Nova Southeastern University
College of Computing and Engineering

Assignment 3

CISC 670 Artificial Intelligence

Fall 2022

Due date: 12/2/2022 11:59PM ET

Total points: 100

Note: Please include your name and the "Certification of Authorship" (located in Canvas) form in EVERY document you submit. Thanks.

Part 1. Text Reading:

Decision Trees (Chap. 19 Sec 19.3), Reasoning with uncertainty (Chap. 12, Sec 13.1, 13.2, Sec 14.1, 14.2, 14.3), Sec 19.8
Unsupervised Learning (Chapter 20 Section 20.3.1, Chapter 21, Section 21.7.1, course slides)
Logic (Chapter 7, 8, 9, course slides)

Course Slides

Part 2. Problems:
(**Note**: Please include any external reference materials other than the textbook. Use the APA format where appropriate.)

Problem 2.1: Decision Tree [30 points]

For this question you need to refer to the decision tree section in the Course Slides (Module 2, file "cisc670f22_3.PDF") posted in Canvas.

One major issue for any decision tree algorithm is how to choose an attribute based on which the data set can be categorized, and a well-balanced tree can be created. The most traditional approach is called the ID3 algorithm proposed by Quinlan in 1986. The detailed ID3 algorithm is shown in the slides. The textbook provides some discussions on the algorithm. For this problem, please follow the ID3 algorithm and manually calculate the values based on a data set similar to (but not the same as) the one in the course slides. This exercise should help you get deep insights on the execution of the ID3 algorithm. Please note that concepts discussed here (for example, entropy, information gain) are very important in information theory and signal processing fields. The new data set is shown as follows. In this example one row was removed from the original set and all other rows remain the same.

Following the conventions used in the slides, please show a manual process and calculate the following values: *Entropy(S)*, $Entropy(S_{weather=sunny})$, $Entropy(S_{weather=windy})$, $Entropy(S_{weather=rainy})$, *Gain (S, weather), Gain (S, parents)* and *Gain (S, money)*. Based on the last three values, which attribute should be chosen to split on?

Please show detailed process how you obtain the solutions.

Weekend	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping

Problem 2.2: Bayes Theorem [20 points]

A quality control manager has used algorithm *C4.5* to come up with rules that classify items based on several input factors. The output has two classes -- Accept and Reject. Test results with the rule set indicate that 4% of the good items are classified as Reject and 3% of the bad items classified as Accept.
Historical data suggests that 2% of the items are bad. Based on this information, what is the conditional probability that:

(i) An item classified as Reject is actually good?
(ii) An item classified as Accept is actually bad?

Please show detailed process how you obtain the solutions.

Problem 2.3: K-means clustering [30 points]

For this question you need to refer to the k-means clustering algorithm in the Course Slides (Module 3, file “cisc670f22_6.PDF”) posted in Canvas.

Point	x	y
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Assume we have the above dataset that shows 7 points in a 2-dimensional space, with x coordinates shown in the *x* column, and y coordinates shown in the *y* column. Assuming the number of clusters is set to 2 (i.e., k=2), and distances among points are measured by Euclidean distance. Based on observation, a good choice of the initial centroids (cluster centers) are point 1 (1.0, 1.0) and point 4 (5.0, 7.0) (these two points are relatively further away from each other). Your job is to run the k-means algorithm and answer the following questions.

(i) After the initial centroids are assigned (shown above), the first step in a clustering process is to check the rest of the points and assign them to one of the two clusters. Here we assume that the centroid remain unchanged during the first round (iteration). Determine cluster memberships (which points belong to Cluster 1 and which belong to Cluster 2) for each of these 7 points after the first iteration. Show intermediate results how you obtain the solution.

(ii) Based on the results obtained from step (i), recalculate centroids (the mean vectors) for the two clusters. Then recalculate the distance from each point to its centroid. Show your results. Are there any points need to change their cluster memberships? If so, what are these points?

(iii) Continue the process shown in step (ii) for another iteration. Do you observe any changes in cluster memberships? If so, what are these points? If not, will there be further changes if we continue the process with more iterations?

Notes:

- It would be easier to solve the problem if you can draw a picture and show changes of centroids on a 2-dimensional system.
- Since this is a tiny dataset, the entire process can be computed by hand.

Problem 2.4: Logic

For this question you need to refer to the discussion on logic in the Course Slides (Module 4, file "cisc670f22_5-Logic.PDF") posted in Canvas.

(i) Look at the following sentences written in first-order logic. Explain these sentences in plain English. Based on the definitions on validity and satisfiability, are these statements valid? If not, are they satisfiable? **[10 points]**

$\forall x\ \exists y\ Loves(x, y) \Leftrightarrow \exists x\ \forall y\ Loves(x, y)$
$\forall x\ Loves(x, movie) \Leftrightarrow \neg\exists x\ \neg\ Loves(x, movie)$
$\exists x\ Loves(x, movie) \Leftrightarrow \neg\forall x\ \neg\ Loves(x, movie)$
$\neg\forall x\ \neg\ Loves(x, movie) \Leftrightarrow \forall x\ Loves(x, movie)$

(ii) Formalize the following statements. Following the procedure shown in the slides, use the deduction proof in which the first three statements are facts, and the last line is the conclusion:

- Every young and healthy person likes football.
- Every active person is healthy.
- Some NSU students are young and active.
- Therefore, some NSU students like football.

Use $Y(x)$ for "x is young," $H(x)$ for "x is healthy," $A(x)$ for "x is active," $B(x)$ for "x likes football," and $N(x)$ for "x is an NSU student".

Note: this simple question can also be proved by forward/backward reasoning. However, you are required to use refutation resolution. **[10 points]**