



LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600 034

M.Sc. DEGREE EXAMINATION – COMPUTER SCIENCE

FIRST SEMESTER – NOVEMBER 2014

CS 1822 - DATA MINING

Date : 31/10/2014
Time : 01:00-04:00

Dept. No.

Max. : 100 Marks

Part A

Answer all the questions:

(10 x 2 = 20 Marks)

1. Write down the steps involved in KDD.
2. Mention any four parts of a genetic algorithm.
3. Stress the need for outlier mining.
4. Draw the structure of a Perceptron.
5. Differentiate similarity and distance measures.
6. Specify the use of a Spanning tree in data mining.
7. How association rules mining are useful for a doctor?
8. State the significance of incremental rules.
9. Draw the hierarchy of text mining.
10. Define the terms: a) Disjoint b) Covered by

Part B

Answer all the questions:

(5 x 8 = 40 Marks)

11. a) Illustrate the use of EM algorithm to find the missing data items in a set {1,5,10, 4} with an initial guess of 4 where $n=6$. (Or)
b) Explain how activation functions help neural networks to do mining.
12. a) Illustrate the use of Rule based algorithms for finding grades of students based on their marks (Or)
b) Determine prediction for the class values 0 as short and 1 as medium, for the heights given below: {(1.6, 0), (1.9,1), (1.88, 1) (1.7, 0), (1.85, 1), (1.6, 0), (1.7, 0), (1.8, 1), (1.95, 1), (1.9,1), (1.8, 1), (1.75, 1)}.
13. a) Use Single Link Technique to cluster cities {A, B, C, D, E} based on the distance between them as given below:

Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

(Or)

- b) Explain PAM algorithm with an example.

14. a) Determine Large Item Sets for the set of transactions $t_1=\{\text{Pencil, Book, Rubber}\}$, $t_2=\{\text{Pencil, Rubber}\}$, $t_3=\{\text{Pencil, Box, Rubber}\}$, $t_4=\{\text{Pen, Pencil}\}$ and $t_5=\{\text{Pen, Box}\}$ where support =30% and confidence =50% (Or)

b) Compare all the association rule algorithms and their metrics.

15. a) Discuss a method to draw a Directed Acyclic Graph (DAG) for the web pages visited during an integer time stamp as given below: $\langle(A, 1), (B, 2), (C, 2), (A, 7), (C, 10), (B, 10), (C, 12), (A, 12), (C, 13), (D, 14), (C, 14), (C, 14), (E, 20)\rangle$ (Or)

b) Elaborate the use of spatial mining primitives and the related data structures in flood relief activities.

Part C

Answer any two questions.

2 x 20 = 40 Marks

16. a) Explain in detail about predictive approaches with suitable example. (10 Marks)

b) Elaborate the design of distance-based algorithms in classification. (10 Marks)

17. a) Illustrate the use of K-means clustering algorithm for the data $\{2, 4, 10, 12, 3, 20, 30, 11, 25\}$ when $k=2$ (10 Marks)

b) Explain how Apriori-Gen algorithm generates rules with suitable example. (10 Marks)

18. a) Describe in detail about web usage mining. (10 Marks)

b) Determine the splitting attribute for the credit card authorization problem given below:

(10 Marks)

Id	Age	Has_job	Own_house	Credit rating	Class
1	Y	F	F	Fair	No
2	Y	F	F	Good	No
3	Y	T	F	Fair	Yes
4	Y	T	T	Fair	Yes
5	Y	T	T	Fair	No
6	M	F	F	Fair	No
7	M	F	F	Good	No
8	M	F	T	Good	Yes
9	M	T	T	Excellent	Yes
10	M	F	T	Excellent	Yes
11	O	F	T	Excellent	Yes
12	O	F	T	Good	Yes
13	O	T	F	Good	Yes
14	O	T	F	Excellent	Yes
15	O	F	F	Fair	No

Age Y- Young, M- Medium and O-Old, Has_Job/ Own_house T- True and F – False
