CSE 572 Data mining Project 1

This is an individual project

Due Date: February 10th , 2019 (total points 110)

Input: Five cell arrays:

a) The first cell array has tissue glucose levels every 5 mins for 2.5 hrs during a lunch meal
The data starts from 30 mins before meal intake an continues up to 2 hrs after the start of meal consumption
There are several such time series for one subject.
b) The second cell array has time stamps of each time series in the first cell array
c) The third cell array has insulin basal infusion input time series at different times during the 2.5 hr time interval
d) The fourth cell array has time stamps for each basal or bolus insulin delivery time series.
e) The fifth cell array has insulin bolus infusion input time series at different times during the 2.5 hr time interval

Some facts about the data:

Each cell array is an array of time series each of which can have varying lengths.

Each subject has multiple such time series but the total number of time series data for each subject may vary.

You have data from 5 subjects

The insulin input may not be every 5 mins hence the insulin time series length may vary significantly

The time stamp which has the highest insulin delivery is the time at which the meal was logged.

Tasks:

a) Extract 4 different types of time series features from only the CGM data cell array and CGM timestamp cell array (10 points each) total 40
b) For each time series explain why you chose such feature (5 points each) total 20
c) Show values of each of the features and argue that your intuition in step b is validated or disproved? (5 points each) total 20
d) Create a feature matrix where each row is a collection of features from each time series. SO if there are 75 time series and your feature length after concatenation of the 4 types of featues is 17 then the feature matrix size will be 75 X 17 (10 points)
e) Provide this feature matrix to PCA and derive the new feature matrix. Chose the top 5 features and plot them for each time series. (5 points)
f) For each feature in the top 5 argue why it is chosen as a top five feature in PCA? (3 points each) total 15