Dr Rajinder Sandhu.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT
TEST-3 EXAMINATION (DEC 2018)
B-Tech (7th SEM)

Course Code: 11B1WCI832 Max. Marks: 35
Course Name: INFORMATION RETERIVAL AND DATA MINING Max. Time: 2 Hr
Course Credit: 3

Note: All questions are compulsory. Attempt all parts of one question at one place.

Q. No. 1 Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how following techniques can help: **[5*1=5 Marks] [CO-4]**

i. Data manipulation
ii. Clustering
iii. Classification
iv. Association rule mining
v. Anomaly detection

Q. No. 2 Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Justify your answers briefly. Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity. **[5*1=5 Marks] [CO-1]**

i. Times in terms of AM and PM
ii. Brightness as measured by a light meter
iii. Angles as measured in degrees between 0 to 360 degrees
iv. Medals awarded at the Olympics
v. ISBN number of books

Q. No. 3 Provide advantage and disadvantage of following techniques for anomaly detection: **[1+2+2 Marks] [CO-4]**

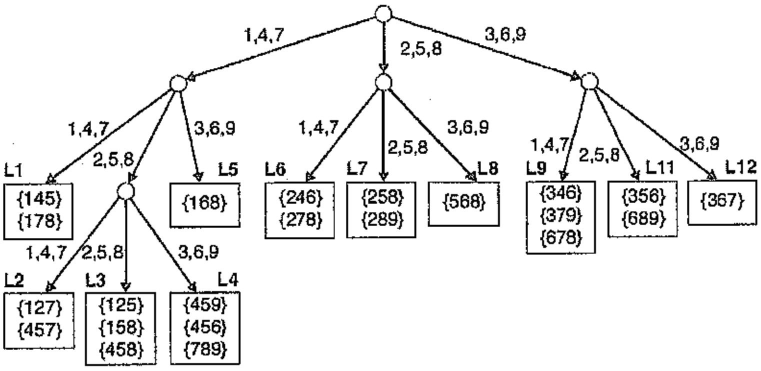
i. Statistical Approach
ii. Distance based Approach
iii. Density based Approach

Q. No. 4 How association rule mining is different from classification and clustering? Explain brute force method for association rule mining with example of shopping basket problem of five transactions and why it is computational expensive to perform? **[2+3 Marks] [CO-2]**

Q. No. 5 The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in Figure below: **[3+2 Marks] [CO-3]**

i. Given a transaction that contains items {1,3,4,5,8},which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?
ii. Use the visited leaf nodes in part (i) to determine the candidate

item- sets that are contained in the transaction {1, 3, 4, 5, 8}.



Q. No. 6 i. Traditional K-means has a number of limitations, such as sensitivity to outliers and difficulty in handling clusters of different sizes and densities, or with non-globular shapes. Comment on the ability of fuzzy c-means to handle these situations. **[5 Marks] [CO-3]**

ii. Provide two examples where clustering is not a good technique to perform in data mining.

Q. No. 7 Consider the training examples shown in Table below for a binary classification problem. **[5 Marks] [CO- 2]**

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

i. Compute the Gini Index for the overall collection of training examples.

ii. Compute the Gini Index for the Customer ID attribute

iii. Compute the Gini Index for the Gender attribute.

iv. Compute the Gini Index for the Car Type attribute using multiway spilt.

v. Which attribute is better, Gender, Car Type or Shirt Size.