GURU GHASIDAS VISHWAVIDYALAY, CENTRAL UNIVERSITY, BILASPUR

END (ODD) SEMESTER EXAMINATION, 2014

Class : MCA-V SEM. Mark : 60

Subject : Data Mining and Data Warehousing(MCA-503) Time : 3 Hours

Q. 1 is compulsory. Answer any four from rest.

<table>
<tr><td></td><td></td><td></td></tr>
<tr><td>Q.1</td><td>(a) What is data mining? Is it a simple transformation of technology developed from databases, statistics and machine learning?<br>(b) Name the databases on which data mining techniques can be applied.<br>(c) What is a time series database. Give an example.<br>(d) What is the meaning of Market-Basket Analysis? Give an example.<br>(e)Differentiate between classification and prediction<br>(f) Differentiate between data ware house and data mart.<br>(g)What is DMQL?How it is different from SQL.<br>(h)Define no coupling, loose coupling, semitight coupling and tight coupling<br>(i) How dimensionality reduction is different from data reduction?<br>(j)What is holistic measure? Give examples.</td><td>2 x 10</td></tr>
<tr><td>Q. 2</td><td>(a)Explain how the evolution of database technology led to Data Mining.<br>(b)Describe the steps involved in data mining when viewed as a process of knowledge discovery.</td><td>5+5</td></tr>
<tr><td>Q. 3</td><td>Briefly compare the following concepts<br>(a)Snowflake schema, fact constellation, starnet query model<br>(b)Data cleaning, data transformation, refresh<br>(c)Enterprise warehouse, data mart</td><td>3+3+4</td></tr>
<tr><td>Q. 4</td><td>Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows<br><br>Age Frequency<br>1-5 200<br>5-15 450<br>15-20 300<br>20-50 1500<br>50-80 700<br>80-110 44<br>(a)Compute an approximate median value for the data.<br>(b) Plot a histogram using the above data.</td><td>5+5</td></tr>
<tr><td>Q. 5</td><td>Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result<br><table><tr><td>Age</td><td>23</td><td>23</td><td>27</td><td>27</td><td>39</td><td>41</td><td>47</td><td>49</td><td>50</td></tr><tr><td>%fat</td><td>9.5</td><td>26.5</td><td>7.8</td><td>17.8</td><td>31.4</td><td>25.9</td><td>27.4</td><td>27.2</td><td>31.2</td></tr><tr><td>Age</td><td>52</td><td>54</td><td>54</td><td>56</td><td>57</td><td>58</td><td>58</td><td>60</td><td>61</td></tr><tr><td>%fat</td><td>34.6</td><td>42.5</td><td>28.8</td><td>33.4</td><td>30.2</td><td>34.1</td><td>32.9</td><td>41.2</td><td>35.7</td></tr></table>(a) Calculate the mean, median and standard deviation of age and %fat<br>(b) Draw the boxplots for age and %fat</td><td>4+3+3</td></tr>
</table>

| | | |
|---|---|---|
| | (c) Draw a scatter plot based on these two variables. | |
| Q.6 | Suppose a group of 12 sales price records has been sorted as follows:<br>5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215<br>Partition them into three bins by each of the following methods :<br>(a)equal-frequency partitioning<br>(b)equal-width partitioning<br>(c) Clustering | 4+3+3 |
| Q.7 | (a)What is attribute subset selection? Why it is important? What are the different methods available for doing attribute subset selection?<br>(b)What is curse of dimensionality? Discuss the different methods available for dimensionality reduction? | 5+5 |
| Q.8 | (a)Write an algorithm for k-nearest neighbor classification.<br><br>(b)From the following table shows the midterm and final exam grades obtained for students in a database course<br>X — Midterm Exam / y — Final Exam<br>72 — 84<br>50 — 63<br>81 — 77<br>74 — 78<br>94 — 90<br>86 — 75<br>59 — 49<br>83 — 79<br>65 — 77<br>33 — 52<br>88 — 74<br>81 — 90<br><br>(i)Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.<br>(ii) Predict the final exam grade of a student who received an 86 on the midterm exam. | 5+5 |