Report of Assignment 1

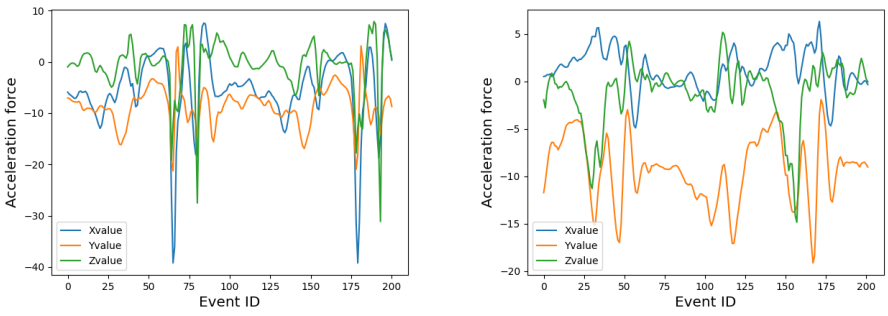
Kai Li

03/04 2021

1 Part A

1.1 Question-a

The plotted figures for UserA_1 and UserB_1 are presented in Figure 1.



(a) UserA_1 (b) UserB_1

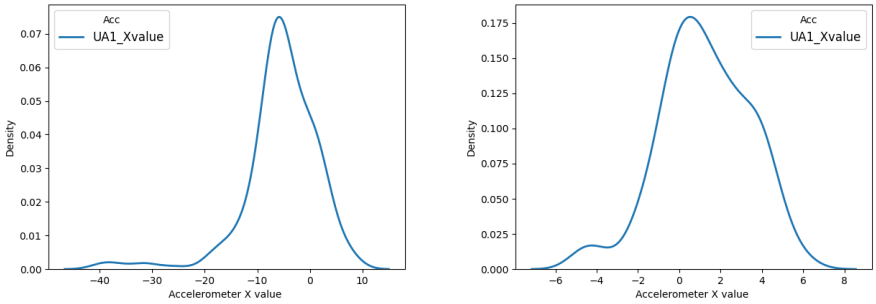
Figure 1: Two users' X/Y/Z value from DataSamples

1.2 Question-b

The density curves for UserA_1 and UserB_1 are presented in Figure 2.

1.3 Question-c

From Figure 1, it can seen that about every 110 events, the X/Y/Z values of both two users have a significant drop. From Figure 2, it can be found that both two users' X value follows the Gaussian distribution, with most of the data points fell in [-10, 0] for userA_1 and [-2, 4] for userB_1.



(a) UserA_1 (b) UserB_1

Figure 2: Two users' density of X-value from DataSamples

2 Part B

2.1 Question-a

The selected features are mean/standard/max/min/sum values of the X axes.

2.2 Question-b

Please find the code in the attached file.

2.3 Question-c

- a) The chosen features can summarize the data distribution at a high level, they are the typical statistical features.
- b) Other features can be the Unique values and Number of peaks/valleys.
- c) They can't work for categorical type of data and Nominal type of data.

3 Part C

3.1 Question-a

I will choose **Euclidean** Distance and **Manhattan** Distance. Euclidean Distance represents the shortest distance between two points while Manhattan Distance is the sum of absolute differences between points across all the dimensions.

3.2 Question-b

Please find the results in *Distance.csv*.

3.3 Question-c

The distance results show that the distance of intra-user is always much smaller than that of inter-user. This can be explained by the density curves in Figure 2, because both user's X-value are high centralized while their dense points are different.

3.4 Question-d

For *Age* and *Height*, since they are non-categorical data, we can use **Minkowski** Distance or **Mahalanobis** Distance to measure the distance. For *Gender*, *Lifestyle*, *Handedness*, *Likes sunny days*, they are categorical data, we can use *Heterogeneous Euclidean Overlap Metric* to measure the distance.

4 Part D

4.1 Question-a

We use the same features as Section 2.

4.2 Question-b

By measuring the distance of the test samples to the trained feature of userA and userB with two distance functions, namely **Euclidean** Distance and **Manhattan** Distance, the result shows all 3 samples belong to userB.