the data values has begun and the data conforming to that falls within the window 2. Therefore, the intuition is correct, and can be considered as a good feature.

Task d

Feature Matrix

After removing the feature vector for every row from the CGM sensor data, a feature matrix for a subject are developed.

- While there are 20 features, each feature vector has a length of **20.**
- All **20-length** feature vector consists of the following:
  ○ the **5** polynomial coefficients - [a_1, a_2, a_3, a_4, a_5]
  ○ the statistical measures that were extracted from the excursion window are mean, time gap, and, standard deviation
  ○ Top **4** Fast Fourier Transform values
  ○ Feature values from the CGM velocity = **8**. Summing up to a total of **20** feature values.

- The size of the feature matrix is calculated by **a x 20.** Here **a** is the number of diverse 2.5-hour duration time series data for a subject.

Task e

Principal Component Analysis (PCA)

To perform PCA on the feature matrix obtained as part of task d. By applying PCA on a matrix, decomposes the covariance of a matrix and yields the principal components along with their order of importance of variance.

- Since PCA is prone to issues like scaling, I normalized the features before executing PCA. To achieve this, I used the Standard scaler function from scikit-learn.

- Using the **PCA** function from the python **scikit-learn** library to perform PCA.

- Then by applying the feature matrix of size **n x 20** to the 5 principal components PCA function. Here **n** denotes the number of different 2.5-hour duration time series data for a subject.

- The output of the PCA is:

    - Variance Explanation: This describes the amount of variance described by each component.

    - Components matrix: The columns resemble the 20 original features whereas the rows corresponds the 5 principal components

Task f

Top Five Features in PCA

PCA with K components attempts to project the data onto top K principal components which produce the highest variance thus decreasing the repetition in the features. By the inspiration of it, the features created from PCA must have the highest variance.

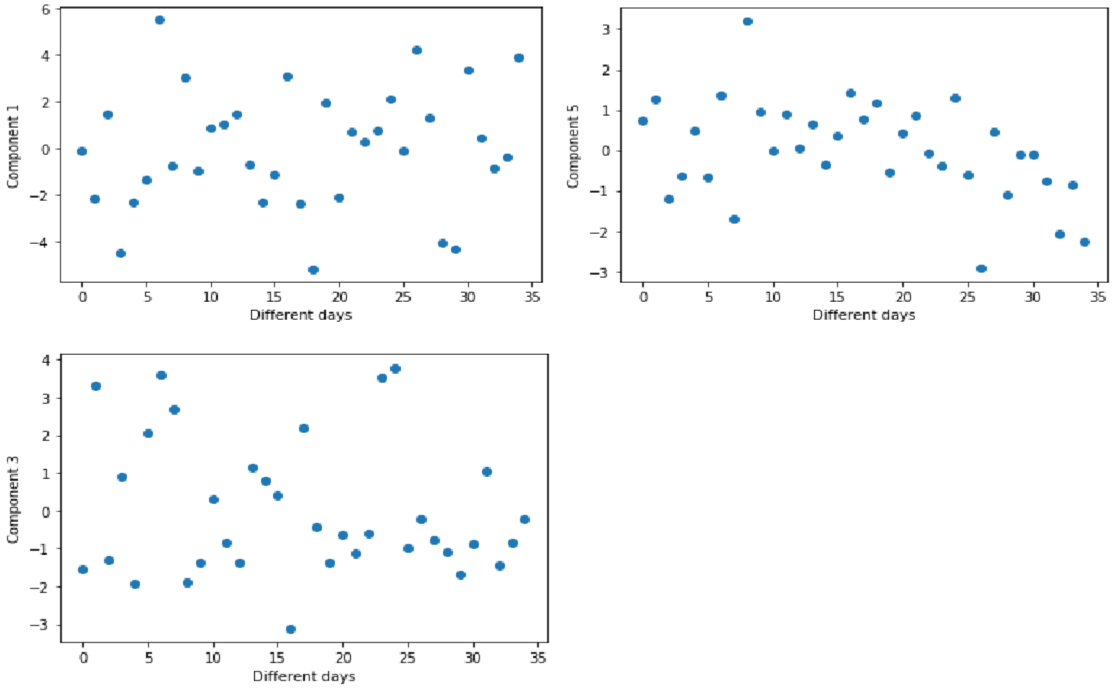


Figure6: Top 1, 3, 5 components of Principal component analysis

As from **figure6**, in the component 1 image, the values are lightly distributed thus providing more variance. The component 3 image, the distribution to be a higher concentrated than the 1st image hence providing a more variance data. Likewise, the values in the Component 5 image are densely arranged than Component 1 and Component 3 images indicating lesser variance than Component 3 and Component 1 images.

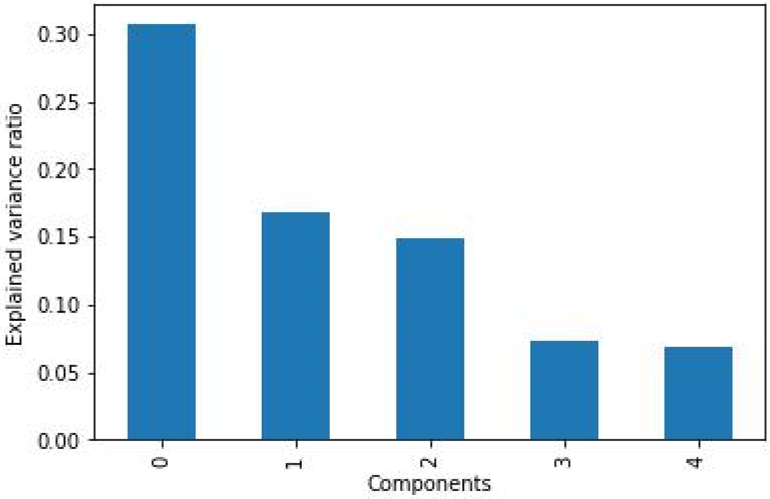


Figure7: Variance ratio for each of the 5 principal components

From the image above, observing the ratio of the variance is important. This graph indicates that the 5 principal components are adequate to preserve approximately 75% of the variance of the data. The first principal component has 30 feature of information.

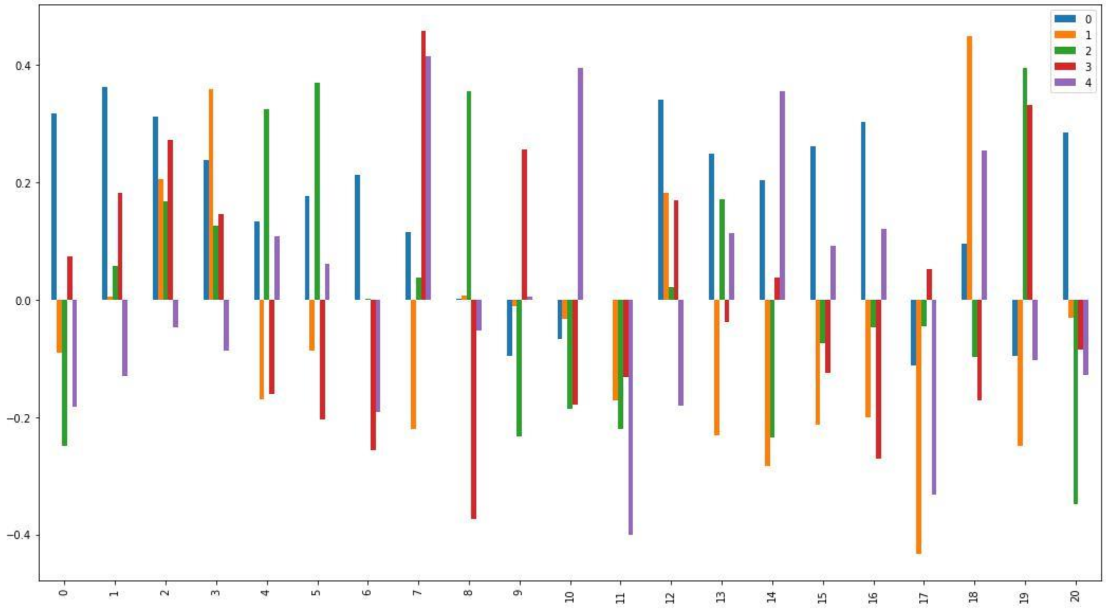


Figure8: eigenvectors weight with features.

Another observation is that while calculating PCA we project the original values on the top K principal components. This is identical to taking a weighted average of original data where the weights are come from the values of the eigenvector’s principal components. The bar plot describes for each of the eigenvectors as the individual weights of original features. Each value from a component are shown by unique color in the bar plot and the weights are represented by the length of the bar. The higher the weight value implies that the feature contributed to create such a principal component or an eigenvector.