CSE 572 Data Mining Assignment 3

Due Date: 18th October, 2017

Points Possible 120

This project involves feature extraction and feature selection aspects of Data Mining. You will be working with the raw sensor data that you collected and annotated in your Assignments 1 and 2. In your raw data there are 17 data streams: a) 3 from accelerometer, b) 3 from gyroscope, c) 3 from orientation, and d) 8 from EMG sensors.

Task 1: (20 points)

Write a matlab code that uses the annotations from the Assignment 2 and segments the raw data into two separate classes: a) eating, and b) non-eating actions. These two classes can be stored in two separate csv files. In the each csv file you can store the time series for an action columnwise and each row indicates a given sensor. Append multiple actions in rows. For example the eating action csv file should look like the following:

Eating Action 1 Acc X 2 3 4 5 1 5 1 6 2 7 8 3 2 1 3 ----------------

Eating Action 1 Acc Y 2 3 4 5 1 5 1 6 2 7 8 3 2 1 3 ----------------

|

|

Eating Action 2 Acc X 2 3 4 5 1 5 1 6 2 7 8 3 2 1 3 ----------------

|

|

Task 2: Feature extraction (50 points)

In this task, you should select and implement five existing feature extraction methods such as Fast Fourier Transform, Discrete Wavelet Transform, a set of statistical features (min, max, avg, std, RMS, energy function), etc. The five types of feature extraction methods can be chosen by you. The aim is to use features that show clear distinction between eating and non-eating actions.

For each type of feature extracted do the following things,

a) write an explanation on how the feature is extracted.

b) Write an intuition on why you use such a feature

c) Write a matlab code to extract that feature from each time series stored in the csv files created in task 1.

d) Generate two plots: i) features extracted from all eating actions, and ii) features extracted from all non-eating actions. For multiple eating or non-eating actions you can choose to overlap the plots. This will give you a better idea of potential patterns in the features.

e) Discuss whether your initial intuition about the features that you selected holds true or not.

Task 3: Feature Selection (50 points)

This step involves reduction of the feature space and keeping only those features which show maximum distance between the two classes. We will use Principal Component Analysis technique discussed in class for this purpose. The PCA code is already available in Matlab, hence there is no need to PANIC! Just use it.

Subtask 1: Arranging the feature matrix

You know PCA only takes one matrix. How will you arrange all sensors and their corresponding features into a single matrix such that the eigenvectors of the covariance matrix directly makes sense to your data set? This means that if the PCA results gives you a eigen vector then the new feature matrix can be obtained by simply multiplying the eigen vector with the old feature matrix. (You might need two matrices corresponding to the two classes)

Write your logic of feature matrix arrangement.

Subtask 2: Execution of PCA

Use Matlab’s PCA function to run PCA on your feature matrix. Show all the eigen vectors in a plot.

Subtask 3: Make sense of the PCA eigen vectors

Write an explanation on the reason why the eigen vectors turned out the way they did.

Subtask 4: Results of PCA

Create the new feature matrix. Generate two plots: i) features extracted from all eating actions, and ii) features extracted from all non-eating actions. For multiple eating or non-eating actions you can choose to overlap the plots.

Subtask 5: Argue whether doing PCA was helpful or not. May be compare the plots generated from subtask d of task 2 and subtask 4 of Task 3.

Submit Code and PDF file with your explanations in Blackboard.