Roll No. .....

Total Pages : 07

017606

May 2024

B.Tech. (EEIOT) (Sixth Semester)

Data Mining (PEC-CS-DS-601)

Time : 3 *Hours*] [*Maximum Marks :* 75

Note : It is compulsory to answer all the questions (1.5 mark each) of Part A in short. Answer any *four* questions from Part B in detail. Different sub-parts of a question are to be attempted adjacent to each other.

Part A

1. (a) What is the relation between data warehousing and data mining ? **1.5**

   (b) A data set for analysis includes only one attribute X : **1.5**

   X : X = {7, 12, 5, 8, 5, 9, 13, 12, 19, 7, 12, 12, 13, 3, 4, 5, 13, 8, 7, 6}

   (i) What is the mean of the data set X ?

   (ii) What is the median ?

   (iii) Find the standard deviation for X.

(c) Briefly explain the concept of association rule mining. **1.5**

(d) What is Clustering and how is it used in data mining ? **1.5**

(e) Define the term 'classification' in the context of machine learning and data mining. **1.5**

(f) What is the role of outliers in data mining, and how are they typically handled ? **1.5**

(g) Discuss the need of human intervention in data mining process. **1.5**

(h) What is the significance of cross-validation in evaluating data mining models ? **1.5**

(i) What is the primary goal of data mining ? **1.5**

(j) What steps you would follow to identify a fraud for a credit card company ? **1.5**

Part B

2. (a) Compare and contrast partitioning methods and hierarchical methods in cluster analysis. Provide examples of scenarios where one method might be more suitable than the other. **10**

(b) For training a binary classification model with five independent variables, you choose to use neural networks. You apply one hidden layer with three neurons. What are the number of parameters to be estimated ? (Consider the bias term as a parameter) **5**

3. (a) What is the confidence measure in association rule mining and how is it used to evaluate the strength of discovered rules ? **5**

(b) Explore the methodologies for processing stream data and the design considerations for stream data systems. Discuss the differences between batch processing and stream processing and provide examples of applications where real-time processing of streaming data is essential. Explain the challenges associated with mining data streams and how these challenges are addressed in modern stream data systems ? **10**

4. Define Genetic Algorithm with its steps. Suppose a genetic algorithm uses chromosomes of the form $x = abcdefgh$ with a fixed length of eight genes. Each gene can be any digit between 0 and 9. Let the fitness of individual x be calculated as :

$f(x) = (a + b) - (c + d) + (e + f) - (g + h)$

and let the initial population consist of four individuals with the following chromosomes :

$x_1 = 65413532$

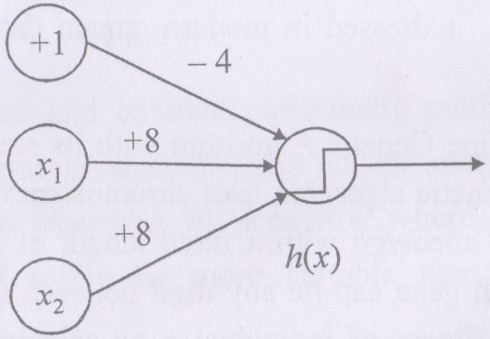
$x_2 = 87126601$

$x_3 = 23921285$

$x_4 = 41852094$

Evaluate the fitness of each individual, showing all your workings and arrange the min order with the fittest first and the least fit last. **15**

5. (a) You are given the following neural networks which take two binary valued inputs x_1, x_2 $\{0, 1\}$ and the activation function is the threshold function ($h(x) = 1$ if $x > 0$; 0 otherwise). Which of the following logical functions does it compute ? **5**

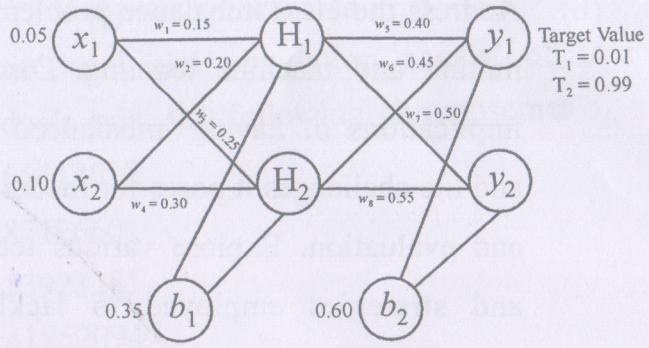


(b) Address the class imbalance problem in data mining and machine learning. Discuss the implications of having imbalanced datasets and the challenges it poses for model training and evaluation. Explore various techniques and strategies employed to tackle class imbalance, including resampling methods, ensemble approaches, and algorithmic adjustments. Illustrate with real-world examples where addressing class imbalance is crucial for model performance. **10**

6. (a) How does frequent pattern mining adapt to the challenges posed by streaming data ? Provide insights into the methodologies and techniques used for discovering frequent patterns in dynamic data streams. **10**

(b) In the context of transactional patterns, explain how market basket analysis can be applied in a retail setting to uncover valuable insights about customer purchasing behavior.

7.



Consider the example given above where :

Input Values

$x_1 = 0.05$

$x_2 = 0.10$

Initial Weights :

$w_1 = 0.1$

$w_2 = 0.205$

$w_3 = 0.25$

$w_4 = 0.30$

$w_5 = 0.40$

$w_6 = 0.45$

$w_7 = 0.50$

$w_8 = 0.55$

Bias Values

$b_1 = 0.35$ $b_2 = 0.60$

Target Values

$T_1 = 0.01$

$T_2 = 0.99$

Calculate h_1, h_2, y_1 and y_2 using forward pass. Use Sigmoid function in hidden layer and output layer.