University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CSE 012/CS 059
Fall 2014

Assignment 2

The deadline for this part is at the beginning of the class on December 16th. For late submissions the late policy on the page of the course will be applied. The details for the turn-in are on the page of the course. Turn in the requested material as well as all the code that you have written, and some instructions on how to run your code.

Question 1

In the course textbook (Introduction to Data Mining by Tan, Steinbach, Kumar) except for the clustering algorithms that we described in class, there are also the algorithms CLARANS, BIRCH, ROCK, CHAMELEON, DENCLUE, and CURE (also described in the free online textbook Mining Massive Datasets by Rajaraman and Ullman). Select one of these algorithms, and describe the main idea in your own words in 2-3 paragraphs. You can read the corresponding paper if you need additional details.

Question 2

You are given a set of N objects that we want to cluster. Instead of a pairwise distance matrix you are given an $N \times N$ matrix, where for each pair (x, y) you are given a value +1/-1 (or true/false) for whether this pair of objects should be place in the same cluster, or a different one. You should propose a clustering methodology for this problem. Your methodology should have the following two steps:

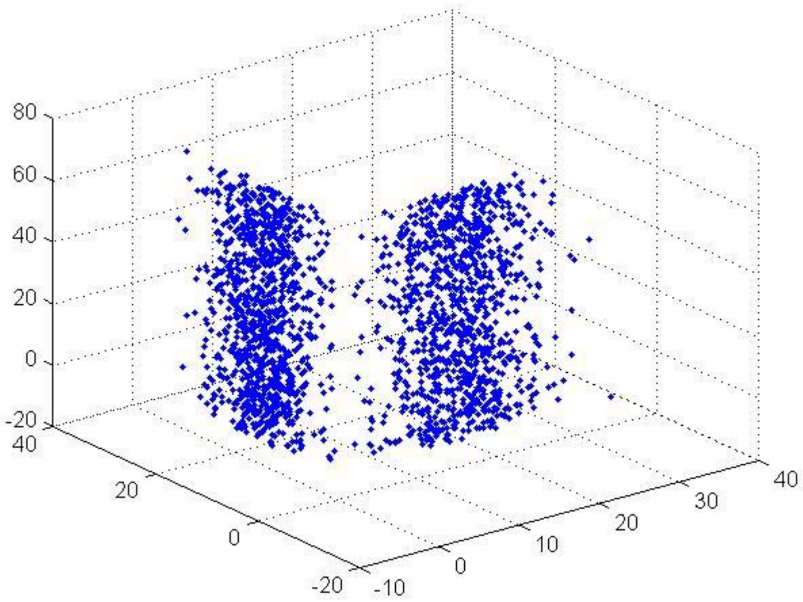
1. Define a function Q which given a clustering C of the N objects into k clusters $C = \{c_1, , c_2, \dots , c_k\}$ it computes the “quality” $Q(C)$ of the clustering. Justify your choice.
2. Propose an algorithm that tries to find the clustering C with the best quality Q(C). If necessary you can assume that the number of clusters k is given as input to the algorithm.

The goal of this question is to make you think about how to approach a “new” problem. Trivial solutions that demonstrate lack of effort will receive low marks.

Question 3

For this question it is recommended that you use MATLAB.

You are given the file “3d-data.txt” (and “3d-data.mat” to be loaded directly to MATLAB) which contains 2000 three-dimensional points. The three-dimensional plot of the points is shown in the figure below.



You should answer the following questions:

1. What do you observe in the figure above? Is there an obvious clustering structure in the data?
2. Apply the k-means algorithm to the data. Create a three-dimensional plot with the different clusters having different colors. What do you observe? Explain why.
3. Take the first two left singular vectors of the data matrix and apply the k-means algorithm on the points in the reduced dimension. Create a three-dimensional plot of the three-dimensional points with the different clusters having different colors. What do you observe? Explain why.
4. Implement the EM algorithm described in class (for three-dimensional points), and apply it to the three-dimensional data. Again, create a three-dimensional plot with the different clusters having different colors. What do you observe? Explain why.

Technical details:

- To implement EM for multidimensional data, you need to compute all the quantities described in class for all dimensions
- The following MATLAB commands may be useful:
  - help <command>: gives information about a specific command
  - D = load ‘data.txt’: loads the ascii data to a matrix D
  - load ‘data.mat’: loads the mat file and creates the matrices that were saved in the mat file
  - kmeans: computes the k-means clustering of a matrix of data
  - find(X==1): returns a vector with the indices of the entries in X that have value 1.
  - scatter3(D(:,1),D(:,2),D(:,3),’.b’): computes the above plot. The last argumen specifies that we want points (‘.’) in blue (‘b’) color.
  - D(x,:): Returns a matrix with all the rows of the matrix D with index in x

- svds: computes the Singular Value Decomposition
- hold: “holds” a figure so that the next command plots on the same figure (the most recent one).

Question 4

Use the dataset in Questions 2 of Assignment 1b and produce a clustering of the courses. To produce the clustering you can use some existing implementation of one of the algorithms we have seen in class, or an implementation of your own. To find the “correct” number of clusters use the techniques we described in class. Examine the clusters and try to understand what each cluster represents. Write a report where you describe the methodology that you used and your findings.