

JK LAKSHMIPAT UNIVERSITY, JAIPUR
INSTITUTE OF ENGINEERING AND TECHNOLOGY
End Term Examination (Regular), November 2017
B. Tech in Computer Science Engineering, Semester VII

Subject Code: CSE 701

Roll No.....

Subject Name: Data Mining and Warehousing

Time: 3 hours

Max. Marks: 100

Instructions to students:

1. Do not write anything other than your roll number on question paper.
 2. Assume suitable data wherever required and mention it clearly.
 3. Writing appropriate units, nomenclature, and drawing neat sketches/schematics/ flow chart wherever, required are an integral part of the answer.
-

Q.1A Objective type multiple choice questions (10 questions @ 1 mark)

10

i. Bias is

- A. A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory
- B. Any mechanism employed by a learning system to constrain the search space of a hypothesis
- C. An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.
- D. None of these

ii. Classification is

- A. A subdivision of a set of examples into a number of classes
- B. A measure of the accuracy, of the classification of a concept that is given by a certain theory
- C. The task of assigning a classification to a set of examples
- D. None of these

iii. Classification accuracy is

- A. A subdivision of a set of examples into a number of classes
- B. Measure of the accuracy, of the classification of a concept that is given by a certain theory
- C. The task of assigning a classification to a set of examples

D. None of these

iv. Cluster is

- A. Group of similar objects that differ significantly from other objects
- B. Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm
- C. Symbolic representation of facts or ideas from which information can potentially be extracted
- D. None of these

v.

A definition of a concept is----if it recognizes all the instances of that concept

- A. Complete
- B. Consistent
- C. Constant
- D. None of these

vi. Data selection is

- A. The actual discovery phase of a knowledge discovery process
- B. The stage of selecting the right data for a KDD process
- C. A subject-oriented integrated time variant non-volatile collection of data in support of management
- D. None of these

vii. Discovery is

- A. It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).
- B. The process of executing implicit previously unknown and potentially useful information from data
- C. An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.
- D. None of these

viii. Learning algorithm refers to

- A. An algorithm that can learn
- B. A sub-discipline of computer science that deals with the design and implementation of learning algorithms
- C. A machine-learning approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.
- D. None of these

ix. Multi-dimensional knowledge is

- A. A class of learning algorithms that try to derive a Prolog program from examples
- B. A table with n independent attributes can be seen as an n -dimensional space
- C. A prediction made using an extremely simple method, such as always predicting the same output.
- D. None of these

x. Prediction is

- A. The result of the application of a theory or a rule in a specific case
- B. One of several possible entries within a database table that is chosen by the designer as the primary means of accessing the data in the table.
- C. Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.
- D. None of these

Q.2 Short answer type questions (4 questions @ 2.5 marks)

10

- i. Differentiate pattern matching and data mining.
- ii. Explain role of big mining in big data analysis.
- iii. Explain concept hierarchy.
- iv. Explain about Measuring Dispersion of Data.

Q.3A Sally measures the pressure of all of tires coming into her garage for an oil change and records the values. Unknown to her, her tire gauge is miscalibrated and adds 3 psi to each reading. According to the definition of noise used by our textbook, is this error introduced by the tire gauge considered noise? Answer "yes" or "no" and justify your answer.

05

- Q.3B How can you convert a decision tree into a rule set? Explain the process. 05
- Q.4A For a two-class classification problem, with a Positive class P and a negative class N, what is : True Positive, False Positive, TN, and FN. 05
- Q.4B An Internet marketer is interesting in segmenting Internet with a clustering tool using the input attributes - top ten search key words used, top 10 URLs, recent 10 online purchases (vendor, product, qty, amt), Internet usage level, heaviest access hour, and heaviest access day of a week. Can we find users with different income level? Why or why not. 05
- Q.5A If {1, 2, 3} and {2, 3, 4} are the only large 3 itemsets, identify for each one of the following sets if it is or is not a large itemset ,or you cannot be certain if it is a large itemset or not. 05
- | | | |
|------|--------------|-----|
| i. | {1} | Yes |
| ii. | {1, 2} | Yes |
| iii. | {1, 4} | No |
| iv. | {1, 2, 3, 4} | No |
| v. | {1, 3, 4} | No |
- Q.5B Assume that the confidence of the decision rule (question 5A), 1-> 2, is 100%. Is the confidence of the decision rule, 2-> 1, also 100%? Give an example of data to justify your answer. 05
- Q.6A Compare the pros and cons of decision tree and neural network classification methods. 05
- Q.6B Explain ways to deal with noisy data. 05
- OR
- Q.6B Explain Data Cube Aggregation, and Dimensionality reduction. 05
- Q.7A The task attributes of the four data mining tasks discussed in class are briefly described below: 05
- Association rule and sequential pattern mining - Customer ID, Transaction ID and Item. Classification/prediction - input and the class label attributes
- Clustering mining - input attributes
- The following are the data fields in the data mining server log:
User ID, Session ID, Dataset ID, MiningTask ID, Parameter Value, Accuracy
- Which task will you perform to identify the data mining tasks that tend to be performed in the same session? Describe the attributes you choose and how they are mapped to the data mining task attributes listed above.
- Q.7B Which task will you perform to determine if the Parameter Value level (low, medium or high) and the level of Parameter Value adjustment (small, moderate or large) tend to have a positive or negative impact on Accuracy. Describe the attributes you choose and how they are mapped to the data mining task attributes listed above. 05
- OR

- Q.7B Explain different Partitioning methods in clustering. 05
- Q.8A Explain the process of knowledge discovery from Data in detail. 10
- Q.9A Explain OLAP Servers, ROLAP, MOLAP, HOLAP. 10
- Q.9B Explain Multi-Dimensional Data Model, Data Cubes, Stars, Snow Flakes, Fact Constellations, Concept hierarchy? 10
- OR
- Q.9B What are the principle problems in data mining (classification, association rules, numerical prediction, clustering)? Describe an example for each task!. 10