Athens University of Economics and Business

MSc in Business Analytics

Data Mining – Assignment 1

Deadline: 24/5/2022

Group assignment (groups of up to 2 people).

The assignment corresponds to 25% 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. Moreover, you will be using these results to predict the rating of the customer to the bank. 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

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 to access the opinion of the customer to the bank service and products. The dataset includes 43191 bank customer profiles with 10 attributes each. Below is a description of the available attributes:

Age: The age of the customer.

Job: type of job (admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services).

Marital Status: Married, Single, Divorced.

Education: Primary, Secondary, Tertiary.

Default: If the customer has credit in default (yes/no).

Balance: average yearly balance, in euros.

Housing: if the customer has a housing loan (yes/no).

Loan: if the customer has a personal loan (yes/no)

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

Products: An array containing the bank products (1-20) each customer has.

For any numerical values missing, you should replace them with the average value of the attribute in the dataset (rounded to the nearest integer). The replaced average values calculated should be reported in the pdf.

**2) Compute data (dis-)similarity**

To assess the similarity between the 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 to derive the dissimilarity over all attributes. Depending on the machine used to implement this assignment you should decide whether it 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 implementation of the previous step, you will calculate the 10-NN (**most similar**) customers for the customers with ids listed below (**customer id=line number-1):**

1200, 3650, 10400, 14930, 22330, 25671, 29311, 34650, 39200, 42000

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** |
| 16641 | |
| 12329 | |
| 1247 | |
| 33282 | |
| 25849 | |
| 24715 | |
| 6001 | |
| 31996 | |
| 5914 | |

| | |
|---|---|
| 7894 | |

**4) Customer rating prediction**

For this assignment you will implement a classification algorithm which, for a given customer, will predict his rating (poor, fair, good, very good, excellent) for the bank. In order to implement the classification for a given customer you need to:

1) Calculate the similarities between the given customer and all other customers and compute his 10-nn (most similar) customers. **IMPORTANT: In the similarity calculations for this step you need to exclude the customer rating attribute.**

2) Based only on the 10 most similar customers computed in the previous step, predict the customer rating rank using:

- The average rating rank of the 10 most similar customers (rounded to the nearest integer).
- The weighted average rating rank of the 10 most similar customers (rounded to the nearest integer).

$$\text{Weighted average rating\_rank} = \frac{\sum_{\text{i}=1}^{10}[\text{similarity(i)} * \text{rank}(\text{rating(i)})]}{\sum_{\text{i}=1}^{10}\text{similarity(i)}}$$

Where:

  - rating(i) = the rating of the i-th nearest neighbor (i=1 for the most similar customer)
  - similarity(i) = the similarity of the i-th nearest neighbor with the given customer

3) For the evaluation of your classification algorithm you will use the 50 first records of the bank dataset and predict the rating for them. Then, for all n=50 records calculate the Mean Prediction Error for both prediction methods.

$$\text{Mean Prediction Error} = \frac{\sum_{\text{i}=1}^{\text{n}}|rank(\text{Predicted rating(i)})-\text{rank}(\text{True rating(i)})|}{\text{n}}$$

**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 scores for every given id)**. Also, in your report you should describe how to use your script and its output **for the classification system at step 4 (for the first 50 records of the dataset) for both prediction methods.** Comment on the mean prediction error of both methods and on any other conclusions you have made. The first page of the report should clearly state the names and student ids of the members of the group. Alternatively, you could provide your jupyter notebook.

2) Your code. 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.