6E1552

Roll No. _____ [Total No. of Pages : 7

6E1552

B.Tech. VI Sem. (Main/Back) Examination June- 2022
Information Technology
6IT4-02 Machine Learning

Time : 3 Hours **Maximum Marks : 120**

Min. Passing Marks : 42

Instructions to Candidates:

*Attempt all ten questions from **Part A**, five questions out of **Seven** from **Part B**, and **Four** questions out of **Five** questions from **Part C**.*

Schematic diagrams must be shown wherever necessary. Any data you feel missing suitably be assumed and stated clearly. Units of quantities used/calculated must be stated clearly.

Use of following supporting material is permitted during examination (Mentioned in form No. 205)

Part - A

(Answer should be given up to 25 words only)

All questions are compulsory **(10×2=20)**

1. *"Is Machine Learning just another name for Artificial Intelligence"?* Comment.
2. Write one difference and one similarity between classification and regression.
3. Let C be a candidate item set in C_k generated by the Apriori algorithm. How many length (K-1) subsets do we need to check in the prune step?
4. It is difficult to assess classification accuracy when individual data objects may belong to more than one class at a time. In such cases, comment on what criteria you would use to compare different classifiers modeled after the same data.
5. Differentiate between feature extraction and feature selection.
6. Suppose we clustered a set of N data points using two different clustering algorithms, such as K-means and Gaussian mixtures. In both cases, we obtained five clusters and in both cases, the centers of the clusters are the same. Can three points that are assigned to different clusters in the K - means solution be assigned to the same cluster in the Gaussian mixture solution? Why or why not?

7. In outlier detection by semi - supervised learning, what is the advantage of using objects without labels in the training data set?

8. *Is reinforcement learning an appropriate abstract model for evolution?* Why or why not?

9. Briefly describe an application of Artificial Neural Network that is used for learning to steer an autonomous vehicle.

10. You are provided with data from a music streaming platform. Each of the 450,000 records indicates the songs a user has listened to in the past month. How would you build a music recommendation system? List the steps.

Part - B

(Analytical/Problem solving questions)

Attempt any **five** questions (5×8=40)

1. a. Assume you have roughly classified a set of your previous e - mail messages as junk, unimportant, normal, and important. Describe how a machine learning system may take this as the training set to automatically classify new e-mail messages or unclassified ones.

   b. Give application examples for each of the following cases and also explain them :

      i. An application that uses clustering as a major machine learning function.

      ii. An application that uses clustering as a pre - processing tool for data preparation and other machine learning tasks. (8)

2. a. Why is the naive bayesian classification called "naive"? Explain.

   b. The support Vector Machine (SVM) is a highly accurate classification method. However, SVM classifiers suffer from slow processng when training with a large set of data tuples. Describe how to overcome this difficulty and design a scalable SVM algorithm for efficient SVM classification in large datasets. (8)

3. a. Give an example to show that items in a strong association rule actually may be negatively correlated.

b. Suppose that you are to allocate several Automatic Teller Machines (ATMs) in a given region to satisfy several constraints. Households or workplaces may be clustered so that typically one ATM is assigned per cluster. The cluster may be constrained by two factors :

i. Obstacle objects. (i.e. there are bridges, rivers, and highways that can affect ATM accessibility), and

ii. Additional user - specified constraints such as that each ATM should serve at least 10,000 households.

How can a clustering algorithm such as K - means be modified for quality clustering under both constraints? (8)

4. a. Explain the concept of Principal Component Analysis with a suitable example.

b. Precision and recall are two essential quality measures of a machine learning system. Explain why it is the usual practice to trade one measure for the other. Explain why the F-score is a good measure for this purpose. (8)

5. a. Write the differences among dynamic programming, Monte Carlo, and temporal methods of reinforcement learning. How is policy evaluation performed in Monte Carlo? Explain.

b. How can you perform sentiment analysis of 'multiple - choice questions' using reinforcement learning? Explain. (8)

6. a. What is the Markov decision process? Explain it with a suitable example.

b. Explain various methods of SARSA. (8)

7. a. Explain the role of collaborative - based and content - based recommendation systems in machine learning along with their advantages and disadvantages.

b. How does the collaborative recommendation system differ from a typical classification or predictive modeling system? Explain. (8)

Part - C

(Descriptive/Analytical/Problem Solving/Design questions)

Attempt any **Four** questions **(4×15=60)**

1. The following table shows the midterm and final exam grades obtained for students in a database course. **(4+5+6)**

Midterm Exam (x)	Final Exam (y)
72	84
50	63
82	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52

i. Do x and y seem to have a linear relationship? Explain.

ii. Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

iii. Predict the final exam grade of a student who received a 94 on the midterm exam.

2. The following table consists of training data from an employee database. The data have been generalized. For example, '31.....35' for 'Age' represents the age range of 31 to 35. For a given row entry, 'Count' represents the number of data tuples having the values for Department, Status, Age, and Salary given in that row. **(4+5+6)**

Department	Status	Age	Salary	Count
Sales	Senior	31.....35	46K.....50K	30
Sales	Junior	26.....30	26K.....30K	40
Sales	Junior	31.....35	31K.....35K	40
Systems	Junior	21.....25	46K.....50K	20
Systems	Senior	31.....35	66K.....70K	5
Systems	Junior	26.....30	46K.....50K	3
Systems	Senior	41.....45	66K.....70K	3
Marketing	Senior	36.....40	46K.....50K	10
Marketing	Junior	31.....35	41K.....45K	4
Secretary	Senior	46.....50	36K.....40K	4
Secretary	Junior	26.....30	26K.....30K	6

Let 'Status' be the class label attribute.

i. How would you modify the basic decision tree algorithm to take into consideration the 'Count' of each generalized data tuple? Explain.

ii. Use the algorithm to construct a decision tree from the given data.

iii. Given a data tuple having the values 'Systems', '26.....30', and '46K.....50K' for the attributes Department, Age, and Salary, respectively, what would a Naive Bayesian classification of the status for the tuple be? Explain.

3. A database has five transactions, where minimum support (s) is 60% and minimum confidence (c) is 80%. **(7+8)**

TID	Items Bought
T100	{M,O,N,K,E,Y}
T200	{D,O,N,K,E,Y}
T300	{M,U,C,K,Y}
T400	{M,A,K,E}
T500	{C,O,O,K,I,E}

i. Find all frequent itemsets using Apriori and FP - growth, respectively.

ii. List all the strong association rules (with s and c) matching the following meta - rule, where X_i is a variable representing customers, and $item_i$ denotes variables representing items :

$\forall x \in$ transaction, buys (X, $item_1$) $\wedge$ buys (X, $item_2$) $\rightarrow$ buys (X, $item_3$) [s, c].

4. Suppose that you are a data scientist who has been recruited to help detect fraud during the college admissions process. It requires the process to narrow the focus to fraudulent information submitted in the college application forms, whether it is an inflated GPA, an invented sports achievement, or a fake community service achievement, or other types of forgeries. You will be building a set of fraud detection models. Explain. **(2+2+3+2+3+3)**

i. In this case, which classification methods would you recommend to develop the model : neural networks, random forest, or naive Bayes?

ii. Why did you choose that as your first method?

iii. How would you build this model?

iv. What training data will you need to run that model?

v. Where and how will you obtain the data?

vi. What cross - validation technique would you use on a time series dataset? If needed.

5. a. Suppose that a training set contains only a single sample, repeated 100 times. In 80 of the 100 cases, the single output value is 1; in the other 20, it is 0. What will a backpropagation network predict for this sample, assuming that it has been trained and reaches a global optimum? Explain. **(7)**

b. Suppose you had a neural network with linear activation functions? For each unit, the output is some constant c times the weighted sum of the inputs.**(4+4)**

i. Assume that the network has one hidden layer. For a given assignment to the weights W, write down equations for the value of the units in the output layer as a function of W and the input layer I, without any explicit mention of the output of the hidden layer. Show that there is a network with no hidden units that compute the same function.

ii. Repeat the calculation part (i), this time for a network with any number of hidden layers. What can you conclude about linear activation functions?