Dr. Dimitrios Letsios
Prof. Elizabeth Sklar
Department of Informatics
King's College London

2nd Semester
Friday 14 February 2020

7CCSMDM1 Data Mining

First Coursework

Due: Friday 6 March 2020

The aim of this coursework assignment is to demonstrate understanding of classification and cluster analysis which are among the most important data mining tasks. The coursework (i) is worth 10% of the overall module mark and (ii) will be marked out of 100 marks. The distribution of marks is (i) 50 marks for the first part on **classification** and (ii) 50 marks for the second part on **cluster analysis**. The data sets required for this coursework are provided in the KEATS page of the 7CCSMDM1 module. No need to download them from their original sources. The links to their origins are provided for referencing purposes. The **submission instructions** are presented in the last part of this document.

1 Classification

This part uses the **adult data set** (`https://archive.ics.uci.edu/ml/datasets/Adult`) from the *UCI Machine Learning Repository* to predict whether the income of an individual exceeds 50K per year based on 14 attributes. For this part, the attribute *fnlwgt* should be dropped and the following attributes should be taken into consideration:

Attribute	Description
age	age group
workclass	type of employment
education	level of education reached
education-num	number of education years
marital-status	type of maritals status
occupation	occupation domain
relationship	type of relationship involved
race	social category
sex	male or female
capital-gain	class of capital gains
capital-loss	class of capital losses
hours-per-week	category of working hours
native-country	country of birth

1. *[10 points]* Create a table in the report stating the following information about the adult data set: (i) number of instances, (ii) number of missing values, (iii) fraction of missing values over all attribute values, (iv) number of instances with missing values and (v) fraction of instances with missing values over all instances.

2. *[10 points]* Convert all 13 attributes into nominal using a *Scikit-learn LabelEncoder*. Then, print the set of all possible discrete values for each attribute.

3. *[10 points]* Ignore any instance with missing value(s) and use *Scikit-learn* to build a decision tree for classifying an individual to one of the $<= 50K$ and $> 50K$ categories. Compute the error rate of the resulting tree.

4. *[20 points]* The aim of this question is to investigate two basic approaches for handling missing values. Initially, construct a smaller data set D' from the original data set D, containing (i) all instances with at least one missing value and (ii) an equal number of randomly selected instances without missing values. That is, if the number of instances with missing values is v in D, then D' should contain these v instances and additional v instances without any missing values, which are randomly selected from D. Then, using D', construct two modified data sets D'_1 and D'_2 to handle missing values. In particular,

   - construct D'_1 by creating a new value "missing" for each attribute and using this value for every missing value in D',
   - construct D'_2 by using the most popular value for all missing values of each attribute.

   Train two decision trees with these two data sets and compare their error rates using instances from D for testing. Briefly comment on the obtained results.

2 Clustering

This part uses the wholesale customers data set (`https://archive.ics.uci.edu/ml/datasets/wholesale+customers`) from the *UCI Machine Learning Repository* to identify similar groups of customers based on 8 attributes. For this part of the coursework, the attributes *CHANNEL* and *REGION* should be dropped. Only the following 6 numeric attributes should be considered:

Attribute	**Description**
FRESH	Annual expenses on fresh products.
MILK	Annual expenses on milk products.
GROCERY	Annual expenses on grocery products.
FROZEN	Annual expenses on frozen products.
DETERGENTS	Annual expenses on detergent products.
DELICATESSEN	Annual expenses on delicatessen products.

- *[10 points]* Create a table in the report with the mean $\mu_j = \sum_{i=1}^{m} x_{i,j}$ and range $[x_{j,\min}, x_{j,\max}]$ for each attribute j, where $x_{i,j}$ is the attribute j value of instance i and $x_{\min,j}, x_{\max,j}$ are the minimum and maximum attribute j values among all instances.

- *[20 points]* Run k-means with $k = 3$ and construct a scatterplot for each pair of attributes using Pyplot. Therefore, 15 scatter plots should be constructed in total. Different clusters should appear with different colors in the scatter plot. All scatter plots should be included in the report, using no more than two pages for them.

- *[20 points]* Run k-means for each possible value of k in the set $\{3, 5, 10\}$. Complete the following table with the between cluster distance BC, within cluster distance WC and ratio BC/WC of the set of clusters obtained for each k. Briefly comment on the obtained results.

	$k = 3$	$k = 5$	$k = 10$
BC			
WC			
BC/WC			

Instructions

Every student is expected to:

- Implement this coursework on his/her own. Students may discuss solution strategies between them, but every student must individually write his/her own code and report. Violation of this rule will be considered as an act of misconduct.
- Submit a zip file with a (i) **report** in pdf format answering the questions posed in each part of the coursework, (ii) **Python code** in .py format (**not** .ipynb iPython Notebooks) for generating the answers, and (iii) a **readme** plain text file explaining the answers generated by each source code file.
- **Not** add any source code in the report.
- Submit the zip file on KEATS via the 7CCSMDM1 coursework submission link.