Code No.: 9A05706/R09

Set-1

B.Tech. III Year II Semester Regular and Supplementary Examinations

April/May - 2013

DATA WAREHOUSING AND DATA MINING

(Information Technology)

Time: 3 Hours | Max. Marks: 75

*Answer any **FIVE** Questions*

***All** Questions carry **equal** marks*

- - -

1. (a) What is data mining? Explain its role in knowledge discovery process.
   (b) Discuss concept hierarchy generation for categorical data with examples.

2. (a) Give the three-tier data warehouse architecture. Explain it.
   (b) Explain BUC algorithm for the computation of sparse or iceberg queries.

3. What is a frequent item set? How to find frequent item sets for a transactional database? Explain any one approach with illustrations.

4. (a) Discuss rule quality measures.
   (b) What is the significance of learning rate in back propagation algorithm?
   (c) How to measure the accuracy of a classifier? Explain.

5. (a) Discuss the typical requirements of clustering in data mining. **5.3)**(b) Describe deviation-based outlier detection.

6. (a) Explain Viterbi algorithm.
   (b) Discuss mining alternative substructure patterns in graph mining.

7. Describe various types of text databases. What is meant by text mining? Which data mining functionalities are applicable to text databases?

8. (a) How to choose a data mining system? Discuss. (b) Discuss ubiquitous and invisible data mining.

SOLUTIONS TO APRIL/MAY-2013, SET-1, QP

Q1. (a) What is data mining? Explain its role in knowledge discovery process.

Answer : **April/May-13, Set-1, Q1(a)**

Data Mining

For answer refer Unit-I, Q1, Topic: Data Mining (Only First Para).

Role of Data Mining in Knowledge Discovery Process

For answer refer Unit-I, Q3.

(b) Discuss concept hierarchy generation for categorical data with examples.

Answer : **April/May-13, Set-1, Q1(b)**

For answer refer Unit-I, Q32.

Q2. (a) Give the three-tier data warehouse architecture. Explain it.

Answer : **April/May-13, Set-1, Q2(a)**

For answer refer Unit-II, Q17.

(b) Explain BUC algorithm for the computation of sparse or iceberg queries.

Answer : **April/May-13, Set-1, Q2(b)**

For answer refer Unit-II, Q29, Topic: Algorithm.

Q3. What is a frequent item set? How to find frequent item sets for a transactional database? Explain any one approach with illustrations.

Answer : **April/May-13, Set-1, Q3**

Frequent Itemset

A set of items that occurs more frequently together in the data set of a transaction is called frequent itemset.

There are many approaches to find frequent itemsets for a transactional database. Apriori algorithm is one of them which is discussed below with an illustrative example.

For remaining answer refer Unit-III, Q6.

Q4. (a) Discuss rule quality measures.

Answer : **April/May-13, Set-1, Q4(a)**

For answer refer Unit-IV, Q24.

(b) What is the significance of learning rate in back propagation algorithm?

Answer : **April/May-13, Set-1, Q4(b)**

The significance of learning rate (l) in back propagation are,

1. It minimizes the mean squared distance between the class prediction of the network and the actual class label of the samples.
2. It prevents from stucking at a local minimum in decision space and encourages to find the global minimum.

(c) How to measure the accuracy of a classifier? Explain.

Answer : **April/May-13, Set-1, Q4(c)**

For answer refer Unit-IV, Q38.

Q5. (a) Discuss the typical requirements of clustering in data mining.

Answer : **April/May-13, Set-1, Q5(a)**

For answer refer Unit-V, Q1, Topic: Requirements of Clustering.

(b) Describe deviation-based outlier detection.

Answer : **April/May-13, Set-1, Q5(b)**

For answer refer Unit-V, Q44.

Q6. (a) Explain Viterbi algorithm.

Answer : **April/May-13, Set-1, Q6(a)**

Viterbi Algorithm

The viterbi algorithm is used to find the most probable path that leads from one symbol of sequence (x) to the next in the model that generates x.

Consider a sequence x. Here, it is required to find the most probable path in the model generating x. It is likely to happen that many paths can generate x. But the most probable path π^* is the path maximizing the probability of x is the desired one. If L is the sequence of length, then the possible paths will be $|Q|^L$ and Q represents the number of states in the model. Let $V_L(i)$ be the probability defined for the most probable path that accounts for the first i of x and ends in state l. The path π^* can be obtained by computing the probability of the most probable path that accounts for all of the sequence and ends in the end state, $\max_k V_k^{(L)}$ the probability $V_L(i)$ is given by,

$$V_l(i) = e_l(x_i).\max_k(V_L(k)a_{KL})$$

This probability states that,

(i) The most probable path generating $x_1 \ldots\ldots x_i$ and ending in state should emit x_i in state x_l, called emission probability $e_l(x_i)$.

(ii) And should posses the most probable path generating $x_1 \ldots\ldots x_{i-1}$ and ending in state k which is followed by a transition from $k \ldots\ldots l$ state, called transition probability, a_{kl}.

Hence, $V_k(l)$ can be computed for any state (k) recursively inorder to find the probability of the most probable path.

Algorithm

Viterbi decoding algorithm for finding most probable path emitting the sequence of symbol x.

Input

1. Hidden Markov model (defined by a set of states Q and probabilities of transition and emission).
2. A sequence of symbols, x.

Output

The most probable path, π^*

Method

Step 1

Initialize ($i = 0$): $V_0(0) = 1$, $V_k(0) = 0$ for $k > 0$

Step 2

Recursion ($i = 1 \ldots\ldots L$) : $V_l(i) = e_l(x_i) \max_k(V_k(i-1)a_{kL})\ ptr_i(l) = \text{argmax}_k(V_k(i=1)a_{kL})$

Step 3

Terminate: $p(x, \pi^*) = \max_k(V_k(L)a_{k0})$

$$\pi_L^* = \text{argmax}_k(V_k(L)a_{k0})$$

In step 1, initialization is performed in this step, each path begins at state (0) having probability 1. Therefore, for i = 0, the value $V_0(0) = 1$ and the starting probability at any other state will be 0.

In step 2, the recurrence formula is applied for $i = 1 \ldots\ldots L$. At every step of iteration it is assumed that the most likely path for $x_1 \ldots\ldots x_{i-1}$ ending in state k is known $\forall k \in Q$.

From this, the most likely path can be found upto the i^{th} state by maximizing $V_k(i-1)a_{kL}$ over all predecessors $k \in Q$ of l where, $V_i(i)$ can be obtained by multiplying with $\max_k(V_k(i-1)a_{kl})$. Since x_i should be produced from l. The value $V_k(i)$ is stored in the dynamic programming matrix such as $Q \times L$. And the pointers are kept in this matrix inorder to get the path. In step 3, the value $\max_k V_k(L)$ is obtained where the end state of 0 is entered, that leads to the transition probability a_{k0}. Finally, in this step the algorithm gets terminated.

(b) Discuss mining alternative substructure patterns in graph mining.

Answer : April/May-13, Set-1, Q6(b)

For answer refer Unit-VI, Q30.

Q7. Describe various types of text databases. What is meant by text mining? Which data mining functionalities are applicable to text databases?

Answer : April/May-13, Set-1, Q7

Various Types of Text Databases

The main purpose of these databases is to describe objects in word format which include sentences or paragraphs rather than comprehensible keywords. Text databases can be,

1. Highly structured
2. Semi-structured
3. Highly unstructured.

- ❖ Highly structured database are usually implemented with the help of relational database system and include employee's information database.
- ❖ Semi-structured database include HTML web pages and email messages.
- ❖ Highly unstructured lined database include www pages.

Data mining techniques can be applied to text database to reveal hidden compressed description of both text documents and clustering characteristics of text objects. In order to extract the description, conventional data mining techniques are combined with information extraction methods.

Text Mining

For answer refer Unit-VII, Q22.

The data mining functionalities which are applicable to text databases include,

1. Text categorization
2. Text clustering
3. Sentiment analysis
4. Document summarization.

Q8. (a) How to choose a data mining system? Discuss.

Answer : **April/May-13, Set-1, Q8(a)**

For answer refer Unit-VIII, Q5.

(b) Discuss ubiquitous and invisible data mining.

Answer : **April/May-13, Set-1, Q8(b)**

For answer refer Unit-VIII, Q11.