4996

**No of Pages :3** **Course Code : 12XWAC**

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

## PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

## SEMESTER EXAMINATIONS, NOVEMBER 2018

## MSc – SOFTWARE SYSTEMS Semester : 9

## 12XWAC DATA MINING

**Time : 3 Hours** **Maximum Marks : 100**

**INSTRUCTIONS:**

1. Answer **ALL** questions. Each question carries 20 Marks.
2. Subdivision (a) carries 3 marks each, subdivision (b) carries 7 marks and subdivision (c) carries 10 marks each.
3. Course Outcome: Table

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Qn.1 | CO1 | Qn.2 | CO2. | Qn.3 | CO3. | Qn.4 | CO4 | Qn.5 | CO5 |

1. For each of the following activities identify the related data mining task and justify your answer.

- Predicting a species of flower based on the characteristics of the flower.
- Finding customers who purchased laptop also purchases digital camera.
- Identifying data objects that do not comply with the general behavior of data model.

b) (i) What is symmetric binary attribute? How does it differ from asymmetric binary attribute? [3]

(ii) When is a distance measure called a metric? Prove that the Manhattan distance is a metric. [4]

c) Real world data tend to be incomplete, noisy and inconsistent. Explain the process of handling incomplete and noisy data with suitable examples.

2. a) In identifying frequent item sets in a transactional database, we find the following to be the frequent 3-itemsets:

{B, D, E}, {C, E, F}, {B, C, D}, {A, B, E}, {D, E, F}, {A, C, F}, {A,C, E}, {A, B, C}, {A, C,D}, {C, D, E}, {C, D, F}, {A, D, E}. Which among the following 4-itemsets can be frequent? Why?

- {A, B, C, D}
- {A, B, D, E}
- {A, C, E, F}
- {C, D, E, F}

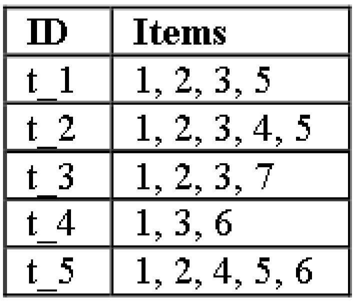
b) (i) Is an association rule symmetric? Justify. [3]

(ii) Draw FP tree for the following transaction database.

| TID | Transaction |
|---|---|
| T1 | A,B,C,D,E,F |
| T2 | B,C,D,E,F,G |
| T3 | A,D,E,H |
| T4 | A,D,F,I,J |

[4]

c) Consider the following transaction database where 1,2,3,4,5 and 6 are items.



| ID | Items |
|---|---|
| t_1 | 1, 2, 3, 5 |
| t_2 | 1, 2, 3, 4, 5 |
| t_3 | 1, 2, 3, 7 |
| t_4 | 1, 3, 6 |
| t_5 | 1, 2, 4, 5, 6 |

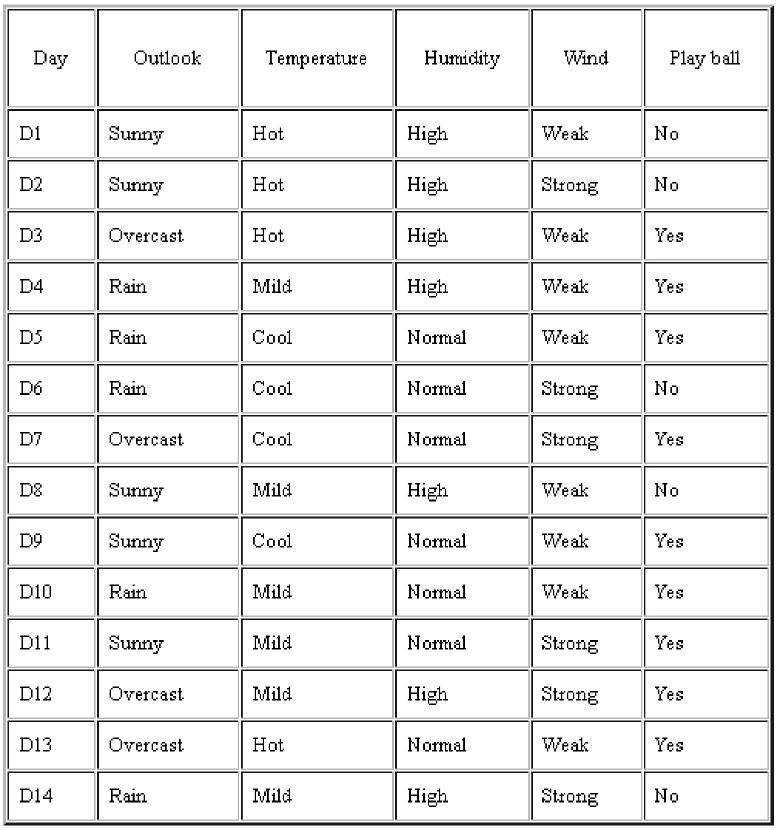
Use apriori algorithm to find all frequent item sets with minimum support 60%. Show your computation steps in detail.

3. a) What are the assumptions made in Naïve Baye's classification? Why is it done?

b) (i) What are the metrics associated with evaluation of classifier? Give an example for each metric. [3]

(ii) What is "bagging" and how is it different from "boosting?" When would you use either of these techniques? [4]

c) Discuss the merits of ID3 classification algorithm when compared with Hunt's classification algorithm. Apply the ID3 algorithm to the following data set and find which attribute is selected as root attribute. (**No need to build the entire decision tree**.)



| Day | Outlook | Temperature | Humidity | Wind | Play ball |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

4. a) Identify the requirements and challenges of clustering algorithms.

b) i) What is a core point in DBSCAN algorithm? How does it differ from border point? [3]

ii) What are the different categories of time series data? How is the trend curve estimated? [4]

c) (i) Outline K-Means algorithm. Cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9) using k-means algorithm. Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). Show the computation for two iterations.

(OR)

(ii) Explain the characteristics of stream data. Also explain how classification of stream data is done using Hoeffding tree algorithm.

5. a) How are failures treated in Map-Reduce process?

b) (i) What is distributed data mining? Mention its applications. [3]

(ii) Describe the structure of CF tree constructed in BIRCH algorithm. [4]

c) Briefly describe and give examples of each of the following applications to data mining: Graph data, Spatial data and Text data.

**/END/**

**FD/RL**