C5-R4:DATA WAREHOUSING AND DATA MINING

NOTE :

1. **Answer question 1 and any FOUR from questions 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours **Total Marks: 100**

1. (a) What is a Pivot Table ? How does it help in analyzing multidimensional data ?

(b) Define Data Cube. How can you convert tables and spreadsheet to data cubes ?

(c) What are outliers ? Explain their usage in the data analysis.

(d) What is an attribute selection measure ? Name some popular attribute selection measures.

(e) Name and briefly explain the various approaches for evaluation of classifiers.

(f) What is the difference between nominal, ordinal, ratio and interval variables ? Give example for each.

(g) Explain the terms Support and Confidence for Association rule mining.

(7 ×4)

2. **(**a) Explain various database architecture used in a data warehouse for parallel processing.

(b) What are Aggregate Tables and what is the need for building an Aggregate Table ? Explain the Fact Constellation Schema and its advantages.

(9+9)

3. (a) What is a Noise ? Explain some data smoothing techniques to remove the noise.

(b) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) as:

13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70

(i) Use min-max normalization to transform the value 35 for age onto the range [0:0; 1:0].

(ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of the age is 12.94 years.

(iii) Use normalization by decimal scaling to transform the value 35 for age.

(6+12)

4. (a) Consider the following data set shown in table where bank officials need to build decision tree to classify bank loan applications by assigning applications to one of the three risk classes that is A, B, C.

Owns Home	Married	Gender	Employed	Credit Rating	Risk Class
YES	YES	MALE	YES	A	B
NO	NO	FEMALE	YES	A	A
YES	YES	FEMALE	YES	B	C
YES	NO	MALE	NO	B	B
NO	YES	FEMALE	YES	B	C
NO	NO	FEMALE	YES	B	A
NO	NO	MALE	NO	B	B
YES	NO	FEMALE	YES	A	A
NO	YES	FEMALE	YES	A	C
YES	YES	FEMALE	YES	A	C

Draw the decision tree for the above table using ID3 algorithm.

(b) What are distances based algorithms for classification ? State K Nearest Neighbors algorithm with its drawback.

(12+6)

5. (a) Consider the following sets of items are given to form clusters: {2,4,10,12,3,20,30,11,25} with K=2.. Using K-Mean Clustering algorithm find the clusters values.

(b) List some techniques for clustering high-dimensional data.

(12+6)

6. (a) Given the following transactional database:

(1) C,B,H
(2) B,F,S
(3) A,F,G
(4) C,B,H
(5) B,F,G
(6) B,E,O

(i) Find all the frequent item sets in the data using Apriori algorithm. Assume the minimum support level is 30% (the set of frequent item sets should be in L1, L2….., and candidate item sets in C1,C2….).

(ii) Find all the association rules that involve only B, C, H (in either left or right hand side of the rule). The Minimum confidence is 70%.

(b) Why do we need correlation analysis in mining association rules ? List various correlation measures which help to mine large data sets.

(12+6)

7. Write short note on the followings:

(a) Time series analysis
(b) Models of OLAP
(c) Multimedia Data Mining

(6+6+6)