Athens University of Economics and Business

MSc in Business Analytics

Data Mining Techniques – Assignment 1

Deadline: 23/6/2019

Group assignment (groups of up to 2 people).

The assignment corresponds to 20% of the total grade of the course.

Discussions between groups are recommended but collaborating on the actual solutions is considered cheating and will be reported.

There will be no extension of the assignment deadline!

Professor: Y.Kotidis (kotidis@aueb.gr)

Assistant responsible for this assignment: I.Filippidou (filippidoui@aueb.gr)

Assignment 1

The goal of this assignment is to implement a simple workflow that will assess the similarity between bank customers and suggest for any input customer a list of his/her 10 most similar other customers. In order to compute the similarity between customers you will have to create the dissimilarity matrix for every given attribute as discussed in lecture "Measuring Data Similarity". In order to fulfill this assignment, you will have to perform the following tasks:

1) Import dataset with bank customers

You will download the bank.csv dataset from moodle. This dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls in order to access if the product (bank term deposit) would be (or not) subscribed. The dataset includes 43192 bank customer profiles with 8 attributes each. The class attribute should be ignored. Full description for the dataset and the attributes are provided in the bank-names.txt file.

2) Compute data (dis-)similarity

In order to measure the similarity between the bank customers you will form the dissimilarity matrix for all given attributes. As described in lecture "Measuring Data Similarity", for every given attribute you must distinguish its type (categorical, ordinal or numerical) and form its dissimilarity matrix. Then you will calculate the average of the computed dissimilarities in order to form the dissimilarity matrix over all attributes.

3) Nearest Neighbor (NN) search

Using the dissimilarity matrix computed from the previous step, you will calculate the 10-NN (most similar) to the customers listed below (customer id=line number in the csv file starting from line 2):

1230, 5032, 10001, 24035, 28948, 35099, 37693, 39543, 40002, 42192

For this task your script must take as input the customer-id and return the list of her 10 nearest neighbors (excluding the given customer)

Assignment handout:

1) A report (pdf) describing in detail any processing and conversion you made to the original data and the reasons it was necessary. The report will also contain examples of how to use your script and its output to the list of customers provided at step 3.
2) The program/script you implemented for calculating the dissimilarity matrix. Implementation can be done in any programming language and should be accompanied by the necessary comments and remarks.
3) The pdf as the required programs/scripts should be uploaded to moodle until the assignment deadline.