Name: Vivek Hemantsing Rajput
UID: 2019110046
class: B.E ETRX (Data Analytics)

Date
Page

Assignment - 1 - 12-11-2022

Q.1 Air Traffic Data

→ Let us consider a set observation recorded in a database

$\therefore A = [\text{Day, season, Fog, Rain}]$ with 20 tuples

$\therefore C = [\text{On Time, Late, Very Late, Cancelled}]$

To find most likely classification for any other unseen instance:

Week Day, Winter, High, None, ??

$\therefore$ Conditional Probability : $\therefore P(y/x_1 \ldots, x_n)$

$$= \frac{P(x_1/y) \cdot P(x_2/y) \ldots P(x_n/y) \cdot P(y)}{P(x_1) \cdot P(x_2) \ldots P(x_n)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(B)}$$

$$= \frac{P(B|A) \cdot P(A)}{P(B)}$$

	Attribute	Class: On-Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	1/2 = 0.5	3/3 = 1	0/1 = 0
	Saturday	2/14 = 0.14	1/2 = 0.5	0/3 = 0	1/1 = 1
	Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
	Holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
Season	Spring	4/14 = 0.29	0/2 = 0	0/3 = 0	0/1 = 0
	Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
	Autumn	2/14 = 0.14	0/2 = 0	1/3 = 0.33	0/1 = 0
	Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0/1 = 0

Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
	Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
Rain	None	5/14 = 0.36	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
	Heavy	1/14 = 0.07	1/2 = 0.5	2/3 = 0.67	1/1 = 1
	Prior Probability	14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

Weekday - Winter - High - None - ???

Case 1 : Class = On-Time : Prior Probability (On-time)
× Weekday (On-Time) - Day
× (Winter (On-Time) - season
× ~~(Foo~~ High (On-time) - Fog
× None (On-Time) - Rain

= 0.70 × 0.64 × 0.14 × 0.29 × 0.36

= 0.0065

Prior Probability (Late)

Case 2 : Class = Late : 0.10 × 0.5 × 1 × 0.5 × 0.5

= 0.0125

Prior Probability (Very Late)

Case 3 : Class = Very Late : 0.15 × 1 × 0.67 × 0.33 × 0.33

= 0.0109

prior probability (Cancelled)

Case 4 : Class = Cancelled : 0.05 × 0 × 0 × 1 × 0

= 0.000

Highest Probability Occurs for the case 2 (Late).
Therefore, Case 2 is the strongest ;
Hence correct classification is Late.
So, When the day is Weekday, Season is Winter, Fog is High, Rain is None, Class mostly like to 'Late'

Page

Q.2

1. Expected Frequencies

The expected e_{ij} frequency which can be computed as

$$e_{ij} = \frac{count\ (A=a_i) \times count\ (B=b_j)}{N}$$

where N = no. of data tuples.

count (A = ai) = no. of tuples having value ai for A

count (B = bj) = no. of tuples having value bj for B

Now,

e_{11} = [count (male) × count (fiction)] / N

∴ e_{11} = [300 × 450] / 1500 = 90

∴ e_{12} = [1200 × 450] / 1500 = 360

∴ e_{12} = [300 × 1050] / 1500 = 210

∴ e_{22} = [1200 × 1050] / 1500 = 360

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

O_{ij} = Observed frequency
e_{ij} = Expected frequency

m & n are no. of rows & no. of colums

Gender / preferred read	Male	Female	Total
Fiction	250 (90)	200 (360)	450
Non-fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Ho: Preferred Reading & gender are independent of each other.

Ha: Preferred Reading & gender are not independent each other.

Noted that in any row, the sum of the expected frequencies must equal the total observed frequency for that row

$$\chi^2 = \Sigma \frac{(\text{Observed} - \text{Expected})}{\text{Expected}}$$

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

$$= 284.44 + 121.90 + 71.11 + 30.48$$

$$\boxed{\therefore \chi = 507.93}$$

For this given table,

	Male	Female	Sum (row)
fiction	250 (90)	200 (360)	450
Non-fiction	50 (210)	1000 (840	1050
sum (col.)	300	1200	1500

the degree of freedom are $\overset{(m-1)}{(2-1)}\overset{(n-1)}{(2-1)} = 1$.
For 1 degree of freedom.
Value χ^2 with degree of freedom 1 and 0.01 significance level from the standard statistical table is 6.635 or 10.828 (taken from the table of upper % points of χ^2)
Our receive value is above this value. Therefore we can reject the hypothesis that gender & preferred reading are independent and
Conclude that the two attributes are strongly correlated for given group of people.