USN | | | | | | | | | | |

16IS63

Sixth Semester B.E. Semester End Examination, May/June 2018-19

DATA MINING

Time: 3 Hours | Max. Marks: 100

Instructions: 1. *Unit I and Unit II are compulsory.*
2. *Answer any one full question from remaining units*
3. *Use suitable examples wherever needed.*

UNIT – I (Compulsory)

			L	CO	PO	M
1	a.	What is data warehouse? List and explain the general guidelines for implementing data warehouse.	(1)	(1)	(1)	(10)
	b.	Explain the process of ETL with an example	(2)	(1)	(1)	(10)

UNIT – II (Compulsory)

			L	CO	PO	M
2	a.	What is an attribute? List and explain different attribute types	(2)	(2)	(1)	(05)
	b.	Describe data preprocessing. Summarize the different strategies/ techniques available for data preprocessing	(2)	(2)	(1)	(10)
	c.	Compute the cosine similarity for the given document vectors X= (3,2,0,5,0,0,0,2,0,8) Y= (1,0,0,0,0,1,0,1,0,2)	(3)	(2)	(2)	(05)

UNIT – III

			L	CO	PO	M
3	a.	Summarize how to extract association rules efficiently from a given frequent dataset.	(3)	(3)	(2)	(10)
	b.	Explain the FP growth algorithm with an example	(3)	(3)	(1)	(10)

OR

4 a. For the transaction data set,

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

Construct the FP tree and explain the steps followed in constructing it. (3) (3) (3) (05)

b. Explain how rule generation is done in Apriori algorithm with a pseudocode for the same (2) (3) (1) (10)

c. Discuss the concept of Support & Confidence (2) (3) (1) (05)

Note: L (Level), CO (Course Outcome), PO (Programme Outcome), M (Marks)

			L	CO	PO	M
		UNIT – IV				
5	a.	What is decision tree? Construct a decision tree for mammal classification problem.	(3)	(3)	(2)	(08)
	b.	Consider the training examples shown in Table for a binary classification problem.				

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

a. Compute the Gini index for the overall collection of training examples.
b. Compute the Gini index for the Customer ID attribute.
c. Compute the Gini index for the Gender attribute.
d. Compute the Gini index for the Car Type attribute using multiway split.
e. Compute the Gini index for the Shirt Size attribute using multiway split.
f. Which attribute is better, Gender, Car Type, or Shirt Size?
g. Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

(3) (3) (2) (08)

c. Write a note on Gain ratio (2) (2) (1) (04)

OR

6 a. Explain the k- nearest neighbor algorithm. Explain 1,2,3- nearest neighbor of an instance (3) (3) (1) (10)

b. Illustrate Hunt's algorithm to induce decision tree with an example (3) (4) (3) (10)

			L	CO	PO	M
		UNIT –V				
7	a.	List and explain the features of cluster analysis	(3)	(3)	(1)	(10)
	b.	With a neat diagram explain taxonomy of cluster analysis methods	(2)	(3)	(1)	(10)

OR

8 a. Explain K- means algorithm with an example (3) (3) (1) (10)

b. Write a note on Quality and validity of cluster analysis methods (2) (3) (2) (10)

USN | | | | | | | | | | |

16IS63

Sixth Semester B.E. Fast Track Semester End Examination, July/August 2019

DATA MINING

Time: 3 Hours | Max. Marks: 100

Instructions:
1. *Unit I and Unit II are compulsory*
2. *Answer any one full question from remaining units*
3. *Use suitable examples wherever needed.*

UNIT – I (COMPULSORY)

			L	CO	PO	M
1	a.	Define data warehouse. List and explain the implementation steps for building a data warehouse	(2)	(1)	(1)	(10)
	b.	Describe OLAP. Explain the general characteristics of OLAP	(2)	(1)	(1)	(10)

UNIT – II (COMPULSORY)

			L	CO	PO	M
2	a.	What is data mining? Explain the different data mining tasks with suitable examples.	(2)	(2)	(1)	(10)
	b.	Write a note on applications of data mining in different fields.	(2)	(2)	(1)	(06)
	c.	Compute the cosine similarity for the given document vectors X= (2,2,1,2,3,4,2,3,2,0) Y= (2,3,2,2,1,2,1,0,1,2)	(3)	(2)	(3)	(04)

UNIT - III

			L	CO	PO	M
3	a.	What is frequent itemset? Explain how frequent item set is generated using FP growth algorithm	(3)	(3)	(1)	(10)
	b.	Illustrate the application of apriori algorithm in generation of frequent itemset	(3)	(3)	(1)	(10)

OR

			L	CO	PO	M
4	a.	How rules can be generated using apriori algorithm? Explain with suitable snippets	(3)	(3)	(3)	(10)
	b.	For the transaction data set,				

TID	Items
1	{milk}
2	{milk, bread}
3	{milk, bread, butter}
4	{bread, butter}
5	{butter, diaper}
6	{bread, butter, diaper}
7	{milk}
8	{diaper}
9	{bread, diaper}
10	{bread}

Construct the FP tree for the above transaction data set. Also generate the frequent itemset. (3) (3) (3) (10)

UNIT - IV

			L	CO	PO	M
5	a.	Discuss hunt's algorithm for decision tree induction with suitable example	(3)	(2)	(2)	(10)
	b.	Explain k nearest neighbor algorithm with suitable example	(3)	(2)	(3)	(10)

Note: L (Level), CO (Course Outcome), PO (Programme Outcome), M (Marks)

OR

6 a. Describe Bayesian theory. With suitable mathematical expressions, explain naïve classificaition. (3) (4) (3)

b. Explain Bayesian Classifiers for the data set

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

(3) (3) (2)

L CO PO

UNIT –V

7 a. With a neat diagram, explain the clustering types (2) (4) (1)

b. Explain how the clustering methods are evaluated in terms of quality and validity (2) (3) (2)

OR

8 a. Write a note on
(a) Partitional methods
(b) Hierarchical methods (2) (3) (2)

b. List and explain the desired features of cluster analysis (3) (3) (1)

Sixth Semester B.E. Makeup Examination, May/June 2018-19 **16IS63**

DATA MINING

e: 3 Hours Max. Marks: 100

Instructions: 1. *Unit I and Unit II are compulsory*
2. *Answer any one full question from remaining units*
3. *Use suitable examples wherever needed.*

UNIT – I (Compulsory)

			L	CO	PO	M
a.	What is ODS? Explain typical structure of ODS with a neat diagram		(2)	(1)	(1)	(08)
b.	Differentiate between ODS and DW		(2)	(1)	(1)	(05)
c.	Explain the implementation steps for building a data warehouse		(2)	(1)	(1)	(07)

UNIT – II (Compulsory)

			L	CO	PO	M
a.	What is data mining? Explain the KDD process with a neat diagram		(2)	(1)	(1)	(08)
b.	List and explain the different data mining tasks with a suitable examples		(2)	(2)	(1)	(08)
c.	Compute the Jaccard's coefficient and simple matching coefficient for the binary vectors. X= (1,0,0,1,0,1,0,1,0,1), Y= (1,1,1,1,0,0,0,1,0,1)		(3)	(2)	(2)	(04)

UNIT - III

			L	CO	PO	M
3 a.	What is frequent itemset generation? Explain frequent itemset generation of Apriori algorithm with a pseudocode.		(3)	(2)	(2)	(10)
b.	List and explain the factors that affects computational complexity of Apriori algorithm		(2)	(2)	(1)	(10)

OR

4 a.	Illustrate different methods for generating frequent item set.		(2)	(2)	(1)	(10)
b.	Write a note on (a) Maximal frequent itemset (b) Closed itemset		(2)	(2)	(1)	(10)

UNIT - IV

			L	CO	PO	M
5 a.	Explain Bayesian Classifiers for the data set		(3)	(3)	(2)	(10)

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

b.	Explain how decision tree can be constructed using hunt's algorithm. Use mammal classification as an example.		(3)	(2)	(2)	(10)

Note: L (Level), CO (Course Outcome), PO (Programme Outcome), M (Marks)

OR

8 6 a. Consider the training examples shown in Table for a binary classification problem.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

a. Compute the Gini index for the overall collection of training examples.
b. Compute the Gini index for the Customer ID attribute.
c. Compute the Gini index for the Gender attribute.
d. Compute the Gini index for the Car Type attribute using multiway split.
e. Compute the Gini index for the Shirt Size attribute using multiway split.

b. Explain the characteristics of decision tree induction (3) (3) (2) (10)

(2) (4) (2) (10)

UNIT –V

L CO PO M

7 a. What is agglomerative clustering? Explain in detail with an example

b. List and explain the cluster analysis methods (2) (3) (1) (10)

(2) (3) (1) (10)

OR

8 a. Explain the density based clustering method in detail

b. Explain how the clustering methods are evaluated in terms of quality and validity (2) (3) (1) (10)

(2) (3) (2) (10)