Indian Institute of Information Technology, Allahabad

C1- Examination (Feb 2023)

Paper: Data Mining

B.Tech. (IT), VIth Semester **Max. Marks : 25** **Duration: 1.5 Hours**

Course Instructor: Prof. O.P. Vyas, Prof. Vrijendra Singh & Dr. Manish Kumar

---

Ques 1: Choose the correct option and write it with brief explanation

[01+01+01+01+02 = 06 Marks]

i) _____computes the difference between entropy before the split and average entropy after the split of the dataset based on given attribute values.

- a) Information gain
- b) Gini ratio
- c) Pruning

ii) Choose the correct statement from below –

- a) A decision tree is a graphical representation of all the possible solutions to a decision based on certain conditions.
- b) Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- c) A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogenous (mutually exclusive) classes.
- d) All of the above.

iii) Our use of association analysis will yield the same frequent itemsets and strong association rules whether a specific item occurs once or three times in an individual transaction. **(T/F)**

iv) Suppose that X, Y, and Z are random variables. X and Y are positively correlated and Y and Z are likewise positively correlated. Does it follow that X and Z must be positively correlated? **(Yes/No)**

v) Why does lift have a bigger role than confidence in Association rules?

Ques 2: A database has 4 transactions, shown below **[03+04 = 07 Marks]**

TID	Date	items_bought
T100	10/15/0	{K, A, D, B}
T200	10/15/0	{D, A, C, E, B}
T300	10/19/0	{C, A, B, E}
T400	10/22/0	{B, A, D}

Assuming a minimum level of support min_sup = 60% and a minimum level of confidence min_conf = 80%:

(a) Find all frequent itemsets (not just the ones with the maximum width/length) using the Apriori algorithm. Show your solved steps (just showing the final answer is not acceptable). For each iteration show the candidate and acceptable frequent itemsets.

(b) List all of the strong association rules, along with their support and confidence values, which match the following metarule, where X is a variable representing customers and item i denotes variables representing items (e.g., "A", "B", "C", "D", "E", "K" etc.).

$\forall x \in transaction, buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)$

Ques 3: Attempt the following with detailed explanation. **[04+08 = 12 Marks]**

(a) Define decision tree. Explain the basic algorithm for classification by decision tree induction. What is the crucial point root attribute and internal node attributes selection.

(b) Consider the following customer database for All Electronics stores. Using information gain measure, construct the decision tree up to two levels.

RID	age	income	student	credit_rating	Class:buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	No	excellent	no