BE899

Total No. of Printed Pages:4

**B.E (Computer) Semester-VIII (Revised Course 2007-08)**
**EXAMINATION AUGUST 2020**
**Elective-III: Data Mining**

**[Duration : Two Hours]** **[Total Marks :60]**

**Instructions**:-
1. Answer THREE FULL QUESTIONS with ONE QUESTION from ANY THREE MODULES.
2. Make suitable assumptions wherever necessary.

## Module-I

1. a) Discuss whether or not each of the following activities is a data mining task. (4)
   i. Sorting a student database based on student identification numbers.
   ii. Monitoring the heart rate of a patient for abnormalities.

   b) Explain binarization in the context of data preprocessing. (5)

   c) For the following vectors, x and y, calculate the indicated similarity or distance measures. (6)
   x = (1, 1, 1, 1), y = (2, 2, 2, 2)
   i) Cosine ii) Correlation iii) Euclidean

   d) Explain the difference and similarity between classification and regression. (5)

2. a) Explain the effects of curse of dimensionality in a data mining system. Explain Principle Component Analysis (PCA) in detail. (6)

   b) Define (6)
   i) Data Mining
   ii) Knowledge Discovery

   State the difference between the two. Draw a complete labeled diagram of a typical data mining system.

   c) Suppose that the data for analysis include the attribute age. The age values for the data tuples are: (8)
   13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 40, 45, 46, 52, 70.
   i) Use smoothing by bin means to smooth the above data, using the bin depth of 3.
   ii) Use z-score normalization to transform the value 35 for age.
   iii) How can you determine outliers in the data?

iv) Which methods of data reduction is efficient in your view and why?

## Module-II

3. a) Suppose a hospital tested the *age* and *bodyfat* data for 18 randomly selected adults shown in Table 1: (8)

Table 1

| Age | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 | 52 | 54 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| %Fat | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 | 34.6 | 42.5 |

| 54 | 56 | 57 | 58 | 58 | 60 | 61 |
|---|---|---|---|---|---|---|
| 28.8 | 33.4 | 30.2 | 34.1 | 32.9 | 41.2 | 35.7 |

i) Calculate the mean, median and standard deviation of *age* and *%fat*

ii) Draw the boxplots for age and %fat.

b) Write decision tree induction algorithm. Indicate why it is called supervised learning? (7)

c) How does a snowflake scheme differs from a star schema. Name any two disadvantages of snowflake schema. (5)

4 a) Consider the training examples shown in Table 2 for a binary classification problem. (9)

Table 2

| Instance | $a_1$ | $a_2$ | $a_3$ | Target Class |
|---|---|---|---|---|
| 1 | T | T | 1.0 | + |
| 2 | T | T | 6.0 | + |
| 3 | T | F | 5.0 | - |
| 4 | F | F | 4.0 | + |
| 5 | F | T | 7.0 | - |
| 6 | F | T | 3.0 | - |
| 7 | F | F | 8.0 | - |
| 8 | T | F | 7.0 | + |
| 9 | F | T | 5.0 | - |

i) What is the entropy of this collection of training examples with respect to the positive class?

ii) What are the information gains of $a_1$ and $a_2$ relative to these training examples?

iii) For $a_3$, which is a continuous attribute, compute the information gain for every possible split.

b) Draw data warehouse architecture; specifically describe 3-tier data warehouse architecture. (5)

c) What are the ways in which overfitting is handled in decision tree induction? (6)

## Module-III

5 a) State the apriori principle. Write the apriori algorithm. State the advantages and disadvantages of apriori algorithm. (8)

b) What is? (2)
i) Maximal Frequent Itemset.
ii) Closed Frequent Itemset.

c) Consider an example with the set of transactions given in Table 3: (10)

Table 3

| TID | Items Bought |
|---|---|
| 001 | B, M, T, Y |
| 002 | B, M |
| 003 | A, T, S, P |
| 004 | A, B, C, D |
| 005 | A, B |
| 006 | T, Y, E, M |
| 007 | A, B, M |
| 008 | B, C, D, T, P |
| 009 | D, T, S |
| 010 | A, B, M |

Build an FP-tree for the transaction given in Table 3.

6 a) Why is the k-nearest neighbor called as a lazy classifier? Explain with the help of an example by comparing it with the decision tree classifier. (6)

b) You are a data analyst hired by a firm to find the strong association between the items sold by the stationary. The dataset given in Table 4 is provided for the same. Use Apriori algorithm. (8)

Table 4

| Transaction ID | Items Purchased Together |
|---|---|
| 1 | Books, Bag, Pencil, Pen |
| 2 | Books, Pencil, Eraser, Pen |
| 3 | Pen, Pencil, Eraser |
| 4 | Stickers, Beads, Glue |
| 5 | Glue, Scissors, Pen, Pencil |
| 6 | Books, Pen, Pencil |
| 7 | Books, Pencil |

| | |
|---|---|
| 8 | Ruler, Glue, Pencil, Eraser, Pen |
| 9 | Pen, Pencil, Eraser |

Consider a minimum support count of 3 and a minimum Confidence of 60%

c) What do you understand by the following terms? Provide Suitable examples (6)
   i) Rule Based rule ordering
   ii) Class Based rule ordering

## Module-IV

7 a) Explain with example three types of outliers. (6)

b) Consider the task to cluster into 3 clusters, the points given in Table 5: (10)

Table 5

| Point | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| X | 2 | 2 | 8 | 5 | 7 | 6 | 1 | 4 |
| Y | 10 | 5 | 4 | 8 | 5 | 4 | 2 | 9 |

The distance function is Euclidean Distance. The initial cluster centers are A1, A4 and A7 respectively. Use k-Means to perform clustering

c) Differentiate between the following wrt strengths and limitations: (4)
   i) k-means and k-mediods clustering
   ii) supervised and unsupervised learning

8 a) Write the DBSCAN algorithm and use diagrams to define the following: (8)
   i) Directly density reachable points
   ii) Density reachable points
   iii) Core point
   iv) Density connected points

b) Explain the three basic approaches to anomaly detection (6)
c) Write and explain the bisecting k-means algorithm. (6)