Mark Scored:

KATHMANDU UNIVERSITY
End Semester Examination
February/March, 2019

Level : B. E./ B. Sc.
Year : IV
Course : COMP 482
Semester: I

Exam Roll No. :
Time: 30 mins.
F. M. : 10

Registration No.:
FEB 2 4 2019
Date :

SECTION "A"
[20 Q. × 0.5= 10 marks]

Tick [√] the most appropriate answer.

1. Data mining is best described as the process of
   [ ] identifying patterns in data
   [ ] deducing relationships in data
   [ ] representing data
   [ ] simulating trends in data

2. Data used to build a data mining model is _____
   [ ] validation data [ ] training data [ ] test data [ ] hidden data

3. Supervised learning and unsupervised learning both require at least one _____
   [ ] hidden attribute [ ] output attribute [ ] input attribute [ ] categorical attribute

4. Supervisedlearning differs from unsupervised learning in that supervised learning requires _____
   [ ] at least one input attribute
   [ ] input attributes to be categorical
   [ ] at least one output attribute
   [ ] ouput attriubutes to be categorical

5. Which of the following is not a characteristic of a data warehouse?
   [ ] contains historical data
   [ ] designed for decision support
   [ ] stores data in normalized tables
   [ ] promotes data redundancy

6. A nearest neighbor approach is best used _____
   [ ] with large-sized datasets
   [ ] when irrelevant attributes have been removed from the data
   [ ] when a generalized model of the data is desireable
   [ ] when an explanation of what has been found is of primary importance

7. Which statement is true about prediction problems?
   [ ] The output attribute must be categorical
   [ ] The output attribute must be numeric
   [ ] The resultant model is designed to determine future outcomes
   [ ] The resultant model is designed to classify current behavior

8. Which of the following statement is true about outliers?
   [ ] Outliers should be identified and removed from a dataset
   [ ] Outliers should be part of the training dataset but should not be present in the test data
   [ ] Outliers should be part of the test dataset but should not be present in the training data
   [ ] The nature of the problem determines how outliers are used

9. Which of the following situations hold true for unstable data mining algorithm?
[ ] test set accuracy depends on the ordering of test set instances
[ ] the algorithm builds models unable to classify outliers
[ ] the algorithm is highly sensitive to small changes in the training data
[ ] test set accuracy depends on the choice of input attributes

10. Given a rule of the form IF X THEN Y, rule *confidence* is defined as the conditional probability that _____
[ ] Y is true when X is known to be true
[ ] X is true when Y is known to be true
[ ] Y is false when X is known to be false
[ ] X is false when Y is known to be false

11. Association rule *support* is defined as _____
[ ] the percentage of instances that contain the antecendent conditional items listed in the association rule
[ ] the percentage of instances that contain the consequent conditions listed in the association rule
[ ] the percentage of instances that contain all items listed in the association rule
[ ] the percentage of instances in the database that contain at least one of the antecendent conditional items listed in the association rule

12. Which statement is true about the *K-Means* algorithm?
[ ] All attribute values must be categorical
[ ] The output attribute must be cateogrical
[ ] Attribute values may be either categorical or numeric
[ ] All attributes must be numeric

13. The choice of a data mining tool is made at _____step of the KDD process.
[ ] goal identification
[ ] data preprocessing
[ ] creation target dataset
[ ] data mining

14. Attibutes may be eliminated from the target dataset during _____ step of the KDD process.
[ ] creation target dataset
[ ] data transformation
[ ] data preprocessing
[ ] data mining

15. Which of the following steps of the KDD process model deals with noisy data?
[ ] Creation target dataset
[ ] Data transformation
[ ] Data preprocessing
[ ] Data mining

16. A common method used by some data mining techniques to deal with missing data items during the learning process is _____
[ ] replace missing real-valued data items with class means
[ ] discard records with missing data
[ ] replace missing attribute values with the values found within other similar instances
[ ] ignore missing attribute values

17. Which of the following is not a characteristic of a data warehouse?
[ ] contains nonvolatile data
[ ] is subject oriented
[ ] supports data processing, collection and management
[ ] stores data to be reported on, analyzed and tested

18. A variation of the star schema that allows more than one central fact table is _____
[ ] snowflake schema
[ ] linked strar schema
[ ] distributed star schema
[ ] constellation schema

19. Which clustering algorithm initially assumes that each data instance represents a single cluster.
[ ] agglomerative clustering
[ ] conceptual clustering
[ ] K-Means clustering
[ ] expectation maximization

20. With Bayes theorem, the probability of hypothesis *H*— specified by *P(H)*— is referred to as _____
[ ] a priori probability
[ ] a conditional probability
[ ] a posterior probability
[ ] a bidirectional probability

KATHMANDU UNIVERSITY
End Semester Examination
February/March, 2019

FEB 2 4 2019

Level : B. E./ B. Sc.
Year : IV
Time : 2 hrs. 30 mins.

Course : COMP 482
Semester: I
F. M. : 40

SECTION "B"
[6 Q. × 4 = 24 marks]

Attempt *ANY SIX* questions.

1. What is a *data warehouse*? Explain different characteristics of a *data warehouse*.

2. Compare *OLTP* and *OLAP* systems.

3. Differentiate between *star* and *snowflake* schema.

4. What is a *classification* problem? How does it differ from the clustering problem?

5. How do you clean the data? Explain different ways you follow to clean data *for missing values* and *for noisy data*.

6. How are *association rules* mined from large databases? List two interesting measures for *association rules*.

7. Define *outliers*. Explain various *outlier detection* approaches.

SECTION "C"
[2 Q. × 8 = 16 marks]

Attempt *ANY TWO* questions.

8. Provide justifications whether each of the following activities is a *data mining* task or not.
   (a) Dividing the customers of a company according to their profitability.
   (b) Monitoring the heart rate of a patient for abnormalities.
   (c) Predicting the future stock price of a company using historical records.
   (d) Extracting the frequencies of a sound wave.

9. For each of the following problems align the most suitable approach among *supervised learning*, *unsupervised learning*, and *data query* with proper justifications.
   i. What is the average weekly salary of all female employees under forty years of age?
   ii. Do meaningful attribute relationships exist in a database containing information about credit card customers?
   iii. Do single men play more golf than married men?
   iv. Determine whether a credit card transaction is valid or fraudulent.

10. Define the term *Interestingness Measures* in data mining. State and explain different criteria to determine whether a *pattern* is interesting. Also explain the importance of *visualization* techniques during data analysis.