KATHMANDU UNIVERSITY
End Semester Examination [C]
11, April, 2023

Level : B.Sc.
Year : IV
Time : 2 hrs. 30 mins.

Course : COMP 484
Semester : I
F.M. : 40

SECTION "B"
[6Q. × 4 = 24 marks]

Attempt *ANY SIX* questions.

1. Define Concept learning. What are the most specific and most general hypotheses?
2. Explain the working mechanism of the FIND-S algorithm.
3. Explain Entropy and Information Gain and their role in Decision Tree classification.
4. Define a perceptron and its ability to classify linearly separable examples.
5. Elaborate on the two difficulties that may arise when you have limited dataset.
6. Explain the features of Bayesian Learning.
7. Explain the terms – "sample complexity", "computational complexity", "mistake bound".

SECTION "C"
[2Q. × 8 = 16 marks]

Attempt *ANY TWO* questions.

8. A Pathology Lab is performing a Test of disease say "D" with two results "Positive" & "Negative." They guarantee that their test result is 99% accurate: if you have the disease, they will give test positive 99% of the time. If you don't have the disease, they will test negative 99% of the time. If 3% of all the people have this disease and test gives "positive" result, what is the probability that you actually have the disease?
9. Distinguish between "lazy" and "eager learning" with suitable examples. How does the k-nearest neighbor algorithm work?
10. Consider the following set of training examples to train a robot janitor to predict whether or not an office contains a recycling bin.

	Status	**Floor**	**Department**	**Office Size**	**Recycling Bin?**
1	Faculty	Four	CS	Medium	Yes
2	Faculty	Four	EE	Medium	Yes
3	Student	Four	CS	Small	No
4	Faculty	Five	CS	Medium	Yes

Assume that each of the attributes considers just the values that appear on the table.

a. What is the size of the set of instances for this example?
b. What is the size of the hypothesis space?
c. Give a sequence of S and G boundary sets computed by the CANDIDATE-ELIMINATION algorithm if it is given in the sequence of examples above in the order in which they appear on the table.

KATHMANDU UNIVERSITY
End Semester Examination [C]
11, April, 2023

Marks Scored:

Level : B.Sc.
Year : IV

Course : COMP 484
Semester : I

Exam Roll No. :

Time: 30 mins.

F. M. : 10

Registration No.:

Date :

SECTION "A"
[20Q. × 0.5 = 10 marks]

Encircle the most appropriate alternative from each set of choices.

1. Identify the kind of learning algorithm for "facial identities for facial expressions".
   a. Prediction
   b. Recognition patterns
   c. Recognizing anomalies
   d. Generating patterns

2. What is the application of Machine Learning methods to a large database called?
   a. Big data computing
   b. Internet of Things
   c. Data Mining
   d. Artificial Intelligence

3. Among the following identify the one in which dimensionality reduction reduces.
   a. Performance b. Entropy c. Stochastics d. Collinearity

4. Which of the following machine learning algorithm is based upon the idea of bagging?
   a. Decision Tree
   b. Random Forest
   c. Classification
   d. Regression

5. The most significant phase in genetic algorithm is _____.
   a. Mutation
   b. Selection
   c. Fitness Function
   d. Crossover

6. Choose the general limitations of the backpropagation rule among the following:
   a. Slow convergence
   b. Scaling
   c. Local Minima Problem
   d. All of the above

7. A feature F1 can take a certain value: A, B, C, D, E or F, which represents the grades of students from a college. Which of the following statement is true in the following case?
   a. Feature F1 is an example of a nominal variable
   b. Feature F1 is an example of an ordinal variable
   c. Feature F1 is an example of ratio variable
   d. It does not belong to any of the above categories

8. Which of the following is an example of deterministic algorithm?
   a. PCA
   b. K-Means Clustering
   c. KNN
   d. None of the above

9. Which of the following hyper-parameter(s), when increased may cause the random forest to overfit the data?
   I. Number of trees
   II. Depth of tree
   III. Learning rate
   Options:
   a. I b. II c. III d. I, II and III

10. Suppose you want to develop a machine learning algorithm that predicts the number of views on the articles in a blog. Your data analysis is based on features like author name, number of articles written by the same author etc. Which of the following evaluation metrics would you choose in that case?

I. Mean Square Error
II. Accuracy
III. F1-score

Options:

a. I b. II c. III d. I, II and III

11. Below are the 8 actual target variable values in the training file [0,0,0,1,1,1,1,1]. What is the entropy of the target variable?

a. $-(5/8\log(5/8) + 3/8\log(3/8))$
b. $5/8\log(5/8) + 3/8\log(3/8)$
c. $3/8\log(5/8) + 5/8\log(3/8)$
d. $5/8\log(3/8) - 3/8\log(5/8)$

12. Let's say you are working with categorical features, and you have not looked at the distribution of the categorical variable in the test data. You want to apply one-hot encoding on the categorical features. What challenge may you face if you have applied one-hot encoding on a categorical variable of the training dataset?

a. All categories of the categorical variables are not present in the test dataset.
b. The frequency distribution of categories is different in the train compared to the test dataset.
c. Train and Test always have the same distribution.
d. Both "a" and "b"

13. Let us say that you are using activation function X in hidden layers of a neural network. At a particular neuron for any given input, you get the output as "-0.0001". Which of the following activation function could X represent?

a. ReLU
b. tanh
c. SIGMOID
d. None of these

14. Imagine you are solving a classification problem with a highly imbalanced class. The majority class is observed 99% of the time in the training data. Your model has 99% accuracy after taking the predictions on the test set. Which of the following is true in such a case?

I. The accuracy metric is not a good idea for imbalanced class problems.
II. The accuracy metric is a good idea for imbalanced class problems.
III. Precision and recall metrics are good for imbalanced class problems.
IV. Precision and recall metrics aren't good for imbalanced class problems.

Options:

a. I and III b. I and IV c. II and III d. II and IV

15. In ensemble learning (i.e., bagging and boosting), you aggregate the predictions for weak learners so that the ensemble of these models will give a better prediction than the prediction than the prediction of the individual machine learning models.
Which of the following statements is/are true for weak learners used in the ensemble model?

I. They don't usually overfit.
II. They have high bias, so they cannot solve complex learning problems
III. They usually overfit.

Options:

a. I & II
b. I & III
c. II & III
d. None of the above

16. In linear regression, we try to _____ the least square errors of the model to identify the line of best fit.
    a. Change
    b. Maximize
    c. Minimize
    d. None of the above

17. _____ is the result of overestimating or underestimating the importance of a particular parameter or hyperparameter in machine learning.
    a. Machine bias
    b. Segmentation fault
    c. Both of the above
    d. None of the above

18. Which of the following cannot be achieved by using Machine Learning?
    a. Classify respondents into groups based on their response pattern
    b. Proving causal relationships between variables
    c. Forecast the outcome variable into the future
    d. Accurately predict the outcome using supervised learning algorithms

19. Which of the following is TRUE about unsupervised Machine Learning?
    a. A semi-autonomous Machine Learning where researchers control some parts of the modeling process
    b. Unsupervised learning comprises algorithms with no pre-existing outcomes
    c. A fully autonomous Machine Learning with no human interference
    d. Learning algorithms with no control over quality of their predictions

20. What is the Elbow method?
    a. An approach to estimating 'black-box' predictions in supervised learning
    b. A method used to determine the optimal number of clusters in Unsupervised learning, for example, K-Means clustering.
    c. A method of forecasting in Machine Learning
    d. A way of assessing the fit of a Machine Learning Algorithm