Athens University of Economics and Business

MSc in Business Analytics

Data Mining – Assignment 1

Deadline: 24/5/2020

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 supermarket customers and suggest for any input customer a list of his/her 10 most similar other customers. In order to calculate the similarity between customers you will first have to compute the dissimilarity 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 and pre-process the dataset with customers

You will download the groceries.csv dataset from moodle. This dataset contains demographic characteristics of supermarket customers along with a list of groceries. In specific the dataset includes 10000 supermarket customer profiles with the following attributes:

Customer ID: The unique id of the customer.

Age: The age of the customer.

Sex: Male-Female.

Marital Status: Married, Single, Divorced.

Education: Primary, Secondary, Tertiary.

Annual Income: The annual customer income.

Customer Rating: The rating of the supermarket from the customer (Poor, Fair, Good, Very Good, Excellent).

Persons in Household: Number of persons in the household.

Occupation: The occupation of each customer (retired, housemaid, unemployed, management, entrepreneur, blue-collar, self-employed, services, technician).

Groceries: A list of the customer groceries.

For any numerical missing values, you should replace them with the average value of the attribute (keeping the integer part of the average).

2) Compute data (dis-)similarity

In order to measure the similarity between the bank customers you could form the dissimilarity matrix for all given attributes. As described in lecture "Measuring Data Similarity", for every given attribute you first distinguish its type (categorical, ordinal, numerical or set) and then compute the dissimilarity of its values accordingly. For set similarity use the Jaccard similarity between sets. Then, you can calculate the average of the computed dissimilarities in order to form the dissimilarity over all attributes. Depending of the machine used to implement this assignment you should decide whether is feasible to compute the dissimilarity matrices or have the computations performed on-the-fly for a pair of customers.

3) Nearest Neighbor (NN) search

Using the dissimilarities computed as discussed in the previous step, you will calculate the 10-NN (**most similar**) customers for the customers with ids listed below:

73, 563, 1603, 2200, 3703, 4263, 5300, 6129, 7800, 8555

For this task your script must take as input the customer-id and return the list of her 10 nearest neighbors (**most similar**), along with the corresponding **similarity score.**

An example of the script output for customer id =1 follows:

10 NN for Customer 1	
Customer ID	**Similarity Score**
7749	
7931	
9514	
628	
6918	
4230	
3148	
4647	
2105	
8050	

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 (10-NN and the corresponding similarity score for every given id)**. The first page of the report should clearly state the names and student ids of the members of the group.
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 and the required programs/scripts should be uploaded to moodle until the assignment deadline. You should create a compressed (e.g. zip/tar) file containing the report, your code and any other files required for executing your script (you do not need to include the original dataset). The name of the compressed file should include the student ids of the members of the group.