DA ASSIGNMENT - I

MANSI DWIVEDI 2019140016 BE IT

Date

Page

Q1) ~~(a)~~

- There we can use the Naive Bayes classifier to accurately map the tuple into a class.
- This is because we can find the most likely classification
- Using Naive Bayes we can try to find the maximum likelihood.
- The formula for Naives Bayes theorem is given as:

$$P(y \mid x_1, x_2, x_3 \cdots x_n) = \frac{P(x_1/y) \cdot P(x_2/y) \cdot P(x_3/y) \cdots P(x_n/y)}{P(x_1) \cdot P(x_2) \cdot P(x_3) \cdots P(x_n)}$$

- For the given data we have:

Total tuples = 20.

(i) $P(\text{on time}) = \frac{14}{20} = 0.7$ [Since there are 14 instances when class was on time]

(ii) $P(\text{Late}) = \frac{2}{20} = 0.1$ [similarly since there are 2 instances when class was late]

(iii) $P(\text{Very late}) = \frac{3}{20} = 0.15$ [3 instances when class was very late]

(iv) $P(\text{Cancelled}) = \frac{1}{20} = 0.05$ [1 instance when the class itself was cancelled]

Date
Page

Now lets find the prior / conditional probabilities for each attribute.

(i) For Days:

For Days:	On time	Late	Very late	Cancelled.
Weekdays	9/14	1/2	3/3	0/1
Holiday	2/14	1/2	0/3	0/1
Saturday	2/14	0/2	0/3	1/1
Sunday	1/14	0/2	0/3	0/1

(ii) For Seasons:

For Seasons:				
Spring	4/14	0/2	0/3	1/1
Winter	2/14	2/2	2/3	0/1
Summer	6/14	0/2	0/3	0/1
Autumn	2/14	0/2	1/3	0/1

(iii) For Fog:

For Fog:				
None	5/14	0/2	0/3	0/1
High	4/14	1/2	1/3	1/1
Normal	5/14	1/2	2/3	0/1

(iv) For Rain

For Rain				
None	6/4	1/2	1/3	0/1
Heavy	2/14	0/2	2/3	1/1
Slight	6/14	1/2	0/3	0/1

Now finding probability for each case:
Set [Weekday, Winter, Fog = High, Rain = None)

Case 1) Class was 'On Time'.
P(on time / Instance) = P(on time) × P(weekday / ontim
× P(winter / on time) × P(High / on
× P(None / on time)

$= 0.7 \times \frac{9}{14} \times \frac{2}{14} \times \frac{4}{14} \times \frac{6}{14} = \boxed{0.007872}$

Case 2) Similarly when class is 'Late'.

$P(\text{late} / \text{Instance}) = P(\text{late}) \times P(\text{weekday} / \text{late}) \times P(\text{winter} / \text{late}) \times P(\text{High} / \text{late}) \times P(\text{None} / \text{late})$

$= 0.1 \times \frac{1}{2} \times \frac{2}{2} \times \frac{1}{2} \times \frac{1}{2} = \boxed{0.0125}$

Case 3) Similarly when class is 'Very late'

$P(\text{very late} / \text{Instance}) = P(\text{Very late}) \times P(\text{weekday} / \text{very late}) \times P(\text{winter} / \text{very late}) \times P(\text{High} / \text{very late}) \times P(\text{None} / \text{very late})$

$= 0.15 \times \frac{3}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} = \boxed{0.01111}$

Case 4) Finally when class is 'Cancelled'

$P(\text{Cancelled} / \text{Instance}) = P(\text{cancelled}) \times P(\text{weekday} / \text{cancelled}) \times P(\text{winter} / \text{cancelled}) \times P(\text{High} / \text{cancelled}) \times P(\text{None} / \text{cancelled})$

$= 0.5 \times \frac{0}{1} \times \frac{0}{1} \times \frac{1}{1} \times \frac{0}{1} = 0.$

Hence we can see that the highest probability occurs for the case of ~~class~~ ~~being~~ starting late. Thus we can say that when the day is a weekday, season is winter, fog is high with no rainfall the class is most likely to be $\boxed{\text{late}}$.

Date
Page

Q.2) Sample size $= n = 1500$.

Let's ~~def~~ state our null and alternate hypothesis

H_0 : Preffered reading and gender are not correlated or are independent

H_a : Preffred reading and gender are not independent of each other.

Let us perform chi-square test to test our hypothesis

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

O_{ij} = Observed frequency
e_{ij} = expected frequency
m = no. of rows
n = no. of columns

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= \frac{(160)^2}{90} + \frac{(-160)^2}{210} + \frac{(-160)^2}{360} + \frac{(160)^2}{840}$$

$$= 284.444 + 121.9047 + 71.1111 + 30.47619$$

$$= 507.93639.$$

Now here degree of freedom is given as $(m-1)(n-1)$

since $m = n = 2$

$\therefore$ $df = (2-1)(2-1) = 1$

From the chi square table we can see that the value corresponding to 1 degree of freedom and a significance level of 0.01 is $\boxed{6.635}$.

Since obtained value is greater than 6.635
[507.93639 > 6.635] we reject the null hypothesis

Hence we can conclude that preffered reading and gender are strongly correlated to each other.