Scheme of Valuation/Answer Key

Scheme of evaluation (marks in brackets) and answers of problems/key

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

EIGHTH SEMESTER B.TECH DEGREE EXAMINATION, MAY 2019

Course Code: CS402

Course Name: DATA MINING AND WAREHOUSING

Max. Marks: 100 Duration: 3 Hours

PART A

Answer all questions, each carries 4 marks. Marks

1 Helps in decision making-2 marks (4)

Explanation- 2marks

2 Any four differences – 4 marks (4)

3 Necessity – 2marks (4)

List any four of the following, (2 marks)

(Smoothing, aggregation, generalization,normalization,attribute construction)

4 Classification- 2 marks, justification – 2marks (4)

5 Use of LR-2 marks, Explanatuion -2 marks (4)

6 Significance – 4 marks (4)

7 Support-2 marks (4)

Confidence – 2 marks

8 Equation – 2 marks (4)

$$distance = \sum_{i=0}^{n-1} |(x[i] - y[i])|$$

Answer – 2 marks ANS:11

9 Any 2 reasons – 2 marks each (4)

10 Any 2 differences – 2 mark each (4)

PART B

Answer any two full questions, each carries 9 marks.

11 a) Diagram – 2 marks (5)

Explanation – 3marks

b) i) 2 marks Ans: 0, 0.125, 0.25, 0.5, 1 (4)

ii) 2 marks ANS: -0.94868, -0.63246, -0.31623, 0.31623, 1.58114

12 a) Fact table -2 marks (See below) (6)

Dimension table -4 marks

b) Query – 3 marks (3)

You first will need to slice on the condition game == 'GM Place'. Secondly you will need to slice on date.year == '2010'. This will give you all the charges for GM Place in 2010. Next we slice to spectator.type == 'student'. Lastly we sum all the charges in the display phase (pivot).

OR

The specific OLAP operations to be performed are:.

Roll-up on date from date id to year.

Roll-up on game from game id to all.

Roll-up on location from location id to location name.

Roll-up on spectator from spectator id to status.

Dice with status="students", location name="GM Place", and year = 2010.

13 a) Brief description on the following: Cleaning, integration & transformation, reduction, discretization (2.25 marks each) (9)

PART C

Answer any two full questions, each carries 9 marks.

14 Classification method – 5 marks (9)

Correct answer – 4 marks

P(Male)=4/8=.5

P(Female)=4/8=.5

Person=argmaxP(person)P(Height=6/Person)P(weight=130/Person)P(Footsize=8/Person)

P(Person=Male)P(Height=6/Male)P(weight=130/Male)P(Footsize=8/Male)=.5*3/4*0/4*1/4 =0

P(Person=Female)P(Height=6/Female)P(weight=130/Female)P(Footsize=8/Female)=.5*1/4*2/4*2/4 =.0313

Therfore, Class is 'Female'.

15 Classification method – 5 marks (9)

Correct answer – 4 marks

Hamming Distance:

Let Qn be {pepper = false, ginger =true, chilly = true}

Dist(A,Qn) =1+0+0=1

Dist(B,Qn) =1+1+1=3

Dist(C,Qn) =0+0+0=0

Dist(D,Qn) =0+0+1=1

Dist(E,Qn) =1+1+1=3

Since it is 3NN, 3 nearest neighbors are taken, ie. A=false,C=false and D=true.

Majority voting is applied, and hence class is liked=false.

16 a) Two difference- 1.5 marks each (3)

b) Algorithm – 3 marks (6)

Explanation – 3 marks

PART D

Answer any two full questions, each carries 12 marks.

17 a) FP tree construction –3 marks (8)

FP generation – 5 marks

Frequent patterns generated:

I2 I5:2, I1 I5:2, I2 I1 I5:2

I2 I4:2

I2 I3:4 I1 I3:2 I2 I1 I3:2

I2 I1 :4

b) Identifying 3 association rules - 4 marks (4)

Strong association rules:

I1&I5 -> I2

I2 & I5 -> I1

I5 -> I1 & I2

18 a) Explanation – 8 marks (8)

b) Any two advantages – 2 marks each (4)

19 a) Explanation – 4 marks (6)

Drawbacks – 2 marks

b) Equation 3 mark..solution 3 mark (6)

$$TF(d,t) = \begin{cases} 0 & \text{if } freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & \text{otherwise.} \end{cases}$$

$$IDF(t) = \log \frac{1+|d|}{|d_t|}, \quad \text{thus} \quad TF\text{-}IDF(d,t) = TF(d,t) \times IDF(t).$$

where d is the document collection, and d_t is the set of documents containing term t.

$TF(D3, T4) = 1+(\log(1+\log 6)) = 1.249$, $IDF(T4) = \log \frac{(1+4)}{3} = 0.2201$

$TF\_IDF = TF*IDF = 1.249*0.2201 = 0.2749$

****