DA - Assignment 1

Khushi Patni
2019120047
BE EXTC

Problem 1 - Air Traffic Data

There are four attributes A = [Day, Season, Fog, Rain]
The categories of classes are C = [On Time, Late, VLate, Cancel]

	Attribute	On Time	Late	Very Late	Cancelled
DAYS	Weekday	9/14 = 0.64	1/2 = 0.5	3/3 = 1	0
	Saturday	2/14 = 0.14	0	0	1/1 = 1
	Sunday	1/14 = 0.07	0	0	0
	Holiday	2/14 = 0.14	1/2 = 0.5	0	0
SEASON	Spring	4/14 = 0.28	0	0	1/1 = 1
	Summer	6/14 = 0.42	0	0	0
	Autumn	2/14 = 0.14	0	1/3 = 0.33	0
	Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0
FOG	None	5/14 = 0.35	0	0	0
	High	4/14 = 0.28	1/2 = 0.5	1/3 = 0.33	1/1 = 1
	Normal	5/14 = 0.35	1/2 = 0.5	2/3 = 0.67	0
RAIN	None	6/14 = 0.42	1/2 = 0.5	1/3 = 0.33	0
	Slight	6/14 = 0.42	1/2 = 0.5	0	0
	Heavy	2/14 = 0.14	0	2/3 = 0.67	1/1 = 1
PRIOR PROBABILITY		14/20 = 0.7	2/20 = 0.1	3/20 = 0.15	1/20 = 0.05

Instance -

Weekday Winter High None ???

CASE I - Class - On Time

$P_{ontime} = 0.7 \times 0.64 \times 0.14 \times 0.28 \times 0.42$

$= 0.00737$

CASE II - Class - Late

$P_{Late} = 0.1 \times 0.5 \times 1 \times 0.5 \times 0.5$

$= 0.0125$

CASE III - Class - Very Late

$P_{VLate} = 0.15 \times 1 \times 0.67 \times 0.33 \times 0.33$

$= 0.0109$

CASE IV - Class - Cancelled

$P_{cancel} = 0.05 \times 0 \times 0 \times 1 \times 0$

$= 0$

Using the formulla,

P(yes) = P(yes) P(weekday | yes) P(Winter | yes) P(High | yes) P(None | yes)

As the probability of class - Late is 0.0125 the greatest The instance - Weekday, Winter, High, None will fall unde

Late category

Problem 2 · Statistical Learning

H0 : Preferred reading and Gender are not correlated.
H1 : Both are correlated

No. of people = 1500

	male	female	Total
fiction	250(90)	200(360)	450
non. fiction	50 (210)	1000(840)	1050
Total	300	1200	1500

Correlation analysis

$$\chi^2 \text{ (chi-squared) test} = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

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

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

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

$$= 507.93$$

For this 2×2 table, the degree of freedom is $(2-1)(2-1) = 1$
For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828.
Since the value which we computed is above this, we can reject this hypothesis that gender and preferred reading are independent.
Concluding that, both attributes are strongly correlated.