Assignment 1

Data Mining - CSE 572

Spring 2019

Guided By:

Prof. Ayan Banerjee

Ira A. Fulton School of Engineering

Arizona State University

Team Members:

Debarati Bhattacharyya (dbhatt14@asu.edu)

Oindrila Das (odas2@asu.edu)

Sanchari Datta (sdatta24@asu.edu)

Anupriya Gupta (agupt224@asu.edu)

1. Introduction

This project is an attempt to develop a computing system that can understand human activities. We have been provided data for eating activity which is in turn mixed with other unknown activities, which will be referred to as non-eating. The goal is to identify the eating activities and discard the noise. The main source of the data is the real-world Mayo wristband sensors. This data will be pre-processed in a way so that useful features for each activity can be extracted and selected to analyze the data. In doing so, the different concepts learnt as part of the data mining course will be applied to build our solution.

2. Data Collection

Our professor and his team collected the sensor data at the Impact Lab at Arizona State University and shared this data with us for this assignment. However, as per our understanding, we could reach the subsequent conclusion for the collected data. The data is collected using two sources: 1) Mayo wristband data and 2) Video recording data. In the first case, the user is wearing the wristband and performing eating actions periodically interspersed with non-eating (unknown) actions. The gestures would be identified by the sensors using his/her hand movements. From the data provided, it is evident that there are four sensors Orientation, Accelerator, EMG Sensors and Gyroscope. From the movement of the hand, the sensors capture different data components like rotations etc. which can be collectively used for identifying the gesture.
Apart from this, the ground truth has been established by the video recording, where the user can be seen performing the eating actions. This data is provided to us as the ground truth where frame numbers have been specified which denote the beginning and end of an eating action.

3. Data Cleaning and Organization

The data is provided in two sections: ground truth and MyoData. The ground truth consists of frame numbers where an eating action starts and ends. MyoData consists of data collected from 41 users. For each user, there are 2 different recordings since the eating activity was done using fork and spoon. Fork and spoon folders consist of data from the IMU, EMG sensors and video information files. The IMU file has a sampling rate of 50Hz and consists of 11 columns: UNIX timestamp, Orientation X, Orientation Y, Orientation Z, Orientation W, Accelerometer X, Accelerometer Y, Accelerometer Z, Gyroscope X, Gyroscope Y, and Gyroscope Z. On the other hand, the EMG file has a sampling rate of 100Hz and consists of 9 columns: UNIX timestamp, EMG 1, EMG 2, EMG 3, EMG 4, EMG 5, EMG 6, EMG 7, and EMG 8. The video file consists of the average frame rate and number of frames. First, all the files are formatted and aligned as per requirement. The IMU and EMG data of all the users from the fork set are

aggregated and kept under a common folder and used as a dataset. The video data is taken at 30 samples per second. Thus, all the three datasets have data collected at different sampling rate. Hence, all the data needs to be synced to properly segregate eating and non-eating gestures from the given data. Therefore, the IMU and EMG data are normalized by subtracting each column with their minimum value and subsequently dividing them by the difference between the maximum and minimum values. The normalized IMU and EMG data are then synced with the video data by using the MATLAB “interp1(x, Y, xi)” command. This command returns a vector yi containing elements corresponding to the elements of xi and is determined by interpolation within vectors x and Y. The vector x specifies the points at which the data Y is given. The terms x Y and xi corresponds to Unix timestamp, normalized data, timeline to which it needs to be converted respectively in the MATLAB code. This command interpolates the corresponding IMU and EMG data to 30 samples per second video data. All the files are named such that it represents the Unix timestamp that is synced with and indicates the starting time of video.

Frame annotations from all users are taken from the ground truth and saved in a folder to separate eating and non-eating data present in sensor. A file “ground_truth.csv” is created from the given data which contains duration, number of frames that help to convert frame numbers to time stamps in IMU and EMG data. Users and files are mapped and recorded in a file named “mappedIds.csv”.

A path directory is setup to create a new directory to form eating and non-eating csv files, simultaneously two feature matrices for eating and non-eating are created and stored in csv files. The interpolated data is concatenated in vertical columns to be further used for feature extraction.

4. Feature Extraction

We have performed our analysis on Eating and Non-Eating activities of several users. As discussed earlier, the given datasets have uneven sampling frequency and thus the linear interpolation technique has been used for every feature to make it even. The methods selected and implemented for feature extraction are as below:

1. Mean
2. Max
3. Standard Deviation
4. Root Mean Square
5. Fast Fourier Transform
6. Entropy

Now, we will look at the individual features:

1. Mean

As per mathematical definition, Mean is a calculated "central" value of a set of numbers. The statistical mean refers to the mean or average that is used to derive the central tendency of the given data. It is determined by adding all the data points in a population and then dividing the total by the number of points [1]. The formula for the statistical mean can be expressed as:

$$\boldsymbol{mean = \frac{1}{n}\sum_{i=1}^{n} a_i}$$

For example, to calculate the mean of orientation's x value as a feature, the input will be the orientation signal's x value as input for each eating activity. So, orientation x mean would be one feature of the eating data, with n values for n data points.

2. Max

In mathematics, max defines a point at which a function's value is greatest [2]. For a given signal or data set, max returns the peak value of that signal or the highest value of the dataset. For example, if orientation x is given as an input to max function, the highest value from all the data points in that signal will be returned. This value is the feature of the orientation x.

3. Standard Deviation

In statistics, the standard deviation (SD, also represented by the lower-case Greek letter sigma σ or the Latin letter s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values [2]. In other words, it tells us how far the data points are located from the mean. The formula for SD can be expressed as:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

In the above equation, N denotes the number of samples taken, xi is the value of each sample and $\overline{x}$ is the mean or average of the random variables x1, x2, x3…..xN. Like the mean calculation, SD accepts the attribute of a signal as input and outputs a single value.

4. Root Mean Square

The RMS value of a set of values (or a continuous-time waveform) is the square root of the arithmetic mean of the squares of the values, or the square of the function that defines the continuous waveform [2]. The formula can be expressed as:

$$x_{\mathrm{rms}} = \sqrt{\frac{1}{n}\left(x_1^2 + x_2^2 + \cdots + x_n^2\right)}$$

If the waveform is a continuous function f(t) defined over the interval $T_1 \leq t \leq T_2$ then the corresponding formula is:

$$f_{\mathrm{rms}} = \sqrt{\frac{1}{T_2 - T_1}\int_{T_1}^{T_2} [f(t)]^2\, dt},$$

5. Fast Fourier Transform

A fast Fourier Transform (FFT) computes the discrete Fourier transform (DFT) of a sequence, or its inverse (IDFT). Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. It is an algorithm that samples a signal over a period of time (or space) and divides it into its frequency components. These components are single sinusoidal oscillations at distinct frequencies each with their own amplitude and phase [2]. FFT takes one signal as input, and generates the frequencies corresponding to the peak values as output which can be used as data features.

6. Entropy

Entropy is a measure of randomness. In other words, it is a measure of unpredictability [3]. The formula for Entropy is expressed as [4]:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

where pi denotes the discrete probability distribution of a dataset.

5. Features Extracted:

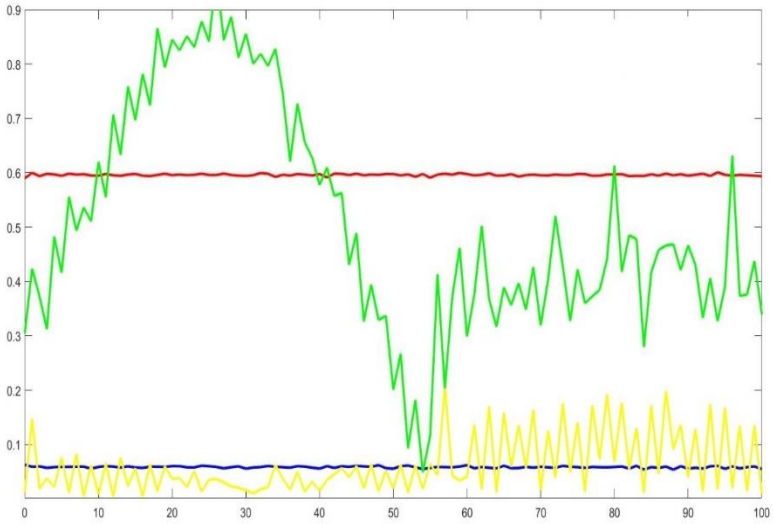
A. Mean of Accelerometer along x-axis:

- **Intuition**

  The acceleration of a human hand will change in the z direction when a person is eating. Thus, we are focusing on the mean values along the x-axis to understand and differentiate an eating activity from the non-eating one. Non-eating activity will have a higher value in the horizontal direction as the user can move his hand in any direction.

- **Conclusion**

  From the graphs that we have plotted, the mean of accelerometer along x-axis for the eating activity has lower values than that of non-eating activity. Thus, the intuition is satisfied. The color convention is as follows:
  Red – Non-eating accelerometer x
  Blue – Eating acc x
  Yellow – Eating acc x mean
  Green – Non-eating acc x mean



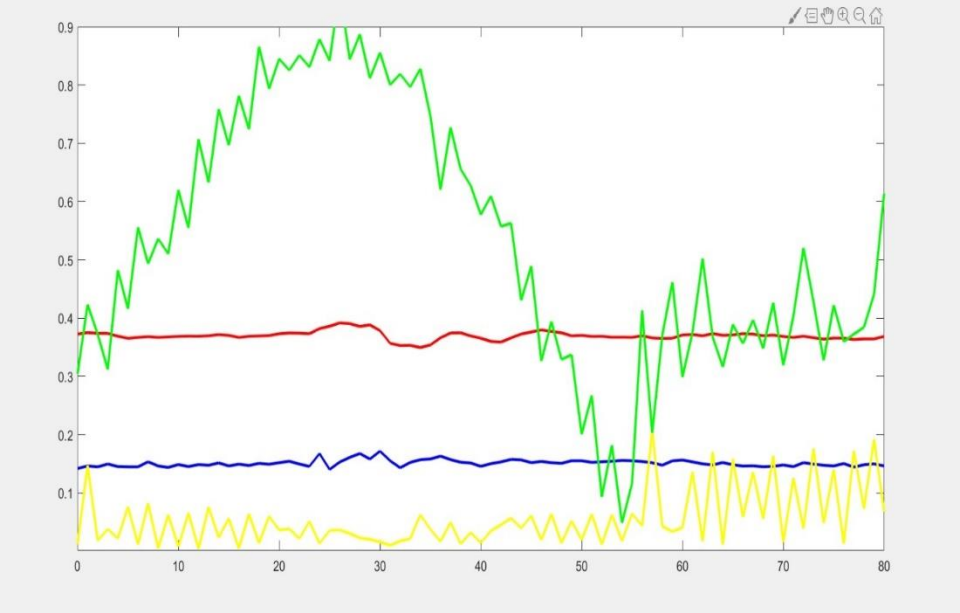
B. Maximum of accelerometer along x axis:

- **Intuition**

  The eating action generally occurs in the y and z direction but not in x.

- **Conclusion**

From the graphs that we have plotted, the maximum of accelerometer values is less for eating activity in the x direction than the non-eating activity. Thus, the intuition is satisfied. The red and green show non-eating actions and blue and yellow are for the eating actions.



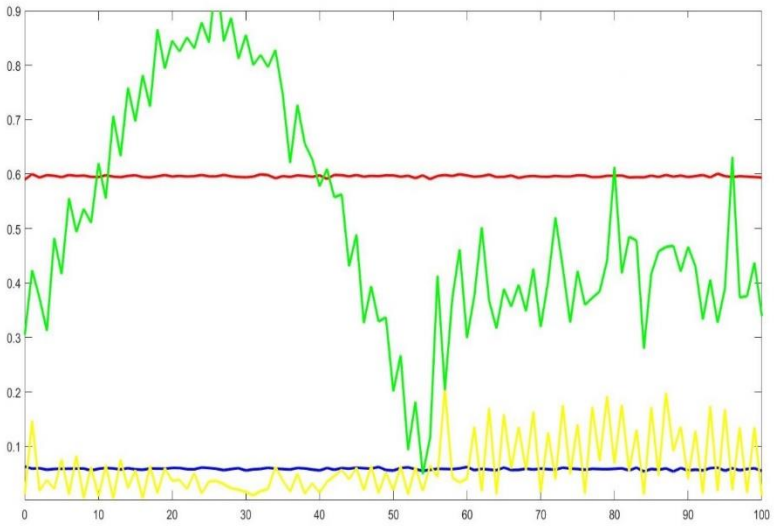
C. Standard Deviation of Accelerometer along x and z direction:

- **Intuition**

  The standard deviation gives us the spread of the values around the mean. The accelerometer values show different range for different activities. Thus, it is easier to distinguish eating and non-eating activities using this parameter. Non-eating activities will have higher acceleration values along x-axis as compared to eating activities. Also, eating actions will have higher variation along z-axis as compared to non-eating activities.

- **Conclusion**

  From the graphs that we have plotted, accelerometer z standard deviation does not vary much with the eating or non-eating activities because it is predominated by acceleration due to gravity. Hence, the distinction is not clear. But there is clarity in the distinction along the x-axis as discussed above. Thus, the intuition is satisfied. This graph is for x-axis. Green and Red are for non-eating while yellow and blue show the eating ones.



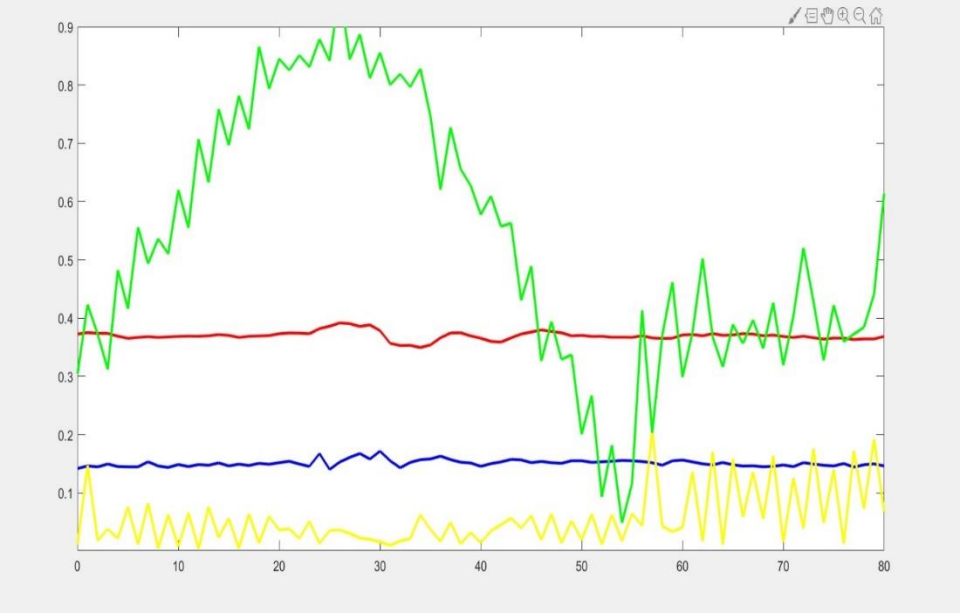
D. Standard Deviation of Gyroscope along x, y and z direction:

- **Intuition**

  Gyroscope attributes give rotation and orientation of the hand. Since, eating activity will have a lesser spread as compared to non-eating ones thus we should get a clear distinction between these two.

- **Conclusion**

  From the graphs that we have plotted, we can see a clear distinction in the standard deviation patterns of the gyroscope along y axis for eating and non-eating activities. Eating activities have a lesser standard deviation than the non-eating ones. Thus, the intuition is satisfied. Eating actions are denoted by blue and yellow lines.



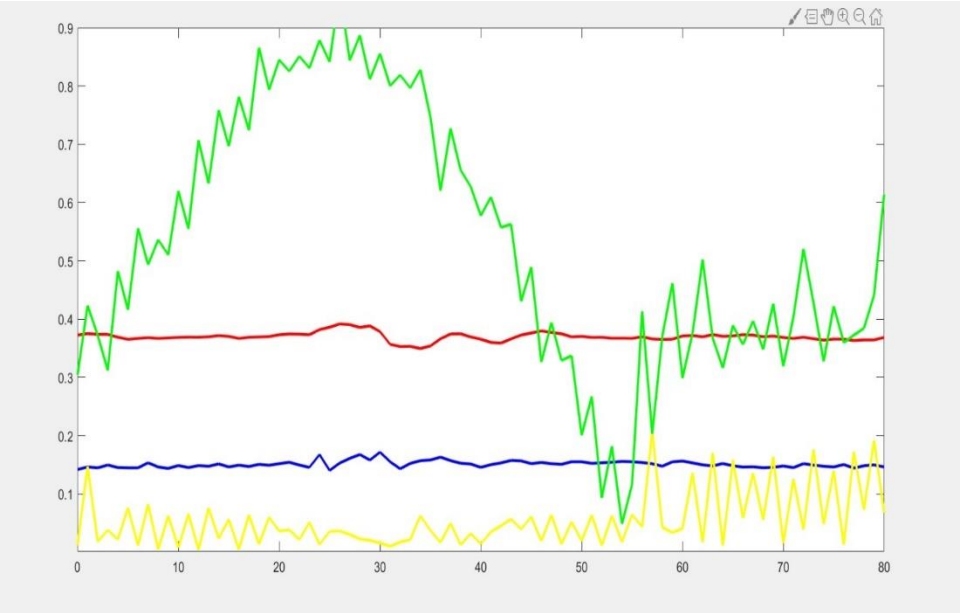
E. Root Mean Square of Orientation along x-axis:

- **Intuition**

  The eating activities will have a lower range of orientation changes along the x-axis as compared to the non-eating ones.

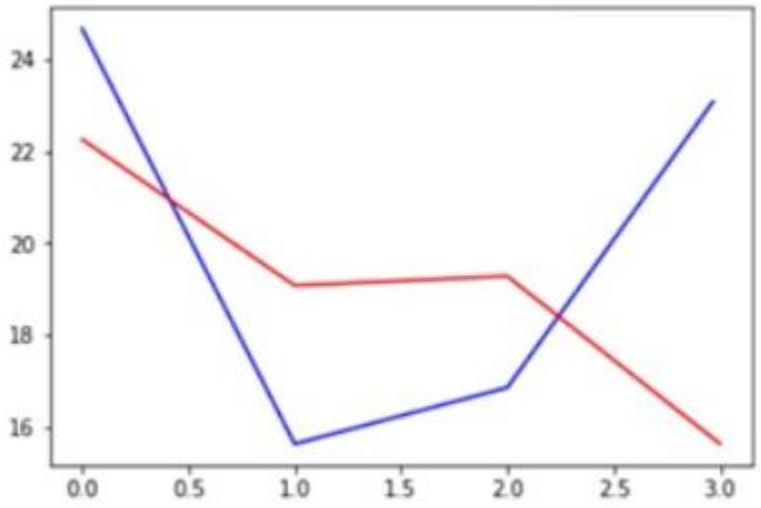
- **Conclusion**

  From the graphs that we have plotted, the non-eating activities have a larger rms value and is clearly distinct from the eating activities. Thus, the intuition is true. Non-eating actions are denoted by green and red.



F. Fast Fourier Transform of EMG data samples:

From the graphs that we have plotted, the time domain data samples are converted to frequency domain. FFT values are collected for each EMG signal and shows a clear distinction between the activities as it maps the signal to a new frequency space. Blue denotes the eating action and red signifies non-eating.



6. Feature Selection:

This is the final section of the project where we have used the Principal Component Analysis feature extraction method on the feature matrix that we have generated, to find and select the distinct features. It reduces the dimension of feature matrix and selects top k-latent semantics to show the variance of different activities. In this case, we have applied PCA on a mxn feature matrix where 'm' number of rows specify m number of activities and 'n' number of columns specify n number of features extracted as part of the feature extraction phase.

- **Arranging the feature matrix:**

Six features are extracted which are shown as columns of the feature matrix. The rows of the feature matrix are the eating and non-eating activities. Two separate feature matrices are generated for eating and non-eating activities. We have applied PCA to understand the directions along which the distinctions can be seen properly thus resulting in dimensionality reduction.

- **PCA Execution:**

MATLAB PCA code is used to run PCA on the feature matrix. Co-Variance matrix when multiplied with a random vector gives the direction of greatest

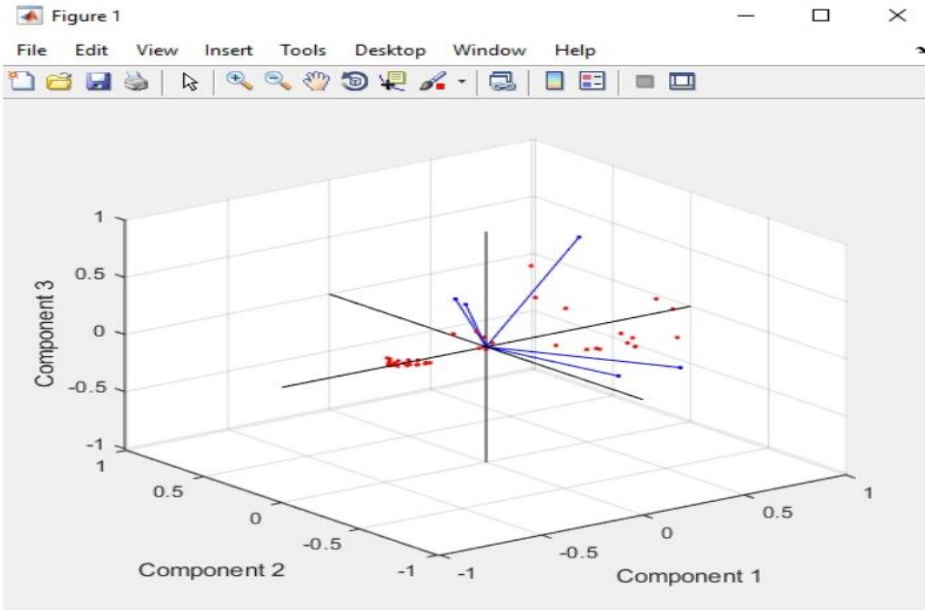
variance. Eigen vectors are the vectors whose directions remain constant when we multiply co-variance matrix by a vector. Thus, it shows the highest variance.

- **Make sense of PCA eigen vectors:**

As we multiply eigen vectors with the data point it gives us vectors of different magnitude in the same direction. Thus, the PCA gives us the direction of greatest variance.

- **Results of PCA:**

The dimensionality of feature matrix gets reduced after performing PCA.



- **Argument for PCA: Helpful or not**

Yes, its helpful. Because there were several features in the original matrix and we reduced the dimensionality using PCA to select the principal components which show variance in overall eating and non-eating data. After comparing the plots, it can be concluded the complexity of data is reduced by PCA while keeping the data intact.

References:

[1] https://www.techopedia.com/definition/26136/statistical-mean
[2] https://en.wikipedia.org/wiki/
[3] https://www.quora.com/What-is-meant-by-entropy-in-machine-learning-contexts
[4] https://www.saedsayad.com/decision_tree.htm