Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division

Second Semester 2019-2020

M.Tech (Data Science and Engineering)

Mid-Semester Exam (EC-2 Make-up)

Course No. : DSECLZC415

Course Title : Data Mining

Nature of Exam : Open Book

Weightage : 30%

Duration : 90 minutes

Date of Exam : 4/07/2020 (AN), 2:00 pm to 3:30 pm

No. of Pages = 3

No. of Questions = 4

Note to Students:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. **All parts of a question should be answered consecutively. Each answer should start from a fresh page.**
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1:

a) Consider the following ordered list of observations of a variable. Answer the following:
25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 41, 42, 42, 99 **[2+1+2+2=7 marks]**

1) What is the five-number summary for the given data?
2) Draw boxplot.
3) Identify the outliers if any.
4) Explain, how do the outliers affect the measures of central tendency (Mean, and Median) of data? Comment using the given data set.

Answer:

1) Five number summary is 25 33 35 40.5 99 - it is acceptable if Q3 is computed as 41
2) Box plot to represent a)
3) IQR = 7.5 (or 8). Hence observations that are away from the box by more than 12 on both sides, are potential outliers. Here 99 meets the criteria.
4) Mean for the data = 39; Median=35. Mean is impacted by value of the outliers; Median is impacted by count of outliers. Since number of outliers are likely very small compared to dataset size, we can safely assume Median is more robust compared to Mean in the presence of outliers

b) As many data mining algorithms cannot handle missing values, analyst sometimes remove all observations (rows) that contain missing values before the analysis. Give two potential disadvantages of this procedure.

[2M]

2 marks Any two valid points from below list:

1) Too many data objects have missing value, leads to more removal, thus unreliable analysis.
2) Even a partially specified data objects contain some information that might be lost if removed.
3) Removed data object may be critical to evaluation in the context.
4) Removal may lead to inaccuracy in the model created.
5) Removed data objects may exhibit a valid cluster of data or separate sample which is required for the analysis.

c) Consider the following marks details of some students

Name	Marks
Abishek	*20*
Ramesh	30
Vinod	55
Rahul	75
Anu	95
Kavita	40
Latha	45
Pravin	57
Sonu	32
Sneha	65

Use min-max normalization to transform the above marks onto the range [50, 70].

Solution:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Name	Marks	Normalized Marks
Abishek	*20*	50
Ramesh	30	52.67
Vinod	55	59.33
Rahul	75	64.67
Anu	95	70
Kavita	40	55.33
Latha	45	56.67
Pravin	57	59.87
Sonu	32	53.2
Sneha	65	62

1 mark for the correct formula
2 marks for all correct normalised values

Question 2 *:* **[4+1+1=6 marks]**

The following table shows the house area(in sq meters) and house rent(in thousand rupees) obtained for a metropolitan city.

Area	172	150	181	174	194
Rent	42	35	46	40	50

(a) Assuming linear relationship, Use the method of least squares to get an equation for the prediction of rent based on the area.

(b) Predict the rent of a house with 180 sq. meter area.

(c) Do you notice any limitations with the solution?

a) *Mean(Area) = (172+150+181+174+194)/5 = 174.2; Mean(Rent) = (42+35+46+40+50)/5 = 42.6. Using Equations for linear regression, we get gradient*

w1 = ((-2.2)(-.6)+(-24.2)(-7.6)+(6.8)(3.4)+(-.2)(-2.6)+(19.8)(7.4))/((-2.2)^2+(-24.2)^2+(6.8)^2+(-.2)^2+(19.8)^2) = 367.28/1028.8 = 0.3455

*and intercept w0 = Mean(rent) – 0.3455*Mean(Area) = 42.6 – 0.3455*(174.2) = -17.57*

*The equation for predicting price based on area is y = -17.57 + 0.3455*Area.*

b) *we get Rent = -17.57 + 0.3455*180 = 44.61, so predicted rent is 44.61 thousands.*

c) *The range and size of data is low/There seems to be other variables (172~42, 174~40)*

[1.5 marks for the correct values of W0 and W1 each, total 3 marks]

[1 mark for the equation]
[1 mark for 2 b)]
[1 mark for limitation]

Question 3 [1.5*4 = 6 marks]

Suppose you are a Data scientist. You are building a Classifier that can predict whether a person is likely to default or Not based on certain parameters/attribute values. Assume, the class variable is "Default" and has two outcomes, {"yes", "no"}

- Own_House = Yes, No
- Marital Status = Single, Married, Divorced
- Annual Income = Low, Medium, High
- Currently Employed = Yes, No

Suppose a rule-based classifier produces the following rules:

1. Own_House = Yes → Default = Yes
2. Marital Status = Single → Default = Yes
3. Annual Income = Low → Default = Yes
4. Annual Income = High, Currently Employed = No → Default = Yes
5. Annual Income = Medium, Currently Employed = Yes → Default = No
6. Own_House = No, Marital Status = Married → Default = No
7. Own_House = No, Marital Status = Single → Default = Yes

Answer the following questions. Make sure to provide a brief explanation or examples to illustrate the answer.
(a) Are the rules mutually exclusive?
(b) Is the rule set exhaustive?
(c) Is ordering needed for this set of rules?
(d) Do you need a default class for the rule set?

Answers:
(a) No. The instance of Own_House = Yes, Marital Status = Single will trigger the first two rules.
(b) No. The instance of Marital Status = Divorced, Own_House = No, Annual Income = High, Currently Employed = Yes is not covered by any of the rules.

(c) Yes because a record can match two or more rules that give conflicting predictions about the class.
For example, the instance Own_House =Yes, MaritalStatus=Divorced, AnnualIncome=Medium, CurrentlyEmployed=Yes will trigger rule 1 (prediction: Default = Yes) and rule 5 (prediction: Default=No). If you do not tell the system to prefer one rule to another (i.e., order them), the system will not know how to classify the instance.
(d) Yes, since the rules are not exhaustive.

[0.5 mark for yes/no in each answer, 1 mark for the explanation/example]

Question 4:

Based on the information given in the table below, find most similar and most dissimilar persons among them. Apply min-max normalization on income to obtain [0,1] range. Consider Profession and City as nominal, and Age as an ordinal variable with ranking order of [Youth, Middle-Aged, Senior]. Give equal weight to each attribute. **[6 marks]**

Name	Profession	Age	City	Income
Sam	Doctor	Middle Aged	Mumbai	80000
John	Engineer	Youth	Delhi	50000
Mary	Politician	Senior	Bangalore	70000
Sonam	Doctor	Middle Aged	Delhi	89000
Sujoy	Carpenter	Middle-aged	Gurgaon	20000

Solution

After normalizing income and quantifying Age, we get

Name	Profession	Age	City	Income
Sam	Doctor	0.5	Mumbai	0.87
John	Engineer	0	Delhi	0.43
Mary	Politician	1	Bangalore	0.72
Sonam	Doctor	0.5	Delhi	1
Sujoy	Carpenter	0.5	Gurgaon	0

After normalizing income and quantifying native place, we get
d(Sam, John) = 1+0.5+1+ 0.44 = 2.94
d(Sam, Mary) = 1+0.5+1+ 0.15= 2.65
d(Sam, Sonam) = 0+0+1+0.13 = 1.13
d(Sam, Sujoy)= 1+0+1+.87 = 2.87
d(John, Mary) = 1+1+1+0.29= 3.29
d(john, Sonam)=1+0.5+0+0.57 = 2.07
d(john, sujoy) = 1+0.5+1+ 0.43 = 2.93
d(mary, sonam) = 1+0.5+1+0.28 = 2.78

d(mary, sujoy) = 1+0.5+1+0.72 = 3.22
d(sonam, sujoy) = 1+0+1+1 = 3

Most similar – Sam and Sonam; Most dissimilar – John and Mary

[1 mark for normalising Income attribute]
[1 mark for converting Age attribute to Numeric]
[3 marks for calculating distances]
[0.5 marks for the most similar and 0.5 marks for the most dissimilar persons]