Assignment 1

(COMP3605 - Introduction to Data Analytics, 2022-2023)

Date Available: Tuesday, September 27, 2022
Due Date: 11.50 PM, Wednesday, October 12, 2022
Total Mark: 100 marks

Answer ALL Questions

INSTRUCTIONS
1. Type or write your answers neatly.
2. Show all working of your answers.
3. Your solutions must be your own. You must not share your working or solutions with your peers.
4. You are not permitted to copy, summarize, or paraphrase the work of others in your solutions.
5. Submit your answers in a single zipped file named A1_ID.zip to the email comp3605@gmail.com, where ID is replaced with your student ID. The file A1_ID.zip contains
- a single PDF file containing all of your typed, handwritten, and screenshots answers.
- a signed and dated UWI Plagiarism Declaration indicating that the work submitted is your own.

Question 1
You are given a training dataset *D* shown in the table below for a binary classification problem, where the attributes MP = Magazine Promotion, WP = Watch Promotion, CCI = Credit Card Insurance, and Gender.

The class-labeled training dataset *D* for credit card customers

Customer ID	MP	WP	CCI	Gender	LIP
1	Yes	No	No	Male	No
2	Yes	Yes	Yes	Female	Yes
3	No	No	No	Male	No
4	Yes	Yes	Yes	Male	Yes
5	Yes	No	No	Female	Yes
6	No	No	No	Female	No
7	Yes	Yes	Yes	Male	Yes
8	No	No	No	Male	No
9	Yes	No	No	Male	No
10	Yes	Yes	No	Female	Yes

The class label attribute LIP (Life Insurance Promotion) of a cardholder has two different values {Yes, No}. Given the test instance $\boldsymbol{X}$ = (MP = Yes, WP = Yes, CCI = *No*, Gender = Female). What the class label will a naive Bayesian classifier predict for the given test instance $\boldsymbol{X}$?

[**Total mark**: **40**]

PLEASE TURN TO THE NEXT PAGE

Question 2

You are given a training data set *D* shown in the table below for a binary classification problem. The class label attribute Play has two different values {Yes, No}.

The Class-Labeled Training Data Set *D*

ID	Outlook	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rainy	Mild	High	Weak	Yes
5	Rainy	Cool	Normal	Weak	Yes
6	Rainy	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rainy	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rainy	Mild	High	Strong	No

a. Use a multiway split to compute the information gain for the attribute **Temperature**.

[10 marks]

b. Compute the gain ratio for the attribute **Humidity** using *Gain*(**Humidity**) = 0.1518.

[10 marks]

c. Use a binay split to compute the Gini index for the attribute **Outlook** and the splitting subset {Sunny, Rainy}.

[10 marks]

[**Total mark**: **30**]

Question 3

Consider a training data set *D* that contains $p = 50$ positive examples and $n = 100$ negative examples. Suppose that we are given the following two candidate rules.

Rule R_1: covers $p_1 = 45$ positive examples and $n_1 = 15$ negative examples,

Rule R_2: covers $p_2 = 4$ positive examples and $n_2 = 1$ negative example.

Which rule is better according to

a. the *accuracy* metric? [5 marks]

b. the *coverage* metric? [5 marks]

c. the *FOIL_Gain* metric? Assume that the initial rule R_0: {} → + covers $p_0 = 50$ positive examples and $n_0 = 100$ negative examples. [10 marks]

d. Laplace metric? [10 marks]

[**Total mark**: **30**]

End of Assignment 1