CSE572
Feb 10th, 2020
Wei Xin
1207050893

a) Extract 4 different types of time series features from only the CGM data cell array and CGM timestamp cell array.

Code is provided in project_1.m.
Four different types are covariance, polynomial coefficient, glucose difference, speed of tissue glucose level.

b) For each time series explain why you chose such a feature.

Feature #1 - covariance of tissue glucose levels every 10 min
Covariance is a measure of how much time stamp and glucose level vary together. It's similar to variance, but where variance tells us how a single variable varies, covariance tells us how two variables vary together.

Feature #2 - get the polynomial coefficient of the tissue glucose levels from the CGMSeriesLunchPat forms
There is a "normr" variable which is the norm of the residuals. It is used to constant the goodness of your fit. The smaller the amount, the better the fit.

Feature #3 - calculate the difference of every time duration (10 min) of the tissue glucose level
This can give us an intuition of how a patient performs during a specific time period to better evaluate the glucose level and when he or she should take pills

Feature #4 - find the speed of tissue glucose level of each test per person at every two timestamps
This gives us the speed of a patient's glucose level either decreasing or increasing. We don't care about the number of negative or positive so it is a matter of absolute values. Also, an intuition of when a patient should take pills and control his glucose level. I divided the value by (-100000) because the time stamp is 10^6 so I just want to control the values between 0 and 1 for a better look in the graph.

c) Show values of each of the features and argue that your intuition in step b is validated or disproved?

Feature #1 - covariance = [0.0111, 0.0183, 0.0536, 0.0481, 0.0306, 0.0586, 0.0219, 0.0036, 0.0128, 0.0179, 0.0844, 0.0379, 0.0145, 0.0071, 0.0383]

As we can see, the covariances are between 0.0036 to 0.0844 and these numbers are very close to each other so there is a very tight relationship between time and glucose level. It is validated.

Feature #2 - polynomial coefficient is [0.0001, -0.0114, 0.3336, -3.9072, 9.4128, 248.7819] And the normr value is 33.8556 which means that the polyfit is not as fit as what we want because the number is so big. There is a huge difference. It is validated.

Feature #3 - diff = [-2, -28, -26, -27, -7, 3, 3, -32, 6, -9]. From this we can tell that every 10 min there is an increasing or decreasing on the glucose level and if its absolute value is large, then there is a need to have food intake or reduce food intake. Otherwise, it might be in the time that there is no food intake. The values are validated because there are large values and small values.

Feature #4 - speed_array = [0.737280340576329 0.720000042915347 0.711359827748579 0.671040039997103 0.639360059545046 0.601920035877230 0.581759859130417 0.564480033645632 0.547200050961976 0.515520030727388 0.486720029010774 0.472319885630635 0.466560027809145 0.466560043451790 0.463680027637483 0.466559887025384 0.475200028324129 0.483840045061116 0.492480029354097 0.483840028839113 0.420479898183370 0.397440023689271 0.391680036478046 0.397440023689271 0.403199902367616 0.408960024375917 0.406080037819151 0.394560023517610 0.383040022830965 0.371519910038731]. We can tell by the numbers that the speed of glucose value goes up during eating and goes down after eating. Somehow, the speed goes up again after about 1.5 hours of eating. However, the numbers remain around 0.4 so we can tell the patient is normal. The values are validated because it obeys the rule of glucose level.

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 features is 17 then the feature matrix size will be 75 X 17.

Code is provided in project_1.m
My feature matrix size is 61 * 18

e) Provide this feature matrix to PCA and derive the new feature matrix. Choose the top 5 features and plot them for each time series.

Code is provided in project_1.m

f) For each feature in the top 5 argue why it is chosen as a top five feature in PCA?

The PCA space which is principal component scores are the representations of X in the principal component space. Rows of score correspond to observations, and columns

correspond to components. The principal component variances are the eigenvalues of the covariance matrix of X.

The generated 18 dimensional processed data is stored in the score. It is the analysis of the original data from the series csv files, and then the data obtained in the new coordinate system. We ranked the 18-dimensional data in descending order of contribution. (That is, when the coordinate system is changed, the 18-dimensional data is sorted). Therefore, the top 5 features contribute the most among all 18 dimensional components.

There are 3 charts in the code which represents the first five features' eigenvectors, scores, and all the features' values. The first five features contribute almost 98% of the whole data. As we can tell from the charts, the first feature fluctuates the most, meaning it's variance is the largest, this corresponds to the first test sample's covariance, correlation coefficient, difference, and the speed of the glucose level. At some point, the values of the above 4 features are the largest, which makes the piece of data valuable.

From then on, the second top feature contains three of the four components, which might be correlation coefficient, difference, and the speed. The third top contains two of the four components, which might be the difference and the speed. The fourth contains only one of the components. The very last one contains the least important feature. These features can obtain almost all the valuable data we want to analyze. From here, I think the top N features can have this equation:

N = Number of features mined from the original data + 1

In this case, we mined 4 different features so we chose the top 5 features extracted from the PCA.