JIS College of Engineering

(An Autonomous Institute)

Block 'A', Phase-III, Kalyani, Nadia, Pin – 741235

FRONT PAGE

TEST – I, EVEN Semester Examination 2020-21

BRANCH	Computer Science & Engineering
SEMESTER / YEAR	6th sem / 3rd Year
REGISTRATION NUMBER	181230110092
UNIVERSITY ROLL NUMBER	12318010309 2
SUBJECT OF EXAMINATION	Data Warehousing and Data Mining (CS606A)
DATE OF EXAMINATION	16/04/21
FULL SIGNATURE OF THE CANDIDATE	Souvik Pal

INSTRUCTION TO THE EXAMINEES

1. Download and print this page for each examination. Fill it up and attach on the top of the answer script. Use A4 size paper. Leave back side of the front page blank.

1. Use A4 size paper to write your answers. Write answers in own hand writing.

2. Specify page number at the top of each page of the answer script.

3. Write branch name, roll number & subject name and put full signature at the bottom of each page of the answer script.

4. Do not forget to attach the front page. In absence of duly filled in front page, answer script will be treated as incomplete and will not be considered for evaluation.

5. Send the answer script along with the filled in front page to respective department.

Group A

1) a) Naive Bayes classifier are a collection of classification algorithm based on Bayes Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle i.e, every pair of features being classified is independent of each other.

b) Variable : weather
Sunny = 1
rainy = 0

Variable : Car
working = 1
broken = 0

Variable : Class
go-out = 1
Stay-home = 0

weather	car	class
1	1	1
0	0	1
1	1	1
1	1	1
0	0	0
0	0	0
1	1	0
1	0	0
0	0	0

Branch - CSE
Roll - 12318070309 2
Subject - Data warehousing & Data Mining
Signature - Souvik pal

Class Probabilities

We can calculate the class probabilities for classes 0 and 1 as follows

$$P(\text{class} = 1) = \frac{\text{count}(\text{class} = 1)}{\text{count}(\text{class} = 1) + \text{count}(\text{class} = 0)}$$

$$= \frac{5}{10} = 0.5$$

$$P(\text{class} = 0) = \frac{\text{count}(\text{class} = 0)}{\text{count}(\text{class} = 1) + \text{count}(\text{class} = 0)}$$

$$= \frac{5}{10} = 0.5$$

Conditional Probabilities

Weather input variable :-

$$P(\text{weather} = \text{sunny} \mid \text{class} = \text{go-out}) = \frac{\text{count}(\text{weather} = \text{sunny} \;\&\; \text{class} = \text{go-out})}{\text{count}(\text{class} = \text{go-out})}$$

$$= \frac{4}{5} = 0.8$$

$$P(\text{rainy} \mid \text{goout}) = \frac{\text{count}(\text{rainy} \;\&\; \text{go-out})}{\text{count}(\text{go-out})}$$

$$= 1/5 = 0.2$$

Branch - CSE Subject - Data warehousing & Data mining

Roll - 12318070309 2 Signature - Souvik pal

$$P(\text{sunny} \mid \text{stayhome}) = \frac{\text{count}(\text{sunny} \& \text{stayhome})}{\text{count}(\text{stayhome})}$$

$$= \frac{2}{5} = 0.4$$

$$P(\text{rainy} \mid \text{stayhome}) = \frac{\text{count}(\text{rainy} \& \text{stayhome})}{\text{count}(\text{stayhome})}$$

$$= \frac{3}{5} = 0.6$$

Car input Variable

$$P(\text{working} \mid \text{go-out}) = \frac{\text{count}(\text{working} \& \text{go-out})}{\text{count}(\text{go-out})}$$

$$= \frac{4}{5} = 0.8$$

$$P(\text{working} \mid \text{stayhome}) = \frac{\text{count}(\text{working} \& \text{stayhome})}{\text{count}(\text{stayhome})}$$

$$= 1/5 = 0.2$$

$$P(\text{broken} \mid \text{go-out}) = \frac{\text{count}(\text{broken} \& \text{go-out})}{\text{count}(\text{go-out})}$$

$$= 1/5 = 0.2$$

$$P(\text{broken} \mid \text{stayhome}) = \frac{\text{count}(\text{broken} \& \text{stayhome})}{\text{count}(\text{stayhome})}$$

$= 4/5 = 0.8$

Make Prediction with Naive Bayes

Let's take the first record from our dataset and use our learned model to predict which class we think it belongs

weather = sunny, car = working

go-out = P(sunny | go-out) * P(working | go-out) * P(go-out)

$= 0.8 \times 0.8 \times 0.5$

$= 0.32$

Stayhome = P(sunny | stayhome) * P(working | stayhome) * P(stayhome)

$= 0.4 \times 0.2 \times 0.5$

$= 0.04$

We can see that $0.32 > 0.04$, so we predict "go-out" for this instance which is correct (Ans)

2) a) K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning Technique. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. KNN can be used for Regression as well as for classification but mostly it is used for the classfication problems KNN is a non-parametric algorithm which means it does not make any assumption on underlying data.

b) $X1 = 8.09$, $X2 = 3.36$

Let $K = 10/2 + 1 = 6$

Euclidean Distance $= \sqrt{(x_H - H)^2 + (x_W - W)^2 + \ldots}$

(x_H: observed value, H: Actual value)

$D(X, i) = \sqrt{(8.09 - 3.93)^2 + (3.36 - 2.33)^2}$

$= 4.29$

$D(X, ii) = \sqrt{(8.09 - 3.11)^2 + (3.36 - 1.78)^2}$

$= 5.22$

Branch - CSE Subject - DWM

Roll - 12318070309 2 Signature - Sourik Pal

$D(X, iii) = \sqrt{(8.09 - 1.34)^2 + (3.36 - 3.36)^2}$

$= 6.73$

$D(X, iv) = \sqrt{(8.09 - 3.58)^2 + (3.36 - 4.67)^2}$

$= 4.69 \rightarrow N6 \rightarrow 0$

$D(X, v) = \sqrt{(8.09 - 2.28)^2 + (3.36 - 2.86)^2}$

$= 5.83$

$D(X, vi) = \sqrt{(8.09 - 7.42)^2 + (3.36 - 4.69)^2}$

$= 1.49 \rightarrow N3 \rightarrow 1$

$D(X, vii) = \sqrt{(8.09 - 5.74)^2 + (3.36 - 3.53)^2}$

$= 2.36 \rightarrow N4 \rightarrow 1$

$D(X, viii) = \sqrt{(8.09 - 9.17)^2 + (3.36 - 2.51)^2}$

$= 1.37 \rightarrow N2 \rightarrow 1$

$D(X, ix) = \sqrt{(8.09 - 7.79)^2 + (3.36 - 3.42)^2}$

$= 0.31 \rightarrow N1 \rightarrow 1$

$D(X, x) = \sqrt{(8.09 - 7.93)^2 + (3.36 - 0.789)^2}$

$= 2.57 \rightarrow N5 \rightarrow 1$

No of 1's = 5

No of 0's = 1

Branch - CSE Subject - DWM

Roll - 12318070309 2 Signature - Souvik pal

Therefore,

$(X1 = 8.09, X2 = 3.36)$ will belong to class 1 (Ans)

Group - B

7) 13, 15, 16, 16, 19, 20, 20, 21,
22, 22, 25, 25, 25, 25, 30, 33,
33, 35, 35, 35, 35, 36, 40,
45, 46, 52, 70

i) Mean = $\frac{13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70}{27}$

= 29.96 (Ans)

Median = $\frac{27+1}{2}$ = 14th term

= 25 (Ans)

ii) Mode = 25, 35

~~1st quartile = 1/4~~

If there are two numbers that appear same number of times then the data has two modes. This

is called bimodal.

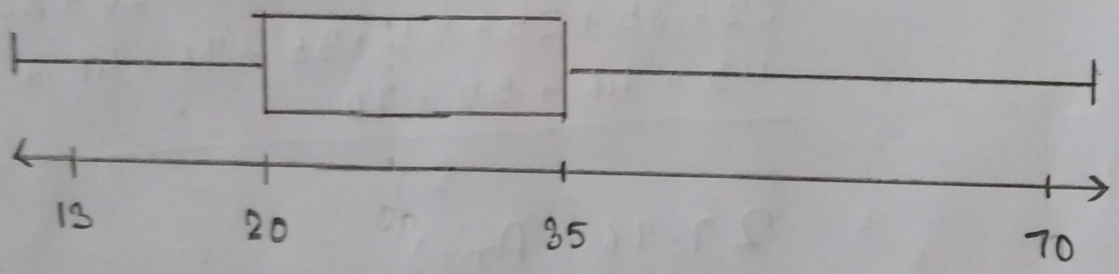
iii) 1st Quartile = This is the median of
{ 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25 }

$Q_1 = 20$ (Ans)

3rd Quartile = This is the median of
{ 30, 33, 33, 35, 35, 35, 36, 40, 45, 46, 52, 70 }
35

$Q_3 = 35$ (Ans)

v)



Minimum value = 13
Maximum value = 70
1st quartile = 20
3rd quartile = 35

iv) (13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25), 25,
(30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)
The five number summary of the data is

Minimum = 13

Q1 = 20

median = 25

Q3 = 35 (Ans)

maximum = 70

6) Apply Euclidean Distance :-

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

$\therefore$ distance $= \{(1.5-1.4)^2 + (1.7-1.6)^2\}^{1/2}$

$= \{(0.01) + (0.01)\}^{1/2} = 0.14142$ (Ans)

How cosine similarity :-

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\|$$

$$\cos(d_1, d_2) = ((1.5 * 1.4) + (1.7 * 1.6)) / ((1.5)^2 + (1.7)^2 * ((1.4)^2 + (1.6)^2))$$

$= 9.9999$

Similarly for x_2, x_3, x_4 and x_5

P.T.O

Branch - CSE Subject - DWM

Roll - 12318070309 2 Signature - Souvik Pal

	Euclidean	Cosine Similarity
x_1	0.1414	0.9999
x_2	0.6708	0.99575
x_3	0.2828	0.99997
x_4	0.2236	0.99903
x_5	0.6083	0.96536

The values produce the following ranking of data points based on similarity :-

Euclidean Distance = x_1, x_4, x_3, x_5, x_2

Cosine Similarity = x_1, x_3, x_4, x_2, x_5

Branch- CSE Subject - DWM

Roll - 12318070309 2 Signature - Souvik pal