VANCOUVER ISLAND UNIVERSITY
CSCI 479 — MIDTERM EXAMINATION
18 October 2022, 10:00 — 11:20

Duration: 80 Minutes **Instructors:** H. Liu

TO BE ANSWERED IN BOOKLETS

Instructions

- Students must count the number of pages in this examination paper before beginning to write, and report any discrepancy immediately to the invigilator.
- This examination paper consists of 3 pages.
- This is a CLOSED BOOK examination. You are allowed to bring one piece of letter-sized and double-sided notes.
- Calculators are NOT permitted.
- Remember to state any assumptions and show rough work.
- Note carefully the weight of each question, and answer appropriately.
- Attempt all questions. All questions relate to material covered in the lectures, labs and assignments.

1. (10 Marks) Why are the decision tree building algorithms called "greedy", while the similarity based classifier algorithms called "lazy"?

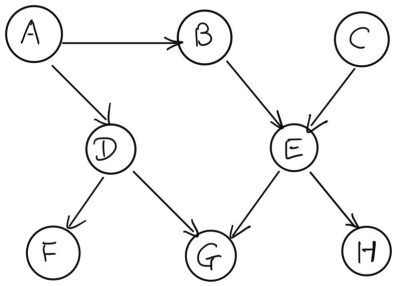
2. (10 Marks) A team of computer scientists were given a set of data collected in an application field to build a predictive model. After careful study of the data, they built a decision tree using the entropy-based information gain to select the "best" descriptive attribute in each step, and applied the post-pruning technique to avoid the overfitting problem.
   When the decision tree model was presented to the customer company, one of the company's member, who is an expert in the application field, exclaimed: "How can the attribute A be selected as the first attribute used in the root node? I've worked in this field for 30 years, and have never noticed any relationship between the attribute A and the target attribute."
   Based on this feedback, should the team of computer scientists go back to re-build the decision tree and take extra precaution to make sure that the attribute A isn't selected in the process?
   What if the expert claimed that attribute A was actually artificially constructed based on the target attribute values, should the team go back to re-build the decision tree and simply exclude the attribute A from being used entirely?

3. (10 Marks) Suppose there are two arrays of the same size (N), A and B, where A[i] stores output of a probability based predictive model predicting how likely (between 0 to 1) that the data item i belongs to class `true` (0 means that its class label absolutely can't be `true`, while 1 means that its class label is absolutely `true`), while B[i] stores the actual class label (string `true` or `false`) of the data item i.
   Write an algorithm that calculates the following percentage values against the total number of data items (N):

   - true positive (both model prediction and class label are `true`)
   - true negative (both model prediction and class label are `false`)
   - false positive (model prediction is `true` but class label is `false`)
   - false negative (model prediction is `false` but class label is `true`)

4. (10 Marks) In your own words, explain the concept of over-fitting, and why it is undesirable in machine learning.
   Explain two methods that can avoid over-fitting from happening in any of the information based learning, similarity based learning or probability based learning applications.

5. (10 Marks) Given the following Bayes Belief Network structure, where the letters inside the nodes represent variable names:



   Suppose that each variable can be either true or false, that is, each variable has exactly two possible values, answer the following questions:

   (a) List the variables that should be included in the Markov blanket if variable E is treated as the target attribute.

   (b) For each variable you listed in the previous step (including the variable E), present its conditional probability table. Use symbols to represent the actual probability values.

   (c) Given a data item with the following variable assignments:
   {A = true, B = false, C = true, D = true,
   F = false, G = true, H = true},
   how can we predict whether variable E is more likely to be true or to be false?

   ====== END OF EXAM QUESTIONS ======