Roll No: --------------------------

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2017

Program: B.Tech. (CSE) OGI **Semester – V**
Subject (Course): Data Warehousing and Data Mining in Energy Sector **Max. Marks : 100**
Course Code: CSEG335 **Duration : 3 Hrs**
No. of page/s:2

Section-A **[5 x 4 = 20 Marks]**

Q1. How does classification work in data mining? How is (numeric) prediction different from classification?

Q2. Discuss different OLAP operations.

Q3. What is difference between data warehouse and data mart? Discuss.

Q4. What are the essential steps of data mining?

Q5. Discuss the steps of decision tree classification.

Section-B **[10 x 4 = 40 Marks]**

Q6. List and describe the five primitives for specifying a data-mining task.

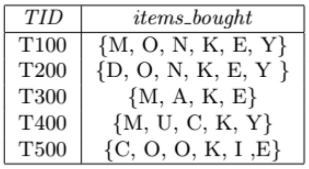
Q7. Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: no coupling, loose coupling, semi tight coupling, and tight coupling. State which approach you think is the most popular, and why.

Q8. Outliers are often discarded as noise. However, one person’s garbage could be another’s treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Taking fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.

Q9. Discuss issues to consider during data integration.

Section-C **[20 x 2 = 40 Marks]**

Q10. Using the transactions mentioned in table below:



TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

(a) Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the e–ciency of the two mining processes.

(b) List all of the *strong* association rules (with support *s* and confldence *c*) matching the following metarule, where *X* is a variable representing customers, and $item_i$ denotes variables representing items (e.g., *“A", “B"*, etc.):

$$\forall x \in transaction,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3) \quad [s, c]$$

Q11. Suppose that you are to allocate a number of automatic teller machines (ATMs) in a given region so as to satisfy a number of constraints. Households or places of work may be clustered so that typically one ATM is assigned per cluster. The clustering, however, may be constrained by two factors: (1) obstacle objects(i.e., there are bridges, rivers, and highways that can afiect ATM accessibility), and (2) additional user-specifled constraints, such as each ATM should serve at least 10,000 households. How can a clustering algorithm such as k-means be modifled for quality clustering under both constraints?

Roll No: -------------------------

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2017

Program: B.Tech. (CSE) OGI **Semester – V**
Subject (Course): Data Warehousing and Data Mining in Energy Sector **Max. Marks : 100**
Course Code: CSEG335 **Duration : 3 Hrs**
No. of page/s:1

Section-A **[5 x 4 = 20 Marks]**

Q1. What is difference between data, information and knowledge?
Q2. What do you understand by outlier analysis?
Q3. Define the terms: Discrimination and characterization.
Q4. Write the steps involved in data mining when viewed as a process of knowledge discovery.
Q5. Explain in brief how the evolution of database technology to data mining? What is the relevance of metadata in this evolution?

Section-B **[10 x 4 = 40 Marks]**

Q6. What is the difierence between discrimination and classiflcation? Between characterization and clustering?
Q7. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):
(a) Compute the Euclidean distance between the two objects.
(b) Compute the Manhattan distance between the two objects.
Q8. What do you understand by back propagation? Discuss
Q9. What are the features of a data warehouse? Explain each feature by taking an example.

Section-C **[20 x 2 = 40 Marks]**

Q10. Compare the advantages and disadvantages of eager classiflcation (e.g., decision tree, Bayesian, neural network) versus lazy classiflcation (e.g., k-nearest neighbor, case-based reasoning).
Q11. Data cubes and multidimensional databases contain categorical, ordinal, and numerical data in hierarchical or aggregate forms. Based on what you have learned about the clustering methods, design a clustering method that finds clusters in large data cubes effectively and efficiently.