**POSSESION OF MOBILE IN EXAM IS UFM PRACTICE**

**Name Enrollment No**

**Jaypee Institute of Information Technology**

**T1 Examination, 2019**

**B.Tech -III Year**

**Course code: 16B1NCI635 Max. Marks: 20**

**Course Name: Data and Web Mining Max. Time: 1Hr**

**Note: Attempt all Questions:**

**Q1 :**[CO1] **[Marks 5]** A genetics engineer was attempting to cross a tiger and a cheetah. She predicted a phenotypic outcome of the traits she was observing to be in the following ratio 4 stripes only: 3 spots only: 9 both stripes and spots. When the cross was performed and she counted the individuals she found 50 with stripes only, 41 with spots only and 85 with both. According to the Chi-square test, did she get the predicted outcome? (For degree of freedom 1, 2 and 3 at 0.05 significance level chi-square values are: 3.841, 5.991 and 7.815)

**Q2 :[CO3] [Marks 4+2]**Consider a data having attributes Color, Height and Width, and the class can be either yes or no.

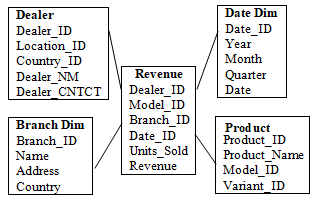
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Example | Color | Height | Width | Class |
| A | Red | Short | Thin | NO |
| B | Blue | Tall | Fat | YES |
| C | Green | Short | Fat | NO |
| D | Green | Tall | Thin | YES |
| E | Blue | Short | Thin | NO |

1. Design a decision tree based on ID3.
2. For the decision tree classifier, verify the results for the tuples (Green , Short, Thin,Yes) (Blue,short, fat, yes) and Determine the classifier accuracy.

**Q3 :[CO2][Marks 3]**Consider a schema of retail chain Bigmart. Perform the OLAPoperations considering the base cuboidas (dealer,year,branch, product).

a) List the revenue generated by all products in the year 2018.

b) List the total sales done by each dealers in January 2018.



|  |  |
| --- | --- |
| D1 | Computer electronic device computations. |
| D2 | Computer electronic device logical arithmetic computations. |
| D3 | CPU processing logical arithmetic computations. |

**Q4 :[CO1][3+1 marks]** Suppose we query an IR system for the query "Computer logical computations". The database collection consists of three pre-processed documents (D = 3) shown below:

1. Which document will be fetched using term-frequency document vectors incidence matrix?
2. Compute supremum distance between document D1 and D2.

|  |
| --- |
| 27, 26, 4, 21, 24, 34, 7, 15, 28, 29, 21, 9. |

**Q5 :[CO1][2 Marks]** Consider the data for attribute“weight”:

1. Partition the data using equal width partitioning. . Number of intervals = 3.
2. Normalize the weight = 34 using min-max method (min=1, max=2)

Q6 :[CO3] **[Marks 3]** Below is a table showing how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you’ve written an IR system that for this query returns the set of documents {4, 5, 6, 7, 8}.

1. Calculate precision, recall, and F1 of your system if a document is considered relevant only if the two judges agree.
2. Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant.

|  |  |
| --- | --- |
| Q7 :[CO4]**[Marks 4]** The daily expenditure on the food and the entertainment of six students are shown in given table. Use agglomerative hierarchical clustering based on longest distance to form the final customer stores. Plot the dendogram.  Q8 **:[CO1][Marks 4+2+2]** The Pants Pizza Parlour sells pizzas with optional toppings: pepperoni, pineapple and pickled onion. Every day this week you have tried a pizza (A to E) and kept a record of which you liked:  a) Show how the naive Bayes classifier would classify {pepperoni = true, pineapple = true, pickledOnion= false}.  b) Are pineapple = true and pickledOnion = true conditionally independent given liked = false? Show your working.  c) Show how the 2NN classifier would classify {pepperoni = false, pineapple = false, pickledOnion= true}; |  |

Q9 :[CO1]**[Marks 2]** Assume you apply DBSCAN to the same dataset, but the examples in the dataset are sorted differently. Will DBSCAN always return the same clustering for different orderings of the same dataset? Give reasons for your answer.

Q10:[CO1]**[Marks 3]**Try placing the following faces into groups using any of the criteria: glasses, gender, and smile? Which method would you apply to justify. Show the results.

|  |  |  |  |
| --- | --- | --- | --- |
| **Summary of the face characteristics** | | | |
| **case** | **sex** | **glasses** | **smile** |
| 1 | m | y | y |
| 2 | m | y | n |
| 3 | f | n | y |
| 4 | f | n | n |
| 5 | m | y | y |
| 6 | f | n | y |
| 7 | m | y | y |

**Q11 [CO6][Marks 4+4]** Given below is the utility matrix representing the ratings on a 1-5 star scale of nine items, I1 to I9, by three users U1, U2, U3. Compute the following from the given data:

1. What can be the missing predicted rating for item 6 (I6) in recommendation system.
2. If we consider I1 to I9 in above matrix as terms and U1 to U3 represents three documents. Therefore, according to the matrix document U1 contains item I1 4 times i.e. term frequency. Now which document will be retrieved using TF-IDF scoring method if we search the query containing following terms: “I2, I4, I8, I9, I8”.

|  |  |
| --- | --- |
| TID | Items |
| 1 | P,Q,V |
| 2 | Q,R,S |
| 3 | P,R,S,T |
| 4 | P,S,T,T |
| 5 | P,Q,R |
| 6 | P,Q,R,S |
| 7 | Q |
| 8 | Q,R,T |
| 9 | V |

**Q12 [CO5] [Marks 5+2+3]** A database has 9 transactions. Let min support = 30% & min confidence = 65%.

1. List the frequent k-itemset for the largest k using FP-growth (FP tree) and
2. Compare the results of part (a) using Apriori algorithm.
3. Which items X and Y can be placed on the same shelf, so that buyers of one item would be prompted to buy the other?

**Q13: [CO6] [Marks 3+3]** If we can get 1,000 pages pointing to our home page, but only have one link leaving from homepage to one of the incoming page.

1. What will be the page rank of all pages according to random surfer model? Damping factor = 0.15, iterations = 3.
2. What will be the page rank of all pages according to HITS? Iterations = 2

|  |  |  |
| --- | --- | --- |
| Student id | Quiz 1 | Quiz 2 |
| 1 | 8 | 6 |
| 2 | 6 | 5 |
| 3 | 5 | 6 |
| 4 | 4 | 5 |
| 5 | 9 | 8 |
| 6 | 7 | 9 |

**Q14: [CO3+CO4] [Marks 2+5+2+2] a)** Suppose we retrieve 11 documents and every odd numbered retrieved document is relevant. Find out the precision and recall of information retrieval system.

**b)** A group of six students whose performance based on marks obtained in quiz1 and quiz2 is given in table. Now this group is required to be sent in two teams for inter college competition. Extract the best two team members. Apply the appropriate clustering technique.

**c)** Prove that in DBSCAN for a fixed *MinPts* value and 2 neighbourhood thresholds, a cluster with respect to e1and *Minpts* must be subsets of a cluster C’ with respect to and *Minpts*.

**d)** Nine actual values of target variable in the training file is as follows: [0,0,1,1,0,1,0,1,1].

What is the entropy of the target variable?