Data Mining
Information Systems Department
Faculty of Computers and Information
Cairo University

Assignment 2

Clustering

Instructions:

1. Assignment should be done individually; copies will be graded to zero.
2. Total grade is 5 marks.
3. No late submissions are allowed.
4. Discussion will be held with Eng. Dina Mohamed as per the below distribution:

Group	Office Hours
IS_DS (G1 & G2)	Monday 3-Decamber-2018 (9.30 – 11)
IS_DS (G3 & G4)	Tuesday 4-Decamber-2018 (9.30 – 11)
IS_DS (G5)	Monday 3-Decamber-2018 (11:00 – 12:30)
IS_IT	Monday 3-Decamber-2018 (9.30 – 11)
IS_CS	Tuesday 4-Decamber-2018 (9.30 – 11)
Pre-master	Tuesday 4-Decamber-2018 (9.30 – 11)

5. Each student can choose any of the two problems in the next pages.

Problem 1

Description:

- Consider the Course Evaluation dataset in Course evaluation file.xlsx, it contains course evaluation scores provided by 150 students each row represents scores provided by one of the students for each question in the survey.
- Write a program in any programming language you prefer to group the students based on the similarity of their answers on the survey**.**
- You should use k-means algorithm to cluster the students to k clusters
- Number of clusters (k) will be provided from the user as an input.
- Initial centroid should be choosing randomly.
- You should use Euclidean distance as your distance function.
- You should detect outlier data (if exists).
- The final output of your program should show k lists of students and show outlier student's records.

Problem 2

Description:

- Consider Sales dataset in Sales.xlsx file, it contains weekly purchased quantities of 200 over products over 30 weeks.
- Write a program in any programming language you prefer to group the products based on the similarity of their purchased quantity over all 30 weeks.
- You should use k-means algorithm to cluster the products to k clusters
- Number of clusters (k) will be provided from the user as an input.
- Initial centroid should be choosing randomly.
- You should use Manhattan distance as your distance function.
- You should detect outlier data (if exists).
- The final output of your program should show k lists of products and show outlier product's records.