COMP7103 Assignment 1

Due date: Oct 18, 2023, 11:59pm

Question 1 Jaccard Coefficients [15%]

Table 1 displays a collection of transaction data. Represent the data using inverted lists in the format of $item \rightarrow \{a\ list\ of\ transaction\ ids\}$. Subsequently, calculate the similarities between each pair of items using **Jaccard Coefficients**.

TID	Items
1	$Bread$
2	$Beer, Bread, Diaper$
3	$Beer, Coke, Diaper$
4	$Beer, Bread, Diaper$
5	$Coke, Diaper$

Table 1 Transaction items

Question 2 Metric Axioms [20%]

Table 2 presents the distance matrices for 4 distance measures: d_1, d_2, d_3, and d_4, applied to 4 distinct data objects. For each of these distance measures, explain why they cannot be considered metrices based on the **Metric Axioms**. Be sure to clearly state the properties that the distance measures fail to satisfy.

$d_1(A,B)$		B			
		O_1	O_2	O_3	O_4
A	O_1	0	1	2	4
	O_2	1	0	1	2
	O_3	2	1	0	1
	O_4	4	2	1	0

$d_2(A,B)$		B			
		O_1	O_2	O_3	O_4
A	O_1	0	0	1	2
	O_2	0	0	1	2
	O_3	1	1	0	1
	O_4	2	2	1	0

$d_3(A,B)$		B			
		O_1	O_2	O_3	O_4
A	O_1	0	1	2	3
	O_2	1	0	1	2
	O_3	1	1	0	1
	O_4	2	2	2	0

$d_4(A,B)$		B			
		O_1	O_2	O_3	O_4
A	O_1	0	1	3	6
	O_2	1	0	4	2
	O_3	3	4	0	5
	O_4	6	2	5	0

Table 2 Distance matrixes of 4 distance measures d_1, d_2, d_3, and d_4 on 4 different data objects

Question 3 Classification [30%]

Consider a training dataset displayed in **Table 3** for a binary classification problem[1], where attribute **BP** is an ordinal attribute with values in the order of $\{Low, Normal, High\}$, and all other attributes are nominal. Construct a **classification tree** using entropy as the impurity measure, allowing only binary split and applying a pre-pruning criteria of information gain < 0.2. Show all your steps.

Record ID	DT	BP	HP	Class
1	$Timely$	Low	$Inept$	Yes
2	$Timely$	$Normal$	Apt	No
3	$Timely$	$Normal$	Apt	No
4	$Timely$	$Normal$	Apt	No
5	$Timely$	$Normal$	$Inept$	No
6	$Timely$	$Normal$	$Inept$	Yes
7	$Timely$	$Normal$	$Inept$	Yes
8	$Early$	$Normal$	Apt	No
9	$Early$	$High$	Apt	Yes
10	$Early$	$High$	$Inept$	Yes
11	$Late$	Low	Apt	No
12	$Late$	$Normal$	Apt	No

Table 3 Training dataset

Question 4 Cross validation [20%]

Table 4 presents an extract from the Lens dataset[2], in which all attributes are nominal. Answer the following questions.

Record ID	Age group	Prescription	Astigmatic	Class
1	$30s$	M	Yes	$Hard$
2	$30s$	M	No	$Soft$
3	$30s$	H	Yes	$None$
4	$30s$	H	No	$Soft$
5	$40s$	M	Yes	$Hard$
6	$40s$	M	No	$None$
7	$40s$	H	Yes	$None$
8	$40s$	H	No	$Soft$

Table 4 Extract of the Lens dataset

a) Construct a classification tree using the entire dataset, using **classification error** as the impurity measure. When multiple attributes yield the same lowest classification error, prioritize splitting at the attribute **Age group** first, followed by **Prescription**, and finally **Astigmatic**.
   i. Show the resulting decision tree.
   ii. Compute the corresponding training error.

b) To evaluate the decision tree built using the method in part a), a two-fold cross-validation process is performed. The dataset is divided into two sets of records $\{1,2,7,8\}$ and $\{3,4,5,6\}$.
   i. Show all trees constructed in the process.
   ii. Compute the corresponding test error, show your steps.

[1]Extracted from the Caesarian Section Classification Dataset: https://archive.ics.uci.edu/dataset/472/caesarian+section+classification+dataset
[2]Lens Dataset: http://archive.ics.uci.edu/dataset/58/lenses

Question 5 Weka [15%]

Download the **HCV data** dataset: https://archive.ics.uci.edu/dataset/571/hcv+data and read the description. A copy of the dataset and an extract of the description are also available on Moodle. Answer the following questions.

a) Prepare an ARFF file for the dataset pre-processed with the following operations:
   i. Remove record ID.
   ii. Make **Category** the class attribute by moving it to the last column.
   iii. Replace all missing data "**NA**" with "**?**" in the ARFF file (without quote).

   Show all sections in the ARFF file before "**@DATA**". You do not need to submit the ARFF file.

b) Provide a screenshot of the Plot Matrix (i.e., the **Visualize** tab in Weka) of attributes **ALB**, **ALP**, **ALT**, **CHOL**, and **PROT** in Weka.

c) Apply the **PrincipalComponents** filter to the dataset in Weka, with **maximumAttributeNames** set to 3 and **maximumAttributes** set to 6. Then, provide a screenshot of the **Attribute** section of the **Preprocess** tab in Weka, showing the resulted set of attributes.

d) With the dataset resulted in part c), use **CVParameterSelection** in Weka with the **J48** algorithm, selecting the value of **C** from 5 values between 0.1 and 0.5. Choose **10-fold cross-validation** for both the test options and the options in **CVParameterSelection**. Give all classifier output before "**=== Stratified cross-validation ===**".

e) Using the dataset obtained in part a). Use **FilteredClassifer** meta-classifier in Weka to build a classifier with all settings described in part c) and d). Show screenshots of all settings you have made to achieve this (I.e., screenshots of all option windows with title "**weka.gui.GenericObjectEditor**").

f) Are there any differences between the two models built, as well as the corresponding evaluation results in part d) and e)? If so, explain why they are different.

Submission

Please save/scan your work as a **PDF file** and submit it on Moodle before the deadline.