Assignment 1

(COMP3605 – Introduction to Data Analytics)

Date Available: Monday, September 24, 2018
Due Date: 11.50 PM, Sunday, October 14, 2018
Total Mark: 100 marks (weighted 10% out of 100%)

Part I [60 marks]
You are given the training data set *D* shown in the table below for a binary classification problem. The class label attribute has two different values {C0, C1}.

The Class-Labeled Training Data Set *D*

Customer ID	Gender	Car_Type	Shirt_Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

1. [20 marks] Compute the **information gain** (used by ID3) for the Gender, Car_Type, and Shirt_Size attributes.
Hint: You can use the following formulas detailed in the lecture notes of Topic 1: Classification Basics.

$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$, (bits), $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$, $Gain(A) = Info(D) - Info_A(D)$

2. [20 marks] Compute the **gain ratio** (used by C4.5) for the Gender, Car_Type, and Shirt_Size attributes.
Hint: You can use the following formulas detailed in the lecture notes of Topic 1: Classification Basics.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)},\ SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

3. [20 marks] Use a binary split to compute the **Gini index** (used by CART) for the attributes Gender, Car_Type, and Shirt_Size. For the attribute Car_Type, the splitting subsets {Family, Luxury} and {Sports} are used. For the attribute Shirt_Size, the splitting subsets {Small, Medium} and {Large, Extra Large} are used.
Hint: You can use the following formulas detailed in the lecture notes of Topic 1: Classification Basics.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,\ p_i = |C_{i,D}| / |D|,$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2),\ \Delta Gini(A) = Gini(D) - Gini_A(D)$$

Part II [40 marks]

Write a complete Python program named A1P2.py to compute the information gain (used by ID3) for the attributes such as Gender, Car_Type, and Shirt_Size. You can use the training data set *D* and formulas given in Part I. Your program A1P2.py contains the following functions.

1. [20 marks] The Python function `calEntropy(dataSet)` to calculate the information gain (also called entropy) of the input data set `dataSet`. The `dataSet` can be the training data set *D* or the partitions D_1, D_2, ..., D_n of *D*.

2. [10 marks] The Python function `dataPartition(dataSet, attIdx, v)` to split the input data set `dataSet` (e.g., the given training data set *D*) into the subsets D_1, D_2, ..., D_n. The parameter `attIdx` is the index of a splitting attribute. For example, the indices of the splitting attributes Gender, Car_Type, and Shirt_Size are 0, 1, and 2. The parameter `v` is one of the possible values of a selected splitting attribute. For instance, for the selected splitting attribute Car_Type indexed at 1, `v` ∈ {Family, Sports, Luxury}.

3. [10 marks] The Python function `computeInfoGains(dataSet)` to compute the information gains of the input data set `dataSet`. The `dataSet` is the given training data set *D*. For example, the function `computeInfoGains()` calculates the information gains for the Gender, Car_Type, and Shirt_Size attributes.

Assignment Requirements

- For Part I, use Microsoft Word to type your answers and save the file as A1P1.docx.
- For Part II, use the training data set *D* given Part I to test your program A1P2.py. You can write a function `loadDataSet()` to load a training data set that can be stored as a text file or a `.cvs` file.
- When running your program, the program should display necessary computed results so that the correctness of your functions can be verified. *Hint*: Use the calculations produced in Part I to check your functions.

Submission

1. At the top of your files (e.g., A1P1.docx, A1P2.py), you should include the following information.

```
/*
Full Name:
Student ID:
Email:
Course Code:
*/
```

2. Submit your assignment files (e.g., A1P1.docx, A1P2.py) zipped into the file named A1_ID.zip to Ms. Shellyann via the email ssooklal27@gmail.com, where ID is your student ID.

3. Late submission penalty: 10% per day, up to five days.

End of Assignment 1