B.C.S.E. FINAL EXAMINATION, 2017
(2ND SEMESTER)
DATA MINING TECHNIQUES AND APPLICATIONS

Time : Three Hours Full Marks 100

Answer Question 1 and any 4 questions from the rest
All questions carry equal marks
Any subpart of a question carry equal marks unless otherwise specified

Q1.A) Choose the most appropriate answer :-

(i) Which amongst the following is a part of data preprocessing :
- (a) Data selection
- (b) Data mining
- (c) Pattern Evaluation
- (d) Knowledge presentation

(ii) Noisy data can be cleaned by
- (a) Correlation Analysis
- (b) Aggregation
- (c) Regression
- (d) Normalization

(iii) Histograms are utilized for
- (a) Data reduction
- (b) Clustering
- (c) Sampling
- (d) Measuring central tendency of data

(iv) A sub-cube is defined by performing a selection on one dimension through the operation
- (a) Roll-up
- (b) Dice
- (c) Pivot
- (d) Slice

(v) Finding Frequent Itemsets involving no candidate generation uses
- (a) The Apriori algorithm
- (b) Vertical data format
- (c) Dynamic itemset counting technique
- (d) FP-growth algorithm

(vi) Which one amongst the following adopts Lazy-learning Classification
- (a) Decision Tree Induction
- (b) Statistical Classifiers
- (c) *k*-Nearest-Neighbor Classifiers
- (d) Neural Networks

(vii) Which of the following is not a measure of correlation ?
- (a) all_confidence
- (b) Kulczynski
- (c) cosine
- (d) F-score

(viii) One method for increasing the accuracy of a predictor is
- (a) Holdout Method
- (b) Cross-validation
- (c) Bootstrap
- (d) Bagging

(ix) The k-Means clustering algorithm belongs to the
- (a) Partitioning methods
- (b) Hierarchical methods
- (c) Density-based methods
- (d) Model-based methods

(x) Which one is a grid based outlier detection method :
- (a) OPTICS
- (b) STING
- (c) CELL
- (d) DENCLUE

[Contd.]

B) Match the entries in the two groups :

Group X		Group Y
(i)	Subset of a Data Warehouse	(a) Decision tree induction
(ii)	Cluster Analysis	(b) MOLAP tools
(iii)	Data reduction	(c) Outlier Detection
(iv)	Attribute Selection	(d) Nonmetric similarity
(v)	Operational metadata	(e) Data cube aggregation
(vi)	Multidimensional Data Models	(f) Support Vector Machines
(vii)	Correlation from Association	(g) Lift measure
(viii)	Bayesian Belief Networks	(h) Data Mart
(ix)	Vector Objects	(i) Statistical Classifier
(x)	"Essential" training tuples	(j) History of migrated data

Q2. (a) Describe the different measures for data characteristics such as
(i) central tendency of data
(ii) dispersion of data and its related plot

(b) Elaborate some graphical methods for
(i) summarizing distribution of an attribute
(ii) determining the relationship between two numeric attributes

Q3. (a) What is a data cube ? Explain with reference to lattice of cuboids in terms of illustrative data involving multiple dimensions. 10

(b) Explain how the above data can be
(i) stored in a Data Warehouse using one of the standard schemas. 5
(ii) expressed using the Data Mining Query Language. 5

Q4. (a) How are association rules generated ? Illustrate with an algorithm.

(b) Is drinking habit related to profession? Find by performing correlation analysis on categorical attributes using χ^2 measure based on observed data represented by the following 2 x 2 contingency table :

Profession->	Researchers		Non-Researchers		TOTAL
Preferred Drink	Observed Frequency	Expected Frequency	Observed Frequency	Expected Frequency	
Coffee	250		200		
Cola	30		1000		
TOTAL					

Complete the above table and compute the χ^2 value to establish / reject the hypothesis posed by the initial question.
Note that the degrees of freedom for this table is (2-1)(2-1)=1 and the χ^2 value needed to reject the hypothesis at the 0.001 significance level is at most 10.828.

[Contd.]

Q5. Consider the learning task represented by the following training tuples where the target attribute *PlayTennis*, which can have values *yes* or *no* for different Saturday mornings, is to be predicted based on other attributes of the morning :

Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Draw a decision tree after calculating the information gain for each attribute as per data given in the above table. Discuss the logic of the algorithm and theory behind the technique employed to select attributes.

Q6. Describe the principle of one standard technique for the following :

(a) Hierarchical Clustering.

(b) Density-based Clustering.

Q7. Write short notes on any <u>four</u> of the following :

(a) Binning and its utility

(b) Data warehouse architecture

(c) Closed frequent itemsets and max-itemsets

(d) Case-based Reasoning techniques

(e) Outliers and their types

--- XXX ---