VANCOUVER ISLAND UNIVERSITY
CSCI 479 — FINAL EXAMINATION
13 December 2022, 13:00 — 16:00

Duration: 180 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 6 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)

   (a) What's the disadvantage if the data collection used in a distance based learning project includes too many irrelevant descriptive attributes for the data items?

   (b) Describe one method that can be used in the data pre-processing stage to determine whether a descriptive attribute of the data items is relevant to a classification model building learning project.

2. (10 Marks) The following paragraphs describes the process of building a machine learning classification model to predict a stock's performance in the stock market:

   > For each stock whose price rose more than 10% in one day in the period of year 2000 to year 2010, collect the stock's closing prices in the previous 10 days just before the day its price rose more than 10%.
   >
   > Then, for the given stock, collect its past 10 days' closing prices (up to today's price), using all the collected data and a distance/similarity based classification method (such as K Nearest Neighbours algorithm), try to predict whether the stock will rise dramatically tomorrow.

   What's the flaw(s) in the above proposed process of building and using a similarity/distance based classification model to make a reasonable prediction?
   Is it possible to fix the flaw(s)? Why or why not?

3. (10 Marks) A company collects its existing retail facilities' information, in order to build a classification model to evaluate future locations for a new store facility.
   For each existing store facility and its surrounding area, the following information is collected:

   - geographic information;
   - population and population density;
   - age distribution of the population;
   - medium family income and family income distribution;
   - ethnic composition;
   - local weather description;
   - whether the facility is inside a mall;
   - the number of other similar facilities nearby;
   - average customer evaluation score of the facility;
   - average daily number of customers visiting the facility;
   - the ratio of average yearly sales amount to the labor cost of the facility;
   - the yearly sales growth rate of the past three years;

   Your tasks:

   (a) Which type of classification model is the best to be used to evaluate a future location to decide whether the company should build a new facility in the location? Why?

   (b) Suppose that the company decided to build the classification model you selected in your answer to the previous question, and the data collected in the question description would be used, which attribute(s) can **not** be used as the descriptive attribute(s) when building and training such a model? Why?

   (c) Which attribute(s) should be excluded entirely in this machine learning project? Why?

4. (10 Marks)

   (a) Given a set of data items with multiple descriptive attributes and a target attribute with 3 possible class labels, what's the assumption that the data set must satisfy so that a Naive Bayes Classifier can be built using the given data set?

   (b) If the assumption is not satisfied by the given dataset (i.e., Naive Bayes Classifier can't be built and used), can we still use the probability based learning method?
   If your answer is yes, describe the general steps how we can build a probability based classifier. If your answer is no, explain what's the main obstacles that prevent us from building a probability based classifier.

5. (10 Marks)

   (a) Why is it so important that we should keep evaluating a classification model after it is deployed?

   (b) Measuring the distribution of the model output is usually more accurate than other methods in model evaluation. Why it is more common to use the distribution of model input to evaluate a model after its deployment in reality?

6. (10 Marks)

   (a) Why is the cluster validation usually more difficult than the classification model evaluation?

   (b) What's the difference between external index and internal index measurements in cluster validation?

   (c) How is internal index measurement usually done in cluster validation?

7. (10 Marks) Typically, the association analysis task is divided into two steps: 1) find all frequent itemsets from the data and 2) generate association rules from each frequent itemset.
   Explain **why and how** A-Priori algorithm can be used in both steps of the association analysis.

8. (10 Marks) There are some mature statistical methods to find outliers in a set of data values. The disadvantage of these statistical methods is that they can only be applied on data with single attribute.
   Is it possible to extend the statistical methods to data with multiple attributes by treating each attribute as an independent collection of values?
   If you answer is yes, describe such a method. (Note that you don't need to describe the statistical method itself, just describe how to extend it to multiple attribute data sets.) If you answer is no, explain what is the main obstacles to make such an extension impossible.

9. (10 Marks) The following table shows the frequent single items found in the data collection. They are already sorted according to their frequency in descending order from left to right.

A	B	C	D	E	F

   and the following table shows part of the transaction database, where each transaction is already pre-processed so that it contains only frequent items and its items are already arranged in descending order of their frequencies from left to right.

Transaction
{A, B, C, D, E}
{A, B, C, D, F}
{A, C}
{A, B, C}
{A, B, E}
{B, C, D}
{B, D}
{A, C, D}

   Your task: Draw the resulting FP-tree complete with the header table, the counts (of each node) and the dashed links based on the provided transactions in the above table.

10. (10 Marks) There are 500 patient in the test data to be classified into 5 categories of "definitely positive", "possibly positive", "not sure", "possibly negative" and "definitely negative", while each patient has a known target value of either "positive" or "negative".
A human expert group made diagnosis on these 500 patients, and generated the following confusion matrix:

	positive	negative
definitely positive	49	1
possibly positive	60	40
not sure	21	79
possibly negative	8	42
definitely negative	0	200

After applying a trained classification model on the same set of patients, the model's confusion matrix is shown below:

	positive	negative
definitely positive	58	2
possibly positive	55	25
not sure	21	79
possibly negative	3	57
definitely negative	1	199

Your task:
Describe a method to compare the classification model's performance against the expert group's performance.
(Note that you don't need to provide a comparison result, but need to describe how a comparison can be done to give a result.)

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