Assignment 3 Data Mining CSE 572

Given: Meal Data of 5 subjects

Amount of carbohydrates in each meal

Todo:

a) Extract features from Meal data
b) Cluster Meal data based on the amount of carbohydrates in each meal

First consider the given Meal data. Take the first 50 rows of the meal data. Each row is the meal amount of the corresponding row in the mealDataX.csv of every subject. So mealAmountData1.csv corresponds to the first subject. The first 50 rows of the mealAmountData1.csv corresponds to the first 50 rows of mealDataX.csv in Assignment 2.

Extracting Ground Truth: Consider meal amount to range from 0 to 140. Discretize the meal amount in bins of size 20. Consider each row in the mealDataX.csv and according to their meal amount label put them in the respective bins. There will be 8 bins starting from 0, >0 to 20, 21 to 40, 41 to 60, 61 to 80, 81 to 100, 101 to 120, 121 to 140.

Now ignore the mealAmountData. Without using the meal amount data use the features in your assignment 2 to cluster the mealDataX.csv into 8 clusters. Use DBSCAN or KMeans. Try these two.

Report your accuracy of clustering based on SSE and supervised cluster validity metrics.

Grading: I will give you a set of Meal data that is not included in the training set.

50 points for developing a code in Python or Matlab that takes the dataset and performs clustering

20 points for developing a code in Python or Matlab that implements a function to take a test input and run the clustering algorithm to provide the clustering result.

30 points will be evaluated on the SSE results obtained by your machine. This will be compared against class average to determine the final score.

Example:

O –	1 , 6, 9 10	1 – 3,4,9,11,12,15 → >0 <= 20
>0 <= 20	3,4,5, 11, 12 13	2 – 1, 2, 10 → 0
>20 <= 40	2, 7,8, 14 15	3 – 5, 6, 7,8,14 → >20 <=40

Classification error → supervised cluster validity metric

2 + 1 + 2 = 5

Error 5/15 = 33.33 %

Test script that does KNN classification choose K, choose distance metric

Given a test data, calculate distances of the test data from each of your training data point.

Then do a K majority based classification

DBSCAN and K means

Test data input has 100 rows of mealDataX.csv

Output

Matrix

1	2
1	2
2	4
2	5
3	2
2	2
2	2
4	4
6	6
7	7
5	5

Output class numbering start from 1 and go till 8