Total No. of Questions : 8]

SEAT No. : □

P3110

[Total No. of Pages : 3

[5154]-677

B.E.(Computer Engineering)
DATA MINING TECHNIQUES AND APPLICATIONS
(2012 Pattern) (Semester-I) (410444D) (End Sem.) (Elective-I)

Time : 2½ Hours] *[Max. Marks : 70*

Instructions to the candidates:

1) *Answer Q1) or Q2), Q3) or Q4), Q5) or Q6), Q7) or Q8).*
2) *Neat diagrams should be drawn wherever necessary.*
3) *Figures to the right side indicate full marks.*
4) *Assume suitable data, if necessary.*

Q1) a) What are the different data normalization methods? Explain them in brief. **[6]**

b) Consider the training examples shown in the table below for a binary classification problem. **[6]**

Instance	A1	A2	Class
1	T	T	Yes
2	T	T	Yes
3	T	F	No
4	F	F	Yes
5	F	T	No
6	F	T	No
7	F	F	No
8	T	F	Yes
9	F	T	No

i) What is the entropy of this collection of training examples with respect to the 'Yes' class

ii) What are the information gains of A1 and A2 relative to these training examples?

c) Explain with suitable example the frequent item set generation in Apriori algorithm. **[8]**

OR

P.T.O.

Q2) a) What is data preprocessing? Explain the different steps in data preprocessing. **[6]**

b) Explain with example K-Nearest-Neighbor Classifier. **[6]**

c) Explain the following terms: **[8]**

i) Support count

ii) Support

iii) Frequent itemset

iv) Closed itemset.

Q3) a) What are interval-scaled variables? Describe the distance measures that are commonly used for computing the dissimilarity of objects described by such variables. **[8]**

b) What is meant by complete link hierarchical clustering? **[6]**

c) Consider the following vectors x and y. x=[1,1,1,1] y=[2,2,2,2].

Calculate:

i) Cosine Similarity

ii) Euclidean distance. **[3]**

OR

Q4) a) Explain with suitable example K-medoids algorithm. **[8]**

b) Differentiate between the following: **[6]**

i) Partitioning and hierarchical clustering

ii) Centroid and average link hierarchical clustering

iii) Symmetric and asymmetric binary variables.

c) How the Manhattan distance between the two objects is calculated? **[3]**

Q5) a) What is Web content mining? Explain in brief. **[7]**

b) Assume 'd' is the set of documents and 't' is the term. Write the formulas to determine. **[8]**

i) Term frequency freq(d, t)

ii) Weighted term frequency TF(d, t)

iii) Inverse document frequency IDF(t)

iv) TE-IDF measure TF-IDF(d, t)

c) What is Web crawler? **[2]**

OR

Q6) a) Compare the different text mining approaches. **[9]**

b) Explain the following terms: **[8]**

i) Recommender system

ii) Inverted index

iii) Feature vector

iv) Signature file.

Q7) a) Explain with neat diagram systematic machine learning framework. **[8]**

b) Write short notes on: **[8]**

i) Big data

ii) Multi-perspective decision making.

OR

Q8) a) What is reinforcement learning? Explain. **[8]**

b) Write short notes on: **[8]**

i) Wholistic learning

ii) Machine learning

⦿ ⦿ ⦿