CSC 240/440 - 2017 Spring: Homework 2

Hand in the hardcopy before the class on Mar. 2

Requirement

Due to the request from some students, the homework is posted online right now, but would be updated probably every week until it is formally released. (A) or (G) indicates questions for all or just graduate students. Undergraduate students are not required to do (G) questions, but they can get bonus points from that. **Please hand in the hardcopy of your homework before the class.**

The homework must be completed individually. However, you are encouraged to discuss the general algorithms and ideas with classmates in order to help you answer the questions. If you work with one or more other people on the general discussion of the assignment questions, please record their names over every question they participated.

However, the following behaviors will receive heavy penalties (lose all points and apply the honest policy explained in syllabus)

- explicitly tell somebody else the answers;
- explicitly copy answers or code fragments from anyone or anywhere;
- allow your answers to be copied;
- get code from Web.

Please also indicate how many late days you want to apply to your submission (Check the late policy in the Syllabus). All late submission without indicating late days or running out the late days cannot be accepted. For medical reasons, if the homework in time cannot be submitted on time, you have to submit the certificate with your homework.

1 (A) 2 points

In a given classification problem there are three classes only: class 1, class 2, and class 3. There are three features: A, B, and C:

$$A \in \{\text{red, blue}\}, \quad B \in \{\text{true, false}\}, \quad C \in \{1, 2, 3\}.$$

Given the following possible logical statement or rule, draw a corresponding decision tree.

IF$((A = \text{red}$ **AND** $C = 1)$ **OR** $(A = \text{blue}$ **AND** $C \in \{1, 2\}$ **AND** $B = \text{false}))$
THEN class label = class 1
ELSE IF $((A = \text{red}$ **AND** $C = 2\})$ **OR** $(A = \text{blue}$ **AND** $C = 3)$
OR $(A = \text{blue}$ **AND** $C = 1$ **AND** $B = \text{true}))$
THEN class label = class 2
OTHERWISE class label = class 3.

2 (A) 3 points

Consider the training examples shown in Table 4.7 of the textbook for a binary classification problem.

1. Compute the Gini index and Entropy for the overall collection of training examples and for the Car Type attribute using multiway split.
2. Which attribute is best according to Gini index?
3. Explain why Customer ID should not be used as the attribute test condition.

3 (A) 4 points

Read the textbook about the generation error rate and finish Question 8 on Textbook page 201.

4 (A) 4 points

Read the textbook about the classifier evaluation and finish Question 9 on Textbook page 202.

5 (A) 3 points

Read the textbook about the classifier evaluation and finish Question 10 on Textbook page 203.

6 (A) 2 points

Let X be a binomial random variable with mean Np and variance $Np(1-p)$. Show that the ratio X/N also has a binomial distribution with mean p and variance $p(1-p)/N$.

7 (G) 2 points

Show that the entropy of a node never increases after splitting it into smaller successor nodes. (Hint: you may want to apply Jensens inequality.)

8 (A) Probability Game (Monty Hall problem): 3 points

The **Monty Hall problem** is a brain teaser, in the form of a probability puzzle, loosely based on the American television game show. Let us Make a Deal and named after its original host, Monty Hall. The problem was originally posed in a letter by Steve Selvin to the American Statistician in 1975 (Selvin 1975a), (Selvin 1975b). It became famous as a question from a reader's letter quoted in Marilyn Savant's "Ask Marilyn" column in Parade magazine in 1990 (Savant 1990a). More background can be found in `http://en.wikipedia.org/wiki/Monty_Hall_problem`. You will have more fun if you do not check this link first.

The following question is a variant to the classic version.

Suppose you are on a game show, and you are given the choice of 100 doors: Behind one door is a car; behind the others, goats. (Our assumption is that car is much more valuable than the goat. The car is what you desire.) You pick a door, say No. 1, and the host, *who knows what is behind the doors*, opens a door which has a goat, say No. 100. Now the host offers you three options

(A) Insist your option door No. 1 and open it;

(B) Randomly select a door from No. 1 to No. 99 and open it;

(C) Randomly select a door from No. 2 to No. 99 and open it.

If the car is behind the door you decide to open, you get the car. The question is which option you will choose? Please provide the probabilities that you can get the car by three options (A), (B), and (C) respectively.

9 (A) 4 points

This is a programming assignment, here we use the Iris dataset (which you used in the last homework assignment). Please write your own program (you are not allowed to call existing functions to automatically generate the decision tree) to implement the following:

1. Load the data from the text file.
2. Implement a Decision Tree as a classifier to detect iris type. Have the depth of your tree and the impurity method ((Gini Index, Entropy, or Misclassification error)) be the input.
3. Report your training accuracy of your classifiers of using three different impurity methods.
4. In your readme you should address what method to compute node impurity you used (Gini Index, Entropy, Misclassification error).

You should upload the code to Blackboard and print out any figures as hard copy (hand in it together with your answers for other questions).