CIS 735 Machine Learning for Security

Syracuse University

Spring 2021

Homework Assignment 1

Due: March 5, 5 pm (Upload on the course Blackboard system)

Instructions: You can use any programing language that you are comfortable with to meet the task requirements. Tasks require visualization and data analysis capabilities.

Submission Requirements: Submit a single zip file named with your SU ID "<SUID>.zip". The zip file must contain the following.

a) A single PDF document with plots and answers for tasks in Part A, B, C and D.

b) Code files and any other supporting material used in preparation of your submission.

About the Data: The zipped file "DataCIS735_Assignment1.zip" contains the data samples for this assignment. These are sourced from SU-AIS BB-MAS dataset[1], the entire dataset can be accessed at http://dx.doi.org/10.21227/rpaz-0h66. (not required for the scope of this assignment)

There are five samples each, for "User A" and "User B" (total: ten files), in the "DataSamples" folder. There are three test-samples that are labelled "Test1", "Test2", and "Test3", in the "TestSamples" folder. The "Demographics.csv" file contains demographic information about user A and B.

Each sample consists of about two seconds of **walking data** (this is known as windowing approach, data provided is already split into two second windows) collected by the phone in user's pocket. For this assignment we will only use the data from the accelerometer.

The accelerometer files have five columns, "EID": event ID (Integer); "Xvalue", "Yvalue", "Zvalue": the acceleration force in m/s^2 on x, y and z axes respectively, excluding the force of gravity (Float); and "time": the timestamp of the data point (String in date-time format with millisecond resolution).

Note: The test-samples do not have real timestamps or EID (all initialized to start with 1). You are free to use any preprocessing steps or additional steps (beyond the requirements), but, have to document them clearly in the PDF submission.

---

[1] "SU-AIS BB-MAS (Syracuse University and Assured Information Security - Behavioral Biometrics Multi-device and multi-Activity data from Same users) Dataset ", IEEE Dataport, 2019. [Online]. Available: http://dx.doi.org/10.21227/rpaz-0h66

Part A (10 Points): Basics and Visualization

a) Plot two data samples, one from each User. The plots must have accelerometer signals from X, Y, Z axes, a legend and title (the sample file name).
b) In a separate graph, plot the density curves of the "Xvalue" columns from the two samples.
c) Provide a summary of your observations from the plots. Summary need not be highly technical in nature; this is a practice to articulate insights from visualizing data.

Part B (30 Points): Feature Design, Description and Extraction (to be performed on all data samples)

a) Design and describe at least 5 features (or more) to be extracted from each (X,Y, Z) columns of the data-samples.
b) Write a program to extract the 5 features (or more) from each data-sample and store them in a single file named "Features.csv" with the following format:

User	Sample no.	<Feature_1>	<Feature_2>	..	..	<Feature_n>

Where, "User" is A or B, "Sample no." is the data-sample file number, and <Feature_1> to <Feature_n> are feature values extracted from the sample (label the feature columns accordingly, for example if feature is "mean" of Xvalue , then label the feature "mean_X")

c) Provide a summary of your observations from tasks in part B:
    a. Why did you choose the features that you did?
    b. What are other features that can be explored for this data?
    c. Do you think the selected features would work for all forms of data, why or why not?

Part C (30 Points): Distance Measures, Measuring Similarities or Dissimilarities

a) List and describe at least 2 distance measures that can be used to measure similarities or dissimilarities between the feature vectors (rows) in "Features.csv".
b) Write a program to implement the two distance measures to measure distances between each feature vector pairs. Store the results in a single file named "Distance.csv" with the following format:

FV1_User	FV2_User	<DistanceMeasure_1>	<DistanceMeasure_2>

Where, “FV1_User” and “FV2_User” columns are the User to whom the Feature Vectors (that are being compared) belong to and “<DistanceMeasure_1>” and “<DistanceMeasure_2>” are the results of applying the distance measure on the two vectors. (label the distance columns accordingly, for example if distance function is “Euclidean”, name the column “Euclidean_Distance”)

c) Briefly explain how the distance measures performed with the help of density curves for intra-user and inter-user distances.

d) If you were to use the demographic information as features. How would you transform them to work with the distance measures? Briefly explain for each column (except “User”) in the “Demographics.csv” file.

Part D (30 Points): A simplistic matching attempt.

a) Using the techniques from Part B, extract features from test-samples “Test1”, “Test2” and “Test3”.

b) Using a distance measure from Part C, classify each test sample as either: “belongs to user A” or “belongs to user B”. Describe your method to arrive at a decision.