Roll No:
(To be filled in by the candidate)

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS,   APRIL - 2014

MSc – THEORETICAL COMPUTER SCIENCE   Semester : 8

09XT83   DATA MINING

Time : 3 Hours   Maximum Marks : 100

INSTRUCTIONS:
1. Answer **ALL** questions from PART - A and Answer any **5** questions from PART - B. Question under PART − C is compulsory.

PART - A   Marks : 10 x 3 = 30

1. List the different usages of Chi square test in data mining.

2. Normalize the following data using z-score measure. What is the nature of the normalized data?

   X = { 72 50 81 74 94 86 }

3. What are the measures for evaluating regression techniques?

4. Suppose that the data for analysis includes the feature Age. The age values are in increasing order. Obtain the five number summary for the following data:

   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

5. How is the strong association rule generated from the frequent item sets?

6. Define Covariance and Correlation. Calculate covariance and correlation for the following students final exam grades (Y) and midterm grade (X) in the course and interpret the results.
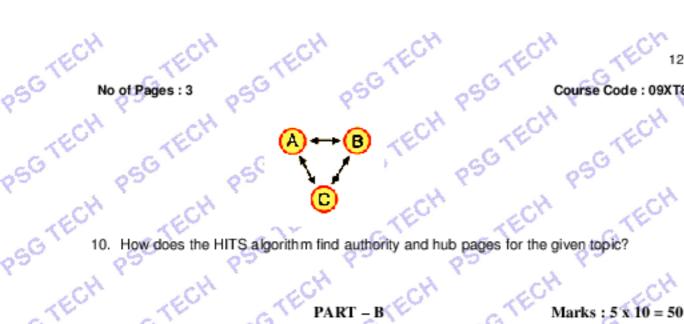
X	72	50	81	74	94	86	59	83	65	33	88	81
Y	84	63	77	78	90	75	49	79	77	52	74	90

7. What is the intuition behind finding the *Eigen vectors* for the *Covariance matrix* in PCA?

8. K-Medoids clustering algorithm guarantees global optimum solution. How? Analyse the pros and cons of K-medoids clustering.

9. How does the Google search engine calculate the page rank? Find the page rank for the nodes in the web graph below.

10. How does the HITS algorithm find authority and hub pages for the given topic?

PART – B   Marks : 5 x 10 = 50

11. Explain in detail about ID3 algorithm with necessary prerequisite. Construct a decision tree for the following data using ID3 algorithm:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

12. Explain the different performance measures for classification. Design the algorithm for generating ROC curve for the given data. Plot the same for the following data. Assume min = 0.4 max = 1.0 increment = 0.1 Y = actual label Y' = prediction of the algorithm and Score = Posterior probability for the positive class.

Y	Y'	Score
+	+	0.99
+	+	0.99
+	+	0.93
+	+	0.86
+	+	0.64
+	+	0.55
-	+	0.68
-	+	0.50
-	-	0.48
-	-	0.44

13. What is meant by outlier? How does it differ from noise? Discuss the different outlier detection techniques for univariate and multivariate data.

14. What is clustering? Explain the different metrics for finding the dissimilarity between two instances. Discuss in detail about any two hierarchical clustering algorithms.

15. Explain in detail about the following feature selection / reduction techniques with mathematical model.

   Entropy (5)

   Principal component analysis (5)

16. What is the objective of linear regression? How does it differ from Classification? Write the model of linear regression with respect to one and two independent variables.

PART – C   Marks : 1 x 20 = 20

17. What is meant by association rule? Discuss in detail about the generation of association rules using   a. Apriori algorithm   b. FP tree Algorithm   (10)

A database has four transactions. Let minimum support = 60% and minimum confidence level = 80%. Find all the frequent item sets and generate association rules using Apriori and FP Growth algorithm. (10)

TID	Items
1	K A D B
2	D A C E B
3	C A B E
4	B A D

/END/