18CSE355T Data Mining and Analytics [DE1]

Question Bank for UNIT-2

1. A data set for analysis includes only one attribute X:

   X={ 7,12,5,8,5,9,13,12,19,7,12,12,13,3,4,5,13,8,7,6}

   a. What is the mean of the data set X?

   b. What is the median?

   c. Find the standard deviation for X.

2. Define Frequent sets, confidence, support and association rule.

3. What do you mean by Market Basket analysis and how it can help in a supermarket?

4. List two interesting measures for association rules. (OR) Rule support and confidence are two measures of rule interestingness.

5. Explain whether association rule mining is supervised or unsupervised type of learning.

6. Name some variants of Apriori Algorithm.

7. How can support and confidence be misleading metrics? Explain using an example how other interestingness measures are used to determine the strength of an association rule.

8. What do you mean by FP- Growth? What advantages does it have over Apriori Algorithm? What are the various steps involved in an FP-Growth algorithm?

9. What is over fitting and what can you do to prevent it?

10. Discuss the importance of Association Rule Mining.

11. The heights of players of a school's basketball team are 72",74",70",78",75" and 70". Find the mean height.

12. The batting averages for members of a basketball team are 0.234, 0.256, 0.321, 0.333, 0.290. Find the median batting average.

13. How is association rule mined from large databases?

Transaction ID	Items
100	A,C,D
200	B,C,E
300	A,B,C,E
400	B,E

14. Consider the Data set D. Given the minimum support 2, apply Apriori algorithm on this dataset.

15. Describe example of data set for which Apriori check would actually increase the cost? By describe I mean either show an instance of the data set or describe how would it look like.

16. Assume that each item in supermarket is bought by 1% of transactions. Assume that there are 10 million transactions and that items are statistically independent. Assume mid-sup = 10. What is the expected size of a frequent set? What is the expected number of frequent sets?

17. A database has five transactions. Let min_sup=60% and min_conf =80%

TID	Items_bought
T100	{M,O,N,K,E,Y}
T200	{D,O,N,K,E,Y}
T300	{M,A,K,E}
T400	{M,U,C,K,Y}
T500	{C,O,O,K,I,E}

a. Find all the frequent itemsets using Apriori and FP-Growth respectively. Compare the efficiency of the two mining processes.

b. List all the strong association rules(with support s and confidence c) matching the following metarule, where X is avariable representing customers and $item_i$ denotes variables representing items (e.g. “A,” “B,”):

$$\forall x \in transaction, buys(X,item_1) \wedge buys(X,item_2) => buys(X,item_3)[s,c]$$

18. Suppose that you have data describing the closing prices of the stock you own for the last 1000 days. Suppose you are interested in generating all rules which tell you about chances of your stock going up on a given day provided you know the pattern (up or down) on K preceding days, with some minsup and minconf defined. How would you model this problem as association rule mining problem, is there a way to represent this as transactions with binary attributes like in the supermarket case?

19. Write and explain the algorithm for mining frequent item sets without candidate generation. Give relevant example.

20. Discuss the approaches for mining multi-level association rules from the transactional databases. Give relevant example.