UMBC CMSC 471 01, Sample Exam

Name:_____

Write all of your answers on this closed-book exam which has eight problems that add up to 210 points. You have the two hours to work on this exam. There are two blank pages at the end that you can use. Good luck.

1) True/False Circle T or F for each statement

T F When overfitting occurs, a machine learning model describes random error or noise instead of the underlying relationships. True

T F The ID3 decision tree induction algorithm uses information gain and is guaranteed to find the optimal decision tree consistent with a given training set. False

T F A SVM with a soft margin will allow some positive training examples to be on the negative side of the margin and also some negative training examples to be on the positive side of the margin. True

T F In a zero-sum, two player game there is necessarily always a winner and a loser. False

T F One drawback of the K-means clustering algorithm is that one needs to specify how many clusters the algorithms should find. True

T F The cells in a NumPy array can only hold numbers or strings. False

T F Overfitting occurs when a machine learning model over-generalizes from the its training data. False

T F The recall metric used in a machine learning binary classification task is defined as the ratio of the number of true positives to the sum of the number of true positives and false negatives. True

T F A learning curve in machine learning is a way to evaluate how a system's accuracy varies with the amount of training data. True

T F Information gain is used to determine the network structure in Recurrent Neural Network. False

T F The precision metric used in machine learning is defined as the ratio of the number of true positives to the sum of the number of true and false positive. True

T F The precision metric used in machine learning is defined as the ratio of the number of true positives to the sum of the number of true and false positive. True

T F A learning curve in machine learning shows the trade-off between a system's precision and recall. False

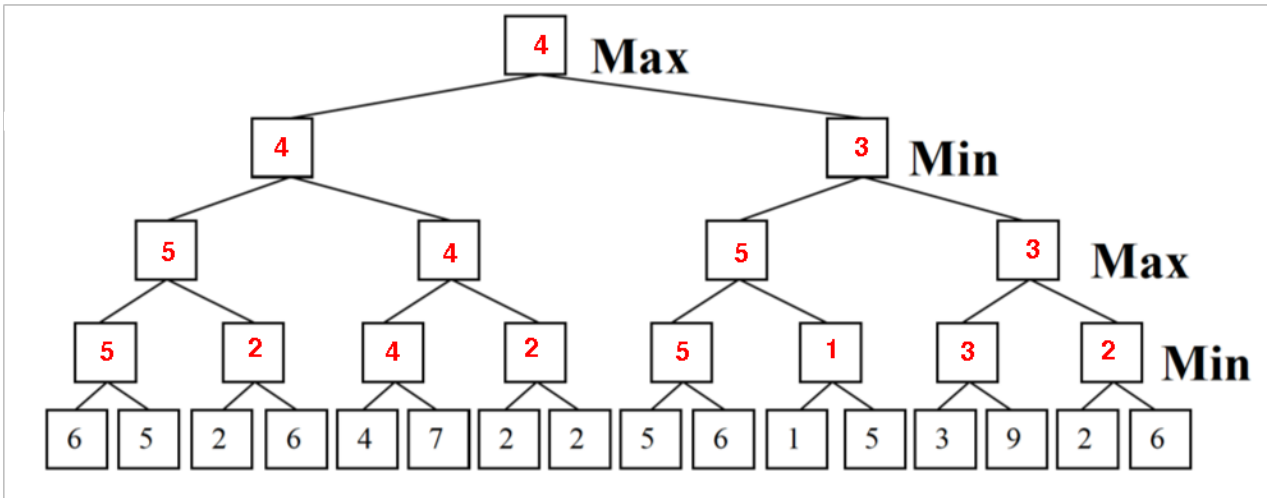
T F A Support Vector Machine classifier can only be used directly to do binary classification. True

T F A single layer perceptron is only capable of learning linearly separable patterns. True

2. Game tree search [25 points]

Consider the game tree below in which the first player is trying to maximize her score and the number at the leaves are the values returned by a static evaluator for the board positions reached.

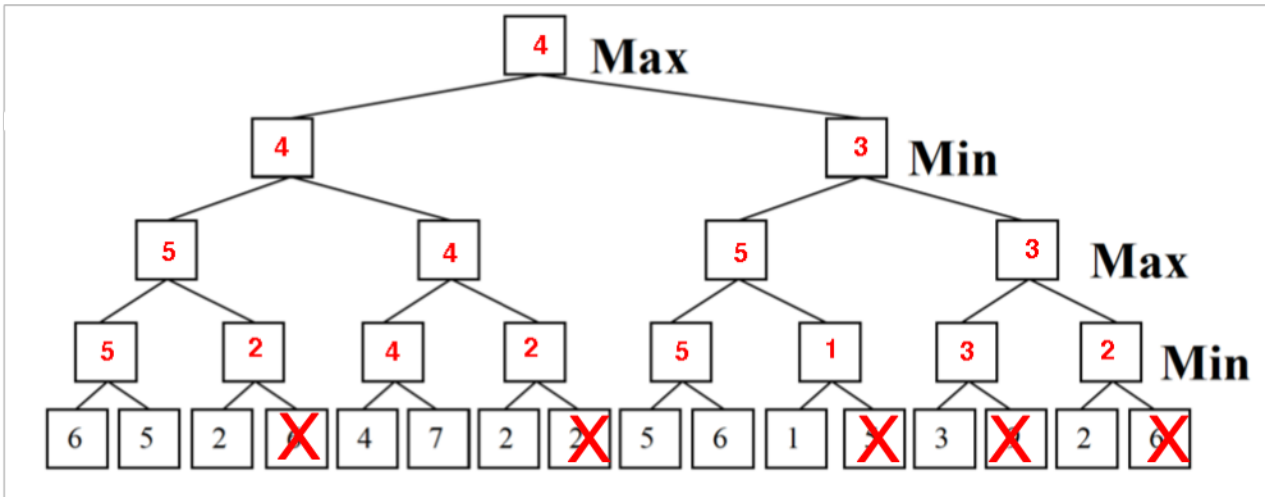
(a) [10] Fill in each box with the value returned by the standard minimax algorithm



(b) [5] Circle the best initial move for the first player: **left** **right**

c) [10] In the copy of this game tree below, fill in each box with the value returned by the standard **alpha-beta algorithm** if the tree is processed from **left to right**. Cross out both leaves and non-leaf nodes that need not be examined or considered.

Comment: Based the grading on the nodes that are crossed out. Don't worry about the numeric values/constraints on the nodes.



3. ML Evaluation

		Correct Value		
		A	**B**	**C**
Guessed Value	**A**	**80**	**9**	**11**
	B	**7**	**86**	**7**
	C	**2**	**8**	**9**

Given the above Cofusion Matrix, answer the following:
Number of True Positives (TP) in Class A
80

Number of False Positives (FP) in Class A
20

Number of True Negatives (TN) in Class A
110

Number of False Negatives (FN) in Class A
9

What is accuracy of Class A?
86.76%

What is the precision of Class A?
0.80

What is the recall of Class A?
0.89

What is the f1 score of Class A?
0.84

4. Decision tree reasoning (35: 10, 15, 10)

Predictive features			*target*
credit	**income**	**debt**	**approve**
Good	Low	Low	Yes
Good	Medium	Low	Yes
OK	High	High	Yes
OK	High	High	Yes
OK	Medium	Low	Yes
OK	High	Low	Yes
Bad	High	Low	No
Bad	Medium	Low	No
OK	Low	High	No
OK	Low	High	No
OK	Low	Low	No
OK	Low	Low	No

The table on the right shows 12 examples of decisions about a credit application (approved or not) based on three variables: credit history, income level and current debt. You plan to use this data to train an ID3-based decision tree to predict if an application will be approved given the values of credit, income and debt.

a) What is the initial entropy of the target variable, i.e., E(approve), which is defined as the sum of the probability of each value times the negative of the log base 2 of that probability. Recall that $\log_2$ of ½ is -1.

1.0

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

b) The ID3 algorithm selects the variable at each level that maximiz- es the information gained. Which attribute would be chosen as the root of the decision tree? (5 pts)

credit

c) Show the entire decision tree that would be constructed by ID3, i.e., by recursively applying in- formation gain to select the roots of sub-trees after the initial decision. If the training data does not determine the predicted value for a case, use a "?" for the decision.(10 pts)

5. Clustering : Follow Question 2 of Assignment 5

6. Neural Network : : Follow Question 1 of Assignment 5

6) Example Short Questions (5 points each):

a) What is Overfitting, and How Can You Avoid It?
b) What is 'training Set' and 'test Set' in a Machine Learning Model? How Much Data Will You Allocate for Your Training, Validation, and Test Sets?
c) How Do You Handle Missing or Corrupted Data in a Dataset?
d) Explain the AUC score and why it can be used to evaluate Machine Learning models?
e) When is it necessary to have higher precision rather than high recall?
f) What is the difference between supervised and unsupervised learning? What is the utility of unsupervised learning?
g) How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?
h) When Will You Use Classification over Regression?
i) What is 'ensemble learning'?
j) Why do we need backpropagation in Neural Network?
k) Why did Multilayer Perceptron evolve over Single Layer Perceptron?