NAME: Please print_____

CISC-483/683 First Midterm Exam

Friday, Oct. 5, 2018

100 points

IMPORTANT NOTE: For problems that involve computation, you must give the formula you are using, fill in the values for the terms in the formula, and explain how you get these values (necessary in case you are wrong and want partial credit) — however, you need not do the additions, subtractions, multiplications, divisions, or logarithms that are needed to compute the final answer. But you must explain how you would get your final answer.

1. SHORT ANSWER

   (a) (10 points) Suppose that you have a dataset with 1000 instances. The predictor attributes are `Credit-Rating` and `Marital-Status`; `Credit-Rating` has three possible values (High, Medium, Low) and `Marital-Status` has four possible values (Single, Married, Widowed, Divorced). The class value is `Decision` and it has three possible values (YES, NO, MAYBE). The training set is to contain 800 instances and the test set is to contain 200 instances.

   What is meant by stratification of the training and test sets? Explain with respect to the above dataset.

   (b) (15 points) Information Gain, δGini, and Gain Ratio are three different methods for selecting a split attribute and building a decision tree. For each of the following, circle the methods that have the listed feature:

      i. Can produce a tree with more than two branches at each node:
      Information Gain δGini Gain Ratio

      ii. Can split more than once on the same nominal attribute on a single path
      Information Gain δGini Gain Ratio

      iii. Tends to produce bushy trees as opposed to very deep trees
      Information Gain δGini Gain Ratio

      iv. Can split more than once on the same numeric attribute on a single path
      Information Gain δGini Gain Ratio

      v. Has a bias toward selecting as split attributes those that have multiple values
      Information Gain δGini Gain Ratio

(c) **CISC-483 ONLY)** (5 points) You have the following confusion matrix for a decision tree that makes college acceptance decisions.

ACCEPT	REJECT	DEFER	← Classified As
10	3	5	ACCEPT
7	11	20	REJECT
1	7	30	DEFER

What is the success rate of the model?

2. Suppose that Sam and Mary have built a model for granting or rejecting an application for a bank loan. They have tested it on 150 test instances and found that it was wrong on 50 of the test instances. Sam claims that he can tell the bank with 92% confidence that his model will be correct at least 60% of the time on new loan applicants.

   (a) (10 points) Can Sam make this claim? Show your work in deciding on your answer.

   (b) (**CISC-683 ONLY)** (5 points) Mary says that being correct on loan applications only 60% of the time is not very good. She says that Sam should lower his confidence level and then he will be able to claim that his success rate on new data will be higher. Is Mary correct that the success rate that Sam can claim with this lowered confidence level will be higher? Explain why or why not.

3. The following is part of a Netflix movie dataset, where `GENRE` and `YEAR` are predictor attributes and `RECOMMEND?` is the class attribute.

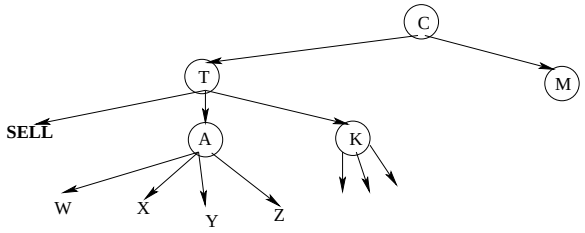
GENRE	YEAR	RECOMMEND?
Action	2016	YES
Action	2010	YES
Comedy	2015	YES
Comedy	2014	NO
Action	2017	YES
Action	2018	NO

You are using top-down entropy-based discretization to discretize the attribute `YEAR`.

(a) (5 points) What are the possible split points that need to be considered and why?

(b) (10 points) Select two possible split points and determine which is the better split point of the two you selected. **Be sure to say which two split points you selected.**

4. Suppose that you have the decision tree shown below, with instances classified as `BUY`, `SELL`, `HOLD`, or `DONATE`.



Of the 70 training instances that reach node A, they are distributed among the leaves **W**, **X**, **Y**, and **Z** as follows:

- 40 instances reach W: 15 are `HOLD`, 20 are `BUY`, and 5 are `SELL`
- 15 instances reach X: 10 are `SELL` and 5 are `BUY`
- 15 instances reach Y: 10 are `DONATE` and 5 are `SELL`
- no instances reach Z:

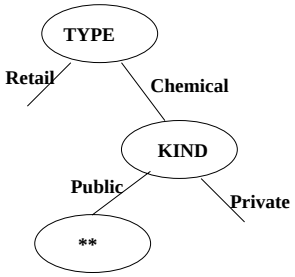
(a) (5 points) Label leaves `W`, `X`, `Y`, and `Z` in the above tree with the Class values (`BUY`, `SELL`, `HOLD`, or `DONATE`) that should be used.

(b) (20 p[oints) You have a pruning dataset, with the following results:

- 40 instances reach W; the actual class for 5 of them is `HOLD`, the actual class for 15 of them is `BUY`, and the actual class for 20 of them is `DONATE`
- 20 instances reach X and the actual class for all of them is `BUY`
- 10 instances reach Y and the actual class for all of them is `DONATE`
- 30 instances reach Z: the actual class for 5 of them is `SELL` and the actual class for 25 of them is `DONATE`

You are using reduced error pruning. Should you prune node **A**? Show your work and explain what you are doing.

5. (20 points) You are building a model for a conglomerate that wants to evaluate companies that they might buy. You have already built the following portion of the decision tree:



Suppose that you have the following dataset, where `SIZE` has the values `Large` and `Small`:

TYPE	SIZE	QUALITY	KIND	BUY?
Chemical	Small	Updated	Private	Yes
Chemical	Large	Updated	Private	No
Chemical	Large	Old	Public	No
Chemical	Large	Old	Public	Yes
Chemical	Large	New	Public	Yes
Chemical	Small	Updated	Public	Yes
Chemical	Small	Updated	Public	Wait
Chemical	Small	Updated	Public	No
Chemical	Small	Old	Public	Yes
Chemical	Small	Old	Public	Yes
Chemical	Small	Old	Public	Yes
Retail	Small	New	Public	Wait
Retail	Small	Old	Public	Wait
Retail	Small	Old	Public	Yes
Retail	Small	Old	Private	Yes

You are using the Gini method to build your decision tree. Consider the node labelled **. You've been told that the value of δGini is $\mathbf{Y}$ if you split on `QUALITY`. Is it better to split on `SIZE` or on `QUALITY`? Show your work and explain how you make the decision.