|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CMR INSTITUTE OF TECHNOLOGY | | | | | | | |  | | | | | | | | |
| Internal Test I – March 2018 | | | | | | | | | | | | | | | | |
| Sub: | Data Warehousing & Data Mining | | | | | | Sub Code: | | 16MCA442 | Branch: | | MCA | | | |
| Date: | 13/03/18 | | Duration: | 90 min’s | Max Marks: | 50 | Sem/Sec: | | B | | | | OBE | | |
| Answer all the five questions having internal choice | | | | | | | | | | |  | | |  |  | |
| PartI-1(a) | | **Explain Knowledge discovery process in data mining.**  The term *Knowledge Discovery in Databases*, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.  The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, sub sampling,  and transformations of that database. | | | | | | | | | [5] | | |  |  | |
| (b) | | **Explain any two data preprocessing steps.**  Ans: **Aggregation**: Combining two or more attributes (or objects) into a single attribute (or object)  **Purpose**  **–** Data reduction  Reduce the number of attributes or objects  – Change of scale  Cities aggregated into regions, states, countries, etc  – More “stable” data  Aggregated data tends to have less variability  **Sampling:** Sampling is the main technique employed for data selection.  – It is often used for both the preliminary investigation of the data and the final data analysis.  Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.  Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.  **Types of Sampling**  Simple Random Sampling  – There is an equal probability of selecting any particular item  – Two types  Sampling without replacement  – As each item is selected, it is removed from the population  Sampling with replacement  – Objects are not removed from the population as they are selected for the sample.  In sampling with replacement, the same object can be picked up more than once  Stratified sampling  – Split the data into several partitions; then draw random samples from each partition | | | | | | | | | [5] | | |  |  | |
| 2 (a) | | **Explain different type of attributes and data sets.**  There are different types of attributes  – Nominal  Examples: ID numbers, eye color, zip codes  – Ordinal  Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall,  medium, short}  – Interval  Examples: calendar dates, temperatures in Celsius or Fahrenheit.  – Ratio  Examples: temperature in Kelvin, length, time, counts | | | | | | | | | [10] | | |  |  | |
| PartII3(a) | | Explain the type of operation that can be performed on data cube. | | | | | | | | | [10] | | |  |  | |
| 4 (a) | | Explain the difference between ROLAP and MOLAP. | | | | | | | | | [5] | | |  |  | |
| (b) | | Write the various schemas for multidimensional databases | | | | | | | | | [5] | | |  |  | |
| PartIII5  (a) | | List the major steps involved in the ETL process. | | | | | | | | | [5] | | |  |  | |
| (b) | | Explain FASMI characteristics of OLAP. | | | | | | | | | [5] | | |  |  | |
| 6(a) | | What is a data ware house? How we can implement a data ware house. | | | | | | | | | [4+6] | | |  |  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Part IV-7(a) | Explain the three tier architecture of Dataware house | [10] |  |  |
| 8(a) | Write the difference between OLAP and OLTP. | [10] |  |  |
| Part V-9(a) | What is data mining. Explain the challenges that motivated the development of data mining. | [6] |  |  |
| (b) | Explain Similarity and Dissimilarity. | [4] |  |  |
|  |  |  |  |  |
| 10(a) | Discuss the data mining task in detail. | [10] |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Part IV-7(a) | Explain the three tier architecture of Dataware house | [10] | CO3 | L2 |
| 8(a) | Write the difference between OLAP and OLTP. | [10] | CO1 | L2 |
| Part V-9(a) | What is data mining. Explain the challenges that motivated the development of data mining. | [6] | CO3 | L2 |
| (b) | Explain Similarity and Dissimilarity. | [4] | CO1 | L1 |
|  |  |  |  |  |
| 10(a) | Discuss the data mining task in detail. | [10] | CO3 | L2 |